Medical Image Analysis 70 (2021) 102004

Contents lists available at ScienceDirect

Medical Image Analysis

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

Challenge Report

Residual cyclegan for robust domain transformation of histopathological tissue slides

Thomas de Bel[a,b,*], John-Melle Bokhorst[a,b], Jeroen van der Laak[a,b,c], Geert Litjens[a,b]

[a] Department of Pathology, Radboud University Medical Center, Nijmegen, The Netherlands
[b] Radboud university medical center, Radboud Institute for Health Sciences, Nijmegen, The Netherlands
[c] Center for Medical Image Science and Visualization, Linkping University, Linkping, Sweden

ARTICLE INFO

Article history:
Received 15 July 2020
Revised 10 February 2021
Accepted 15 February 2021
Available online 18 February 2021

Keywords:
Histopathology
Adversarial networks
Stain normalization
Structure segmentation

ABSTRACT

Variation between stains in histopathology is commonplace across different medical centers. This can have a significant effect on the reliability of machine learning algorithms. In this paper, we propose to reduce performance variability by using -consistent generative adversarial (CycleGAN) networks to remove staining variation. We improve upon the regular CycleGAN by incorporating residual learning. We comprehensively evaluate the performance of our stain transformation method and compare its usefulness in addition to extensive data augmentation to enhance the robustness of tissue segmentation algorithms. Our steps are as follows: first, we train a model to perform segmentation on tissue slides from a single source center, while heavily applying augmentations to increase robustness to unseen data. Second, we evaluate and compare the segmentation performance on data from other centers, both with and without applying our CycleGAN stain transformation. We compare segmentation performances in a colon tissue segmentation and kidney tissue segmentation task, covering data from 6 different centers. We show that our transformation method improves the overall Dice coefficient by 9% over the non-normalized target data and by 4% over traditional stain transformation in our colon tissue segmentation task. For kidney segmentation, our residual CycleGAN increases performance by 10% over no transformation and around 2% compared to the non-residual CycleGAN.

1. Introduction

Central to histopathology is the fixation and staining of tissue slides, highlighting relevant structures. Diagnosis relies on careful examination of these tissue slides under the microscope by a pathologist. Differences in staining are common and can, for example, occur from tissue fixation and processing, staining protocols, and section thickness (Bancroft and Gamble, 2008). With increasing affordability of whole slide scanners and popularity of computer-aided diagnosis systems, more pathology labs are 'going digital' (Stathonikos et al., 2019). Digitization of tissue into whole slide images (WSIs) introduces new sources of variation, such as the age of the slide during scanning, and more directly due to the scanner: differences in digital post-processing (e.g. sharpening filters), scanning resolutions and storage formats (Weinstein et al., 2009). The net effect of all factors (tissue processing, staining and scanning) may result in drastically differing digital tissue slides across centers, even when using routine staining protocols (Fig. 3 & 4) and staining consecutive tissue sections Bejnordi et al. (2016).

Deep learning has recently gained a lot of traction in the medical imaging domain (Litjens et al., 2017), for a large part due to the use of fully convolutional neural networks (Long et al., 2015). Applications are seen in all fields of medical imaging and in different organs, such as breast (Bejnordi et al., 2017), kidney (Hermsen et al., 2019) and prostate (Bulten et al., 2020). Before deep learning algorithms can be introduced in the workflow of the pathologist, they need to achieve reliable performance and be able to deal with all types of dissimilarities, induced by scanners, protocols and labs. While pathologists are generally highly capable of dealing with variety in stainings, deep learning systems have been shown to be dramatically affected in performance (Ciompi et al., 2017). The performance and robustness of a network is difficult to

* Corresponding author.
E-mail addresses: thomas.debel@radboudumc.nl, debel.thomas@gmail.com (T. de Bel).

https://doi.org/10.1016/j.media.2021.102004

change once deployed at a medical centre. If algorithms would be optimized or tuned for a specific center, newly introduced staining protocols or whole-slide scanners could result in algorithm performance degradation. This can be resolved by retraining the algorithm to deal with the changes, but this may be cumbersome and time-consuming.

To a large extent, dissimilarities between and within centers can be accounted for by training deep neural networks with color/stain augmentations or by using training data from multiple centers (Tellez et al., 2018). Augmentation techniques are standard practice in deep learning and of particular importance in the medical domain, since datasets are commonly small due to the scarce nature of the data and requirement of expert knowledge to annotate and provide a ground truth (Janowczyk and Madabhushi, 2016). Geometric operations like elastic distortions (Simard et al., 2003) have previously been shown to be beneficial in histopathology applications when dealing with data scarcity (Ronneberger et al. (2015); Cui et al. (2019)). Other morphological operations such as image scaling, adding Gaussian noise and blurring are commonly used augmentations. The effect of color augmentation on deep learning network performances has recently been studied in (Tellez et al., 2019; 2018; Liu et al., 2017b). However, due to the often linear alterations introduced by these augmentations, it is unclear whether they are able to capture all variations that may occur 'in the wild'. Considering that variation is derived from color as well as high-frequency (i.e. fluctuations in adjacent pixels of the image) differences in texture, artificial augmentation may oversimplify the variability that occurs in real-world tissue stainings.

An alternate strategy to augmentation is to normalize whole-slide images to mimic the data that a network was trained on, alleviating the need for algorithm re-training. A large field of research in medical imaging deals with stain standardization or transformation Tschuchnig et al. (2020). Traditional approaches are often based on stain-specific color matrix deconvolution, using a reference slide to normalize the data to (Macenko et al., 2009; Bejnordi et al., 2016; Reinhard et al., 2001; Vahadane et al., 2016). A drawback of these methods is that they are often specifically tailored to work with the most commonly used H&E staining.

Generative adversarial networks were introduced in Goodfellow et al. (2014). The objective of these networks is to learn the distribution of training data to generate new samples that realistically resemble the original data. This is accomplished by introducing two different network components with opposing goals. A generator component is tasked with creating realistic images in a target domain, either from noise input or from some source database. A second component, the discriminator, is tasked with discerning between the generated images and original images. While the generator is trained to fool the discriminator with realistic images, the discriminator is fed with synthetic and original images. In adversarial training, the optimal outcome is a generator that has converged to the data distribution, with the discriminator confused between real and fake images. Recently, adversarial based approaches for paired image-to-image translation such as UNIT (Liu et al., 2017a), Pix2Pix (Isola et al., 2017) have made their way into the medical domain (Welander et al., 2018; Nie et al., 2018). Cycle-consistent generative adversarial networks (CycleGANs) are another popular stain transformation method in histopathology that have enjoyed a lot of recent interest, reaching state of the art results in a lot of transformation tasks (Shaban et al., 2019; de Bel et al., 2019; Mahmood et al., 2019; Gadermayr et al., 2018; Mercan et al., 2020). CycleGANs allow for targeted unsupervised domain transformation with unpaired data (Zhu et al., 2017), by utilizing the so-called cycle-consistency. This makes them particularly useful in histopathology, where paired data, i.e. the same slide in two different stains, is hard to come by. Moreover, this approach is stain-agnostic, i.e. the same method can be applied to all stains. Concerns on the use of unsupervised transformation methods have been raised in Cohen et al. (2018), who have shown that CycleGANs can introduce a bias when matching source and target domains.

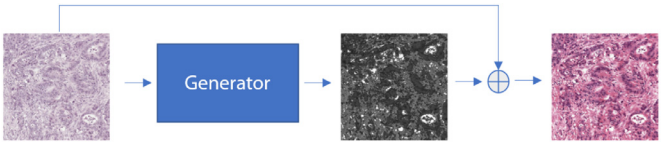

Fig. 1. The generator learns the difference mapping or residual between a source and target domain. The original image is summed with this residual to acquire the mapped image. The discriminator and components are omitted in this example.

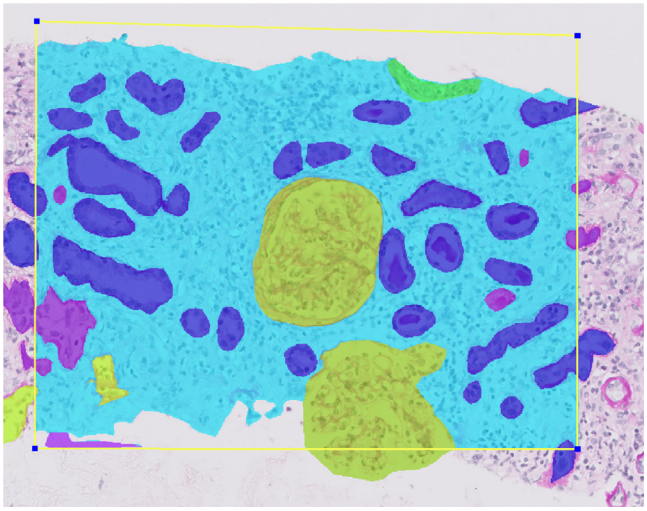

Fig. 2. A sample region of interest in our segmentation datasets. Within the box, all pixels are assigned to a relevant structure. Structures partially outside the ROI are fully annotated. Pixels outside of the tissue are not annotated.

