Stain-Transforming Cycle-Consistent Generative Adversarial Networks for Improved Segmentation of Renal Histopathology

Thomas de Bel[1] T.DEBEL@RUMC.NL
[1] *Department of Pathology, Radboud University Medical Center, Nijmegen, the Netherlands*
Meyke Hermsen[1] MEYKE.HERMSEN@RUMC.NL
Jesper Kers[2] J.KERS@AMC.UVA.NL
[2] *Department of Pathology, Amsterdam University Medical Center, the Netherlands*
Jeroen van der Laak[1] JEROEN.VANDERLAAK@RUMC.NL
Geert Litjens[1] GEERT.LITJENS@RUMC.NL

Abstract

The performance of deep learning applications in digital histopathology can deteriorate significantly due to staining variations across centers. We employ cycle-consistent generative adversarial networks (cycleGANs) for unpaired image-to-image translation, facilitating between-center stain transformation. We find that modifications to the original cycleGAN architecture make it more suitable for stain transformation, creating artificially stained images of high quality. Specifically, changing the generator model to a smaller U-net-like architecture, adding an identity loss term, increasing the batch size and the learning all led to improved training stability and performance. Furthermore, we propose a method for dealing with tiling artifacts when applying the network on whole slide images (WSIs). We apply our stain transformation method on two datasets of PAS-stained (Periodic Acid-Schiff) renal tissue sections from different centers. We show that stain transformation is beneficial to the performance of cross-center segmentation, raising the Dice coefficient from 0.36 to 0.85 and from 0.45 to 0.73 on the two datasets.

Keywords: Deep learning, generative adversarial networks, medical imaging, stain transformation

1. Introduction

Staining of tissue is a central part of histopathology, highlighting tissue structures crucial for diagnosis. The staining process is subject to high variability. Differences can be introduced, among others, by variations in staining protocols between pathology centers and differing whole-slide scanners. Large variety in stainings has been shown to dramatically affect the performance of deep learning image analysis (Ciompi et al., 2017). To a large extent, dissimilarities between centers can be accounted for by training with color/stain augmentations or using data from multiple centers (Tellez et al., 2018). However, it is uncertain whether data augmentations are able to capture all variations that occur 'in the wild' due to the linear nature of many color and stain augmentations. This may be an oversimplification of the variability that occurs in real-world tissue stainings.

Once deep learning algorithms are introduced in the workflow of the pathologist, they need to achieve reliable performance, regardless of the center they are deployed. Using only augmentations, the robustness of a network is unchangeable at test time. Even 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 could only be resolved by retraining the algorithm for

such modifications, which is cumbersome and time-consuming. An alternate strategy is to normalize whole-slide images to mimic the data that a network was trained on, alleviating the need for algorithm re-training.

Most previous work on stain normalization focuses on hand-engineered methods. These methods are typically tuned for a specific stain, for example haematoxylin and eosin (H&E) (Bejnordi et al., 2016; Khan et al., 2014). Recent approaches have used cycle-consistent generative adverserial networks (cycleGANs), and have shown the effectiveness of this architecture when used for stain transformation (Gadermayr et al., 2018; Shaban et al., 2018).

In a GAN setup, a discriminator network **D** is used to adversarially learn a generator **G** a domain mapping $G : X \rightarrow Y$. CycleGANs add to this by introducing an inverse mapping $F : Y \rightarrow X$ and enforce $F(G(X)) \approx X$, to retain structural information while transferring domains. CycleGANs are trained in an unsupervised and unpaired manner and are 'stain-agnostic', i.e. they can be applied for normalization of any stain.

CycleGANs may prove to be a solution to real-world stain variations, by transforming whole-slide images to the exact same stain. This would resolve the need for in-network stain robustness for algorithms that perform, for instance, cancer detection. A deployment setup can be imagined where cycleGANs are trained 'just-in-time' for new stain variations, after which a chain of other networks is executed to perform a variety of tasks (e.g. detection, segmentation, grading) without re-training.

Contributions: In this paper we make several contributions to existing work applying cycleGANs to stain transformation:

- We show that the original cycleGAN architecture benefits from optimization for stain transformation. We introduce several changes to the generator part of the cycleGAN architecture, reducing the amount of parameters of the transformation network. We tune the learning rate, batch size, and add an extra identity loss term that stabilizes training. We show that these changes result in improved stain-transferred images.

