CycleGAN for virtual stain transfer: is seeing really believing?

Jelica Vasiljevic, Zeeshan Nisar, Friedrich Feuerhake, Cédric Wemmert, Thomas Lampert

CycleGAN for virtual stain transfer: is seeing really believing?

Jelica Vasiljević[a,b,c,*], Zeeshan Nisar[a], Friedrich Feuerhake[d,e], Cédric Wemmert[a], Thomas Lampert[a]

[a] *ICube, University of Strasbourg, CNRS (UMR 7357), France*
[b] *University of Belgrade, Belgrade, Serbia*
[c] *Faculty of Science, University of Kragujevac, Kragujevac, Serbia*
[d] *Institute of Pathology, Hannover Medical School, Germany*
[e] *University Clinic, Freiburg, Germany*

Abstract

Digital Pathology is an area prone to high variation due to multiple factors which can strongly affect diagnostic quality and visual appearance of the Whole-Slide-Images (WSIs). The state-of-the art methods to deal with such variation tend to address this through style-transfer inspired approaches. Usually, these solutions directly apply successful approaches from the literature, potentially with some task-related modifications. The majority of the obtained results are visually convincing, however, this paper shows that this is not a guarantee that such images can be directly used for either medical diagnosis or reducing domain shift. This article shows that slight modification in a stain transfer architecture, such as a choice of normalisation layer, while resulting in a variety of visually appealing results, surprisingly greatly effects the ability of a stain transfer model to reduce domain shift. By extensive qualitative and quantitative evaluations, we confirm that translations resulting from different stain transfer architectures are distinct from each other and from the real samples. Therefore conclusions made by visual inspection or pretrained model evaluation might be misleading.

*Corresponding author
Email address: jvasiljevic@unistra.fr (Jelica Vasiljević)

Keywords: CycleGAN, stain transfer, stain normalisation, image-to-image translation, digital histopathology

1. Introduction

Digital pathology has become a rich area of innovation in both clinical application and research. However, its crucial process of staining is known to be prone to high variation [1] due to differences in tissue preparation (exposure time, tissue fixation, section thickness etc), scanner characteristics (sensor, resolution, storage format, etc) or staining protocol. Examples of such variations for the case of kidey pathology are given in Figure 1. These differences can affect automatic systems [2] as they represent a source of domain shift [3]. A pathologist is able to correct for these variations due to experience, however current AI algorithms are not able to use such background knowledge. Thus, standardising the appearance of histological slides has become of great importance from both a diagnostic point-of-view and for the successful development and application of automated systems.

The standardisation is often addressed using computer vision techniques such as virtual staining — artificially changing the appearance of an image after its acquisition. Historically, research has focused on standardising the appearance of one particular stain, i.e. reducing the variation along the rows in Figure 1. This is usually referred in literature as stain normalisation. However, for the sake of better understanding, herein this is referred as intra-stain normalisation. Classical (non-deep) approaches to intra-stain normalisation use stain separation to isolate specific channels and then standardise the colour levels with respect to a reference image [4, 5, 6]. More recent approaches use machine learning or deep learning strategies to standardise image appearance [7, 8]. Nowadays, the problem of virtual staining is typically considered to be

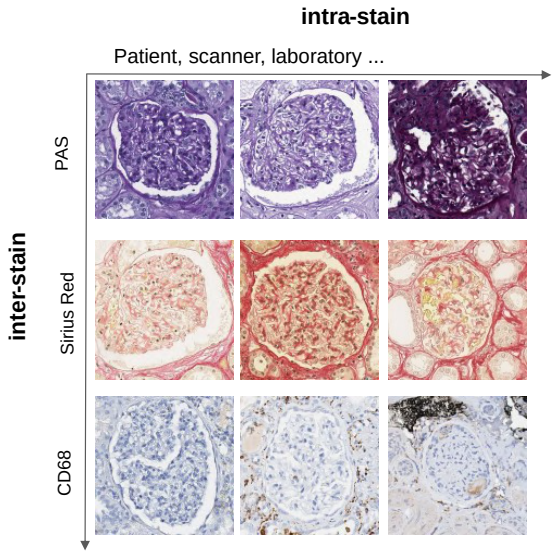


Figure 1: Examples of stain variability in kidney pathology. Intra-stain variation refers to differences in appearance of the same stain due to factors such as exposure time, tissue fixation, section thickness etc. Inter-stain variation refers to differences in appearance of the same tissue structures under different staining protocols e.g. glomeruli have different appearances in different stains.

a style-transfer problem, and a number of successful approaches based on style transfer techniques developed for natural images have been adapted to digital pathology [9, 10, 11, 12]. Their introduction however enabled the possibility to translate between two (or more) physically different stains (i.e. stain translation). With such models, it becomes possible to reduce the variation along the columns of Figure 1. This represents the second type of normalisation studied in this article, and is referred to as inter-stain normalisation. One of the most successfully and widely applied approaches is CycleGAN [13]— an unsupervised and unpaired method that enables virtual staining between two stainings without any additional effort for data preparation.