In this paper we propose a simple yet effective adaptation to CycleGANs, where we move the task of the generator network towards learning the difference mapping, i.e. the residual of the source and target domains. This allows the generator network to focus solely on the domain adaptation, while the morphological integrity is kept in tact as a reference. We hypothesize that learning the residual improves the structural stability of CycleGANs, making them specifically suitable in the field of histopathology. Fig. 1 shows schematically how the generator works in a residual CycleGAN setup.

Introducing stain color transformation into the digital pathology workflow, calls for extensive research into the benefits and challenges that are introduced with the involved techniques. Two non-trivial tasks are chosen for this purpose: kidney tissue segmentation (Hermsen et al., 2019) and colon tissue segmentation (Bokhorst et al., 2018). The two segmentation tasks span five centers and pose a considerable challenge in terms of stain variation, as seen in Figs. 3 and 4. We demonstrate the benefit of CycleGAN-based transformation on top of color augmentations in these applications. With our broad assessment, we hope to establish residual CycleGANs as a reliable technique for applying stain transformation in histopathology.

Summarizing, the following contributions are made:

- We qualitatively and quantitatively compare our approach with other methods, including traditional stain transformation and color augmentation techniques.
- We include data spanning five centers and cover challenging segmentation tasks in two different domains. We trained seg-

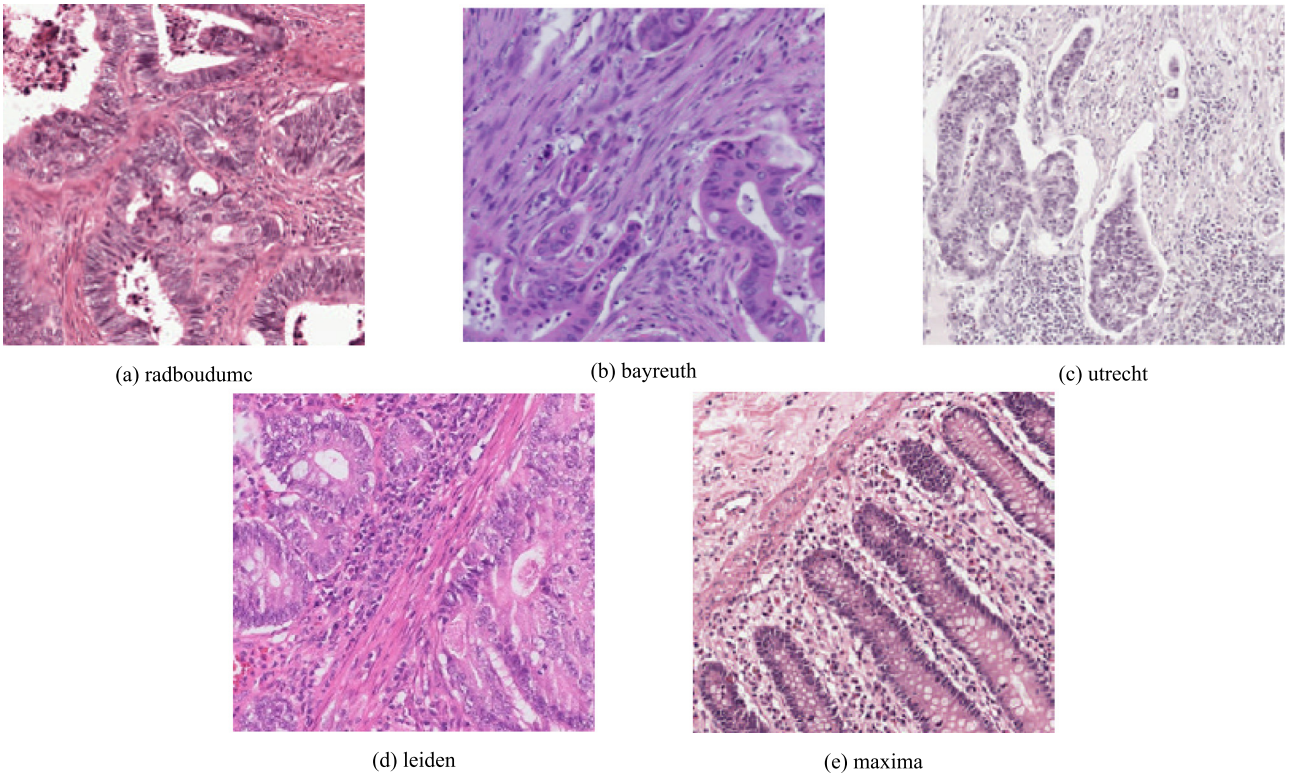

Fig. 3. A sample of colon tissue from the five centers, illustrating the high variety in which H&E-stained slides can vary between centers.

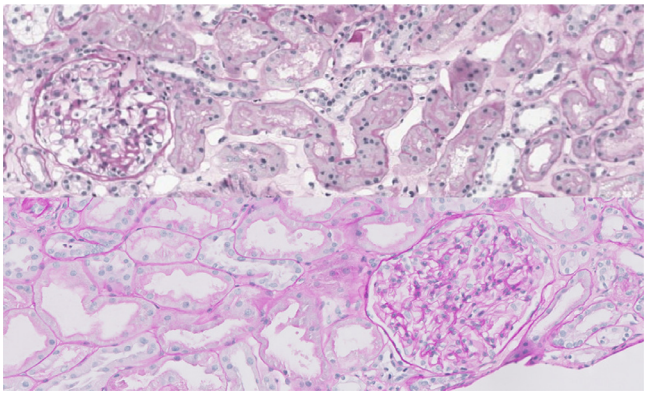

Fig. 4. A sample of PAS-stained kidney tissue from both datacenters. Top image: *radboud_kidney*. Bottom image: *amsterdam*.

mentation network using data from a single center and applied the segmentation network on the other data centers for evaluation.
- While most other work is evaluated solely on the H&E stain, we show that our method works on PAS-stained tissue as well. Arguably, this allows our method to be applied on a larger domain of staining protocols.

We organized the paper as follows. Section 2 details the data that we used in our experiments. Sections 3 and 4 describe the methods and experimental setup. In Section 5 we illustrate our results. In Section 6 and 7 we discuss our results in the context of previous and future work and conclude.

Table 1
Overview of the data that was used in this study. The uses for the different data sets are abbreviated with '*s*' and '*t*', for training the segmentation and the stain transformation, respectively.

center	slide count	ROIs	tissue type	purpose
radboud_kidney	40	80	kidney	*s, t*
amsterdam	10	20	kidney	*s, t*
radboud_colon	34	182	colon	*s*
radboud_c_test	5	20	colon	*t*
bayreuth	4	8	colon	*t*
maxima	5	16	colon	*t*
utrecht	5	14	colon	*t*
leiden	5	15	colon	*t*
total	100	355	/	/

2. Materials

In this paper data from five different centers was used. Images of five centers were used for the colon tissue segmentation task and two centers for the kidney tissue segmentation task. One center provided data for both tasks. An overview of the data is shown in Table 1.

2.1. Colon tissue segmentation & transformation

To train the colon tissue segmentation network, we used data solely from the Radboud University Medical Centre (*radboud_colon*). For the multi-class segmentation task, the following structures were annotated: Tumor, Desmoplastic stroma, Necrosis, Granulocytes, Lymphocytes, Erythrocytes, Muscle, Healthy stroma, Fat, Mucus, Nerve, Healthy glands and Stroma lamina propria. For each slide, up to ten ROIs were identified and annotated. The ROIs comprised an area of up to $1mm^2$ and were fully annotated by an experienced technician using the open-source program ASAP (Automated Slide Analysis Platform). The *radboud_colon* slides were

stained with H&E and digitized using a Panoramic P250 Flash II scanner (3D-Histech) at a spatial resolution of $0.24\mu m/px$. A total of 34 WSIs were included, 26 for training and 8 for validation.

For tissue transformation and subsequent application of the segmentation network, we included up to five slides from four external centers, which are: Institute of Pathology, Bayreuth (*bayreuth*), Mxima MC, Eindhoven (*maxima*), UMC, Utrecht (*utrecht*) and LUMC, Leiden (*leiden*). The data from the four centers will collectively be referred to as *external*. Additionally, five slides from the Radboud University Medical Centre *radboud_c_test* were included and served as a baseline to compare with the other centers. Each slide contained up to four ROIs in which all relevant structures are annotated. All slides were digitized at a resolution of $0.24\mu m/px$, except for the *leiden* dataset, which was captured at $0.5\mu m/px$. The variety of stains is demonstrated in Fig. 3.

