RESEARCH

Normalization of HE-Stained Histological Images using Cycle Consistent Generative Adversarial Networks

Marlen Runz[1,2*], Daniel Rusche[1], Martin R. Weihrauch[3], Jürgen Hesser[2,4,5] and Cleo-Aron Weis[1]

*Correspondence: marlen.runz@medma.uni-heidelberg.de
[1]Institute of Pathology, Heidelberg University, Mannheim, Germany
[2]Mannheim Institute for Intelligent Systems in Medicine, Heidelberg University, Mannheim, Germany
Full list of author information is available at the end of the article

Abstract

Background: Histological images show huge variance (e.g. illumination, color, staining quality) due to differences in image acquisition, tissue processing, staining, etc. The variance can impede many image analyzes such as staining intensity evaluation or classification. Methods to reduce these variances are gathered under the term image normalization.

Methods: We present the application of CylceGAN - a cycle consistent Generative Adversarial Network for color normalization in hematoxylin-eosin stained histological images using typical clinical data including variability of internal staining. The network consists of a generator network G_B that learns to map an image X from a source domain A to a target domain B, i.e. $G_B : X_A \rightarrow X_B$. In addition, a discriminator network D_B is trained to distinguish whether an image from domain B is an original or generated one. The same process is applied to another generator-discriminator pair (G_A, D_A), for the inverse mapping $G_A : X_B \rightarrow X_A$. Cycle consistency ensures that the generated image is close to the original image when being mapped backwards ($G_A(G_B(X_A)) \approx X_A$ and vice versa). We validate the CycleGAN approach on a breast cancer challenge and a follicular thyroid carcinoma dataset for various stain variations. We evaluate the quality of the generated images compared to the original images using similarity measures.

Results: We present qualitative results of the images generated by our network compared to the original color distributions. Our evaluation shows that by mapping images from a source domain to a target domain, the similarity to original images from the target domain improve up to 96%. We also achieve a high cycle consistency for the inverse mapping by obtaining similarity indices bigger than 0.9.

Conclusions: CycleGANs have proven to efficiently normalize HE-stained images. The approach enables to compensate for deviations resulting from image acquisition (e.g. different scanning devices) as well as from tissue staining (e.g. different staining protocols), and thus overcomes the staining variations in images from various institutions.

The code is publicly available at https://github.com/m4ln/stainTransfer_CycleGAN_pytorch. The dataset supporting the solutions is available at https://heidata.uni-heidelberg.de/privateurl.xhtml?token=12493b50-1538-4bdf-aca5-03352a1399a8.

Keywords: Histology Stain Normalization; HE-stain; Digital Pathology; Generative Adversarial Networks; Unpaired Image-to-Image Translation; Style Transfer; Deep Learning

Background

In both histology and surgical pathology, the inherent individual appearance of the considered object on the one hand, or the different staining protocols on the other hand, must be compensated in addition to factors that influence the image acquisition (e.g scanning devices). This demand applies to hematoxylin-eosin (HE) staining being the standard method in pathology but also to all other histochemical and immunohistochemical staining. Regarding HE-staining, solutions and protocols are standardized at first glance. However, even within a single institution, protocols may vary slightly and may not be coordinated with other institutes. Especially when training deep neural networks, for example for image classification, there is a need for stain normalization of images so that models are transferable to other datasets.

The term color normalization is an umbrella term for image processing techniques compensating for effects such as variable illumination, camera setting, etc. This evident request drives an active research. Conventional image processing such as color deconvolution or look-up tables with the need for selecting a reference template slide for normalization are widespread [1, 2, 3, 4, 5, 6]. A particular but quite similar issue is stain quantification [7, 8]. Recent publications investigate in the use of deep learning approaches with GANs and show the benefits compared to the conventional methods [9, 10]. It has also been shown how normalizing images using GANs can highly improve results of image classification [11] or segmentation [12]. Mahapatra et al. [13] integrate self-supervised semantic information such as geometric and structural patterns at different layers to improve stain normalization with CycleGANs. So far, however, the approaches have been applied to study-like conditions, whereas it is not clear to what extent these strategies can be used for typical daily data.

