Medical Image Analysis 94 (2024) 103149

Contents lists available at ScienceDirect

Medical Image Analysis

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

Multi-domain stain normalization for digital pathology: A cycle-consistent adversarial network for whole slide images

Martin J. Hetz, Tabea-Clara Bucher, Titus J. Brinker *

Division of Digital Biomarkers for Oncology, German Cancer Research Center (DKFZ), Heidelberg, Germany

ARTICLE INFO

Dataset link: https://github.com/DBO-DKFZ/multistain_cyclegan_normalization

Keywords:
Stain normalization
Adversarial networks
Domain shift
CycleGAN
Histopathology
Domain adaptation

ABSTRACT

The variation in histologic staining between different medical centers is one of the most profound challenges in the field of computer-aided diagnosis. The appearance disparity of pathological whole slide images causes algorithms to become less reliable, which in turn impedes the wide-spread applicability of downstream tasks like cancer diagnosis. Furthermore, different stainings lead to biases in the training which in case of domain shifts negatively affect the test performance. Therefore, in this paper we propose MultiStain-CycleGAN, a multi-domain approach to stain normalization based on CycleGAN. Our modifications to CycleGAN allow us to normalize images of different origins without retraining or using different models. We perform an extensive evaluation of our method using various metrics and compare it to commonly used methods that are multi-domain capable. First, we evaluate how well our method fools a domain classifier that tries to assign a medical center to an image. Then, we test our normalization on the tumor classification performance of a downstream classifier. Furthermore, we evaluate the image quality of the normalized images using the Structural similarity index and the ability to reduce the domain shift using the Fréchet inception distance. We show that our method proves to be multi-domain capable, provides a very high image quality among the compared methods, and can most reliably fool the domain classifier while keeping the tumor classifier performance high. By reducing the domain influence, biases in the data can be removed on the one hand and the origin of the whole slide image can be disguised on the other, thus enhancing patient data privacy.

1. Introduction

The gold standard of cancer diagnosis is histopathologic investigation of tissue. This involves microscopic examination of dissected and stained tissue to examine signs and characteristics of specific diseases. During dissection, the tissue is fixed in formaldehyde and embedded in paraffin Chatterjee (2014). The following staining of the tissue sections serves to highlight the various structures and cells in the tissue (Ghaznavi et al., 2013). For light microscopy, the tissue sections are routinely stained with hematoxylin and eosin (H&E) (Alturkistani et al., 2015). The staining of the tissue helps the pathologist to make diagnoses based on certain features such as cell morphology or the arrangement of cells (Gurcan et al., 2009). Staining mainly depends on the formulation of the stain and the application time among other pipeline-dependent aspects (Kothari et al., 2014). After the tissue is stained, the slides are reviewed by pathologists. Both diagnosis and tumor grading are performed with the goal of providing prognosis and treatment recommendations (Farahani et al., 2015). With the advancement of technology and the rise of whole-slide imaging, i.e., the digitization of high-resolution microscopic images, digital pathology has rapidly gained importance (Alturkistani et al., 2015). Digitizing the slides allows the use of automated systems for detection, classification or segmentation of a desired entity. Deep Learning (DL) and Convolutional Neural Networks (CNN) emerged as an effective tool in automatic image analysis, using big data to parameterize models that no longer rely on hand-crafted features (LeCun et al., 2015). By using DL-based algorithms, pathologists can be relieved of monotonous and repetitive tasks by supporting them (Echle et al., 2021). DL has already been successfully tested in various clinical tasks, such as tumor detection (Cruz-Roa et al., 2017), mitosis detection (Saha et al., 2018; Albarqouni et al., 2016), grading of cancer (Bulten et al., 2020), predicting the lymph sentinel status (Kiehl et al., 2021), patient survival estimation (Wessels et al., 2022) or tumor subtyping (Wang et al., 2020).

The appearance of slides is highly variable as a result of different scanners, staining techniques, and laboratories (Ciompi et al., 2017). This variability is not a problem for pathologists (Bancroft, 2008), but current deep learning algorithms have significant issues with this (Ciompi et al., 2017). Even very small changes in the image can

* Corresponding author.
E-mail address: titus.brinker@dkfz.de (T.J. Brinker).

https://doi.org/10.1016/j.media.2024.103149
Received 21 November 2022; Received in revised form 11 December 2023; Accepted 20 March 2024
Available online 28 March 2024

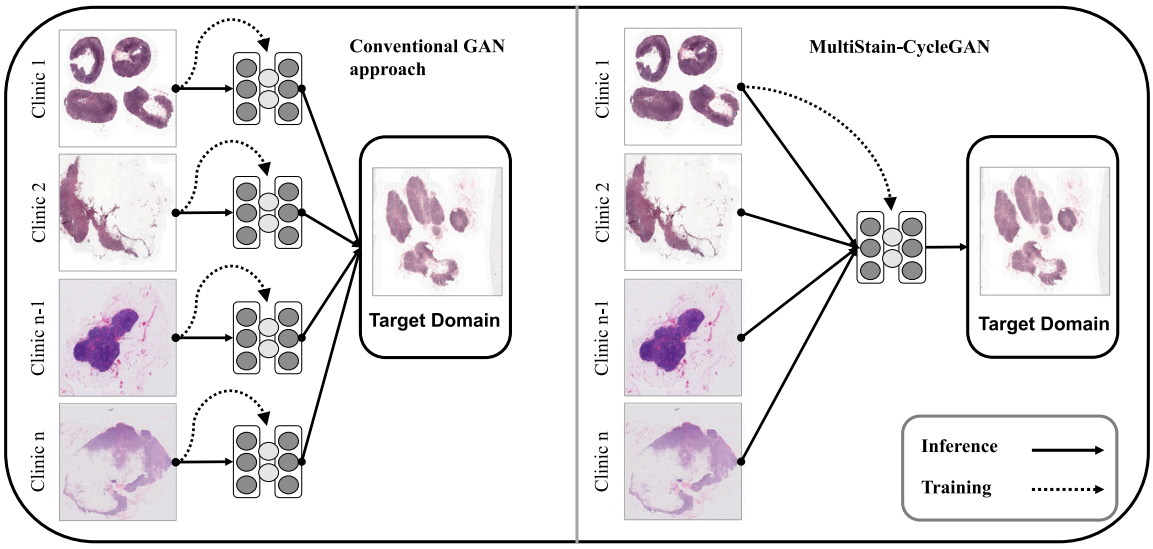

Fig. 1. Left: Stain normalization based on conventional GAN approaches. For each staining a separate model has to be trained to normalize many stainings to the target domain. Right: Stain normalization with MultiStain-CycleGAN, which is trained on one staining and can normalize any H&E staining of the same tissue type. Dotted arrows indicate the data needed for training the respective model, normal arrows show the inference path.

lead to large deviations in performance for DL-based algorithms (Kurakin et al., 2016). This is especially true for domain shifts, where a different input data distribution can lead to reduced performance and in turn potentially harm the patient (Stacke et al., 2021). As early as 1994, Lyon et al. postulated that the standardization of dyes and stains will play an increasingly important role in the future (Lyon et al., 1994). However, not all differences are due to non-standardized processes and thus not all variations are avoidable (Niethammer et al., 2010). In conclusion, every tissue source site e.g. medical center or clinic, has a distinct signature due to biological variations in patients treated at various centers, specimen acquisition, staining, and digitization. On the one hand, this center-specific signature leads to biases in the data, which many algorithms suffer from; on the other hand, it can be used to determine the source of a whole slide image (WSI). The origin of the WSI can then be used to draw conclusions about the patient demographics, such as age, nationality and ethnicity. DL is able to determine the origin of a WSI with high accuracy (Howard et al., 2021). This means that patient data privacy is no longer guaranteed and also enables the misuse of this information. To integrate DL into the work of pathologists and clinicians, methods that make these algorithms robust and stain-invariant must be developed, as well as new approaches that can remove the center-specific signature. Stain normalization is one possible method for achieving this objective.

1.1. Stain normalization

Methods for normalization of histological slides have already been shown to be effective for some applications (Ciompi et al., 2017). These normalization methods transform images x of a domain $\mathcal{X}$ to look like they originated from a target domain $\mathcal{Y}$ or to match the appearance of a template image y. Because of this, the field of stain normalization is a very active field, which tries to map images from a source domain to a target domain. In this regard, Salvi et al. divides the field into the three areas: Global color normalization, color normalization after stain separation, and Deep Learning based normalization (Salvi et al., 2021). Methods based on global color normalization apply procedures that use the statistics of a template image and obtain a transformation from it, such as the global color transformation using Principal Component Analysis (PCA) by Reinhard et al. (2001) or histogram specification proposed by Gurcan et al. (2009). In methods based on stain normalization after stain separation, the individual staining components, usually hematoxylin and eosin, are separated. This technique makes use of the property that the two stains can be linearly separated by a transformation into optical density space (Roy et al., 2018). Each pixel can then be calculated by the product of a stain color appearance matrix, which is acquired by a template image, and the stain density map (Salvi et al., 2021). The estimation of the appearance matrix can be based on singular value decomposition as described by Macenko et al. (2009), prior information (Ruifrok and Johnston, 2001), non-negative matrix factorization (Vahadane et al., 2016) or spectral matching (Tosta et al., 2019). More recent approaches rely increasingly on neural networks. Generative Adversarial Networks (GAN) are used to normalize stains and show promising results (Zanjani et al., 2018; Bentaieb and Hamarneh, 2018). Following the work of Gatys et al. (2016), stain normalization is treated as neural style transfer, where the goal is to give an input image the appearance of a learned distribution of images. GANs are deep generative models which consist of two networks: A generator which generates images and a discriminator which tries to separate the generated images from the actual images corresponding to the distribution of the training data. The training is a minimax game in which the two models compete with each other. The goal of the generator is to generate images from a noise vector z such that the distribution of the generated images $P_G(z)$ corresponds as closely as possible to the distribution of the training data $P_{data}(x)$. The loss function can be described as:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}(x)}\left[\log\left(D_x\right)\right] + \mathbb{E}_{z\sim p_z(z)}\left[\log\left(1 - D\left(G\left(z\right)\right)\right)\right]. \tag{1}$$