2.2. Kidney tissue segmentation & transformation

For the kidney segmentation and tissue transformation we used data from the Radboud University Medical Center, Nijmegen, the Netherlands (*radboud_kidney*) and from a single external center: Academic Medical Center, Amsterdam, the Netherlands (*amsterdam*). The *radboud_kidney* dataset consisted of forty biopsies, stained with periodic acid-Schiff (PAS) according to routine standard procedures. All slides were digitally scanned using the 3D Histechs Panoramic 250 Flash II scanner. The external dataset (*amsterdam*) consists of twenty-four PAS-stained biopsies, digitized with the Philips IntelliSite Ultra Fast Scanner. All slides were scanned at a spacial resolution of $0.25\mu m$ per pixel. For the segmentation task, seven structure classes were included: glomeruli, empty glomeruli, sclerotic glomeruli, distal tubuli, proximal tubuli, atrophic tubuli and arteries. In total, ten *amsterdam* WSIs and forty *radboud_kidney* WSIs were annotated for testing the effectiveness of stain transformation on segmentation performance. Per slide, 1–2 ROIs of an area of up to $2mm^2$ were selected and exhaustively annotated in ASAP. All pixels in the ROIs that were not part of any of the structure classes were added to an eighth 'background' class. All annotations from both centers were produced by a technician with experience in renal pathology and afterwards checked by an experienced nephropathologist. An example of such an ROI is depicted in Fig. 2.

3. Methods

First, we introduce the general building blocks of generative adversarial networks (GANs) and, specifically, CycleGANs. Subsequently, we introduce the residual learning extension to CycleGANs, after which we introduce the architectures used to perform the segmentation tasks in colon and kidney tissue. Last, we introduce (traditional) baseline methods to compare against.

3.1. General concepts

Generative adversarial networks In a basic GAN setup, a generator (G) + discriminator D_1 pair together to produce some domain mapping $G : X \rightarrow Y$ (Goodfellow et al., 2014). Here, our source data is usually sampled from some prior noise distribution $x \in X_{noise}$ (for example a truncated normal distribution). The task of the generator is to resemble a target domain $y \in Y$ as accurately as possible, generating $G(x) = \hat{y}$, where $\hat{y} \approx y$. The domain of Y can be any desired domain, including but not limited to images, sound or temporal data. The discriminator has the binary task to distinguish between data sampled from the generator ($\hat{y}$) and the original distribution y, with $D(\hat{y}) = 0$ and $D(y) = 1$. Conversely, the generator

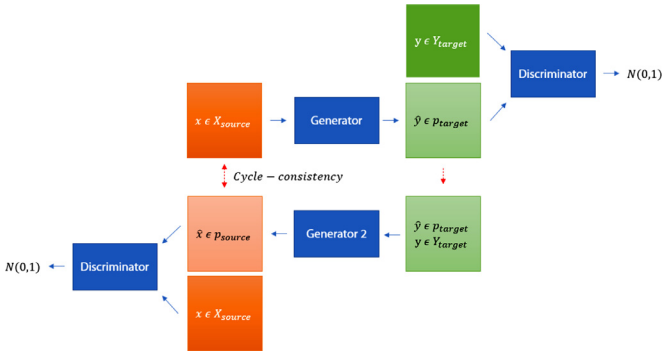

Fig. 5. Architecture of the CycleGAN setup with its generator and discriminator components. The top generator performs the mapping from source to target distribution, resulting in $\hat{y} \in p_{target}$ and vice versa for the bottom generator.

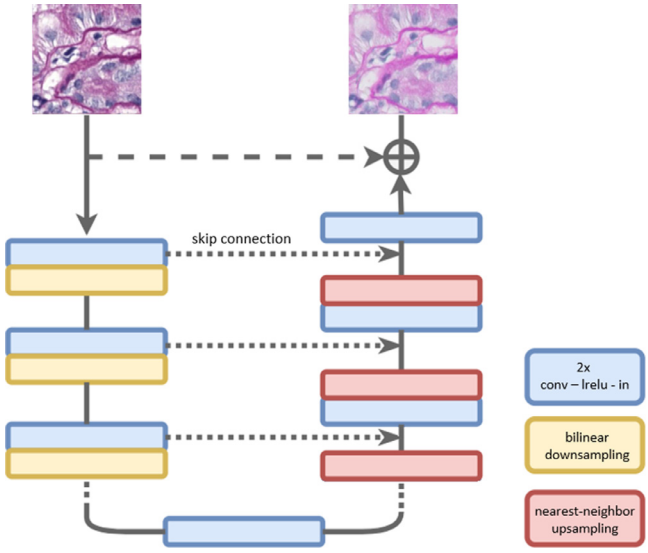

Fig. 6. Architecture of the generator in the residual CycleGAN, closely resembling the standard U-net. The output of the network is summed with the input. While the network can be of arbitrary depth, we used a depth of four convolution-leaky RelU-batch transformation blocks.

tries to fool the discriminator, training for $D(\hat{y}) = 1$. This adversarial game is expressed in the loss function:

$$L_{GAN}(G, D, x, y) = \log D(y) + 1 - \log D(G(x)) \quad (1)$$

GANs are usually trained alternating between optimizing the generator and discriminator (Goodfellow et al., 2014). In practice, GANs are hard to optimize, due to mode collapse or convergence to local optima (Brock et al., 2019; Mao et al., 2017).

Cycle-consistency In theory, a GAN can learn an injective mapping from a source to target domain, replacing $x \in X_{noise}$ with $x \in X_{data}$. However, as long as the discriminator objective is satisfied, the generator can map the source image to any image in the target domain. When dealing with unpaired image-to-image translation, cycle-consistency is used to conserve structural image information (Zhu et al., 2017). This is accomplished by adding a second generator (F) + discriminator D_2 pair and enforcing $F(G(x)) \approx x$ (see Fig. 5). The cycle-consistency loss is expressed with:

$$L_{cycle}(G, F, x, y) = E_x[||G(F(x)) - x||_1] + E_y[||F(G(y)) - y||_1], \quad (2)$$

Here, x and y are sampled from a source and target image distribution, respectively. The cycle-consistency loss term allows CycleGANs to be trained without paired data, while still allowing for a targeted image translation. The original CycleGAN approach includes a cycle-consistency loss term in both directions (as in Eq. 2),

which was shown to result in better performance (Zhu et al., 2017). The architecture of the CycleGAN at a component level can be viewed in Fig. 5.

3.2. Residual cyclegan

We made several changes to the standard CycleGAN setup. Most importantly, we add a skip-connection from the original image to the output of the final layer. This changes the task of the generator network towards learning the residual between domains, instead of rebuilding the image from scratch. While most architectures used in CycleGAN (e.g. U-net, ResNet) approaches already include a lot of skip-connections/shortcuts, the networks are not explicitly forced to keep the structural information intact. We hypothesize that by forcing the network to learn the residual/difference directly, the network is more likely to preserve the structural components, which is especially beneficial in pathology applications. We formulate the target of the residual generator as follows:

$$\hat{x} = 2 \cdot G(x) + x, \tag{3}$$

where G is the generator as in the standard version. As the final tanh activation in our networks produces results in the range of $(-1, 1)$, we multiply the output of the network with a factor two. This theoretically allows for transformation from the maximum (1) to minimum value (-1) and vice versa, if the source and target image domains would require so. Values outside the range $(-1, 1)$ are clipped at inference, but not during training.

We used $L1$-loss to minimize the cycle-consistency error and mean squared error for the discriminators, which was shown to be more stable during training (Mao et al., 2016; Zhu et al., 2017). The discriminator loss function from 3.1 then becomes:

$$L_{GAN}(G, D, x, y) = D(y)^2 + (1 - D(G(x)))^2 \tag{4}$$

Optimization was performed using stochastic gradient descent with Adam, with B_1 set at 0.5 (Kingma and Ba (2014); Zhu et al. (2017)). We empirically set the weight of the cycle-consistency loss at 5.0 and the discriminator loss weight at 1.0.

Generator architecture For our generator network, we adopted a fully convolutional encoder-decoder network inspired by the U-net (Ronneberger et al., 2015). The main convolutional building block consists a 3x3 convolution, followed by a leaky relu activation function (Bai et al., 2018) with leak set at 0.2 and finished with an instance normalization layer (Johnson et al., 2016; Ulyanov et al., 2016). Reflection padding was used with every convolution. In the encoder part of the network, we perform 2x2 bilinear downsampling after every two sets of convolutional blocks. In our implementation, we put 8 convolutional layers in the encoder. For the decoder, we used nearest-neighbour upsampling instead of transposed convolutions, which has been shown to reduce the extent of checkerboard artefacts (Odena et al., 2016). The amount of filters is set at 32 for the first convolution. This amount is doubled after each downsampling layer and halved after each upsampling layer. As with the standard U-net, we put skip-connections between the layers in the encoder and decoder that mirror each other. We hypothesize that these skip-connections are especially important in stain transformation, as high-resolution information may get lost in the deeper layers of the network. All weights are initialized by sampling from a truncated normal distribution, with the standard deviation set at 0.01. Our generator architecture is illustrated in Fig. 6.