- We introduce a novel method for applying cycleGANs to whole-slide images for stain transfer. In short, this method works in a fully convolutional fashion at inference time. By sliding through the whole slide image, using weighted merging of overlapping adjacent tiles to remove tiling artifacts that would occur in regular patch-by-patch application.

- We demonstrate the effectiveness of stain transformation for cross-center tissue segmentation with convolutional neural networks. A segmentation network is trained on a dataset from one center and then applied to a test dataset from a different center. We compare the performance on test dataset with and without stain transformation. We also train the segmentation network with and without augmentations in an attempt to assess how well augmentations capture stain variation.

2. Experiments

2.1. Quantitative Analysis

Central to our method will be the performance of a segmentation network trained on data from the Radboud University Medical Centre (RUMC), Nijmegen, the Netherlands, and tested on data from the Academic Medical Center (AMC), Amsterdam. We refer to the section '3.1' for a detailed

description of the data. The segmentation network is trained twice: once with and once without extensive color and spatial augmentations. We apply both versions of the segmentation network on the AMC dataset to assess the effectiveness of the augmentations. On top of this, we add our cycleGAN stain normalization. The stain normalization is trained with both datasets, to learn the transformation from both RUMC to AMC and vice versa. Again, we apply both the augmented and non-augmented versions of the segmentation network on the AMC dataset, this time after performing stain normalization. This allows us to compare augmenting vs. not augmenting in conjunction with normalization vs. no normalization. Additionally, we train two segmentation networks on the AMC dataset and perform the same four experiments on the RUMC data.

Because we introduce several alterations to the original cycleGAN architecture, we compare the performance of our algorithm with the baseline cycleGAN architecture that was used in the original paper (Zhu et al., 2017).

2.2. Models

Segmentation network: For the segmentation network we used a standard U-net, based on Ronneberger et al. (2015). During training, patches were sampled at roughly 1.0 μm per pixel, with a patch size of $412, 412$. Apart from standard flipping and rotating, extensive color augmentations (e.g. brightness, contrast, HSV color shift) were used in an attempt to enhance the robustness to unseen stains. Figure 1 shows an example patch with the color augmentations that were performed. During training of the segmentation network, the augmentations were randomly combined to induce even more variation.

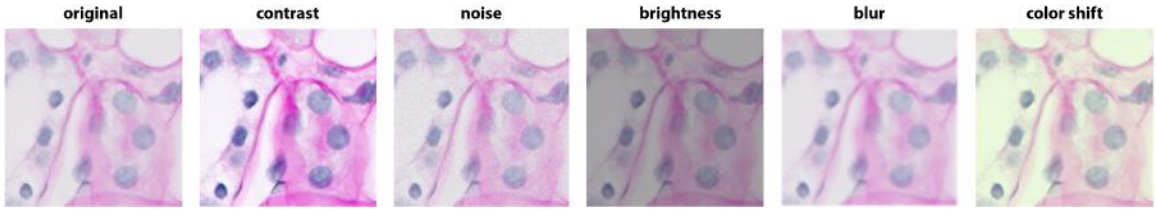

Figure 1: A sample of the color augmentations that were performed during training of the segmentation network.

Stain transformation network: The baseline cycleGAN-architecture is optimized using the cycle-consistency loss. The baseline adopts the generator **G**) of Johnson et al. (2016) for its good results on neural style transfer. Similar **G** networks were also utilized in recent cycleGAN stain transformation approaches (Shaban et al., 2018; Rivenson et al., 2018). For our discriminator **D**, we used the same 70x70 PatchGAN model as the baseline (Isola et al., 2017).

Our first modification to the baseline approach is increasing the batch size to 6 patches and increasing the learning rate to 0.008. As also mentioned in Brock et al. (2018), we hypothesize that more modi are covered in one batch, stabilizing training and reducing the probability of introducing hallucination artifacts (Cohen et al., 2018). To further stabilize training, we add an extra identity loss to the optimization process:

$$L_{identity}(G,F) = E_x[||G(x)-x||_1] + E_y[||F(y)-y||_1], \quad (1)$$

where G and F are the generators of both transformation directions. This loss forces the generators to perform an identity mapping of the input. As this loss discourages the generator to learn the stain transformation, the weight of this loss is gradually decreased to zero in the first 20 epochs. We found that adding this loss stabilizes training by forcing the network to initially look at 'simple' solutions close to the identity function. This prevents divergence to poor local optima in initial phases of the training process. An example of bad convergence is shown in the results section in Figure 6 (d).