Many works attest that CycleGAN-based methods for virtual staining can achieve translations that are visually indistinguishable from real samples [12, 9, 10, 14, 15]. Furthermore, many works propose extensions to the original

CycleGAN architecture [16, 9], its loss function [12] or, with respect to a specific task, extend the training paradigm with additional modules [15, 17]. These works tend to rely on visual inspection to compare several approaches [18, 15], which may be unreliable [10]; or use consecutive slides stained differently [19] to validate the translation. However, such absolute comparisons are also limited since the staining process is prone to high variation and tissue structure will vary between consecutive slides.

As such, it is hard to quantitatively compare two methods or architectural changes. Moreover, assessing the quality of a translation is dependent on the purpose for which it will be used. Although CycleGAN based approaches have great success and the resulting translations look plausible[1], the use of artificially generated images is usually limited to the computer vision domain since in the medical sense, these images can be untrustworthy, e.g. it is known that such approaches can hallucinate features [20, 21] and thus can be unreliable for diagnostic purposes.

Assuming that the translation results in high fidelity, these methods are more often used in the computer vision domain to reduce domain shift [12, 22]; or as a domain augmentation strategy to reduce the need of additional annotations [10, 23]. Since these approaches are becoming more commonplace, and new possibilities are being explored such as multi-stain segmentation [10] or improving tumor classification [24], it is of a great importance to raise awareness of the sensitivity of such methods to some common, and rather small, changes.

As such, this article demonstrates that even the most simple architectural choice in CycleGAN-based models can play an important role in the ability of

[1] The term 'plausible' refers to the fact that an isolated histological image, without knowledge of adjacent sections processed with other staining modalities, and in absence of patient-specific information such as the underlying disease, looks visually correct to a trained expert with regard to the staining characteristics and the morphological appearance of the tissue components.

obtained models to reduce domain shift, even when visual appearance is not affected. Although most models produce plausible translations, i.e. those visually indistinguishable from real samples, the huge performance difference observed in pretrained models when applied to translated images, confirms that the quality of translations differ. In this study, the datasets are chosen to be as representative as possible, containing both histochemical (HC) and immunohistochemical (IHC) stains, and different directions of translations are investigated. In order to limit the number of experimental degrees-of-freedom, the modifications to the original CycleGAN architecture are restricted to the normalisation layer. In the original architecture Instance normalisation is used, in this study this is varied to other approaches commonly found in the literature: Batch, Layer, and Group. We show that the translations obtained by varying the normalisation layer belong to different distributions, and are distinct from those of real samples, causing pretrained models to perform badly.

Furthermore, since manual visual inspection cannot determine a difference in quality between the translations, it follows that visual inspection cannot be used as a validation criteria for virtual staining.

The main contributions of this article are:

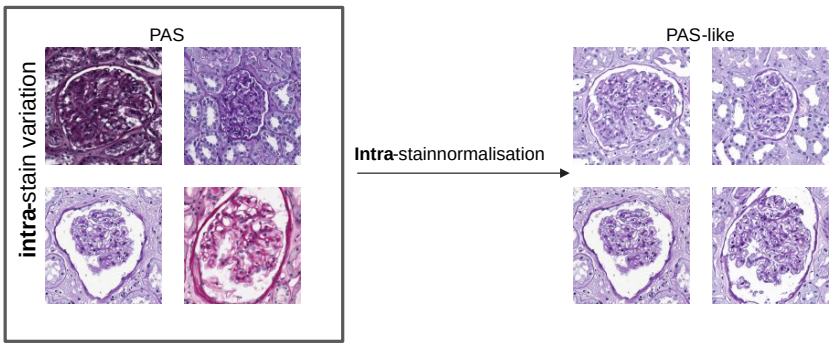
- To demonstrate that relatively small changes in CycleGAN-based methods, such as different normalisation layers, can have a great impact on translation quality, from the perspective of its ability to reduce a domain shift introduced by both inter- and intra-stain variation.
- To better define the limitations of visual inspection when assessing virtual staining.
- To give evidence that physical differences between stains, in addition to architectural choices, can play an important role when applying virtual stain transfer for reducing inter-stain domain shift.