The generators in the residual CycleGAN start off with performing a near-identity mapping, due to layer initialisation with low values and summation of input and output, which results in a near-zero cycle-loss. This caused the network to sometimes be numerically unstable, producing *nans* during back-propagation. To alleviate this problem, we averaged the cycle-consistency loss only over individual pixels with a loss in the 90th-percentile.

As training CycleGANs is unsupervised, there is no direct stopping criterion when training. We let the network converge by starting the learning rate at 0.0001 and statically reducing it by a factor of 2 after every 15 epochs until the learning rate is virtually zero.

Discriminator architecture For the discriminators, we used an architecture inspired by the convolutional "PatchGAN" setup (Isola et al., 2017). Based on a 256×256 input, this network network has a reduced receptive field and produces a 32x32 output, instead of a single value of whether the image is real or not. The discriminator is restricted to smaller parts of the image, which is hypothesized to result in more attention to high-frequency changes in the images (Isola et al., 2017). The discriminator consists of four convolutional layers with 64, 128, 256 and 256 filters, respectively and with kernel size of 4×4 and stride 2. A final convolution is added to reduce the output to a single filter map. All convolutions except for the final one were followed by an instance normalization layer and a leaky RelU, with the leak parameter set at 0.2. Filters were initialized with a truncated normal distribution with standard deviation set at 0.02. We did not put a sigmoid layer after the final convolution to improve convergence speed (Nair and Hinton, 2010).

3.3. Cyclegan variants

We compared the residual setup with a few alternative setups. Firstly, we added a "baseline" comparison, which uses the exact same setup as our residual approach, with the residual connection between input and output removed from the generator. To stabilize training of the baseline CycleGAN, we added an identity loss term:

$$L_{identity}(G, F, x, y) = E_x[||G(x) - x||_1] + E_y[||F(y) - y||_1], \tag{5}$$

where G and F are the generator networks. This identity loss directly forces the network to recreate the input image in the initial epochs, which reduced the chance of the network to get stuck in local optima (de Bel et al., 2019). The weight of the identity loss term was initially set at 3.0 and reduced to zero after 20 epochs, as it directly works against the discriminator loss. The same learning rate schedule was used as in the residual approach.

Secondly, we compared our method with StainGAN, introduced by Shaban et al. (2019). They achieved good results with CycleGAN on pathology data, beating several classical stain normalization approaches (Macenko et al., 2009; Vahadane et al., 2016; Reinhard et al., 2001; Khan et al., 2014). The main differences are that they used ResNet instead of U-net for their generators and they applied (1) instead of (4) for their adversarial loss term. Finally, we experimented with adding spectral normalization, which has been shown to stabilize training of the discriminators (Miyato et al., 2018). Spectral normalization has been shown to improve results in GAN training (Miyato et al., 2018; Brock et al., 2019), but is not yet generally used in CycleGAN applications. We experimented with spectral normalization in all CycleGAN setups (baseline, StainGAN and residual).

3.4. Cyclegan inference

We apply the transformation networks in a tile-by-tile fashion, as memory constraints prohibit application on the WSI in a single run. We used a method based on de Bel et al. (2019), in which tiles are sampled from the WSI with overlap between adjacent tiles. We sampled patches at a size of 1024×1024, moving by 512 pixels per tile. Overlapping sections were blended together according to de Bel et al. (2019) to prevent artifacts at the edges of tiles, which may be introduced by padding in the network layers.

3.5. Segmentation networks

For both segmentation tasks, we used a standard U-net setup (Ronneberger et al., 2015), using 4 max-pooling layers. The networks were individually optimized to improve segmentation performance, through empirically determining the parameters (elaborated below).

Colon tissue We trained the network with patches of 512×512 and a batch size of 5 for 500 epochs at 100 iterations per epoch. The initial learning rate was set at 0.0001, decreased by a factor of 0.5 after a plateau of 10 epochs. Patches were sampled at a resolution of $0.96 \mu m/px$. L2-regularization was applied with a weight of $1 \cdot 10^{-5}$. The first convolutional layers contained 64 features, doubling after each max-pool in the encoder and halving after each transposed convolution in the decoder.

Kidney tissue The setup for the kidney segmentation network was largely the same as the colon tissue segmentation network. The initial convolutional layer was set at 32 features and increased following the same scheme. Patches were sampled at 412×412 a batch size of 6 for 100 epochs at 100 iterations per epoch. The learning rate was initially set at 0.0005 and decreased by a factor of 0.5 after a plateay of 10 epochs. Patches were sampled at $0.96 \mu m/px$. L2-regularization was put at $1 \cdot 10^{-5}$.

3.6. Traditional stain transformation

Several color normalization methods based on template matching are applied as baseline comparisons, specifically those introduced by Vahadane et al. (2016); Macenko et al. (2009); Reinhard et al. (2001). These methods require the WSIs to be stained in H&E, as opposed to the CycleGAN methods, which can be applied to any kind of stain. Furthermore, a template patch is required as a reference to fit the stain distributions. To get a good template, we sampled multiple patches from the *radboudumc* dataset at a size of 2048×2048, before settling on one that produced acceptable transformation based on visual inspection. This reference patch was used to transform all *external* datasets, for the three normalization methods.

Apart from the template matching methods, we implement the approach by Bejnordi et al. (2016), which uses a Lookup-table (LUT) to perform color conversions. Cell nuclei are extracted from sampled patches, using Restricted Randomized Hough Transform, and used to describe the hematoxylin and eosin chromatic distributions. This is performed for both a source and target image. Next, the color distributions of source and target are matched and the color equivalence is stored in a look-up table (LUT). The LUT can be used to transform all colors of a source domain to a target domain. This method has been shown to outperform other older traditional stain normalization methods that are based on template matching (Bejnordi et al., 2016; Zanjani et al., 2018). We utilized the same hyper-parameters that were used in the original paper for generating the LUTs Bejnordi et al. (2016).

3.7. Stain color augmentation

Color augmentation techniques are a widely used alternative to normalization for providing algorithm robustness to unseen stains. Color augmentations are often performed by perturbing the brightness, hue or contrast of the image. We will study if, while training with extensive stain augmentation methods, there is still room for improvement by using stain transformation techniques.

We use a color augmentation method during the training of the segmentation networks specifically tailored to work well with H&E stained tissue (Ruifrok et al., 2001), which was shown to be the best performing augmentation in Tellez et al. (2019). This method works by separating the hematoxylin and eosin channels by means of a standardized color matrix. Both channels are then independently shifted and scaled, and transformed back to the RGB space. This color augmentation was recently shown to work best for robustness to stain variation in an extensive study on augmentations (Tellez et al., 2019). Fig. 7 demonstrates the variety introduced by this augmentation.

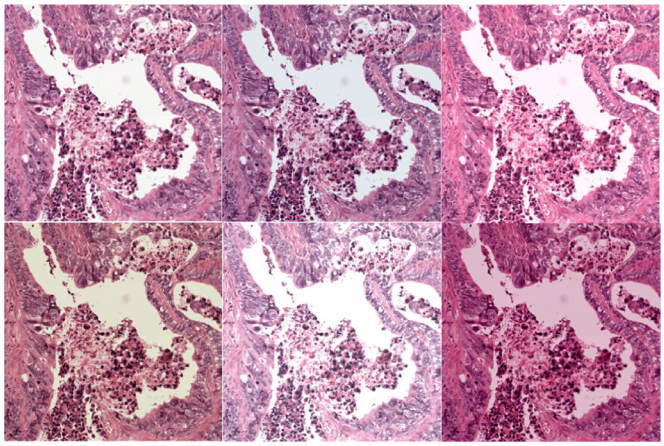

Fig. 7. Samples of color variations generated by the HED color augmentation. The original image is depicted at the top left.

Apart from extensive color augmentation, we used flipping/mirroring, scaling, blur and noise augmentations to make the segmentation networks as robust as possible to unseen variation. These augmentations were allowed to happen in conjunction with each other, resulting in heavily augmented patches. For the color augmentation, we opted to use the *HED-light* variant, in which the intensity ratio parameter is put between $[-0.05, 0.05]$ for all channels in the HED space (Tellez et al., 2019).

4. Experimental setup

4.1. Colon tissue transformation:

In the first application, we train to perform the transformation: G : external $\rightarrow$ radboud_colon. A separate CycleGAN setup was trained to perform transformation from each *external* centre to the *radboud_colon* centre. In this specific task we deal with little data in the *external* datasets. Moreover, we want to compare the CycleGAN approach with the LUT-based approach, which uses a single slide as a template. For fair comparison between all normalization methods, we used a single slide from both the source and target domains during training of all CycleGAN variants in the colon centres. A random slide was picked from each *external* dataset as well as from the *radboud_colon* dataset as a reference. This directly tests the CycleGAN's capabilities to deal with small (in this case single slide) datasets. After training, all colon tissue slides from *external* centers were normalized to look like *radboud_colon* data.