In this work, we investigate the potential and limitation in a clinical setting using a machine learning-based approach with a cycle consistent Generative Adversarial Network (CycleGAN) to learn the mapping from one HE-stain variant to an other. The approach we follow was proposed by Zhu et al. [14]. It learns the image-to-image mapping between two different HE-stained datasets to generate fake images in each image domain. We apply the technique to two independent datasets: the Mitosis-Atypia-14 challenge which provides two image sets of breast cancer tissue scanned with two different devices, and our *HE-Staining Variation* (HEV) dataset, which is follicular thyroid carcinoma slices stained with different protocols. We evaluate the results using the Fréchet Inception Distance (FID) and the Structural Similarity Index Measure (SSIM). Our results show that FID scores between generated and real images from two image domains can be improved by a factor of ≈ 25. We also achieve a high cycle consistency for the backward mapping by obtaining SSIM values bigger than 0.9.

Methods

CycleGAN Formulation

The CycleGAN framework used in this work is based on the implementation from Zhu et al. [14]. It consists of two generator and discriminator pairs each of which learns the mapping from one image domain to the other. Given the image domains

A and B with training images X_A and X_B, the generator G_B learns the mapping from A to B such that $G_B : X_A \rightarrow X_B$, while the generator G_A learns the mapping in reverse direction, i.e. $G_A : X_B \rightarrow X_A$. A discriminator D is a binary classifier. It decides whether a sample is real (1), i.e. given from the training dataset, or fake (0), i.e. produced by the generator. More precisely, discriminator D_B learns to distinguish between real images X_B^{real} and generated ones X_B^{fake}, while in the same way, D_A is trained to discriminate between X_A^{real} and X_A^{fake}.

For training, the objective function to be optimized is modeled by two loss functions: the adversarial loss $\mathcal{L}^{adv}$ [15] and the cycle consistency loss $\mathcal{L}^{cyc}$ [14].

Adversarial Loss

Introduced by Goodfellow et al. [15] the adversarial loss refers to the two-player game between the generator and the discriminator networks. More precisely, for the mapping $G_B : X_A \rightarrow X_B$, the discriminator D_B is trained to classify X_B^{real} and X_B^{fake} correctly, while the generator seeks X_B^{fake} being classified as real by the discriminator. In this way, both, the generator and the discriminator try to fool the other. Zhu et al. [14] use the least-squares loss as objective since it ensures more stability during training and generates higher quality results. Thus, the adversarial loss function is expressed as follows [14]:

$$\min_{G_B}\max_{D_B}\mathcal{L}_B^{adv} = \mathbb{E}_{X_B^{real}}\left[D_B\left(X_B^{real}\right)^2\right] + \mathbb{E}_{X_A^{real}}\left[\left(D_B\left(G_B\left(X_A^{real}\right)\right) - 1\right)^2\right],$$

with $\mathbb{E}$ being the expected value over all instances of X_A^{real} and X_B^{real}. In the same way, we can formulate the adversarial loss for the inverse mapping function $G_A : X_B \rightarrow X_A$, i.e.

$$\min_{G_A}\max_{D_A}\mathcal{L}_A^{adv} = \mathbb{E}_{X_A^{real}}\left[D_A\left(X_A^{real}\right)^2\right] + \mathbb{E}_{X_B^{real}}\left[\left(D_A\left(G_A\left(X_B^{real}\right)\right) - 1\right)^2\right],$$

Thus, the total adversarial loss $\mathcal{L}^{adv}$ is obtained by the sum of both terms $\mathcal{L}_A^{adv}$ and $\mathcal{L}_B^{adv}$.