- To show that generative approaches can be used to indicate whether a divergence from the true stain distribution has taken place or not when virtual staining is performed.

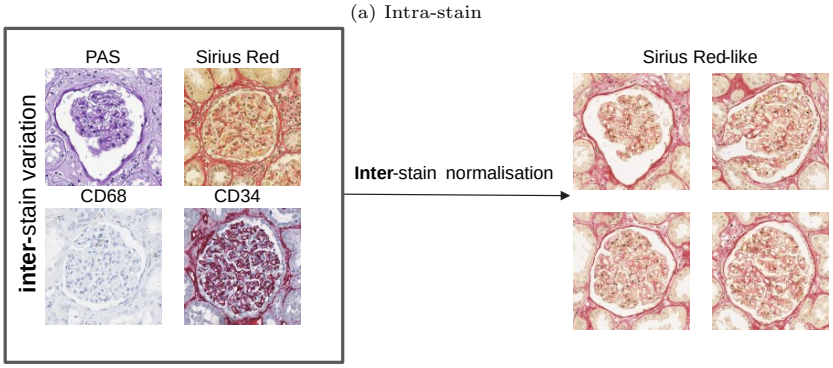
The remainder of this article is organised as follows: in Section 2 literature related to stain transfer, and particularly approaches which are based on the CycleGAN architecture, are reviewed. Section 3 gives a detailed description of the presented study and dataset. Section 4 presents the experimental results. Finally, Section 5 analyses stain translation models in terms of their visual quality, training stability, failure cases and generated data distributions.

2. Related work

Generally, two main sources of domain shift in Digital Pathology can be identified and these are illustrated in Figure 2: intra-stain variability, which represents the visual differences of one particular stain; and inter-stain variability, which is the result of the physical/chemical differences between stainings. Addressing inter-stain variability is of interest when tackling tasks that are solvable across various stains (such as glomeruli segmentation [10]) whereas intra-stain variability is more focussed in that it solves one task for a particular staining. Although from the computer vision viewpoint, intra-stain variation can be considered as a special case of inter-stain variability, the former is more often addressed in the literature since it represents a great obstacle in the reliable development of automated systems dedicated to a specific task (or stain). Such methods aim for a stain normalisation process that can be applied as a pre- of post-processing step. Traditional methods [5, 6, 25] rely on decomposing the image into concentration and colour matrices in order to stain new images using equivalent matrices from a reference image (stain). In addition to being dependent on the choice of a reference image, when it comes to a more challeng-



(a) Intra-stain



(b) Inter-stain

Figure 2: Illustration of the differences between intra-stain and inter-stain normalisation goals. Intra-stain aims to normalise image appearance within one stain (here PAS), while inter-stain normalisation aims for stain translation between different stains.

ing tasks such as inter-stain variation, such methods do not result in plausible outputs [26] (see definition in Introduction, page 4). GAN-based methods [27] have recently achieved great success in this area, particularly those for unpaired image-to-image translation, among which the CycleGAN [13] is the dominant approach [14, 9, 28, 29, 10, 30]. Many works rely on the original CycleGAN method with specific modifications, such as loss function [12], additional modules [15, 17], or changes in architectural design often related to the generator, such as using a U-Net [31] or a ResNet variation [12].

StainGAN [12] is one of the first applications of a CycleGAN for this task. The proposed solution contains the original CycleGAN architecture (ResNet

generator) but replaces the adversarial mean squared error loss function with classification loss [32]. Cai et al. [16] builds upon StainGAN by modifying the generator's architecture to that of a U-Net, and by reverting back to the mean squared error loss. Moreover, several works [33, 34] employ the original CycleGAN architecture and training procedure for normalising the H&E stain. Later, a variant of StainGAN, named StainNet [18], proposed to simplify the model using distillation learning to increase the speed of inference. Shrivastava et al. [15] propose a self-attention variant of the CycleGAN (U-Net-based) for stain normalisation. Mahapatra et al. [17] extend CycleGAN (U-Net-based) with a self-supervised segmentation module in order to incorporate semantic guidance into the translation process. Lahiani et al. [19] incorporate a perceptual embedding loss function in a ResNet-based CycleGAN. Moreover, Liu et al. [35] extend CycleGAN with specifically-designed pathology networks in order to ensure pathology preservation during translation. Bouteldja et al. [36] show that incorporating a pretrained segmentation module into the original CycleGAN architecture can enhance the translation process with semantic guidance, which is shown to be very beneficial in several tasks and stains. Ke et al. [37] add a self-supervised cluster label as additional input to one of the CycleGAN generators, using contrastive learning. On the other hand, de Bel et al. [31] change the goal of a generator from translation between two domains (stains) to predicting the residual between two stains by adding the original image to the output of the generator.