4.2. Kidney tissue transformation:

In our second application, we perform kidney tissue transformation. We sampled from all slides of the *amsterdam* and *radboud_kidney* datasets for training the CycleGANs. In this instance, we can not compare our method to the template matching and LUT-based transformation, which only work on H&E-stained tissue. This allowed us to use more data for the transformation. In this smaller experiment, we mainly test the residual CycleGAN's ability to perform better when data is abundant. As such, we only compare the residual with the basic variant. Because we utilized both source-to-target and target-to-source, a single CycleGAN setup was

Table 2
Colon tissue Dice-coefficient score for all methods. Results for the individual external datacenters are shown, as well as the overall score. 'base' refers to non-normalized tissue. Standard deviations are shown between the parenthesis.

	base	base CGAN	res CGAN	StainGAN	babak	vahadane	macenko	reinhard
radboudumc	0.78	x	x	x	x	x	x	x
utrecht	0.51 (0.15)	0.52 (0.16)	**0.65 (0.12)**	0.51 (0.15)	0.64 (0.05)	0.21 (0.10)	0.34 (0.01)	0.49 (0.04)
leiden	0.59 (0.07)	0.24 (0.09)	**0.74 (0.06)**	0.64 (0.07)	0.55 (0.23)	0.24 (0.07)	0.41 (0.06)	0.37 (0.04)
maxima	**0.68 (0.07)**	0.64 (0.11)	0.65 (0.13)	0.60 (0.12)	0.58 (0.16)	0.48 (0.09)	0.34 (0.13)	0.41 (0.19)
bayreuth	0.50 (0.11)	0.58 (0.10)	**0.59 (0.12)**	0.50 (0.07)	**0.59 (0.08)**	0.34 (0.05)	0.39 (0.06)	0.43 (0.04)
overall	0.57 (0.10)	0.50 (0.12)	**0.66 (0.11)**	0.56 (0.10)	0.59 (0.13)	0.32 (0.08)	0.37 (0.07)	0.43 (0.08)

Table 3
Dice coefficient score for all methods in both centers. 'base' refers to non-normalized tissue. Standard deviations are shown between the parenthesis.

trained on (x), applied to (y)	base	CGAN	res CGAN
radboudumc, amsterdam	0.78 (0.05)	**0.85** (0.03)	**0.85** (0.03)
amsterdam, radboudumc	0.71 (0.08)	0.73 (0.09)	**0.75** (0.07)

Table 4
Comparison of different types of artifacts (discoloration, hallucination, contrast changes) that occur for the different normalization methods. The template matching methods consist of Macenko, Reinhard and Vahadane.

	discoloration	hallucination	contrast
Template-based	+	+	−−
LUT-based	+	+	−
CycleGAN	±	−−	+
StainGAN	±	−	+
res CycleGAN	±	±	+

trained for transformation in both directions, thus using both generators.

Training process: In all applications, patches were sampled on the fly during training by sampling coordinates from a tissue mask generated using a tissue background segmentation algorithm (Bándi et al. (2017)).

4.3. Segmentation networks

Colon tissue segmentation A single segmentation network was trained on the *radboud_colon* dataset and used for application on all *external* centers and the *radboud_c_test* set. We used 34 slides designated for training the segmentation network (Table 1). The data was randomly split into 26 slides for training and 8 slides for validation, checkpointing after every epoch. We used the checkpoint with the lowest loss on the validation set for evaluating the external set.

Kidney tissue segmentation Two separate segmentation networks were trained using 40 and 10 slides from the *radboud_kidney* and the *amsterdam* data, respectively. The training/validation splits were put at 32/8 and 9/1. Cases were randomly assigned to the training or validation set.

4.4. Evaluation metrics

We used the mean pixel Dice coefficient per center to evaluate the performance of the segmentation networks for each transformation technique. We aggregated the Dice coefficients by calculating the weighted average across the different classes. Standard deviations of the Dice coefficients were calculated on a per-slide basis.

5. Results

The Dice-coefficient for each external colon dataset over the fourteen classes for all transformation techniques is shown in Table 2. The scores for the kidney tissue segmentation over all eight classes is shown in Table 3. Samples of the transformation of kidney tissue using the residual CycleGAN are shown in Fig. 11.

A qualitative comparison of the colon segmentation performance between the non-normalized data and our Residual CycleGAN approach for the individual external centers is shown in Fig. 8. Qualitative comparisons of the CycleGAN variant normalisation for the colon are shown in Fig. 9. The different classical approaches are compared in Fig. 10, using the same the same patch samples as in Fig. 9. Adding spectral normalization to the discriminator resulted in a 0, −0.01 and −0.03 difference for the average Dice coefficients, when applying it to the baseline, StainGAN and residual method, respectively. As such, we did not add it in our qualitative comparisons.

In Table 4 we give some intuition as to what kind of artifacts can occur when using the different methods and the frequency with which these artifacts occur. Here, we distinguish between different types of artifacts: discoloration (loss of information in the transformed image by whitening of an area), hallucination, (introduction of structures in the transformed image) and contrast changes (changes in contrast between different tissue structures, not corresponding with target stain). In Table 4, when an artifact occurs often, i.e. approximately more than 50% of the inspected views at $2 \times 2\mu m^2$ pixels, we use '−−'. At more than 25% we use '−', at more than 10% '±' and at less than 10% '+'. To illustrate some of the artifacts that occur, we show some samples in Fig. 12. Finally, we demonstrate that the cycle-consistency keeps up particularly well for our residual approach, compared to the baseline CycleGAN method, in Fig. 13.

6. Discussion

A major bottleneck in applying neural network applications in the medical field, is its fragility to stain variation (Ciompi et al., 2017). Studying and improving robustness of neural networks is an important part of bringing deep learning applications to the clinic. In this study we extensively studied the interaction between segmentation performance and stain normalization using data from five different data centers. We demonstrated the effect of normalization on top of training with extensive augmentation.

Our experiments show that CycleGAN-based stain transformation improves the quality of segmentation algorithms. Using residual CycleGANs resulted in equal or better scores in most centers. We show that residual CycleGANs perform well with medical data. By learning the residual, the morphological/structural information gets passed on directly, leaving the network with only the task to learn the style information. Furthermore, the training process is more stable over different runs, removing the need for the identity loss term that was introduced in (de Bel et al., 2019). Training

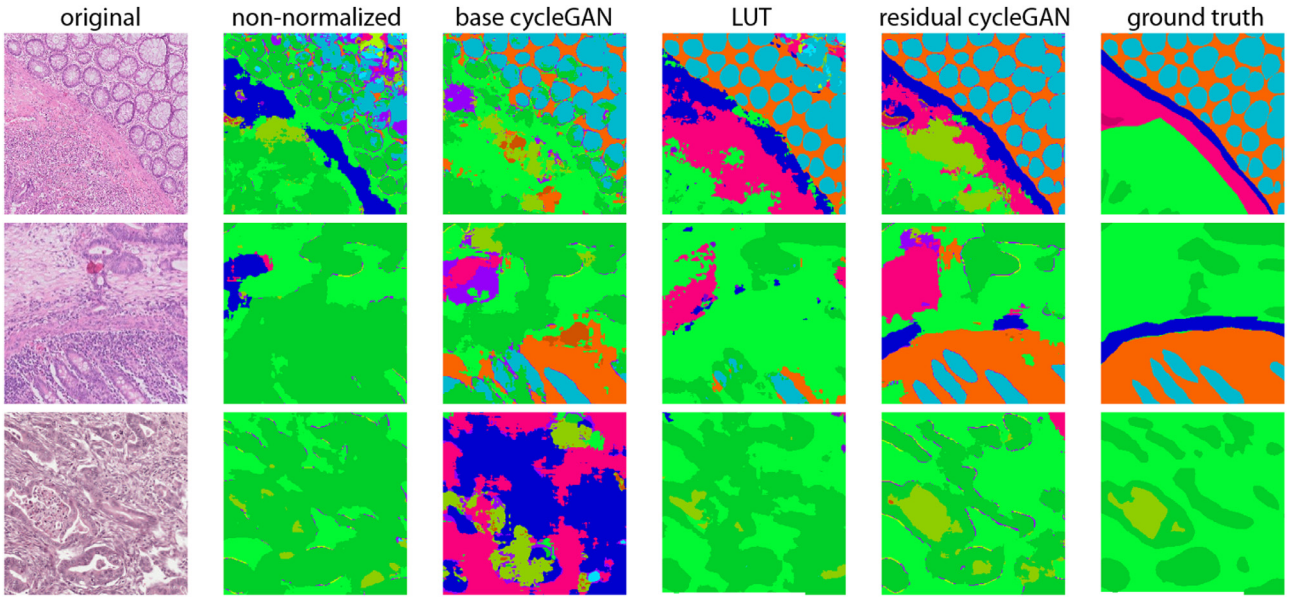

Fig. 8. Samples of the tissue segmentation task with all methods. From left to right: the tissue patch to be segmented, application of the segmentation algorithm without any normalisation method, the base CycleGAN setup, the LUT setup, our residual CycleGAN approach and the ground truth. The top two images are taken from *leiden* dataset, the bottom one from *maxima*. Colors are mapped as follows: healthy glands (blue), healthy stroma (orange), stroma lamina propria (dark blue), muscle (pink), vessel (red), desmoplastic stroma (light green), tumor (dark green), mucus and debris (yellow). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

stability is often a problem in training of GANs (Brock et al., 2019). The residual approach seems to drastically improve on this stability. Adding a residual connection from input to output is a simple yet effective addition for stain transformation of histopathological tissue.