Cycle Consistency Loss

Zhu et al. [14] presented this loss function to enforce that both mapping functions G_A and G_B learned by the generators are inverse functions. In other words, if an image is mapped from one domain to the other domain the backward mapping should bring the image back to its original state. Thus, it must satisfy the cycle $X_A^{real} \rightarrow G_B(X_A^{real}) \rightarrow G_A(G_B(X_A^{real})) = X_A^{rec} \approx X_A^{real}$ and in the same way for $X_B^{real} \rightarrow G_A(X_B^{real}) \rightarrow G_B(G_A(X_B^{real})) = X_B^{rec} \approx X_B^{real}$ for the backward mapping. Therefore, the total cycle consistency is given by:

$$\mathcal{L}^{cyc} = \underbrace{\mathbb{E}_{X_A^{real}}\left[\left\|G_A\left(G_B\left(X_A^{real}\right)\right) - X_A^{real}\right\|_1\right]}_{\mathcal{L}_A^{cyc}} + \underbrace{\mathbb{E}_{X_B^{real}}\left[\left\|G_B\left(G_A\left(X_B^{real}\right)\right) - X_B^{real}\right\|_1\right]}_{\mathcal{L}_B^{cyc}},$$

where $\|\cdot\|_1$ denotes the ℓ_1-Norm.

Hence, the total loss function is:

$$\arg\min_{G_B,G_A} \arg\max_{D_B,D_A} \mathcal{L} = \mathcal{L}^{adv} + \lambda \mathcal{L}^{cyc},$$

with λ being a regularization factor to control the relative importance of both, adversarial and cycle consistency losses.

Figure 1 illustrates the CycleGAN structure for mapping an image from domain A to domain B by $G_B : X_A^{real} \rightarrow X_B^{fake}$ and backwards by $G_A : X_B^{fake} \rightarrow X_A^{rec}$. The discriminator D_B tries to identify if an image is generated X_B^{fake} or real X_B^{real}. During training, the network is optimized by computing the adversarial loss $\mathcal{L}^{adv}$ and the cycle consistency loss $\mathcal{L}^{cyc}$. The same process is done for the reverse direction when a real sample image X_B^{real} is mapped from domain B to domain A, i.e $X_B^{real} \xrightarrow{G_A} X_A^{fake} \xrightarrow{G_B} X_B^{rec}$.

Figure 1 Illustration of the applied CycleGAN architecture for mapping images from domain A to domain B. A real sample image X_A^{real} is mapped to domain B by the generator $G_B : X_A^{real} \rightarrow X_B^{fake}$ and then back to domain A by the generator $G_A : X_B^{fake} \rightarrow X_A^{rec}$. The discriminator D_B differentiates between the generated image X_B^{fake} and a real sample image X_B^{real}. The same process is done for the reverse direction when mapping a real sample image X_B^{real} from domain B to domain A and backwards, i.e $X_B \xrightarrow{G_A} X_A^{fake} \xrightarrow{G_B} X_B^{rec}$. During training, the loss is computed by the adversarial loss $\mathcal{L}^{adv}$ and the cycle consistency loss $\mathcal{L}^{cyc}$.

Experiments

We train the CycleGAN model to map images from one HE-stain to another. Therefore, we choose two datasets with different configuration: **(a)** The Mitos-Atypia-14 challenge dataset in which the HE-stain in images appears different in color and resolution due to different scanning devices. **(b)** Our clinical HEV dataset, which contains images of serial sections that were subjected to different staining protocols.

Mitos-Atypia-14 Dataset

This is a publicly available challenge dataset containing breast-cancer histological images [16]. The tissue was HE-stained and scanned by two different whole-slide image (WSI) scanners: the Aperio ScanScope XT and the Hamamatsu Nanozoomer 2.0-HT. Both devices scan images with different resolutions, the Aperio 1539×1376 pixels and the Hamamatsu 1663×1485 pixels at X20 and X40 magnification. From each scanned set, 7936 tiles are selected for training and 15000 tiles for testing. We resize the images to 1024×1024 pixels and extract image tiles of 256×256 pixels as input to our network.

HE-Staining Variation Dataset

This dataset was collected at the Institute of Pathology, Medical Faculty Mannheim, Heidelberg University. It contains serial sections of a follicular thyroid carcinoma and is stained with the following HE-staining variants: standard protocol (of the Institute of Pathology, Mannheim) HE-stain (henceforth *HE*), intentionally stained too short (henceforth *shortHE*), intentionally stained too long (henceforth *longHE*),

only stained with hematoxylin (henceforth *onlyH*), and only stained with Eosin (henceforth *onlyE*). Figure 2 shows thumbnails from each WSI. For each set, we extract tiles of 256×256 pixels. We collect 10000 and 15000 tiles for the training and testing, respectively. The whole dataset including our training patches is made publicly available [17].