Overall, CycleGAN-based methods are the current state-of-the art for addressing both inter- or intra- stain variation. Some studies include pathologist evaluation in order to confirm the plausibility of the obtained translations, Lo et al. [30] have shown, in a specific setting of renal pathology, that the participating experts were not able to differentiate between real and artificially

produced microscopic kidney biopsy images.

Although all the previously mentioned works consider stain transfer as a style transfer problem, the application of such approaches to the medical domain opens new questions beyond those in the domain of natural images. It is particularly important to address the possibility of misinterpreting images produced by GAN-based models in the medical domain [20, 28, 31]. For example, one possible side-effect of such approaches is hallucination, which has been proven to be dangerous in medical imaging [20]. In the particular case of stain transfer, de Bel et al. [31] and Mercan et al. [28] showed that hallucination can occur during translation, even while producing plausible output. However, the degree, position and physical meaning of hallucination remains an open question. Moreover, Vasiljević et al. [21] showed that in the case of immunohistochemical stainings, the translation can contain information embedded as imperceptible noise that encodes information about the position of stain-specific markers. This information can be perturbed in a way to vary the position and number of such markers. All these findings indicate that assessing the quality of translations using visual inspection may be ill-advised. Thus, recent advances in stain transfer have moved towards disentangled learning approaches [38, 23] and using stain transfer to build models that are more robust to stain variations, e.g. as an augmentation strategy [10, 23].

The state-of-the-art architecture for style transfer [39] advocate the use of Instance normalisation [40] as it has been shown to have beneficial effects on the results. However, other stain translation approaches often deviate from this recommendation. For example, Shrivastava et al. [15] chose to use Batch normalisation after finding that it (empirically) outperforms Instance normalisation when using their self-attention based architecture. Lahiani et al. [19] noted that Instance normalisation forces pixels in the output patch to be dependent on

the statistics of the entire patch, resulting in a visible tailing effect in the final WSI image. Thus, several approaches have been proposed to mitigate this problem when the goal is WSI reconstruction [19, 41]. Nevertheless, when it comes to justifying the choice of architecture and loss function, visual inspection is the default assessment. New stain transfer approaches, although based on style transfer, do not stress the choice of a normalisation layer. Thus, some approaches do not use any normalisation layer [42, 43], while others use Batch normalisation [16, 15], Group normalisation [17] or Instance normalisation [31]. Usually, articles report measures such as Structural Similarity Index (SSIM) [16, 19] in order to justify the benefits of a proposed approach using either ground truths (in case of different scanners) or after non-rigid registration of a consecutive slide. However, it is not guaranteed that such comparisons reflect translation quality since the considered ground truth captures just one modality of all possible variations in that stain and there is no valid 'gold-standard' appearance of a stain.

Through extensive experiments this article shows that a simple design choice, such as the choice of normalisation layer, can play an important role in the quality of the obtained translations from the perspective of reducing domain shift. Most of the time, such modifications do not dramatically affect visual appearance, providing evidence that visual assessment cannot serve as a valid indicator of translation quality. Moreover, it shows that the importance of proper architectural choices are correlated with the biological difference between the stains to be translated.

3. Methods

In order to demonstrate the sensitivity of virtual stain transfer to the underlying architecture, the ubiquitous CycleGAN architecture is taken and different

stain transfer models are created by replacing the normalisation layers in both the discriminators and the generators. To quantitatively validate the obtained translations, their ability to reduce domain shift introduced by inter- or intra-stain variation is measured for the task of glomeruli segmentation. This is achieved using the well-established U-Net [44] architecture (see Appendix Appendix A.2 for training details). It is trained on real images from one stain type, then its performance when applied to images that have been translated to match that stain type is measured.

This approach has already been applied as a validation technique in the literature [9, 10, 31, 17]. In order to be consistent with the literature [9, 10], this is referred to as MDS1 (Multi-Domain Supervised, MDS). Also, for ease of reading, the stain on which the segmentation model is trained in a supervised manner is referred to as the source stain, and the stain that is translated to the source stain during application as the target stain. Using MDS1, the problem of reducing inter- and intra- stain variability, i.e. stain translation and stain normalisation (respectively), will be tackled with datasets chosen to be as representative as possible. More precisely, it contains histochemical (HC) stains: Periodic Acid Schiff's reagent (PAS), Jones Hematoxilin-Eosin (H&E), Sirius Red; immunohistochemical (IHC) stains CD68 and CD34; and variations of PAS from the publicly available dataset AIDPATH [45]. More details about the dataset is given in Section 3.3.