In traditional CycleGANs, both the style and structural information need to be learned. This may be one of the reasons that the basic CycleGAN does not perform well in the colon tissue segmentation case, where we used only a single slide for both the source and target to learn from. The baseline CycleGAN and StainGAN need to be able to both reproduce the structure and change the style domain, while sampling from very little data. The residual CycleGAN effectively retains the structure and is able to focus on stain, which we hypothesize to be the easier task of the two. Our results (Table 2) suggest that the LUT-based approach produces unstable outcomes, producing segmentation performances of up to 10% lower than the non-normalized tissue. The template matching approaches scored lower than the other methods, which is likely due to the lacking quality of the normalization. The wide variety of stains introduced in this paper (as shown in Fig. 3) appears to be challenging for these traditional methods. Table 2 shows that none of the stain transformation methods were capable of increasing the colon segmentation scores to the level of the *radboud_c_test* data, as opposed to the kidney data, where we do improve our results. We hypothesize that this may be due to the *radboudumc* colon slides having a less difficult ground truth. Experiments on paired data, i.e. different stains on the same or consecutive slides, are required to directly compare network performance between synthetic (normalized) stains and original stains.

Interestingly, the non-normalized colon tissue segmentation on the *maxima* dataset performed best. Based on visual inspection we suggest that this is due to the high similarity between the *maxima* images and the *radboud_colon* images used for training the segmentation algorithm (Fig. 3). We hypothesize that adding stain transformation may not be beneficial to network performance when the stains are already similar. This confirms the result of (Tellez et al., 2019), which showed that solely color augmentation worked well without color transformation. Future work may focus on defining criteria to decide whether stain transformation is necessary.

According to Table 3, both the traditional and residual CycleGAN networks show a performance increase in the kidney when compared with using the original stains. The residual CycleGAN provides a benefit on top of the traditional approach in case of the *amsterdam* dataset. Additionally, the performance of the network trained on *radboud_kidney* data performs better. This is not surprising since the amount of data available for training the segmentation network was 4 times larger than the *amsterdam* dataset. We show that our technique can provide a benefit when used on top of stain augmentation. We omitted comparison with a segmentation network that was trained without augmentation. It was previously shown that segmentation does not work on unseen stains without augmentation (Tellez et al., 2019; de Bel et al., 2019).

From Fig. 12, we can clearly see that in the baseline CycleGAN tries to hallucinate new tissue. We see this effect more or less pronounced in both colon tissue and kidney tissue, while the sampling process is exactly the same for all CycleGAN variants. In all cases, masks were created to sample only from tissue locations, which effectively reduces the amount of patches with background. We hypothesize from this that residual CycleGANs are more robust to "out-of-distribution" data. In StainGAN, we sometimes see "checkerboard" patterns, which mainly becomes apparent in background locations, but might also be present in tissue. This may be the main cause for reduced performance, compared to the residual method. Finally, a type of artifact, which we call discoloration, is seen in all CycleGAN approaches. This type of artifact mainly occurs in background locations. When "hazy" textures are seen in the original, this texture is sometimes less pronounced in the converted tissue. This can be seen in Fig. 11. We pose, however, that this type of artifact is the least harmful to segmentation quality.

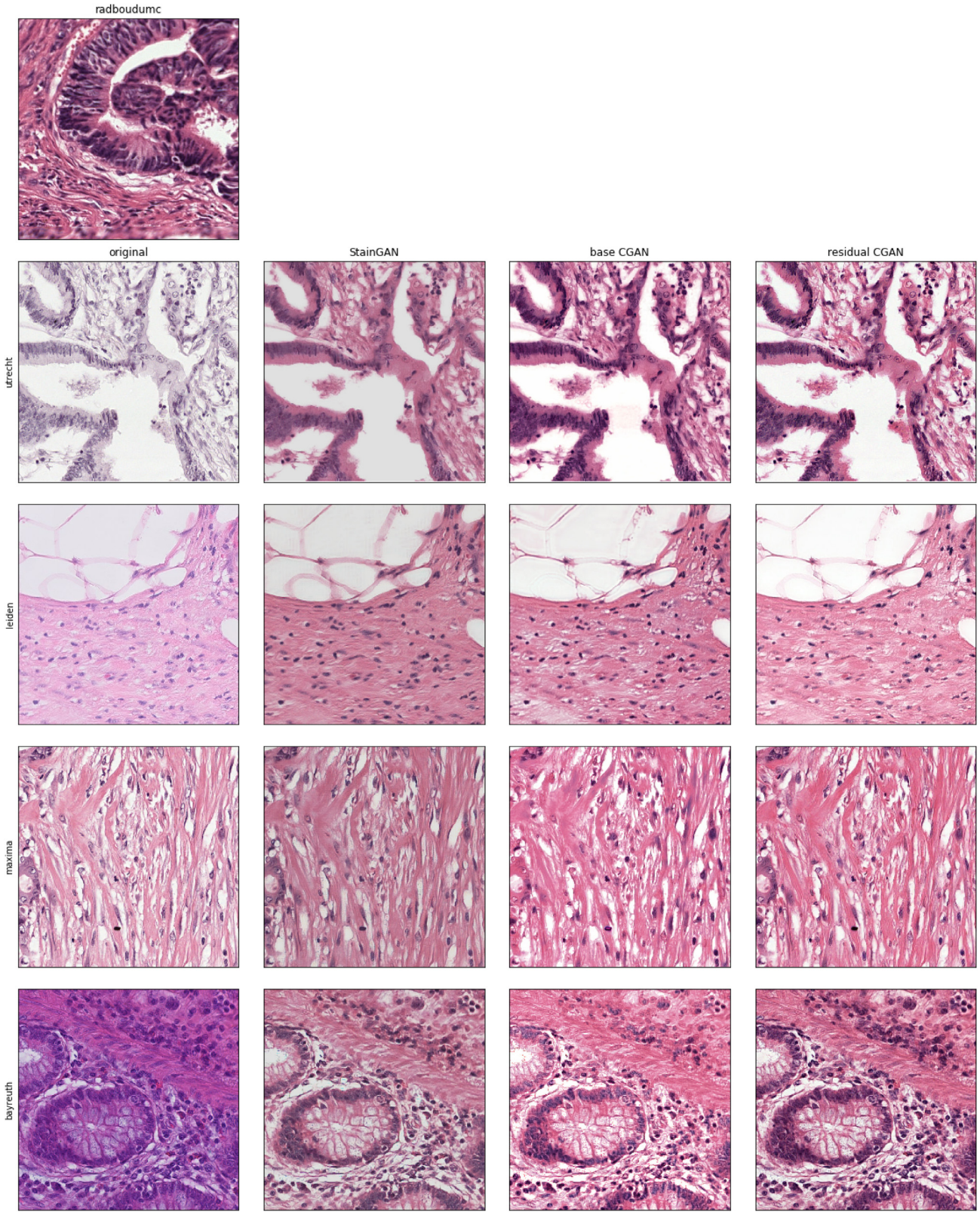

Fig. 9. Samples of colon tissue before and after transformation with the CycleGAN approaches. One sample was picked from each *external* center.