The generator tries to minimize the loss function, while the discriminator tries to maximize it. The models are trained until they reach an optimum (Goodfellow et al., 2014). Isola et al. (2017) extends this approach by passing an image instead of a noise vector z as input to the generator. This allows images to be transferred to another domain, which is essential for the normalization of histological stains. However, the disadvantage of this approach is that image pairs are required. In histopathology, it is rare for the same slide to be re-stained multiple times, so this approach is of limited use. Salehi et al. circumvent this drawback by converting images to gray-scale images, thus generating image pairs synthetically (Salehi and Chalechale, 2020). Due to the missing image pair problem, more work has been directed towards cycle-consistent adversarial networks (CycleGANs), proposed by Zhu et al. (2017), which no longer require image pairs by using cycle-consistency. Approaches based on CycleGAN are particularly suitable

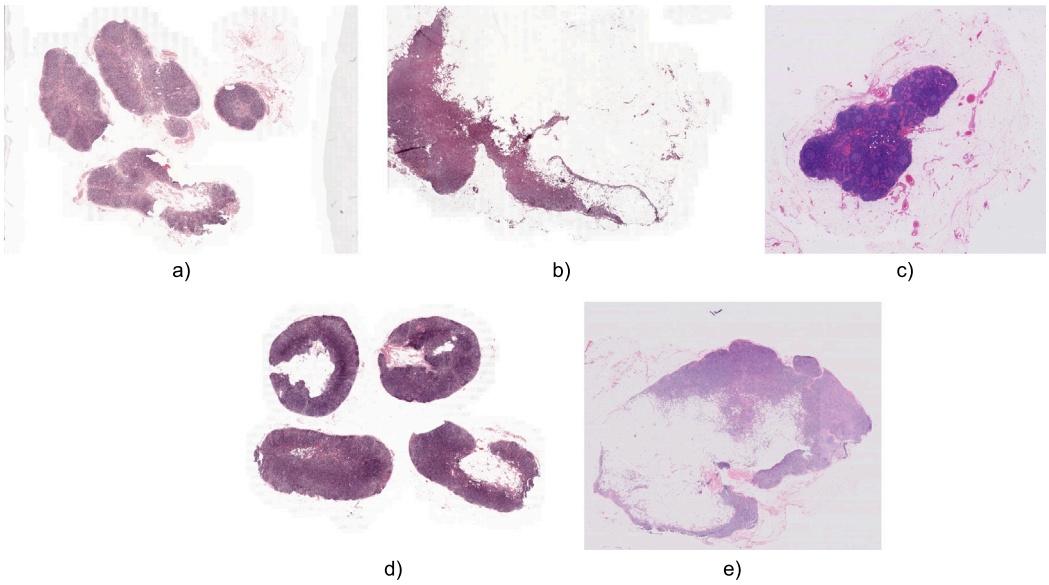

Fig. 2. Example slides for the different domains from the CAMELYON17 dataset. The images (a)–(e) show examples of tissue sections of the different centers with their different stainings: (a) Canisius-Wilhelmina Hospital (CWZ); (b) Rijnstate Hospital (RST); (c) University Medical Center Utrecht (UMCU); (d) Radboud University Medical Center (RUMC); (e) Laboratory of Pathology East-Netherlands (LPON).

for the field of image-to-image translation in histopathology. CycleGAN and its variants have already demonstrated in several studies that they are suitable for stain-to-stain translation or normalization tasks (Shaban et al., 2019; de Bel et al., 2021; Runz et al., 2021; Zhou et al., 2019).

1.2. Multi-domain stain normalization

The majority of previous literature focuses on transfer between two stains, but not on a many-to-one approach, which offers more flexibility in real-world settings. One-to-one approaches require to train a new model whenever stains from a new domain have to be normalized, see Fig. 1. Furthermore, they depend on the local availability of the data to perform the normalization. In privacy-preserving settings such as federated or swarm learning, this is not the case. For this reason, we introduce MultiStain-CycleGAN, an unsupervised multi-domain capable stain normalization method based on CycleGAN. The method presented here follows a comparable approach as described (Tellez et al., 2019) but with the additional reconstruction of the input image to ensure the integrity of the image structure through the cycle consistency condition. In doing so, we reformulate the stain normalization task into an image-to-image translation task, in which heavily augmented input images are subjected to gray-scale conversion and transformed into the desired stain. Color augmentation has already been shown in other work to be a very valuable component in achieving multi-domain normalization (Wagner et al., 2021; Cong et al., 2022). Our proposed network learns to reconstruct images from gray images with different contrasts, which have the appearance of the learned staining and thus perform stain normalization. By transforming the RGB images into gray scale images, the normalization problem is simplified because the input data are closer together. This principle is illustrated in 3. Here we use the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) method as the dimension reduction method. The input space for a tissue type can be covered very well by a preceding color augmentation. The color augmentation is used to broaden the distribution of possible inputs. Our method is tested for several properties including tumor classification, domain classification, image quality of generated images and distribution shift after normalization. Our contributions can be summarized to:

- Developing MultiStain-CycleGAN, a robust deep learning based multi-domain approach for stain normalization without the need for retraining to normalize untrained stainings.
- Achieving low accuracy on tissue source site classification to remove spurious domain factors and thus improving data privacy while generating images with a very high SSIM index and retaining tumor prediction accuracy
- Detailed analysis of our method in comparison with state-of-the-art and widely used baseline normalization methods in terms of their ability to improve a downstream task, their image quality and their ability to disguise the origin of the images, as well as the influence of different data augmentation intensities.

The following parts of this work are structured as follows: Section 2 introduces the datasets and their attributes. In addition, it gives a description of how the data is acquired, preprocessed and stratified. Section 3 explains stain normalization with CycleGAN and how we derived a multi-domain approach from it. In Section 4 a detailed description of how the used models are setup. Section 5 describes the experimental setup and the metrics used. Section 6 lists the results in detail. Afterwards the results are discussed and placed into context in Section 7. Section 8 provides a concluding summary and identifies research gaps for future work.

2. Materials

For our experiments, we use two datasets, the CAMELYON 17 Challenge dataset (Bandi et al., 2019) and the SCC dataset (Wilm et al., 2023), which are both excellent for studying stain normalization. Both datasets offer a clear domain shift and lesion-level annotations. These datasets allow us to make valid statements about the proposed normalization method and the quality of the normalized images.

2.1. Camelyon17 dataset

The Camelyon17 dataset consists of whole slide images originating from five different medical centers, digitized with three different scanners. The five centers are: Radboud University Medical Center (RUMC), Canisius-Wilhelmina Hospital (CWZ), University Medical Center Utrecht (UMCU), Rijnstate Hospital in Arnhem (RST) and Laboratory of Pathology East-Netherlands (LPON). The dataset provides the domain shift we require for our investigation of stain normalization. The dataset encompasses a total of 500 breast tissue WSIs, 50 of them with annotations at lesion level. For our experiments we only use the 50 WSIs with lesion level annotations. A few example slides of the different domains are shown in Fig. 2.

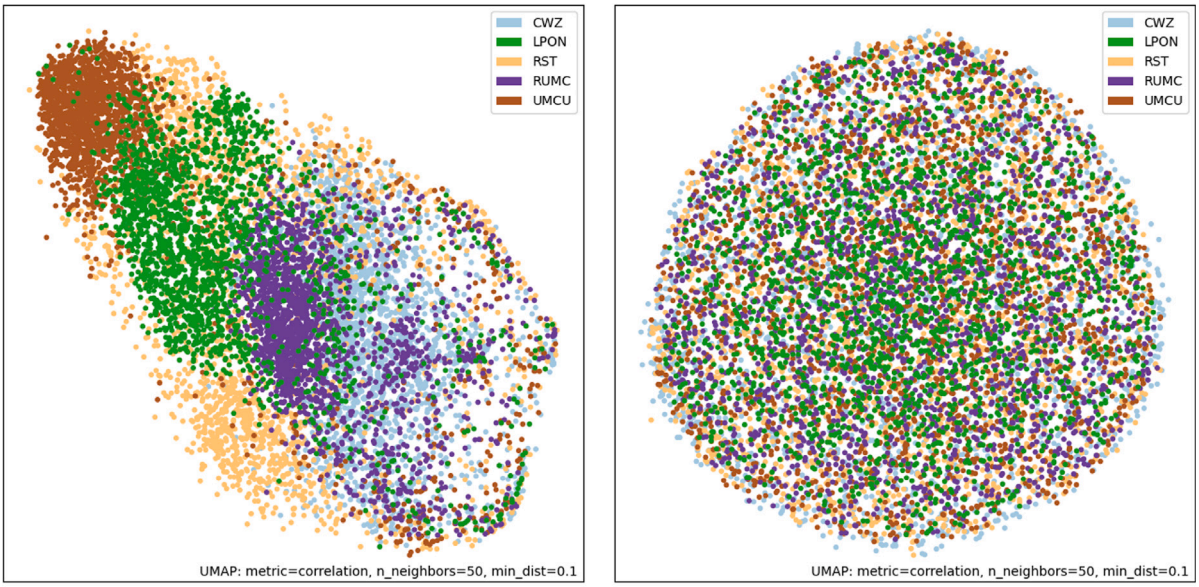

Fig. 3. UMAP plots of randomly selected tiles from the Camelyon17 dataset. These show (a) the distribution of the images in their original RGB format and (b) the distribution of the data after a grayscale conversion. (a) shows a very clear clustering of the individual domains, whereas in (b), due to the gray value conversion, the data points are much closer together and the individual clusters are no longer present.