We also modify both **G** and **F** to follow a U-net-like structure, using ResNet blocks and skip-connections between the encoder and decoder (He et al., 2016). We change the transposed convolutions in the decoder part to nearest neighbours up-sampling layers based on Odena et al. (2016). The width of the first layer starts at 32. The amount of filters is increased by a factor of two after each max-pooling layer and decreased by the same factor after each up-sampling layer. We use a small generator network with U-net depth of three, i.e. three max-pooling layers and up-sampling layers. We also experimented with further reducing the amount of parameters, by lowering the depth of the network to two and one. Both **G** and **D**, with each convolution, use leaky RelU's and instance normalization, which has shown to work well for style transfer (Ulyanov et al., 2016). For cycle-consistency loss we used the $L1$-norm, **D** was optimized with the mean squared error loss. We trained the networks with patches of size 256x256.

Patch sampling: We used tissue masks for sampling from the whole slide images (WSIs) during training of the cycleGAN. The masks were generated by using adaptive thresholding, with a window size of 11. During training, we uniformly sampled patches on-the-fly from the tissue based on the mask. This leads to a high variability where no two patches are exactly the same. Figure 2 shows the use of these masks with randomly sampled seed-points to generate patches.

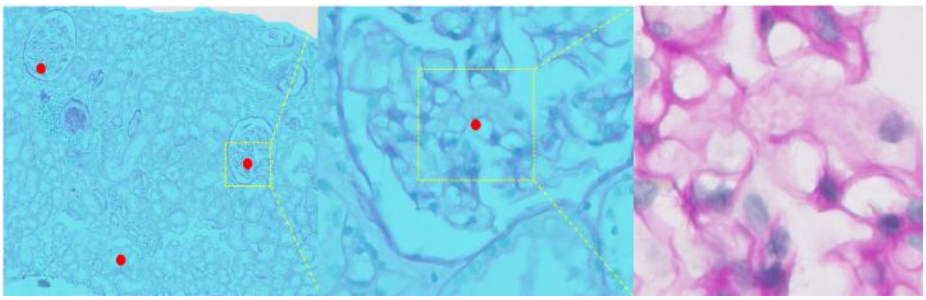

Figure 2: We generate patches on the fly from seed points samples from the mask, overlaid on the image. Taking the the seed point as central pixel, we generate a patch (of 256×256 in our case). These seed-points and patches are generated on-the-fly during training.

3. WSI inference technique

As WSI images are too large to fit directly on a GPU, we perform inference tile-by-tile to obtain the stain transformed whole-slide images. This introduces artifacts between adjacent tiles in the transformed WSI due to instance normalization relying on tile statistics. One option would be to use the running mean and average values obtained during training. This would reduce the quality

of the transformation, as the individual WSIs differ in their color intensities due to stain variations within a dataset. We propose a tile-wise inference method that eliminates the tiling artifacts. First, we increase the input size of our cycleGAN to 2048 pixels during inference. This will reduce the variation in instance normalization tile statics that occur when using small patches. We subsequently crop the network output to 1024 to get rid of the border artifacts introduced by zero-padding. Second, we take overlapping tiles by only shifting 512 pixels to our next tile. These tiles will largely have the same normalization statistics due to the small shift. This is visualized in Figure 3 (a). Last, we weight the pixels in the tiles based on their distance from the center pixel of the tile, to create a smooth transition between overlapping tiles. Here, the weight for a single pixel is based on the following formula:

$$w = min(|x - x_{cp}|, |y - y_{cp}|), \tag{2}$$

where cp stands for the center pixel. The weight map this creates for each tile is visualized in Figure 3 (b). Finally, due to overlap in both the x and y direction, there are four weighted values per pixel. We sum the weighted pixel values and normalize by the sum of the weights to create the final result. The effect of these tiling strategies can be seen in Figure 4.

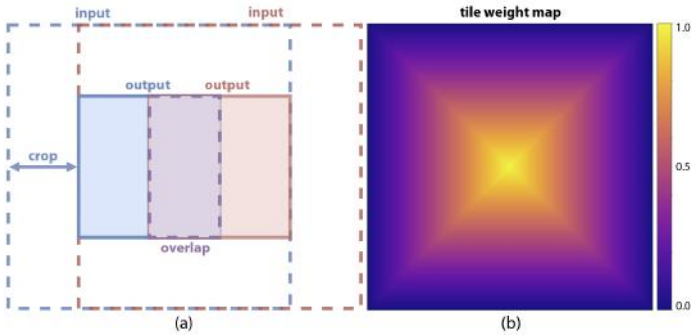

Figure 3: Schematic overview of the WSI inference strategy. (a) Shows the sliding window technique. Inference is performed on a large input, after which half of the image is cropped to remove zero-padding artifacts. We shift the window by half of the output (cropped) size, creating an overlap between tiles. (b) Visualizes the tile weight map.