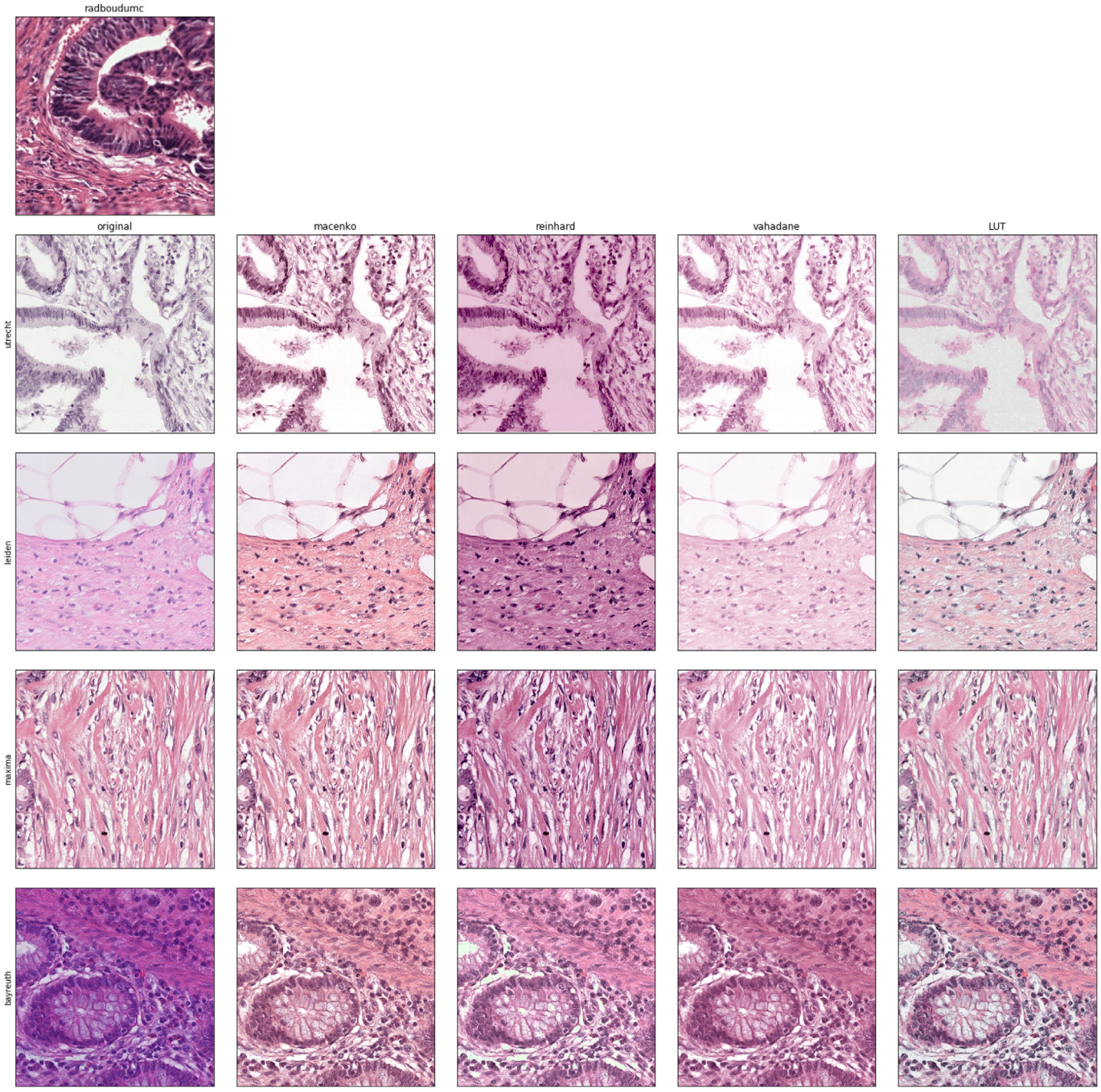

Fig. 10. Samples of colon tissue before and after transformation with the classical (template matching and lookup table) approaches. The top patch is put as a reference of the target *radboudumc* dataset, but was not necessarily used for normalization. One sample was picked from each *external* center.

From Table 4, there is a clear contrast between classic normalization methods and the CycleGANs. Due to the per-pixel normalization of classic methods, hallucinations are never seen. However, the quality of the normalization is lacking, resulting in contrast differences, which is also visible in Fig. 10. This is the main reason for the low normalization scores as seen in Table 2. Conversely, the general quality of normalization in all CycleGAN variants is higher (as seen in Fig. 9), while these methods are prone to hallucinations. As shown in (Cohen et al., 2018), the application of unsupervised domain transfer can introduce tissue artifacts in transformed images, specifically hallucinations, raising concerns for when these techniques are applied in medical diagnosis. From visual inspection and examples (Fig. 12), we demonstrated that our residual CycleGAN is less susceptible to hallucinations when compared to the other CycleGAN variants. This may allow for safe introduction of stain transformation techniques in medical practice. Further work may be focused on testing this assumption, in cases when a large difference in distribution of morphological features exists between the two domains. This will be especially interesting in slide level classification tasks, where a single hallucination might change the classification from 'benign' to 'tumor', for instance.

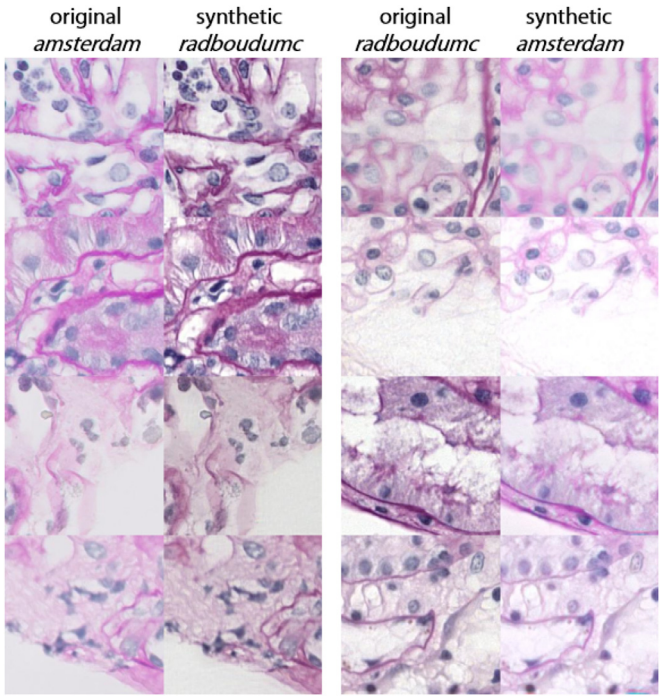

Fig. 11. Random samples of kidney tissue before and after transformation with the residual CycleGAN. Small border artifacts can be observed in the bottom right of the synthetic *amsterdam* images. This is the effect of padding during training. We remove these artifacts using our inference method (section 3.4).

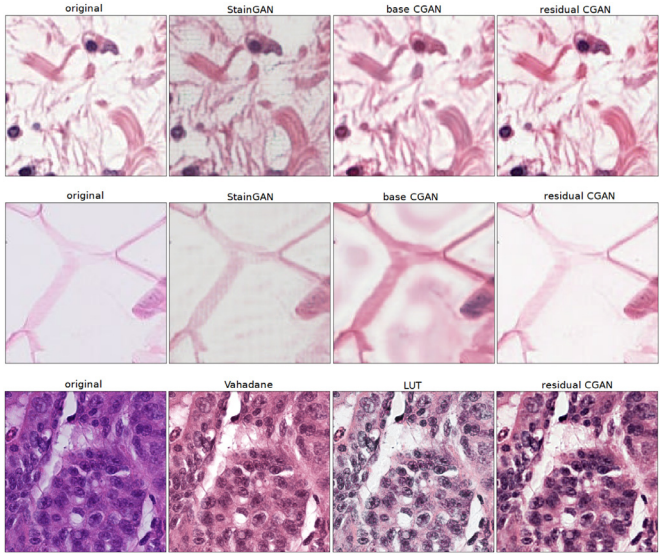

Fig. 12. Three samples of tissue conversions with artifacts. In first row we can observe checkerboard pattern hallucinations in StainGAN. The second row shows the same checkerboard pattern for StainGAN and hallucinations in the base CycleGAN. The final row demonstrates contrast changes in the classic normalization methods (with low contrast between cell nuclei and surrounding tissue in Vahadane and inaccurate colorization in the LUT approach).

The current challenge of normalization methods in histopathological tissue, is the reduction of artifacts. We think that the residual CycleGAN has shown to be a step in the right direction, reducing hallucinations, which is the most harmful type of artifacts. Furthermore, the residual CycleGAN shares a beneficial property with the classical methods: needing little data to perform well. In clinical settings, one might imagine the case of getting a single slide from another hospital for evaluation. Residual CycleGAN would perform well in this case, providing high quality normalization.