Figure 2 Exemplary miniature image of the WSI that forms the HEV dataset. Serial tissue sections from a thyroid tissue with a follicular carcinoma with HE-staining. For every slide the staining protocol is intentionally modified: **A**: Standard protocol at the Institute of Pathology, Medical Faculty Mannheim, Heidelberg University (*HE*) **B**: Shortened staining time (*shortHE*) **C**: Prolonged staining time (*longHE*) **D**: Only hematoxylin-stain (*onlyH*) **E**: Only eosin-stain (*onlyE*)

Training Details

In total, we train five models: first, we train the network on the Mitos-Atypia-14 challenge to learn the mapping between the two image sets X_A and X_B obtained by the scanners Aperio and Hamamatsu, respectively. We then perform further experiments on the HEV dataset, with set A being the standard stained tissue (see Figure 2 **A**) and set B being one of the other stained tissues (see Figure 2 **B-E**). The experiments are summarized in the appendix in Table 1. For our models, we use the same network architecture as described by Zhu et al. [14]. We train each network for 60 epochs in total where the initial learning-rate is set to $2e^{-4}$ and then decreases to zero after every 30 epochs. The regularization factor λ is set to 10 for all experiments. Adam optimizer is used ($\beta_1 = 0.5, \beta_2 = 0.999$) with a batch size of 1. We train and evaluate the models on an NVIDIA Quadro P6000 graphics card.

Evaluation

Two different measures are used to assess the quality of the images generated by the GAN: The Fréchet Inception Distance (FID) and the Structural Similarity Index Measure (SSIM). They are calculated on basis of python code provided from [18] and [19].

FID This metric consists of the Fréchet distance also known as Wasserstein-2 distance computed on the basis of feature vectors. Here, a feature vector is the 2048-sized output of a pre-trained inception v3 model applied on one image. For the whole set of input images we get a sample of feature vectors with m_1 as its collective mean and C_1 as its covariance while for the GAN output images we get m_2, C_2 respectively [20]. The Fréchet distance is then applied to calculate the minimum distance between the means and covariances[21]:

$$d^2\ ((m_1, C_1), (m_2, C_2)) = ||m_1 - m_2||^2\ + Tr\left(C_1 + C_2 - 2\sqrt{C_1 * C_2}\right).$$

For identical images the FID is zero, whereas it increases with noise and disturbances.

SSIM For a given original image x and the corresponding output of the GAN y the features luminance $l(x, y)$, contrast $c(x, y)$ and structure $s(x, y)$ are compared on basis of the respective average, variance and covariance. The product of these components with the weighting factors α, β, γ yields the SSIM:

$$\mathrm{SSIM}(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^{\alpha} \cdot [c(\mathbf{x}, \mathbf{y})]^{\beta} \cdot [s(\mathbf{x}, \mathbf{y})]^{\gamma} .$$

The SSIM scale ranges from 0 to 1 and equals one only for exact identical images. An SSIM close to zero hardly represents similar images [22].

Results

The results of our experiments are presented in the following section. For generators G_A and G_B, image tiles from image domains A and B can be mapped in both directions such that $X_A^{real} \xrightarrow{G_B} X_B^{fake} \xrightarrow{G_A} X_A^{rec}$ and $X_B^{real} \xrightarrow{G_A} X_A^{fake} \xrightarrow{G_B} X_B^{rec}$.

Mitos-Atypia-14

Example results of on the Mitos-Atypia-14 dataset are shown in Figure 3. Columns **A-C** refer to the image tiles scanned by the Aperio scanner (X_A^{real}) being mapped by the generator G_B to produce the corresponding image in the domain of the Hamamatsu scanner (X_B^{fake}) and the reconstruction from mapping the image back to its original domain (X_A^{rec}). The same process is done in the reverse direction for image tiles scanned in domain B being mapped to domain A and backward (columns **D-F**). Each row **1-4** presents another example image.