3.1. Evaluation

Data: We utilize two datasets with periodic acid-Schiff (PAS) stains. The first dataset consists of forty biopsies originating from RUMC. The tissue slides were digitized using the 3D Histech's Panoramic 250 Flash II scanner. The second dataset consists of twenty-four biopsies, stained at the AMC. The slides were scanned with the Philips IntelliSite Ultra Fast Scanner. All slides were scanned at roughly 0.25 μm per pixel. Figure 5 shows an example of RUMC and AMC PAS-stained tissue. We included seven structure classes in our segmentation task: glomeruli, empty glomeruli, sclerotic glomeruli, distal tubuli, proximal tubuli, atrophic tubuli and arteries. All pixels within the regions of interest that did not belong to any category, were put in an eighth background structure class. Ten slides of the AMC dataset and forty slides of the RUMC dataset were annotated

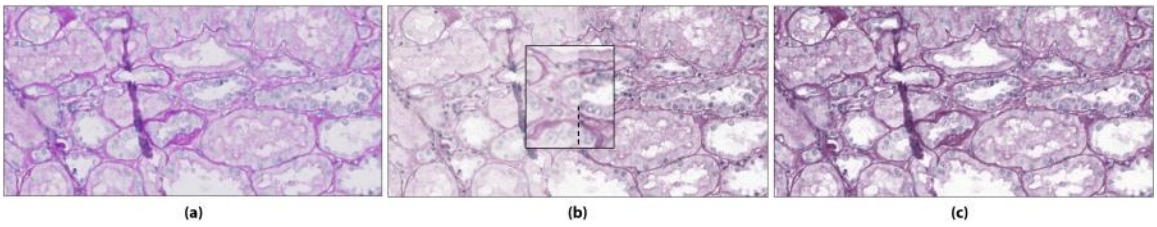

Figure 4: Results of the WSI inference strategy. In (a) the source image. In (b) the result using a naive tile-by-tile strategy. A clear artifact is present between the top and bottom tile. In (c) the tiling artifact is eliminated by using the overlapping tiling strategy.

for testing the effectiveness of stain transformation on segmentation performance. Per slide, 1-2 regions of interest were picked in which the selected renal structures were exhaustively annotated. Annotations for both datasets were made by a technician with experience in renal histopathology and checked by an experienced nephropathologist.

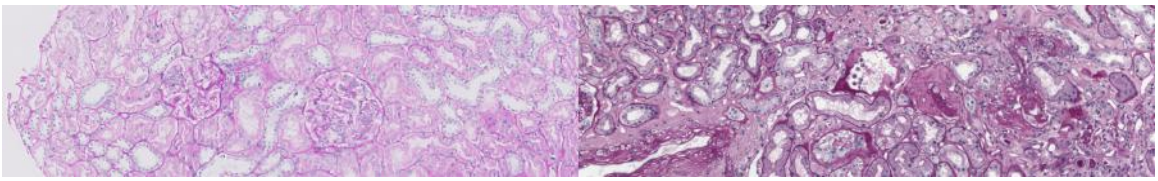

Figure 5: A tissue example of from both data-centers, illustrating the color differences of the stains. Left: AMC. Right: RUMC

Performance measures: We use the structural similarity (SSIM) index to assess the benefit of our changes to the default cycleGAN architecture. SSIM is a perception-based metric that quantifies image degradation as change in structural information (Wang et al., 2004). SSIM is most commonly used in stain conversion approaches where there is a lack of paired tissue, as it compares the structure of images while largely disregarding the color scheme. We used $C_1 = 0.01$, $C_2 = 0.03$ and a window size of seven in our calculations. We use the network with the highest SSIM to perform the stain transformation.

Due to the lack of paired data, we can't use simple statistics like mean-squared difference to assess the quality of transformation. Instead, we compare the color histograms of synthetic and original stained patches. For this we use the Wasserstein distance between the histograms averaged across the RGB channels (Ling and Okada, 2007).