2.2. SCC dataset

The SCC dataset comprises 44 whole-slide images (WSI) of canine cutaneous squamous cell carcinoma, collected from multiple scanners. Each WSI was scanned by five different scanners. This is particularly desirable as it allows the domain shifts caused by each scanner to be considered in isolation. The scanners used are Aperio ScanScope CS2 (CS2), NanoZoomer S210 (NZ210), NanoZoomer 2.0-HT (NZ20), Pannoramic 1000 (P1000) and the Aperio GT 450 (GT450). The annotations include tumor and different skin tissue classes. To ensure consistent evaluation of both datasets, we merged the annotations into two classes: tumor and non-tumor. The resolution of the WSIs is between 0.22 μm/pixel and 0.26 μm/pixel depending on the scanner used. Full-resolution WSIs are available upon request, and a low-resolution version is publicly available. For our work, we use the high-resolution dataset.

2.3. Data preprocessing

The lesion-level annotated slides were used to evaluate tumor classification, classifying the tissue submitting site, image quality and determine the distance between the target and source domains. For the preprocessing of the tile-based approach, we utilized our self-developed publicly available pipeline[1] for the extraction of tiles from WSIs and a subsequent filtering by tissue presence according to Khened et al. (2021). We employed a configuration such that each tile had a spatial extent of 64×64 μm and a resolution of 256×256 pixels. Furthermore, we sampled the tiles with an overlap of 25% for non-annotated and 50% for annotated tissue.

2.4. Normalization network

Our MultiStain-CycleGAN required tiles from two domains for training. In our experiments, we chose 3 normalization paths per dataset. Both directions of normalization are evaluated to ensure the best possible complete picture of the methods investigated. The normalization paths were chosen arbitrarily with the constraint that each domain acts as target and source domain at least once. For the training we used around 200 000 tiles for each domain.

2.5. Deep learning classifiers

For the tumor classifier, we used a 5-fold cross-validation. The folds were stratified according to the class 'tumor' or 'non-tumor'. In training, we only used data from the respective target domain. For the Camelyon17 classifiers, about 200 000 tiles per domain were used. All tiles of a domain were always used for training the tumor classifier.

For the tumor classifiers trained on the SCC dataset, 500 000 tiles were used per classifier, with 50% of the tiles representing non-tumor samples, resulting in a well-balanced tumor-non-tumor ratio.

Also for the training of the domain classifier we decided to use a 5-fold cross-validation, stratified by domains. For the domain classifiers of both datasets, we used about 200 000 tiles, stratified by domains.

Our two testsets consist of 5000 non-tumor and 5000 tumor tiles, which were not part of the training dataset of one of the classifiers, per domain in the respective dataset. This results in a total of 40 000 tiles per testset.

3. Methods

This section introduces the general methods needed for this work such as CycleGAN and stain normalization. Then, we present how we derived MultiStain-CycleGAN and how we perform stain normalization. Furthermore, we explain how we conduct the study. Last but not least, we describe the used performance metrics.

3.1. Cycle-consistent adversarial networks

Cycle-consistent adversarial networks proposed by Zhu et al. learn a mapping from a domain $\mathcal{X}$ to a domain $\mathcal{Y}$ using training data $x_i \in X$ and $y_i \in Y$. Where X and Y denotes the respective images of the domains. In Fig. 4, the two mappings are illustrated with $G : \mathcal{X} \rightarrow \mathcal{Y}$ and $F : \mathcal{Y} \rightarrow \mathcal{X}$. The domain-dependent adversarial discriminators D_x and D_y learn whether the input image is a generated image $G(x)$ or $F(y)$ or a sample x or y from the distribution of the training data (Zhu

[1] https://github.com/DBO-DKFZ/wsi_preprocessing

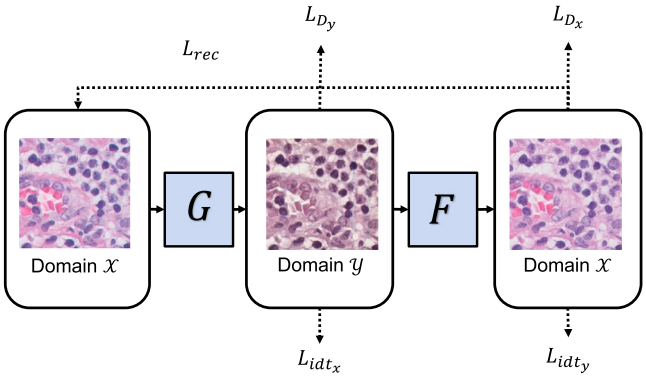

Fig. 4. The principle of image-to-image translation with CycleGAN proposed by Zhu et al. An image from a domain $\mathcal{X}$ is mapped to a domain $\mathcal{Y}$ by a generative model. After the mapping the image will be reconstructed into its original domain and the cycle-consistency loss is computed, enabling unpaired image-to-image translation.

et al., 2017). Here, the objective function contains several loss terms. Among them are adversarial losses L_{D_x} and L_{D_y}, to learn matching the distribution of generated and train data (Goodfellow et al., 2014) and a cycle-consistency loss L_{rec}, which helps to preserve the structure of input images, as well as identity losses L_{idt_x} and L_{idt_y}, which help to keep the color palette close to the input image (Zhu et al., 2017). The two adversarial losses are applied to the two mapping functions G and F. For example, the adversarial loss for the function $G : \mathcal{X} \rightarrow \mathcal{Y}$ with its discriminator D_y can be expressed as follows:

$$L_{GAN}\left(G, D_y, X, Y\right) = \mathbb{E}_{y \sim p_{data}(y)}\left[\log\left(D_y(y)\right)\right] + \mathbb{E}_{x \sim p_{data}(x)}\left[\log\left(1 - D_y\left(G\left(x\right)\right)\right)\right]. \tag{2}$$

Here, the generator G tries to generate images $G(x)$ that appear as if they originate from the distribution Y, while the discriminator D_y tries to distinguish the generated samples $G(x)$ from the real samples y. This process is repeated for the function $F : \mathcal{Y} \rightarrow \mathcal{X}$ with $L_{GAN}(F, D_x, Y, X)$(Zhu et al., 2017). Adversarial training can theoretically learn mappings G and F such that the generated images correspond to the distribution of the respective target domains $\mathcal{X}$ and $\mathcal{Y}$. Furthermore, with a sufficiently large capacity of the model, it is possible to map the same input to several different outputs in the target domain. To reduce the solution space, Zhu et al. suggest that the mapping functions should be cycle-consistent. This behavior is enforced by the cycle consistency loss with:

$$L_{cyc}\left(G, F, X, Y\right) = \mathbb{E}_{x \sim p_{data}(x)}\left[\left\|F\left(G\left(x\right)\right) - x\right\|_1\right] + \mathbb{E}_{y \sim p_{data}(y)}\left[\left\|G\left(F\left(y\right)\right) - y\right\|_1\right]. \tag{3}$$

For certain tasks, including stain normalization, it is useful to add an identity loss which ensures that the mapping is consistent with the color of the input image. The two mapping functions G and F learn the identity function in case a sample from the real distribution represents the input image. This loss is described by:

$$L_{identity}\left(G, F, X, Y\right) = \mathbb{E}_{y \sim p_{data}(y)}\left[\left\|G(y) - y\right\|_1\right] + \mathbb{E}_{x \sim p_{data}(x)}\left[\left\|F(x) - x\right\|_1\right]. \tag{4}$$

The complete objective function is therefore obtained as:

$$\begin{aligned} L\left(G, F, D_x, D_y, X, Y\right) = {} & L_{GAN}\left(G, D_y, X, Y\right) + \\ & L_{GAN}\left(F, D_x, Y, X\right) + \\ & \lambda L_{cyc}\left(G, F, X, Y\right) + \\ & L_{identity}\left(G, F, X, Y\right) \end{aligned} \tag{5}$$

(Zhu et al., 2017).