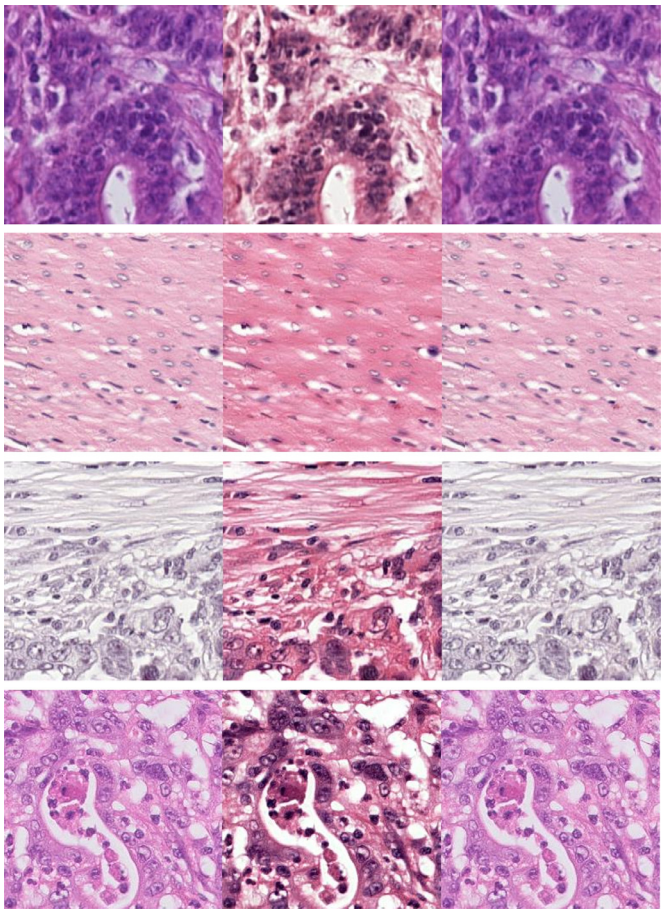

Fig. 13. Random samples from the colon dataset of the original images (left), with the transformed version (middle) and the cycled image (right), demonstrating the quality of the cycle-consistency.

7. Conclusion

We presented a new residual CycleGAN approach for normalizing tissue stainings. We comprehensively quantified the usefulness of different stain transformation and stain augmentation methods using data from five different centers. We have shown that our CycleGAN approach works comparatively well when challenged with data scarcity and benefits segmentation performance when used on top of extensive color augmentation. We recommend using residual CycleGANs for color transformation in histopathological tissue, as it has shown to be well suited for keeping structural integrity.

8. Disclosure

Jeroen van der Laak is a member of the advisory boards of Philips, The Netherlands and ContextVision, Sweden, and received research funding from Philips, The Netherlands and Sectra, Sweden in the last five years.

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.

Acknowledgements

We would like to thank Meyke Hermsen, Milly van de Warenburg and Eric Steenbergen for their contributions to creating and curating the annotations. The Knut and Alice Wallenberg foundation is acknowledged for generous support.

References

Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv preprint arXiv:1803.01271.

Bancroft, J.D., Gamble, M., 2008. Theory and practice of histological techniques. Elsevier health sciences.

Bándi, P., van de Loo, R., Intezar, M., Geijs, D., Ciompi, F., van Ginneken, B., van der Laak, J., Litjens, G., 2017. Comparison of different methods for tissue segmentation in histopathological whole-slide images. In: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017). IEEE, pp. 591–595.

Bejnordi, B.E., Litjens, G., Timofeeva, N., Otte-Höller, I., Homeyer, A., Karssemeijer, N., van der Laak, J.A., 2016. Stain specific standardization of whole-slide histopathological images. IEEE Trans. Med. Imaging 35 (2), 404–415.

Bejnordi, B.E., Veta, M., Van Diest, P.J., Van Ginneken, B., Karssemeijer, N., Litjens, G., Van Der Laak, J.A., Hermsen, M., Manson, Q.F., Balkenhol, M., et al., 2017. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. JAMA 318 (22), 2199–2210.

de Bel, T., Hermsen, M., Kers, J., van der Laak, J., Litjens, G., et al., 2019. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In: Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning; Proceedings of Machine Learning Research, pp. 151–163.

Bokhorst, J.-M., Pinckaers, H., van Zwam, P., Nagtegaal, I., van der Laak, J., Ciompi, F., 2018. Learning from sparsely annotated data for semantic segmentation in histopathology images.

Brock, A., Donahue, J., Simonyan, K., 2019. Large scale GAN training for high fidelity natural image synthesis. In: International Conference on Learning Representations.

Bulten, W., Pinckaers, H., van Boven, H., Vink, R., de Bel, T., van Ginneken, B., van der Laak, J., Hulsbergen-van de Kaa, C., Litjens, G., 2020. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. The Lancet Oncology.

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: Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on. IEEE, pp. 160–163.

Cohen, J.P., Luck, M., Honari, S., 2018. Distribution matching losses can hallucinate features in medical image translation. In: International conference on medical image computing and computer-assisted intervention. Springer, pp. 529–536.

Cui, Y., Zhang, G., Liu, Z., Xiong, Z., Hu, J., 2019. A deep learning algorithm for one-step contour aware nuclei segmentation of histopathology images. Medical & biological engineering & computing 57 (9), 2027–2043.

Gadermayr, M., Appel, V., Klinkhammer, B.M., Boor, P., Merhof, D., 2018. Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 165–173.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. In: Advances in neural information processing systems, pp. 2672–2680.

Hermsen, M., de Bel, T., Den Boer, M., Steenbergen, E.J., Kers, J., Florquin, S., Roelofs, J.J., Stegall, M.D., Alexander, M.P., Smith, B.H., et al., 2019. Deep learning–based histopathologic assessment of kidney tissue. Journal of the American Society of Nephrology 30 (10), 1968–1979.

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.

Janowczyk, A., Madabhushi, A., 2016. Deep learning for digital pathology image analysis: acomprehensive tutorial with selected use cases. J. Pathol. Inform. 7.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: European Conference on Computer Vision. Springer, pp. 694–711.

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

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

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

Liu, M.-Y., Breuel, T., Kautz, J., 2017. Unsupervised image-to-image translation networks. In: Advances in neural information processing systems, pp. 700–708.

Liu, Y., Gadepalli, K., Norouzi, M., Dahl, G.E., Kohlberger, T., Boyko, A., Venugopalan, S., Timofeev, A., Nelson, P.Q., Corrado, G.S., et al., 2017. Detecting cancer metastases on gigapixel pathology images. arXiv preprint arXiv:1703.02442.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

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. IEEE, pp. 1107–1110.

Mahmood, F., Borders, D., Chen, R., McKay, G.N., Salimian, K.J., Baras, A., Durr, N.J., 2019. Deep adversarial training for multi-organ nuclei segmentation in histopathology images. IEEE Trans. Med. Imaging.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., 2016. Multi-class generative adversarial networks with the l2 loss function. arXiv preprint arXiv:1611.04076 5, 00102.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2794–2802.

Mercan, C., Mooij, G., Tellez, D., Lotz, J., Weiss, N., van Gerven, M., Ciompi, F., 2020. Virtual staining for mitosis detection in breast histopathology. In: 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI). IEEE, pp. 1770–1774.

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

Nair, V., Hinton, G.E., 2010. Rectified linear units improve restricted boltzmann machines. In: Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807–814.

Nie, D., Trullo, R., Lian, J., Wang, L., Petitjean, C., Ruan, S., Wang, Q., Shen, D., 2018. Medical image synthesis with deep convolutional adversarial networks. IEEE Trans. Biomed. Eng. 65 (12), 2720–2730.

Odena, A., Dumoulin, V., Olah, C., 2016. Deconvolution and checkerboard artifacts. Distill doi:10.23915/distill.00003.

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

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer, pp. 234–241.

Ruifrok, A.C., Johnston, D.A., et al., 2001. Quantification of histochemical staining by color deconvolution. Analytical and quantitative cytology and histology 23 (4), 291–299.

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). IEEE, pp. 953–956.

Simard, P.Y., Steinkraus, D., Platt, J.C., et al., 2003. Best practices for convolutional neural networks applied to visual document analysis.. Icdar, 3.

Stathonikos, N., Nguyen, T.Q., Spoto, C.P., Verdaasdonk, M.A., van Diest, P.J., 2019. Being fully digital: perspective of a dutch academic pathology laboratory. Histopathology 75 (5), 621–635.

Tellez, D., Balkenhol, M., Otte-Höller, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., et al., 2018. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. IEEE Trans. Med. Imaging 37 (9), 2126–2136.

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.

Tschuchnig, M.E., Oostingh, G.J., Gadermayr, M., 2020. Generative adversarial networks in digital pathology: a survey on trends and future potential. Patterns 1 (6), 100089.

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.

Weinstein, R.S., Graham, A.R., Richter, L.C., Barker, G.P., Krupinski, E.A., Lopez, A.M., Erps, K.A., Bhattacharyya, A.K., Yagi, Y., Gilbertson, J.R., 2009. Overview of telepathology, virtual microscopy, and whole slide imaging: prospects for the future. Hum. Pathol. 40 (8), 1057–1069.

Welander, P., Karlsson, S., Eklund, A., 2018. Generative adversarial networks for image-to-image translation on multi-contrast mr images-a comparison of cyclegan and unit. arXiv preprint arXiv:1806.07777.

Zanjani, F.G., Zinger, S., Bejnordi, B.E., van der Laak, J.A., de With, P.H., 2018. Stain normalization of histopathology images using generative adversarial networks. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE, pp. 573–577.

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.