Last, to assess the segmentation network performance, we calculate the Dice coefficients on the ten annotated slides from the AMC dataset and the forty slides from the RUMC dataset, calculating the weighted average across the different classes. We report the average score over the ten slides, the standard deviation between the slide scores and the highest and lowest scored slides.

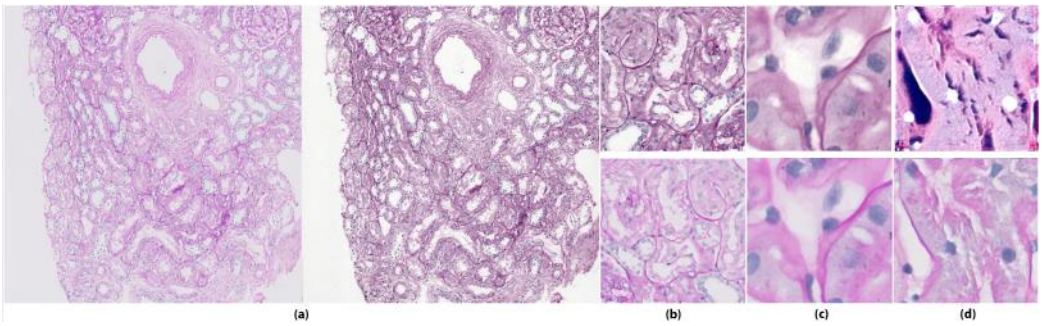

Figure 6: Samples of stain transformation by the network at increasing levels of zoom (a,b,c). At (d) an example of a failure run is given where the network learned to 'invert' the tissue. This problem was solved by adding the identity loss (1).

4. Results

Performance of the stain transformation: We present transformed samples in Figure 6. Additionally, randomly sampled tiles are shown in Appendix A. We report the SSIM values of our networks in Table 1, together with the histogram Wasserstein distances. As our network with 3 max-pooling layers obtained the best SSIM, we used it for the rest of the experiments.

Comparison of segmentation performance with and without stain transfer: The Dice coefficients obtained after segmentation are reported in Table 2. A table with scores per class is added in Appendix C. Qualitative results on the AMC dataset are shown in Figure 7, with additional examples in Appendix B.

5. Discussion

As expected, when applying the network trained without data augmentation on the non-transformed AMC data, the network fails with a Dice coefficient of 0.36. The segmentation network trained with data augmentation was able to achieve good results on the AMC dataset, improving the Dice coefficient from 0.36 to 0.78. However, there is still a performance gain when adding stain transformation on top of augmentation, increasing the average Dice coefficient from 0.78 to 0.85. This might indicate that not all the stain variation can be captured with only data augmentation and that the cycleGAN is better able to model non-linear stain variations. Interestingly, when stain transformation is applied, the average Dice coefficient is the same, regardless of whether the segmentation network was trained with or without augmentation. This provides evidence for when deploying these networks, re-training with data augmentation in case of protocol or scanner changes is not needed for algorithms downstream of the stain transformation cycleGAN.

The segmentation performance on the RUMC dataset shows a similar pattern. As we trained the segmentation networks with very few annotations, the overall scores expectantly turned out lower (Table 2, RUMC coefficients). Using either stain transformation or augmentation increased the Dice coefficient from 0.46 to 0.71. Using both techniques together gives a slight edge, increasing the score to 0.73. This supports our hypothesis that segmentation benefits from both augmentation and

stain transformation, combining non-linear and linear stain variations. Future research with more datasets will turn out whether the AMC dataset was an anomaly considering there was no increase in performance when using both augmentation and stain transformation. Over the two datasets we can conclude that stain transformation is at least as useful as augmentation.

There is no paired data available for our datasets, preventing the use of straightforward performance measures to quantitatively assess transformation quality (e.g. mean-squared difference). Instead, we opted to use the Wasserstein distance between the color histograms of the stains, to show that the color distributions of our transformed AMC slides are similar to the original RUMC slides.

We used the SSIM to show that the structural integrity of the original slides was not tampered with by the cycleGAN, demonstrating that our modifications score slightly better than the original cycleGAN architecture, while using less parameters. We think that the SSIM and Wasserstein distance on color histograms nicely complement each other, where the first quantifies the structure integrity and the second compares the color distributions.

In future work it would be valuable to assess our method on paired data to better quantitatively assess the performance of the stain transformation. This can, for example, be done by performing staining/re-staining. In this approach, a slide is cleared after the initial staining and scanning and then re-stained and scanned at a different site. Furthermore, we would like to investigate whether our approach directly translates to other types of stains, for example H&E or immunohistochemical stains. Finally, comparing different stain transformation methods, both other cycleGAN and classical machine learning approaches, will be an interesting venue to explore in future research.