3.2. Multi-domain stain normalization with MultiStain-CycleGAN

To convert CycleGAN into a many-to-one approach, we performed several modifications. To reduce the space of possible inputs and thus simplify the problem, the images are converted to 3-channel grayscale images. Thus, the new task for the two mapping functions G and F is the reconstruction of the RGB images of the respective domain from the gray-converted images. We hypothesize that if the reconstruction $F(G(x))$ and $G(F(y))$ is sufficiently good, the loss of information due to gray scale conversion will be compensated by the model and thus be valid. In order to increase the variance of the input data and thus improve the generalization ability of the generators, an augmentation function is applied. The augmentation function H then is obtained from the color augmentation and the gray value conversion. This function is essential to later normalize data outside the distribution of the raw training data. Depending on the intensity of the augmentation, more or less information can be lost in the image due to lack of contrast. The task of the generators changes to denoising and recoloring into the target domain. By applying the function H, the images are transformed into the intermediate domain $\mathcal{W}$, which represents the input space (see Fig. 5). The use of the intermediate domain $\mathcal{W}$ allows us to normalize a large variation of input images at inference time. Since the original identity loss task is no longer relevant in this setup, the additional task of reconstructing unaugmented grayscale images was added instead. This additional task allows to compensate for the noisy images that may result from strong contrast augmentations, thus focusing G and F on the normalization of color instead of the denoising task. This domain faithful reconstruction loss of the gray converted input image x', y' and the original image x, y is:

$$L_{idt_{rec}}(G, F, X, Y) = \mathbb{E}_{y \sim p_{data}(y)}\left[\|G(y') - y\|_1\right] + \mathbb{E}_{x \sim p_{data}(x)}\left[\|F(x') - x\|_1\right]. \tag{6}$$

Thus, the complete objective function for our model results in:

$$\begin{aligned} L\left(G, F, D_x, D_y, X, Y\right) = {} & L_{GAN}\left(G, D_y, X, Y\right) + \\ & L_{GAN}\left(F, D_x, Y, X\right) + \\ & \lambda_{cyc} L_{cyc}\left(G, F\right) + \\ & \lambda_{idt} L_{idt_{rec}}\left(G, F\right). \end{aligned} \tag{7}$$

4. Implementation

4.1. Normalization network architecture

We base the model architecture on the original implementation by Zhu et al. (2017), but use a U-Net generator instead of a ResNet generator. G and F are each an adapted U-Net, which has proven to be a very effective architecture in various medical tasks (Ronneberger et al., 2015). For our evaluation, we implemented a tile size of 256 × 256 and a filter count for the innermost U-Net block of 32 proved effective. In the case of a downsampling block, the blocks consist of convolution layers and a leaky ReLu activation function with a slope $a = 0.2$ and an instance normalization layer described by Ulyanov et al. (2016). In the case of an upsampling block, it consists of a transpose convolution layer, a ReLu activation function and an instance normalization layer. The two discriminators are PatchGANs proposed by Isola et al. (2017), their loss can be interpreted as a kind of style loss. In addition, spectral normalization introduced by Miyato et al. (2018) was used as a normalization layer to stabilize the training. The respective discriminator consists of three blocks, each consisting of a convolution layer, a leaky ReLu activation function and a normalization layer. The filter numbers increase quadratically with the depth of the discriminator. The least squares GAN (LSGAN) loss is used as loss, which allows a better image quality for generated images (Mao et al., 2017). To further stabilize the training, and reduce oscillations (Goodfellow, 2016), we implemented an image buffer as proposed by Shrivastava et al. (2017), which includes a history of generated images, of size 50. Since in

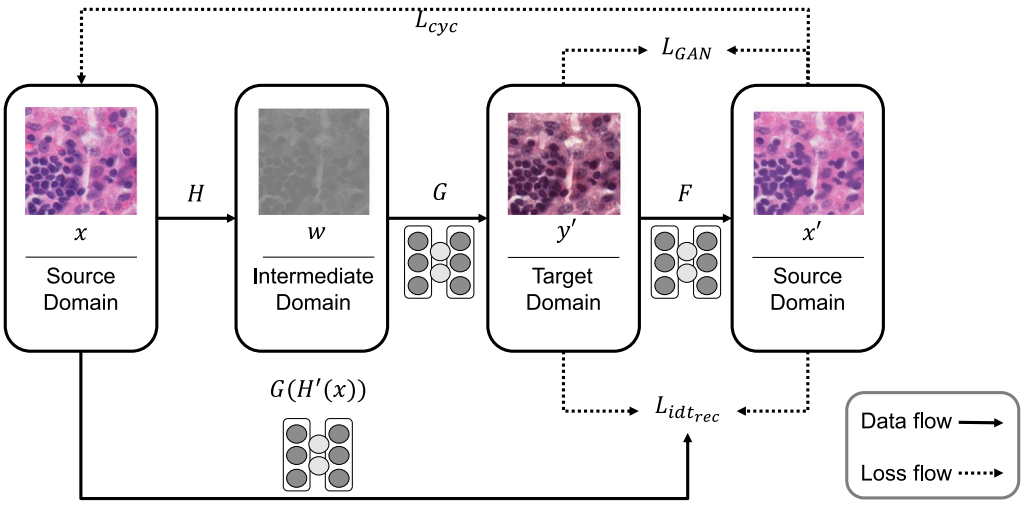

Fig. 5. Overview of the MultiStain-CycleGAN. Images x from a source domain will be mapped to an intermediate domain by a function H. H consists of a color augmentation function and a grayscale conversion. The generator G then transforms the gray image w into the target domain. This process, including projecting y' into the intermediate domain, is repeated for the normalized image y' again, to reconstruct the original image. The second path has been omitted for clarity. Further, instead of feeding the network a real image from the respective domain for calculating the identity loss used by Zhu et al. a reconstruction task of unaugmented gray images $H'(x)$ is done. The intermediate domain allows to normalize any H&E stains without having to re-train the model.

our experiments we noticed a tendency of the discriminator loss to converge to zero, we introduced an update threshold to avoid this, which prevents a gradient update as soon as one of the discriminator losses falls below a threshold value. The threshold was set to 0.1. We choose $\lambda_{cyc} = 10$ and $\lambda_{idt} = 0.5$ as in the original implementation of Zhu et al. For all the architecture related hyperparameters used except the update threshold, we followed the work of Zhu et al. (2017) and Runz et al. (2021). The intensity of the color augmentation used during training was determined empirically on the basis of the augmented images so that the structures were still present.

For our training, we used the ADAM optimizer (Kingma and Ba, 2014) with a learning rate of 10^{-5} with a linear decay over 60 Epochs. We trained the MultiStain-CycleGAN for 120 epochs in total on a Nvidia V100. For the color augmentation we used the following factors: Saturation 0.75, brightness 0.75, contrast 0.5. Since we subject our images to gray conversion, changing the hue value has no effect on the output image. These parameters were determined empirically by visual inspection of test images. They were chosen in such a way that the morphology of the structures is largely preserved.

4.2. Domain classifier

To verify that the respective normalization methods are able to disguise the tissue source site, we use an Xception classifier proposed by Chollet (2017) pre-trained on ImageNet to follow (Howard et al., 2021). Since we use tiles of size 256 × 256 as input for the normalization methods as described in chapter 2, they were brought to the size 224 × 224 using a center crop in order to use the pretrained Xception model. We applied a 5-fold cross-validation to obtain an approximate distribution of accuracy and exclude outliers. Each model was trained with a learning rate of 10^{-4} and a batch size of 32 for 50 epochs. Due to very large performance differences of the discriminator with different levels of color augmentations in training, we decided to use a high intensity of color augmentation. Due to the greater variance in stainings between different clinics than morphological differences as described in Tellez et al. (2019), we follow the color augmentation strategy described in their work. With a wide range of possible augmentations, we cover a large variety of possible stainings. We used color jitter with the parameters brightness: 0.7, contrast: 0.7, saturation: 0.7, hue: 0.5.

Table 1
Normalization paths that are evaluated for the respective data set. For each normalization path, each domain acts as both source and target domain, as the normalization is evaluated in both directions.

Camelyon17	SCC
CWZ ↔ RST	NZ20 ↔ P1000
UMCU ↔ RUMC	NZ210 ↔ NZ20
CWZ ↔ LPON	GT450 ↔ CS2

4.3. Tumor classifier

For tumor classification, we utilized a ResNet18 pre-trained on ImageNet. Analogous to the domain classifier, we trained the classifier with a learning rate of 10^{-4} and a batch size of 32 for 50 epochs. Also with this classifier we could see very large performance differences depending on the intensity of the color augmentation that was used. Thus, we employed the same parameters as described for the domain classifier in 4.2. During training, tumor tiles were drawn more frequently due to the weighting of the sampling process using the frequency of the classes. Analogous to the domain classifier, a center crop was also used for tumor classification in order to achieve the required input dimension of the pretrained ResNet18. All in all 5 models were trained per data set, each in the style of 5-fold cross-validation. Furthermore, different color augmentation intensities were used per model in the training. This results in a total of 100 trained tumor classifiers per data set.

5. Experimental setup

5.1. Normalization

For our experiments, we decided on three different normalization paths per dataset in order to evaluate the investigated methods with regard to their normalization capability. The three paths were chosen so that each domain occurs at least once. The normalization paths for the Camelyon17 and the SCC dataset are shown in Table 1. We trained our MultiStain-CycleGAN to learn the transformation $G : \mathcal{X} \rightarrow \mathcal{Y}$ and $F : \mathcal{Y} \rightarrow \mathcal{X}$ for every normalization path. Due to the huge memory requirements of WSIs, we were forced to perform the evaluation in a tile-by-tile manner. Thus, in order to analyze all the criteria under consideration, for every normalization path the tiles have to be normalized

using the normalization methods under investigation and the respective model. When normalizing after training, no data augmentation is used, the images are only converted to gray scale. To place our method in the context of existing literature, we compare it with other common normalization methods, which can be used in a multi-domain manner. We chose the methods of Macenko et al. (2009), Reinhard et al. (2001) and Vahadane et al. (2016) because of their frequent use in stain normalization in histopathology as well as state-of-the-art GAN-based methods. For the template-based methods, we selected three representative templates per domain shown in Fig. C.14 and Fig. C.15 for both datasets. This leads to three normalized subsets for each of the template-based approaches per domain per dataset. For the Macenko normalization and the Reinhard and Vahadane method, we utilized the implementations from torchstain[2] and staintools[3] respectively. In order to benchmark our method against state-of-the-art GAN-based methods, we compare our method against the Pix2Pix-based approach by Salehi and Chalechale (2020) and the default CycleGAN used for stain normalization by Runz et al. (2021). For the Pix2Pix-based approach, one model was trained per domain and dataset in order to be able to normalize into the respective domain. For the CycleGAN-based approach, one model was trained per normalization path. In total, all metrics were calculated on over 100 normalized subsets.