Figure 3 Results gallery from our experiments on the Mitos-Atypia-14 challenge dataset. Columns **A-C** refer to the image tiles scanned by the Aperio scanner (X_A^{real}) being mapped by the generator G_B to produce the corresponding image in the domain of the Hamamatsu scanner (X_B^{fake}) and the reconstruction from mapping the image back to its source domain (X_A^{rec}), i.e $X_A^{real} \xrightarrow{G_B} X_B^{fake} \xrightarrow{G_A} X_A^{rec}$. The same process is done in the reverse direction for image tiles scanned in domain B, i.e $X_B^{real} \xrightarrow{G_A} X_A^{fake} \xrightarrow{G_B} X_B^{rec}$ (column **D-F**). Each row **1-4** presents another example tissue section.

HE-Staining Variation

Figure 4 presents several test results when mapping a standard stained HE-image X_A^{real} to one of the four stains of domain X_B^{fake}. Each block **A-D** shows another example tissue section. The top row of each block represents an exemplary image tile of the stain to be mapped into (*shortHE, longHE, onlyH, onlyE*), while the bottom row depicts the input image (*HE*) and the corresponding output for each stain.

FID and SSIM Scores

For all five experiments FID scores are shown in Figure 5 **A**. As reference, FID scores of all testing images from datasets A and B (blue) are computed. They range between 31.5 (*MA14*) to 203.68 (*onlyE*). Our experiments reach on average FID scores for real vs. fake of 7.09 (A) and 6.93 (B), while for real vs. rec we obtain an average of 5.76 (A) and 5.58 (B). When mapping images from a source domain

Figure 4 Results gallery from our experiments on the HEV dataset for the mapping $G_B : X_A^{real} \rightarrow X_B^{fake}$.
Here, the input image is from domain A of the standard stained tissue (HE) being mapped to domain B corresponding to the image-sets *shortHE*, *longHE*, *onlyH*, *onlyE*. Each block **A-D** shows another example tissue section. The top row of each block represents an exemplary image tile of the stain to be mapped into, while the bottom row depicts the input image and the corresponding output for each stain.

to a target domain, the FID scores compared to original images from the target domain improve up to 96% (blue vs. orange, red, green and purple). More precisely, for each experiment it is 76.85% ($MA14$), 91.93% ($shortHE$), 89.23% ($longHE$), 95.76% ($onlyH$), 95.57% ($onlyE$). A table with all FID is presented in the appendix in Table 2.

In addition, SSIM scores (see Figure 5 **B** are computed between the real and reconstructed images for each image domain A (blue) and B (orange). Each value refers to the average SSIM for all test images and the bars represent the corresponding standard deviation (SD). For each set A we obtain SSIM scores in the range of $SSIM = 0.94$ ($SD = 0.02$) ($MA14$) and $SSIM = 0.97$ ($SD = 0.01$) ($onlyH$), whereas for set B we obtain scores between $SSIM = 0.96$ ($SD = 0.02$) ($MA14$) and $SSIM = 0.98$ ($SD = 0.01$) ($onlyH$).A table with all SSIM scores is presented in the appendix in Table 3.

Figure 5 Evaluation of our experiments using FID and SSIM scores. A FID scores between real and generated (fake, rec) images. For identical images the FID is zero, whereas it increases with noise and disturbances. **B**: SSIM scores between real vs. rec images. The SSIM scale ranges from 0 to 1 and is close to zero for hardly similar images. A table with all FID and SSIM scores is presented in the appendix in Table 2 and Table 3.

Discussion

Our trained models show compelling results, both visually (Figure 3 and Figure 4) and quantitatively (Figure 5) by obtaining FID scores up to 96% better for images mapped to a target domain. The trained models are able to fully convert to the desired color scheme while preserving the structural contents of the original image due to the cycle consistency constraint leading to SSIM scores greater than 0.9 when mapping generated images back to their source domain.

Some limitations of the model can be seen when mapping images obtained by different scanning devices with varying resolutions. This can cause a loss in structural information despite the consistently good quality of the color normalization. With the HEV data set, the generated images look very realistic compared to the original images in a target domain without any decline in the image content (see Figure 4).