References

Babak Ehteshami Bejnordi, Geert Litjens, Nadya Timofeeva, Irene Otte-Höller, André Homeyer, Nico Karssemeijer, and Jeroen AWM van der Laak. Stain specific standardization of whole-slide histopathological images. *IEEE transactions on medical imaging*, 35(2):404–415, 2016.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Francesco Ciompi, Oscar Geessink, Babak Ehteshami Bejnordi, Gabriel Silva de Souza, Alexi Baidoshvili, Geert Litjens, Bram van Ginneken, Iris Nagtegaal, and Jeroen van der Laak. The importance of stain normalization in colorectal tissue classification with convolutional networks. In *Biomedical Imaging (ISBI 2017), 2017 IEEE 14th International Symposium on*, pages 160–163. IEEE, 2017.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. How to cure cancer (in images) with unpaired image translation. In *International Conference on Medical Imaging with Deep Learning (MIDL 2018) – Abstract track*, 2018.

Michael Gadermayr, Vitus Appel, Barbara M Klinkhammer, Peter Boor, and Dorit Merhof. 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*, pages 165–173. Springer, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint*, 2017.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016.

Adnan Mujahid Khan, Nasir Rajpoot, Darren Treanor, and Derek Magee. 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.

Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE transactions on pattern analysis and machine intelligence*, 29(5): 840–853, 2007.

Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 2016. doi: 10.23915/distill.00003. URL http://distill.pub/2016/deconv-checkerboard.

Yair Rivenson, Hongda Wang, Zhensong Wei, Yibo Zhang, Harun Gunaydin, and Aydogan Ozcan. Deep learning-based virtual histology staining using auto-fluorescence of label-free tissue. *arXiv preprint arXiv:1803.11293*, 2018.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. *arXiv preprint arXiv:1804.01601*, 2018.

David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob van de Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *CoRR*, abs/1607.08022, 2016.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*, 2017.

Table 1: The leftmost table shows the structural similarity between the original AMC slides and the transformed AMC slides. We also report the amount of parameters that the networks used. The rightmost table reports the average Wasserstein distances (WD) of the averaged color histograms. We compare: transformed AMC slides vs. RUMC slides, original AMC slides vs. RUMC, RUMC vs. RUMC.

	Param. (M)	SSIM
cycleGAN-baseline	2.8	0.83
Our approach (depth 1)	0.11	0.73
Our approach (depth 2)	0.45	0.75
Our approach (depth 3)	2.0	0.85

	WD
Conv. AMC vs RUMC	3363
Orig. AMC vs RUMC	14923
RUMC vs. RUMC	4594

Table 2: Dice coefficient of the segmentation of both datasets. Additionally, we show the highest and lowest scored slides and the standard deviation.

Experiment		Dice coefficient AMC				Dice coefficient RUMC			
Augmentations	Stain transformed	Mean	Std	Min	Max	Mean	Std	Min	Max
x	x	0.36	0.21	0.09	0.65	0.46	0.12	0.15	0.78
x	✓	**0.85**	0.06	0.69	0.91	0.71	0.12	0.34	0.87
✓	x	0.78	0.08	0.65	0.87	0.71	0.10	0.44	0.86
✓	✓	**0.85**	0.05	0.72	0.91	**0.73**	0.11	0.37	0.87

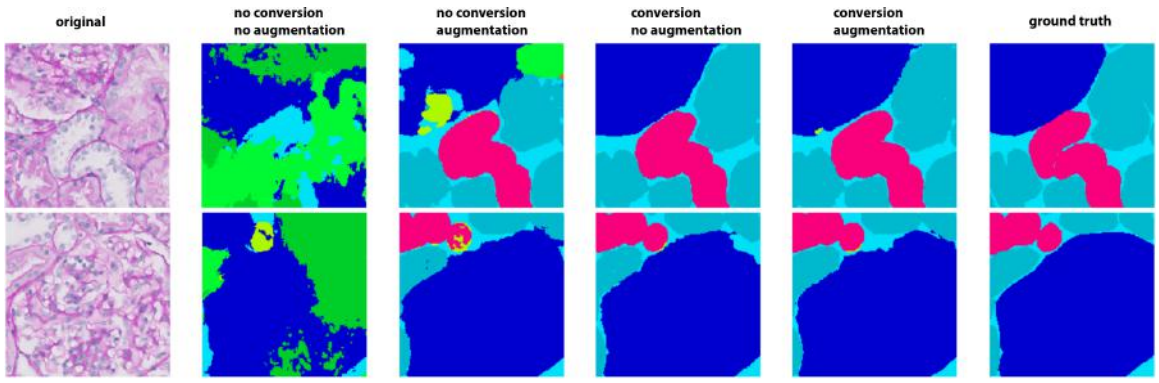

Figure 7: Samples from our segmentation results with and without augmentations and stain transformation.