5.2. Image quality and domain shift metrics

For the evaluation of the image quality and the measurement of the domain shift of our normalization method, we chose the Structural Similarity (SSIM) Index (Wang et al., 2004) and the Fréchet Inception Distance (FID) (Heusel et al., 2017), respectively. These metrics are intended to help evaluate the quality of the generated images and quantify the change in domain shift.

5.2.1. Fréchet inception distance

In order to be able to make a statement about the domain shift before and after normalization and to evaluate different stain normalization methods with respect to their ability to reduce the domain gap, we decided to use the FID described by Heuse et al. The FID is an improvement over the Inception Score proposed by Salimans et al. (2016) in terms of consistency with human perception as the disturbance of the image increases. We chose this metric because of its frequent application in generative tasks and the extensive evaluation of the method (Xu et al., 2018; Lucic et al., 2017). The FID represents the difference between two Gaussians consisting of the features of an inception model. The FID is given by:

$$\mathrm{FID}\left(\mu_X, \mu_Y, C_X, C_Y\right) = \|\mu_X - \mu_Y\| + Tr\left(C_X + C_Y - 2\sqrt{C_X C_Y}\right). \tag{8}$$

Where the mean and covariance μ_X, C_X corresponds to the Gaussian of the generated data and μ_Y, C_Y corresponds to the Gaussian of the real world data. The FID is zero in case of matching images.

Due to the issue described by Liu et al. (2018) that using an ImageNet model to project generated images of domains unrelated to ImageNet into feature space can be ineffective. Following the suggestion of Liu et al. we used the model of Ciga et al. (2022) as a domain-specific encoder, which was trained on several histopathology datasets. The FID correlates with human visual perception and measures of how large the perceived difference is between images from two different distributions. In the stain normalization use case, a high FID means that there is a large domain shift. This domain shift, and thus the FID, should be reduced by stain normalization methods. For calculating the FID we use an open-source implementation[4] and have modified it as described above to use a model trained on histopathological data. We calculated the FID for the normalized testsets by computing the distance between 40 000 random tiles from the respective domain and the normalized tiles of a given normalization path.

[2] https://github.com/EIDOSLAB/torchstain
[3] https://github.com/Peter554/StainTools
[4] https://github.com/mseitzer/pytorch-fid

5.2.2. Structural similarity index

To compare the perceived structure before and after normalization, we utilize the Structural Similarity Index proposed by Wang et al. for our evaluation. The SSIM index compares two images in terms of their similarity and image quality. The SSIM index is 1 in case of two identical images. In contrast to the use of the Mean Squared Error (MSE) for the comparison of images, the SSIM index calculates contrast, structure and luminance separately and then combines them. We chose this metric because it has already been used in stain normalization scenarios (Hoque et al., 2021; Shaban et al., 2019) and due to the high importance of keeping structural features unaffected by normalization. Preserving the structure after normalization is essential, otherwise the morphology of the cells is altered. This can lead to errors in classification, as will be shown in Section 6. The SSIM is given by:

$$\mathrm{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}. \tag{9}$$

Where μ_x and σ_x^2 correspond to the mean and variance of the image x, respectively. σ_{xy} corresponds to the covariance of the images x and y, and C_1 and C_2 are constants to stabilize the denominator. For the calculation of the SSIM index in our experiments, we use the implementation of scikit-image (van der Walt et al., 2014). The SSIM index is calculated using image pairs compared to the FID. Here, for each of the normalized testset, the SSIM index is calculated using the normalized image and the non-normalized image. Thus, for each of the testsets, we estimate a mean value and the associated standard deviation.

5.3. Network-based metrics

Image-based similarity metrics such as FID and SSIM can give clues about the visual similarity of the normalized pictures. However, they do not evaluate how well downstream classifiers perform on the normalized images. As the goal of stain normalization is to at least maintain task performance while ideally obscuring the origin of the slide, we additionally evaluate our approach with network-based metrics. As described in 2, the testsets have a balanced class ratio, which is why we chose accuracy as the metric for both classifiers, as we are interested in performance without considering the cost of misclassification.

6. Results

A summary of the results is given in Table 2. This includes the tumor and domain classifier accuracy, the mean SSIM index and the FID. For the results shown in the table, the metrics for MultiStain-CycleGAN and the standard CycleGAN were averaged over the six normalization paths. For the Pix2Pix-based method and the template-based methods, the metrics of each of the five normalized domains were averaged. Additionally, for the template-based methods, metrics were averaged over the three templates. The results of the classifiers are from the five models of the 5-fold cross-validation, for which the color augmentation intensity “heavy” was used in the training. The performance was also averaged over the five classifiers. An overview of the normalization results of the different methods studied is provided in Appendix A. This shows the original images and the normalized variant in each case for both datasets. Here, the normalization paths studied are RST→CWZ for the Camelyon17 dataset and GT450→CS2 for the SCC dataset. In the following, the results of domain classification will be discussed in more detail.

Table 2
Comparison of different normalization techniques on various metrics for the Camelyon17 and SCC datasets. The table includes the estimated mean domain and tumor classifier accuracy, SSIM index and FID and the standard deviations of the respective metrics. In each case, the metrics were averaged over all normalization paths. The performance of the template-based methods is averaged over the three templates used. The best performance for each metric is highlighted in bold for each dataset. MultiStain-CycleGAN improves tumor classifier accuracy, provides very good image quality and at the same time is able to effectively fool the domain classifier.

	Domain Classifier Accuracy ↓	Tumor Classifier Accuracy ↑	SSIM Index ↑	FID ↓
Camelyon17 Dataset				
Unnormalized	$0.998_{\pm 0.000}$	$0.881_{\pm 0.033}$	1	–
CycleGAN	$0.790_{\pm 0.042}$	$0.870_{\pm 0.034}$	$0.935_{\pm 0.033}$	**26.501**
Pix2Pix	$0.849_{\pm 0.052}$	$0.828_{\pm 0.048}$	$0.917_{\pm 0.064}$	56.592
Macenko	$0.859_{\pm 0.047}$	$0.822_{\pm 0.072}$	$0.858_{\pm 0.023}$	55.371
Reinhard	$0.944_{\pm 0.041}$	$0.839_{\pm 0.047}$	$0.853_{\pm 0.019}$	49.881
Vahadane	$0.869_{\pm 0.031}$	$0.794_{\pm 0.080}$	$0.847_{\pm 0.034}$	53.501
MultiStain-CycleGAN	$\mathbf{0.672_{\pm 0.037}}$	$\mathbf{0.897_{\pm 0.014}}$	$\mathbf{0.947_{\pm 0.006}}$	44.964
SCC Dataset				
Unnormalized	$0.998_{\pm 0.000}$	$0.870_{\pm 0.029}$	1	–
CycleGAN	$0.820_{\pm 0.057}$	$0.864_{\pm 0.024}$	$0.952_{\pm 0.024}$	**17.611**
Pix2Pix	$0.912_{\pm 0.036}$	$0.836_{\pm 0.044}$	$\mathbf{0.957_{\pm 0.007}}$	18.365
Macenko	$0.839_{\pm 0.068}$	$0.816_{\pm 0.043}$	$0.804_{\pm 0.037}$	28.928
Reinhard	$0.912_{\pm 0.037}$	$0.842_{\pm 0.042}$	$0.842_{\pm 0.036}$	23.358
Vahadane	$0.844_{\pm 0.066}$	$0.815_{\pm 0.056}$	$0.835_{\pm 0.034}$	24.841
MultiStain-CycleGAN	$\mathbf{0.609_{\pm 0.152}}$	$\mathbf{0.874_{\pm 0.018}}$	$0.942_{\pm 0.015}$	19.240

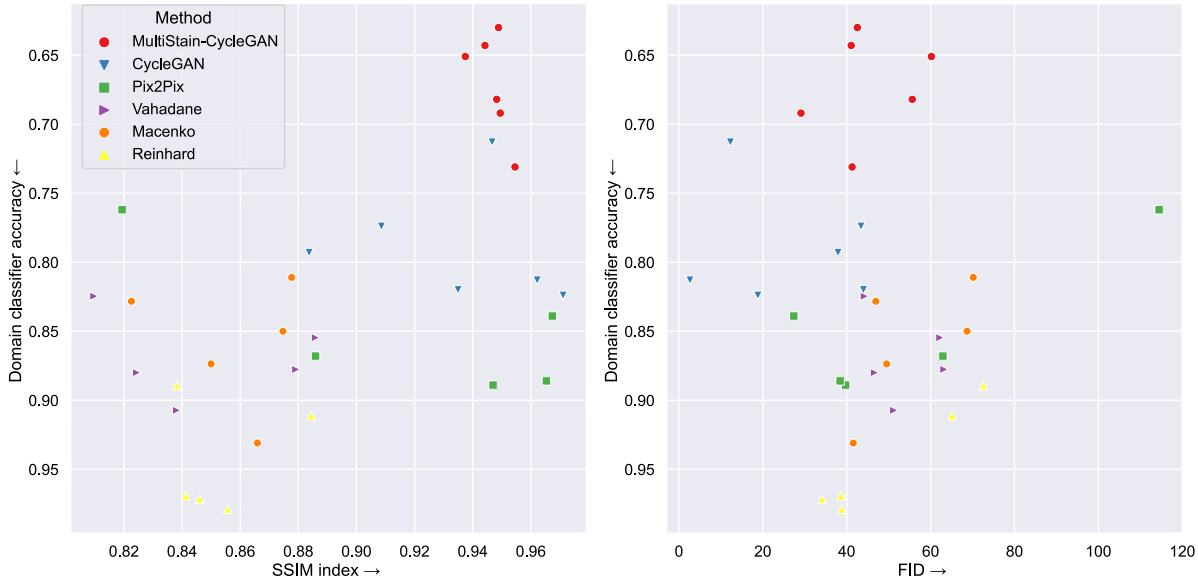

Fig. 6. The behavior of the domain classifier under different normalization methods over the two metrics SSIM index and FID for the Camelyon17 dataset. All normalization paths used are shown for each method. MultiStain-CycleGAN yields images with the highest SSIM index on average and can fool the domain classifier well. The accuracy of the domain classifier in this case hardly depends on the two metrics for all investigated methods.