The CycleGAN approach used here, always learns the mapping between two image stains and can instantly normalize any unseen image if it is within one of the trained stains. For each other staining, the network needs to be retrained from scratch. However, the network is able to learn even from a small amount of images (1000-10000 per set) which can be obtained from a single WSI. In addition, the images do

not have to be labeled or paired to learn the mapping between two domains. The network can learn to add a stain to images which is not present in the source domain, e.g we are able to create a full HE-stained image from an image which has only a single stain (*onlyH*, *onlyE*) or vice verca. This may simplify the manual staining process. How this affects other stains besides HE needs further investigation.

Conclusion

In this work we show that CycleGANs are a powerful tool for normalization of different variants of HE-stains and tissue types. We validated this approach on two datasets covering images from different scanning devices, staining protocols and tissue types. The method has been successfully applied to compensate for variances resulting from image acquisition as well as from tissue staining while preserving structural content of the images. In order to make use of this approach in a clinical manner, the training process should be accelerated, i.e. using transfer learning, an increased batch size and specialized hardware. The method may be added to various image processing frameworks at WSI level to be applied to tasks such as classification or segmentation.

Abbreviations

Abbr.	Abbreviation
HE	Hematoxylin-eosin
GAN	Generative Adversarial Network
FID	Fréchet Inception Distance
SSIM	Structural Similarity Index Measure

Acknowledgements
We thank our project partners Smart In Media AG for the collaboration in this project.

Funding
Our work is supported by the grant ZIM ZF4689501TS9.

Availability of data and materials
The code is publicly available at https://github.com/m4ln/stainTransfer_CycleGAN_pytorch. The dataset supporting the solutions is available at
https://heidata.uni-heidelberg.de/privateurl.xhtml?token=12493b50-1538-4bdf-aca5-03352a1399a8.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute of Pathology, Heidelberg University, Mannheim, Germany. [2]Mannheim Institute for Intelligent Systems in Medicine, Heidelberg University, Mannheim, Germany. [3]Smart In Media AG, Köln, Germany. [4]Interdisciplinary Center for Scientific Computing, Heidelberg University, Heidelberg, Germany. [5]Central Institute for Computer Engineering, Heidelberg University, Heidelberg, Germany.

References

1. Bianconi, F., Kather, J.N., Reyes-Aldasoro, C.C.: Experimental assessment of color deconvolution and color normalization for automated classification of histology images stained with hematoxylin and eosin. Cancers **12**(11) (2020). doi:10.3390/cancers12113337
2. Bukenya, F.: A hybrid approach for stain normalisation in digital histopathological images. Multimedia Tools and Applications **79**(3), 2339–2362 (2020). doi:10.1007/s11042-019-08262-0
3. Vicory, J., Couture, H.D., Thomas, N.E., Borland, D., Marron, J.S., Woosley, J., Niethammer, M.: Appearance normalization of histology slides. Computerized Medical Imaging and Graphics **43**, 89–98 (2015). doi:10.1016/j.compmedimag.2015.03.005
4. Khan, A.M., Rajpoot, N., Treanor, D., Magee, D.: A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution. IEEE Transactions on Biomedical Engineering **61**(6), 1729–1738 (2014). doi:10.1109/TBME.2014.2303294

5. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Xiaojun Guan, Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 1107–1110 (2009). doi:10.1109/ISBI.2009.5193250
6. Bautista, P.A., Yagi, Y.: Staining correction in digital pathology by utilizing a dye amount table. Journal of Digital Imaging **28**(3), 283–294 (2015). doi:10.1007/s10278-014-9766-0
7. Ruifrok, A., Johnston, D.: Quantification of histochemical staining by color deconvolution. Anal Quant Cytol Histol **23** (2001)
8. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Computer Graphics and Applications **21**(5), 34–41 (2001). doi:10.1109/38.946629
9. Ghazvinian Zanjani, F., Zinger, S., Ehteshami Bejnordi, B., van der Laak, J., With, P.: Stain normalization of histopathology images using generative adversarial networks, pp. 573–577 (2018). doi:10.1109/ISBI.2018.8363641
10. Shaban, M.T., Baur, C., Navab, N., Albarqouni, S.: StainGAN: Stain Style Transfer for Digital Histological Images (2018). 1804.01601
11. Swiderska-Chadaj, Z., de Bel, T., Blanchet, L., Baidoshvili, A., Vossen, D., van der Laak, J., Litjens, G.: Impact of rescanning and normalization on convolutional neural network performance in multi-center, whole-slide classification of prostate cancer. Scientific Reports **10**(1), 14398 (2020). doi:10.1038/s41598-020-71420-0
12. de Bel, T., Hermsen, M., Kers, J., van der Laak, J., Litjens, G.: Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In: Cardoso, M.J., Feragen, A., Glocker, B., Konukoglu, E., Oguz, I., Unal, G., Vercauteren, T. (eds.) Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 102, pp. 151–163. PMLR, London, United Kingdom (2019). http://proceedings.mlr.press/v102/de-bel19a.html
13. Mahapatra, D., Bozorgtabar, B., Thiran, J.-P., Shao, L.: Structure Preserving Stain Normalization of Histopathology Images Using Self-Supervised Semantic Guidance (2020). 2008.02101
14. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks (2020). 1703.10593
15. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative Adversarial Networks (2014). 1406.2661
16. MITOS-ATYPIA-14 Grand Challenge. https://mitos-atypia-14.grand-challenge.org/ Accessed 22 Mar 2020
17. Runz, M., Weis, C.-A.: Normalization of HE-Stained Histological Images using Cycle Consistent Generative Adversarial Networks [Dataset]. heiDATA (2021). doi:10.11588/data/8LKEZF. https://doi.org/10.11588/data/8LKEZF
18. w13b3: SSIM-py Structural Similarity (SSIM) index, where the core dependency is NumPy (2019). https://github.com/w13b3/SSIM-py Accessed 20 Dec 2020
19. Seitzer, M.: pytorch-fid: FID Score for PyTorch. https://github.com/mseitzer/pytorch-fid. Version 0.1.1 (2020)
20. Brownlee, J.: How to implement the frechet inception distance (fid) for evaluating gans (2019). https://machinelearningmastery.com/how-to-implement-the-frechet-inception-distance-fid-from-scratch/
21. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium (2018). 1706.08500
22. Zhou Wang, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). doi:10.1109/TIP.2003.819861

Appendix

Table 1 Overview of our experiments

Dataset	Experiment Name	Set A	Set B
Mistos-Atypia-14	*MA14*	Aperio scanner	Hamamatsu scanner
HEV	*shortHE*	standard HE stained	shortened staining time
	longHE		prolonged staining time
	onlyH		only stained with hematoxylin
	onlyE		only stained with eosin

Table 2 FID scores for all experiments between real and generated (fake, rec) images for A and B

FID	*MA14*	*shortHE*	*longHE*	*onlyH*	*onlyE*
X_A^{real} vs. X_B^{real}	31.5017	59.4240	51.4460	119.0061	203.6761
X_A^{real} vs. X_A^{fake}	12.1464	4.5465	6.0007	4.1793	8.5647
X_A^{real} vs. X_A^{rec}	4.0544	4.2877	7.8630	4.0363	8.5685
X_B^{real} vs. X_B^{fake}	10.3222	4.0365	5.3136	7.0321	7.9218
X_B^{real} vs. X_B^{rec}	2.6451	6.3173	2.9931	4.9206	11.0160

Table 3 SSIM scores (SD = standard deviation) for all experiments between real and rec images for A and B

SSIM (SD)	*MA14*	*shortHE*	*longHE*	*onlyH*	*onlyE*
X_A^{real} vs. X_A^{rec}	0.9406 (0.0147)	0.9724 (0.0055)	0.9572 (0.0073)	0.9731 (0.0057)	0.9534 (0.0080)
X_B^{real} vs. X_B^{rec}	0.9606 (0.0148)	0.9760 (0.0063)	0.9702 (0.0098)	0.9763 (0.0056)	0.9648 (0.0107)