Appendix A. Randomly sampled patches of real and artificially stained tissue

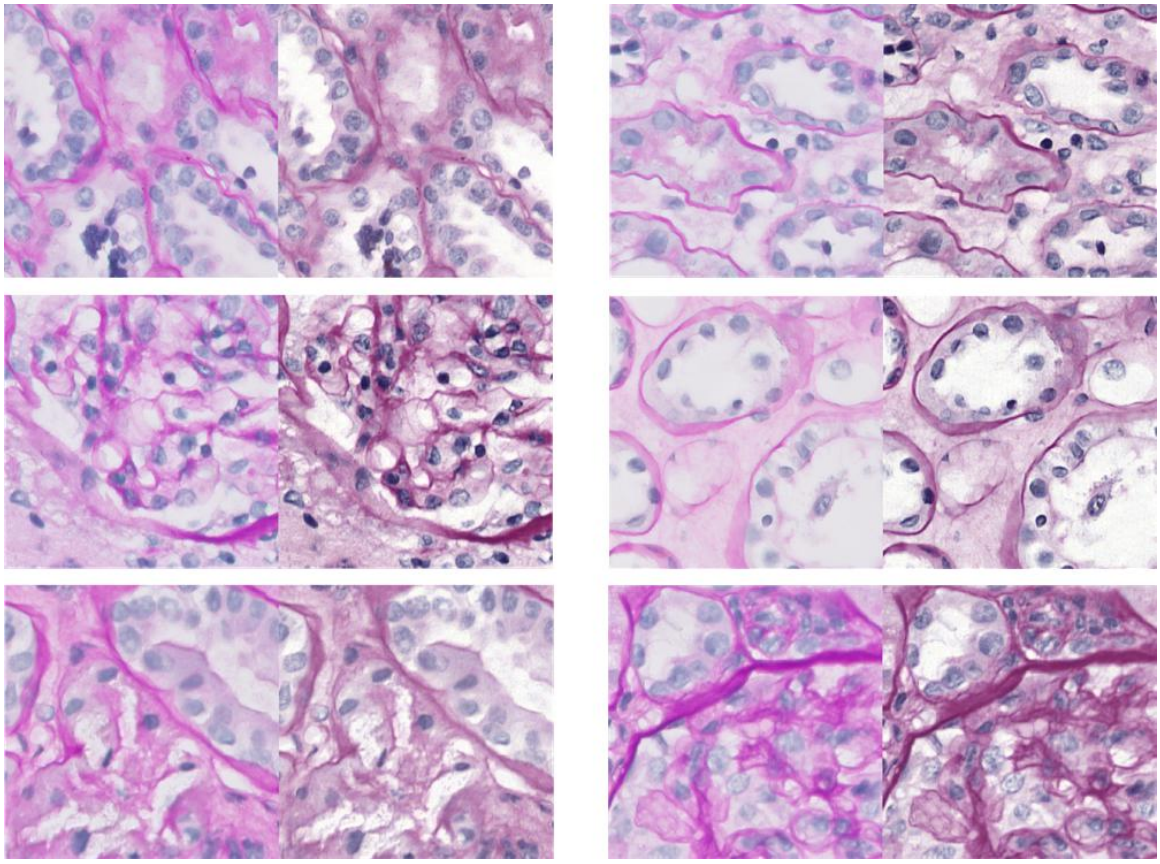

Figure 8: Additional samples of stain transformation. The leftmost image of each tissue pair is from the original AMC stain, the rightmost the synthetic RUMC stain.