6.1. Domain classification

We examined the results of the tissue site classifier, or domain classifier, for the different normalization methods, as well as for the unnormalized case. A detailed overview of the different classification results of the domain classifiers are given the appendix in tables D.5 and. D.6.

The domain could be estimated with very high accuracy of over 0.99 in the unnormalized case for both datasets while using a classifier trained with color augmentation. The Reinhard and Pix2Pix normalization still both show a high accuracy of over 0.9. The lowest accuracy was achieved by MultiStain-CycleGAN with an accuracy of 0.672, followed by the default CycleGAN with an accuracy of 0.79. By using the percentile method, it can be seen that the domain classifier accuracy is significantly lower for our method compared to the other methods tested for both datasets. Fig. 6 shows the accuracy of the domain classifier over the two metrics SSIM index and the FID for the Camelyon17 dataset. For the SCC data set, an analogous plot can be found in Fig. B.12. In the left plot in Fig. 6, as well as in Table 2, it is shown that our proposed method achieves the highest SSIM index on average, with a value of 0.947 for the Camelyon17 dataset. From this plot, no clear dependence of domain classifier accuracy on SSIM index can be derived for our case.

In the right plot in Fig. 6, the accuracy of the domain classifier over the FID is shown. Again, there is no clear dependence of the domain classifier accuracy on the FID. The two CycleGAN based methods exhibit the highest ability to fool the domain classifier. Our method is placed in the middle FID range. The behavior of our method with respect to domain classifier accuracy behaves consistently across the two datasets examined as shown in Fig. B.12.

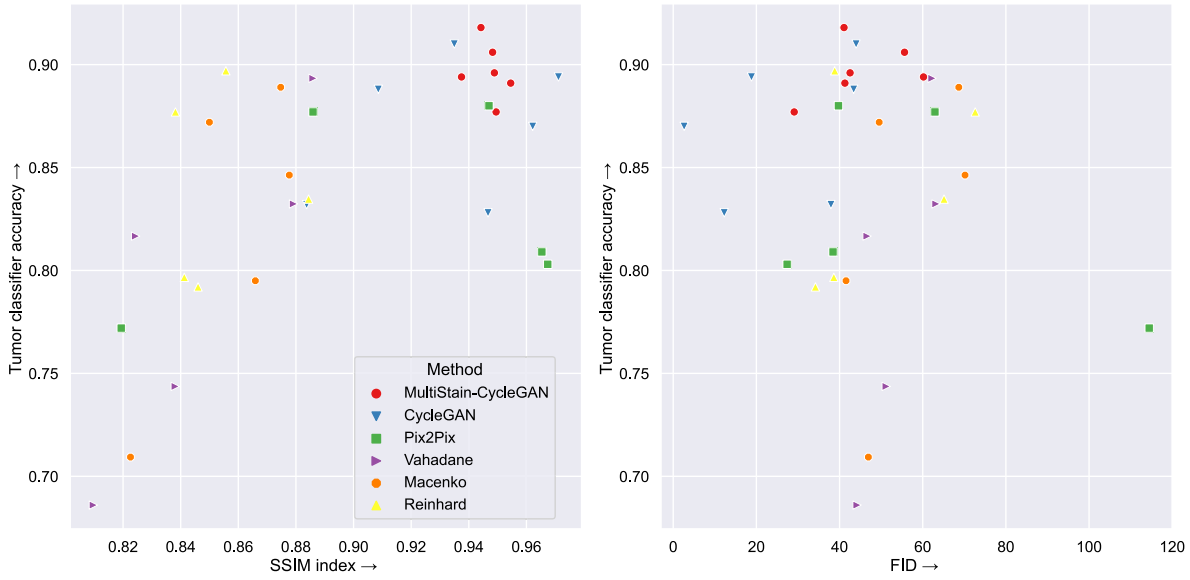

Fig. 7. The dependence of tumor classifier accuracy on SSIM index on the left and FID on the right for the Camelyon17 dataset. All normalization paths used are shown for each method. It can be seen that as the SSIM index decreases, the performance of the tumor classifier tends to decrease as well. This may be due to the loss of contrast and associated structural changes and loss of information in the images. In this case, tumor classifier accuracy shows little dependence on FID.

6.2. Tumor classification

The evaluation of the tumor classifier was performed analogously to the domain classifier. Again, a detailed listing of the results is given in the appendix in Table D.3 and D.4. Tumor classifier accuracy decreased for most methods by using normalization in our experiments compared to the unnormalized case under the condition that moderate to heavy color augmentation was used in the training of tumor classifiers. In the case of no or light color augmentation, the classification in our case can benefit from multiple normalization methods. The advantages of our model show up particularly strongly when little color augmentation is used, but was able to slightly increase performance for one of the two datasets even with heavy color augmentation. This behavior can be observed for both datasets under investigation. Utilizing the percentile method, similar to the domain classifier, reveals that our method does not significantly increase the accuracy of the tumor classifier.

Fig. 7 on the left shows a relatively consistent dependence of tumor classifier accuracy on the SSIM index for the Camelyon17 dataset. Possibly, a better classification can occur as a result of the higher image quality or better preservation of the structures. Furthermore, several methods decrease the accuracy moderately to strongly compared to the unnormalized case. This case can be observed especially when the SSIM index decreases strongly. On the right side of Fig. 7, tumor classifier accuracy over FID is visualized. In our case, there is also no clear behavior regarding tumor classifier accuracy and FID.

7. Discussion

Our evaluation shows that with a drop in the SSIM index, the performance of the downstream task decreases, but the accuracy of the two classifiers show little dependence on the FID. In general, we find that GAN-based methods are superior to template-based methods in our experiments. As shown in Fig. 7, slight saturation effects occur for the SSIM index. It can be seen that the performance of the tumor classification is only significantly reduced when the SSIM index falls below a threshold. Such a trend is not apparent to FID.

We expected a performance drop in the tumor classifier in the event of a falling SSIM index or rising FID. Both of these are indicators of how similar the normalized images are to the unnormalized ones. A lower SSIM index may suggest that the contrast has been changed too dramatically, which in turn affects the appearance of tissue morphology. This is equivalent to severe information loss, which should reduce the classification performance of the tumor classifier. We expected similar trends with an increasing FID. However, these trends are not evident in our results.

Furthermore, there is a consistently high variance in the tumor classifier accuracy for the template-based methods for both datasets examined. This may be due to the strong dependence on the choice of template on the quality of normalization. This behavior is again made clearer by the results of the template-based normalizations in Fig. A.9 and Fig. A.10. Here, by alternating the templates per image, it can be seen clearly how big the influence of the choice of the template is on the normalization quality. Thus, the choice of an inappropriate template can lead to performance losses. GAN-based approaches have an advantage here, since they take the entire training dataset into account and can thus lead to more consistent solutions.

We expected that as the FID increases, the accuracy of the domain classifier will decrease as the images are further away from the source domain and may have a completely altered distribution as some of the images normalized by a template-based method. Due to the loss of essential information, the domain can no longer be estimated by the domain classifier. However, this expectation is not indicated by our results. Thus, in our case, we see the FID only conditionally suitable to quantify domain shifts for histopathological data.

Fig. 8 shows the domain classifier accuracy over the tumor classifier accuracy for the Camelyon17 dataset. This shows a very interesting behavior, namely that methods which have a higher tumor classifier accuracy in our experiments tend to be better able to fool the domain classifier. This is possibly due to very good normalization quality, as the images are so close to the target domain that both the domain classifier has a hard time determining the origin and the images closely resemble the training data of the tumor classifier, thus achieving better classification on domain shifted data. The plot very clearly shows the superiority of MultiStain-CycleGAN, which performs best in both metrics. Furthermore, the metrics show that our method is able to normalize stainings that are not contained within the training set and thus can normalize different domains with one model without the need for retraining. This is supported by the high tumor classifier accuracy despite distribution shifts through the different domains showing how good the ability of our method is to normalize unseen domains in training. A disadvantage of CycleGAN based methods is that data of

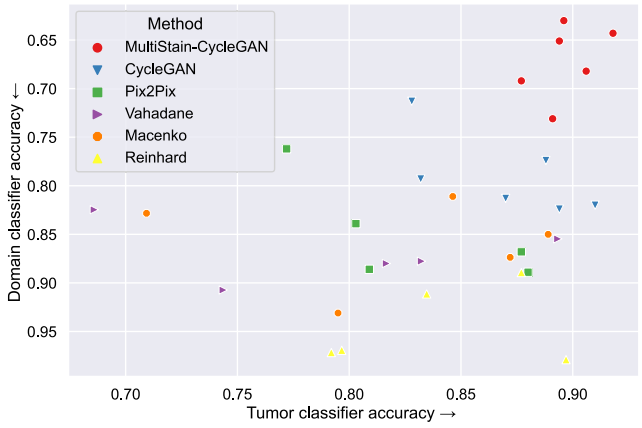

Fig. 8. The accuracy of the domain classifier over the tumor classifier accuracy of the different methods for the Camelyon17 dataset. MultiStain-CycleGAN is the only method capable of both fooling the domain classifier and slightly improving the accuracy of the tumor classifier. The problem of template dependence of template-based methods is well demonstrated by the high variance of tumor classifier accuracy for these methods.

two domains are required in contrast to Pix2Pix or template based approaches.

During training, we did not observe any cases of mode collapse or other instabilities in the two CycleGAN based methods, but a few Pix2Pix based models suffered from mode collapse. This also explains why the variance of the metrics here is very large compared to the other two GAN-based methods examined.

The results in Table 2, but especially the resulting images in Fig. A.9 and Fig. A.10, clearly show that CycleGAN is not capable of satisfactorily normalizing images from unlearned domains. Traces of the source domain can still be clearly recognized in larger domain shifts, which is not the case with our method. We could not observe any significant improvement of the downstream task, in our case tumor classification, in the case of heavy data augmentation. This is partly consistent with the findings of Tellez et al. (2019). Based on the tables D.5 and D.6, it can be seen that it is essential to investigate stain normalization methods considering color augmentation. The results show that when using no to moderate color augmentation, our normalization method can achieve significant improvements, which however become smaller when using strong color augmentation. Thus, the domain classifier can be fooled much less with heavy color augmentation applied in training and the accuracy increases by a huge margin compared to the use of no color augmentation. A potential drawback of our method with respect to the domain classifier is that tiles from the same patient can end up in both the training and test sets. Despite this limitation, it turns out that some of the normalization methods have a very good ability to disguise the origin of the tiles. A similar picture emerges for tumor classification, where our normalization provides an accuracy gain of over 10% in the case of no color augmentation for the Camelyon17 dataset. In the case of medium color augmentation, we still achieve a slightly increased performance, with only a small improvement present in the case of heavy color augmentation.

In the end, we can say that we can partially reproduce the results of Howard et al. (2021) and extend them with a GAN-based approach. Also, our results show that despite the use of commonly used stain normalization methods, the domain can be predicted with high accuracy. Our method is the exception here, as it results in a significant loss of domain classifier accuracy while maintaining high image quality. Due to a good image quality and images which are difficult to assign to a clinic by a classifier, our method allows to contribute to the data protection of patients by pseudonymization on the image level. Pseudonymization at the image level makes it difficult to draw conclusions about the origin of an image and thus about the characteristics of a patient. One possible limitation of our evaluation of image quality is that we only determine the quality of normalization based on comparative metrics, but do not involve human experts. Human experts can be helpful to ensure that no essential structures are lost or hallucinated by the transformation.

8. Conclusion

We have presented MultiStain-CycleGAN, a new approach to stain normalization based on a modified CycleGAN, using an intermediate domain which works for multiple unseen stainings without the need to retrain and is thus multi-domain capable. We have extensively compared our approach with several commonly used normalization methods. We have intensively analyzed the different methods using different metrics and augmentation levels to understand the behavior of the methods under investigation. It has been shown that our method does not suffer from the problems of template-based approaches, and at the same time, in contrast to conventional GAN-based approaches, training a single model is sufficient. Furthermore, our method is best able to fool the domain classifier while providing a very high image quality and high performance in the downstream task. In doing so, this work is a step towards disguising the origin of the tissue sections and reducing bias through the different domains. Continuing this work, other stainings such as IHC can be investigated. Furthermore, more complex downstream tasks should be analyzed, since in the case of our task none of the normalization methods led to a significant improvement in performance while using color augmentation for the downstream classifiers.

CRediT authorship contribution statement

Martin J. Hetz: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Tabea-Clara Bucher:** Project administration, Supervision, Writing – review & editing. **Titus J. Brinker:** Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

A minimal working example of the normalization and code for training the model are available on Github https://github.com/DBO-DKFZ/multistain_cyclegan_normalization.

Acknowledgments

The research is funded by *the Ministry of Social Affairs, Health and Integration* of the Federal State Baden-Württemberg, Germany (grant: AI-Translation-Initiative (“KI-Translations-Initiative”), grant-holder: Titus J. Brinker, German Cancer Research Center, Heidelberg, Germany)

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103149.

References

Albarqouni, S., Baur, C., Achilles, F., Belagiannis, V., Demirci, S., Navab, N., 2016. AggNet: Deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans. Med. Imaging 35 (5), 1313–1321. http://dx.doi.org/10.1109/TMI.2016.2528120.

Alturkistani, H.A., Tashkandi, F.M., Mohammedsaleh, Z.M., 2015. Histological stains: A literature review and case study. Glob. J. Health Sci. 8 (3), 72–79. http://dx.doi.org/10.5539/gjhs.v8n3p72.

Bancroft, J.D., 2008. Theory and Practice of Histological Techniques. Elsevier Health Sciences.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Ehteshami Bejnordi, B., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Cetin, M., Halici, E., Jackson, H., Chen, R., Both, F., Franke, J., Kusters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., Litjens, G., 2019. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. IEEE Trans. Med. Imaging 38 (2), 550–560. http://dx.doi.org/10.1109/TMI.2018.2867350.

Bentaieb, A., Hamarneh, G., 2018. Adversarial stain transfer for histopathology image analysis. IEEE Trans. Med. Imaging 37 (3), 792–802. http://dx.doi.org/10.1109/TMI.2017.2781228.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., de Kaa, C.H.-v., Litjens, G., 2020. Automated deep-learning system for Gleason grading of prostate cancer using biopsies: A diagnostic study. Lancet Oncol. 21 (2), 233–241. http://dx.doi.org/10.1016/s1470-2045(19)30739-9.

Chatterjee, S., 2014. Artefacts in histopathology. J. Oral Maxillofac. Pathol. 18 (Suppl 1), S111–6. http://dx.doi.org/10.4103/0973-029X.141346.

Chollet, 2017. Xception: Deep learning with depthwise separable convolutions. In: Proc. IEEE Conf. Decis. Control.

Ciga, O., Xu, T., Martel, A.L., 2022. Self supervised contrastive learning for digital histopathology. Mach. Learn. Appl. 7, 100198. http://dx.doi.org/10.1016/j.mlwa.2021.100198.

Ciompi, F., Geessink, O., Bejnordi, B.E., de Souza, G.S., Baidoshvili, A., Litjens, G., van Ginneken, B., Nagtegaal, I., van der Laak, J., 2017. The importance of stain normalization in colorectal tissue classification with convolutional networks. In: 2017 IEEE 14th International Symposium on Biomedical Imaging. ISBI 2017, pp. 160–163. http://dx.doi.org/10.1109/ISBI.2017.7950492.

Cong, C., Liu, S., Di Ieva, A., Pagnucco, M., Berkovsky, S., Song, Y., 2022. Colour adaptive generative networks for stain normalisation of histopathology images. Med. Image Anal. 82, 102580. http://dx.doi.org/10.1016/j.media.2022.102580.

Cruz-Roa, A., Gilmore, H., Basavanhally, A., Feldman, M., Ganesan, S., Shih, N.N.C., Tomaszewski, J., González, F.A., Madabhushi, A., 2017. Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent. Sci. Rep. 7, 46450. http://dx.doi.org/10.1038/srep46450.

de Bel, T., Bokhorst, J.-M., van der Laak, J., Litjens, G., 2021. Residual cyclegan for robust domain transformation of histopathological tissue slides. Med. Image Anal. 70, 102004. http://dx.doi.org/10.1016/j.media.2021.102004.

Echle, A., Rindtorff, N.T., Brinker, T.J., Luedde, T., Pearson, A.T., Kather, J.N., 2021. Deep learning in cancer pathology: A new generation of clinical biomarkers. Br. J. Cancer 124 (4), 686–696. http://dx.doi.org/10.1038/s41416-020-01122-x.

Farahani, N., Parwani, A.V., Pantanowitz, L., et al., 2015. Whole slide imaging in pathology: Advantages, limitations, and emerging perspectives. Pathol. Lab. Med. Int. 7 (23–33), 4321.

Gatys, L.A., Ecker, A.S., Bethge, M., 2016. Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, http://dx.doi.org/10.1109/cvpr.2016.265.

Ghaznavi, F., Evans, A., Madabhushi, A., Feldman, M., 2013. Digital imaging in pathology: whole-slide imaging and beyond. Annu. Rev. Pathol. 8, 331–359. http://dx.doi.org/10.1146/annurev-pathol-011811-120902.

Goodfellow, I., 2016. NIPS 2016 tutorial: Generative adversarial networks. arXiv:1701.00160.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Gurcan, M.N., Boucheron, L.E., Can, A., Madabhushi, A., Rajpoot, N.M., Yener, B., 2009. Histopathological image analysis: A review. IEEE Rev. Biomed. Eng. 2, 147–171. http://dx.doi.org/10.1109/RBME.2009.2034865.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S., 2017. GANs trained by a two Time-Scale update rule converge to a local Nash equilibrium. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), In: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc..