Appendix B. Randomly sampled patches from the segmentation results

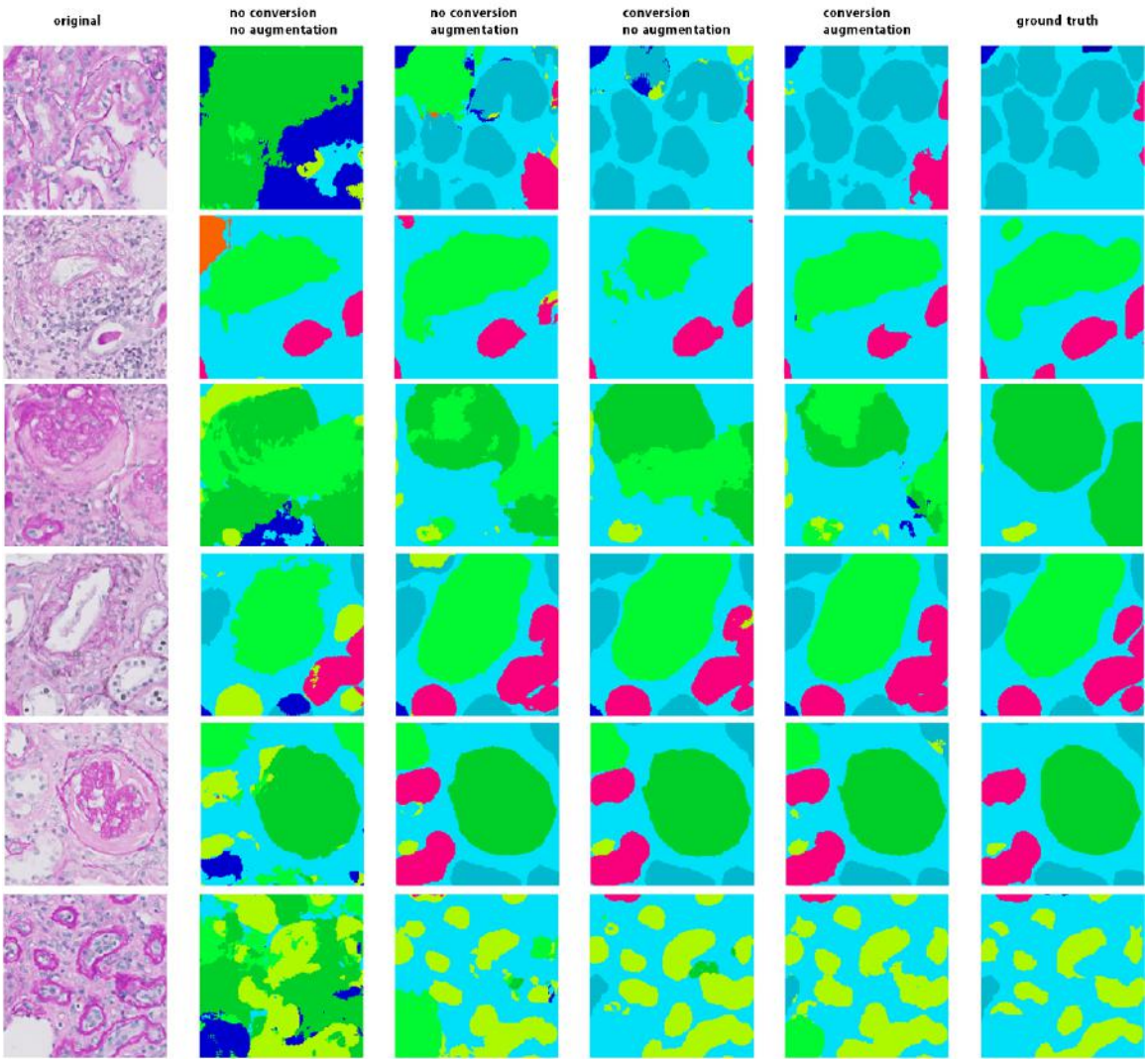

Figure 9: Additional samples of segmentation results.

Appendix C. Performance of the segmentation on the AMC dataset by class

Table 3: Dice coefficients of Sclerotic Glomeruli, Empty Glomeruli and Atrophic tubuli turned out considerably lower due to their low annotation count.

	aug, no conv	no aug, no conv	no aug, conv	aug, conv
Arteries	0.32	0.26	0.50	**0.51**
Atrophic tubuli	0.16	0.12	0.18	**0.19**
Background	0.79	0.49	0**.82**	**0.82**
Distal tubuli	0.64	0.23	**0.73**	0.71
Empty glomeruli	0.17	0.06	**0.24**	0.19
Glomeruli	0.78	0.45	0.88	**0.92**
Proximal tubuli	0.77	0.27	**0.85**	**0.85**
Sclerotic glomeruli	0.13	0.12	**0.32**	0.30