Hoque, M.Z., Keskinarkaus, A., Nyberg, P., Seppänen, T., 2021. Retinex model based stain normalization technique for whole slide image analysis. Comput. Med. Imaging Graph. 90, 101901. http://dx.doi.org/10.1016/j.compmedimag.2021.101901.

Howard, F.M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O.I., Kather, J.N., Cipriani, N., Grossman, R.L., Pearson, A.T., 2021. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nature Commun. 12 (1), 4423. http://dx.doi.org/10.1038/s41467-021-24698-1.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1125–1134.

Khened, M., Kori, A., Rajkumar, H., Krishnamurthi, G., Srinivasan, B., 2021. A generalized deep learning framework for whole-slide image segmentation and analysis. Sci. Rep. 11 (1), 11579. http://dx.doi.org/10.1038/s41598-021-90444-8.

Kiehl, L., Kuntz, S., Höhn, J., Jutzi, T., Krieghoff-Henning, E., Kather, J.N., Holland-Letz, T., Kopp-Schneider, A., Chang-Claude, J., Brobeil, A., von Kalle, C., Fröhling, S., Alwers, E., Brenner, H., Hoffmeister, M., Brinker, T.J., 2021. Deep learning can predict lymph node status directly from histology in colorectal cancer. Eur. J. Cancer 157, 464–473. http://dx.doi.org/10.1016/j.ejca.2021.08.039.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv:1412.6980.

Kothari, S., Phan, J.H., Stokes, T.H., Osunkoya, A.O., Young, A.N., Wang, M.D., 2014. Removing batch effects from histopathological images for enhanced cancer diagnosis. IEEE J. Biomed. Health Inform. 18 (3), 765–772. http://dx.doi.org/10.1109/JBHI.2013.2276766.

Kurakin, A., Goodfellow, I., Bengio, S., 2016. Adversarial examples in the physical world. http://dx.doi.org/10.1201/9781351251389-8/adversarial-examples-physical-world-alexey-kurakin-ian-goodfellow-samy-bengio, arXiv:1607.02533.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. http://dx.doi.org/10.1038/nature14539.

Liu, S., Wei, Y., Lu, J., Zhou, J., 2018. An improved evaluation framework for generative adversarial networks. arXiv:1803.07474.

Lucic, Kurach, Michalski, et al., 2017. Are GANs created equal? A large-scale study. Adv. Neural Inf. Process. Syst..

Lyon, H.O., De Leenheer, A.P., Horobin, R.W., Lambert, W.E., Schulte, E.K., Van Liedekerke, B., Wittekind, D.H., 1994. Standardization of reagents and methods used in cytological and histological practice with emphasis on dyes, stains and chromogenic reagents. Histochem. J. 26 (7), 533–544. http://dx.doi.org/10.1007/BF00158587.

Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E., 2009. A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: from Nano to Macro. pp. 1107–1110. http://dx.doi.org/10.1109/ISBI.2009.5193250.

Mao, X., Li, Q., Xie, H., Lau, R.Y.K., Wang, Z., Smolley, S.P., 2017. Least squares generative adversarial networks. ICCV, http://dx.doi.org/10.1109/iccv.2017.304.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426.

Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations.

Niethammer, M., Borland, D., Marron, J.S., Woosley, J., Thomas, N.E., 2010. Appearance normalization of histology slides. Mach. Learn. Med. Imaging 58–66. http://dx.doi.org/10.1007/978-3-642-15948-0_8.

Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P., 2001. Color transfer between images. IEEE Comput. Graph. Appl. 21 (5), 34–41. http://dx.doi.org/10.1109/38.946629.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. MICCAI 2015, Springer International Publishing, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Roy, S., Kumar Jain, A., Lal, S., Kini, J., 2018. A study about color normalization methods for histopathology images. Micron 114, 42–61. http://dx.doi.org/10.1016/j.micron.2018.07.005.

Ruifrok, A.C., Johnston, D.A., 2001. Quantification of histochemical staining by color deconvolution. Anal. Quant. Cytol. Histol. 23 (4), 291–299.

Runz, M., Rusche, D., Schmidt, S., Weihrauch, M.R., Hesser, J., Weis, C.-A., 2021. Normalization of HE-stained histological images using cycle consistent generative adversarial networks. Diagnostic Pathol. 16 (1), http://dx.doi.org/10.1186/s13000-021-01126-y.

Saha, M., Chakraborty, C., Racoceanu, D., 2018. Efficient deep learning model for mitosis detection using breast histopathology images. Comput. Med. Imaging Graph. 64, 29–40. http://dx.doi.org/10.1016/j.compmedimag.2017.12.001.

Salehi, P., Chalechale, A., 2020. Pix2Pix-based Stain-to-Stain translation: A solution for robust stain normalization in histopathology images analysis. In: 2020 International Conference on Machine Vision and Image Processing. MVIP, pp. 1–7. http://dx.doi.org/10.1109/MVIP49855.2020.9116895.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training GANs. Adv. Neural Inf. Process. Syst. 29.

Salvi, M., Acharya, U.R., Molinari, F., Meiburger, K.M., 2021. 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, 104129. http://dx.doi.org/10.1016/j.compbiomed.2020.104129.

Shaban, M.T., Baur, C., Navab, N., Albarqouni, S., 2019. Staingan: Stain style transfer for digital histological images. In: 2019 IEEE 16th International Symposium on Biomedical Imaging. ISBI 2019, pp. 953–956. http://dx.doi.org/10.1109/ISBI.2019.8759152.

Shrivastava, Pfister, Tuzel, et al., 2017. Learning from simulated and unsupervised images through adversarial training. In: Proc. Estonian Acad. Sci. Biol. Ecol..

Stacke, K., Eilertsen, G., Unger, J., Lundstrom, C., 2021. Measuring domain shift for deep learning in histopathology. IEEE J. Biomed. Health Inform. 25 (2), 325–336. http://dx.doi.org/10.1109/JBHI.2020.3032060.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., van der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. 58, 101544. http://dx.doi.org/10.1016/j.media.2019.101544.

Tosta, T.A.A., de Faria, P.R., Servato, J.P.S., Neves, L.A., Roberto, G.F., Martins, A.S., do Nascimento, M.Z., 2019. Unsupervised method for normalization of hematoxylin-eosin stain in histological images. Comput. Med. Imaging Graph. 77, 101646. http://dx.doi.org/10.1016/j.compmedimag.2019.101646.

Ulyanov, D., Vedaldi, A., Lempitsky, V., 2016. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022.

Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N., 2016. Structure-Preserving color normalization and sparse stain separation for histological images. IEEE Trans. Med. Imaging 35 (8), 1962–1971. http://dx.doi.org/10.1109/TMI.2016.2529665.

Wagner, S.J., Khalili, N., Sharma, R., Boxberg, M., Marr, C., de Back, W., Peng, T., 2021. Structure-Preserving multi-domain stain color augmentation using Style-Transfer with disentangled representations. In: Medical Image Computing and Computer Assisted Intervention. MICCAI 2021, Springer International Publishing, pp. 257–266. http://dx.doi.org/10.1007/978-3-030-87237-3_25.

van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T., scikit-image contributors, 2014. Scikit-image: Image processing in Python. PeerJ 2, e453. http://dx.doi.org/10.7717/peerj.453.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: From error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. http://dx.doi.org/10.1109/tip.2003.819861.

Wang, X., Chen, H., Gan, C., Lin, H., Dou, Q., Tsougenis, E., Huang, Q., Cai, M., Heng, P.-A., 2020. Weakly supervised deep learning for whole slide lung cancer image analysis. IEEE Trans. Cybern. 50 (9), 3950–3962. http://dx.doi.org/10.1109/TCYB.2019.2935141.

Wessels, F., Schmitt, M., Krieghoff-Henning, E., Kather, J.N., Nientiedt, M., Kriegmair, M.C., Worst, T.S., Neuberger, M., Steeg, M., Popovic, Z.V., Gaiser, T., von Kalle, C., Utikal, J.S., Fröhling, S., Michel, M.S., Nuhn, P., Brinker, T.J., 2022. Deep learning can predict survival directly from histology in clear cell renal cell carcinoma. PLoS One 17 (8), e0272656. http://dx.doi.org/10.1371/journal.pone.0272656.

Wilm, F., Fragoso, M., Bertram, C.A., Stathonikos, N., Öttl, M., Qiu, J., Klopfleisch, R., Maier, A., Breininger, K., Aubreville, M., 2023. Multi-scanner canine cutaneous squamous cell carcinoma histopathology dataset. In: Bildverarbeitung Für Die Medizin 2023. Springer Fachmedien Wiesbaden, pp. 206–211. http://dx.doi.org/10.1007/978-3-658-41657-7_46.

Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., Weinberger, K., 2018. An empirical study on evaluation metrics of generative adversarial networks. arXiv:1806.07755.

Zanjani, F.G., Zinger, S., Bejnordi, B.E., van der Laak, J.A.W.M., de With, P.H.N., 2018. Stain normalization of histopathology images using generative adversarial networks. ISBI 2018, http://dx.doi.org/10.1109/isbi.2018.8363641.

Zhou, N., Cai, D., Han, X., Yao, J., 2019. Enhanced Cycle-Consistent generative adversarial network for color normalization of H&E stained images. In: Medical Image Computing and Computer Assisted Intervention. MICCAI 2019, Springer International Publishing, pp. 694–702. http://dx.doi.org/10.1007/978-3-030-32239-7_77.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232.