In the following, the term 'translation quality' refers to an architecture's ability to produce images that reduce domain shift when deep models are used for solving a segmentation task. These artificially produced images for stain X are X-like (see Figure 2, X is represented on the left side (real images) and X-like (translated) images are on the right side), they are not real representations of the stain and thus they cannot replace the real staining process or be directly

used for diagnostic purposes.

Two sets of complementary hypotheses concerning what can affect the performance measured in this setting are identified as follows.

1. Pretrained model

    1.1. Short-cut learning [46] in pretrained models: a model makes a decision based on some source dataset characteristics that are not necessarily related to the given problem. Thus, if the translated images do not contain the shortcut characteristics, the pretrained model will not perform well.

2. Stain transfer.

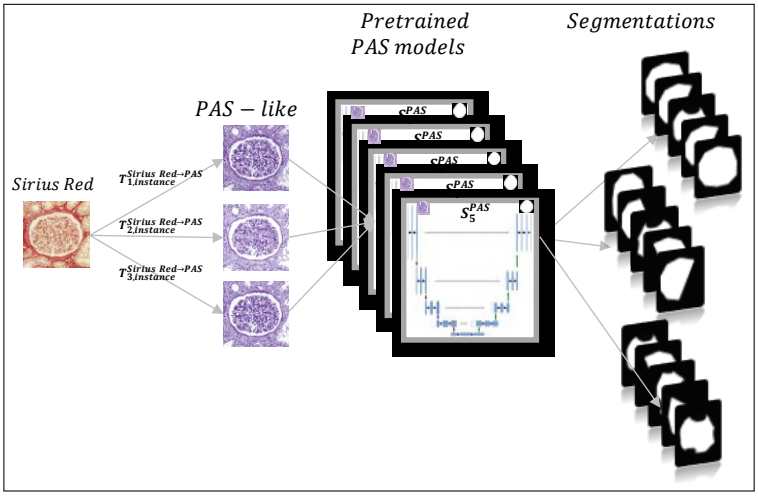
    2.1. Stain transfer model: the model's ability to produce accurate translations between the target and source stains should impact downstream task performance.

    2.2. Direction of translation: some stain translation directions may be harder (e.g. translation from a general purpose stain to a specific stain).

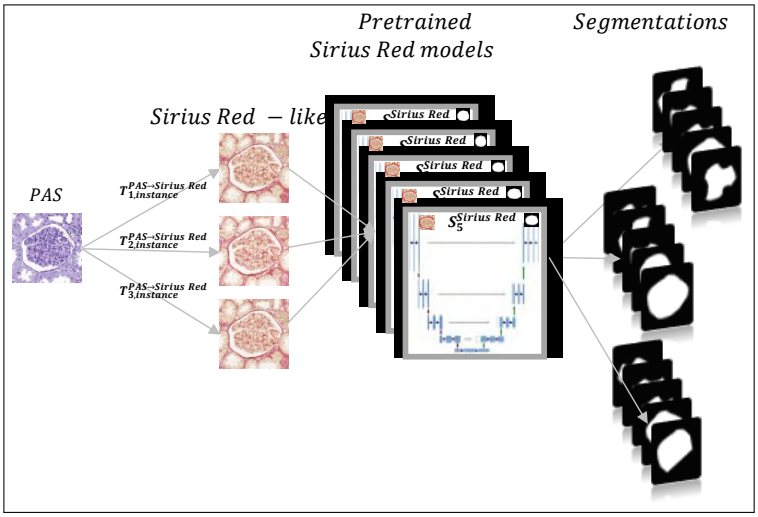
In order to test these hypotheses, several experiments are conducted, as illustrated in Figure 3 (stains taken for illustration are PAS and Sirius red, translation model has Instance normalisation layer. The same experiments are performed for other combinations).

In the case of inter-stain variability, this analysis is performed from two perspectives:

- Target to PAS (Figure 3a): Five PAS segmentation models are trained (S_1^{PAS}, S_2^{PAS}, ..., S_5^{PAS}) and their performance are evaluated on translations from four other stainings. For each normalisation layer and



(a) Target to PAS experiments illustration



(b) PAS to Target experiments illustration

Figure 3: Experiments design illustration: (a) Target to PAS – models trained on PAS data are evaluated on different target stains and translation models; (b) PAS to Target – models trained on target stain are evaluated on different translations from PAS stain.

each stain, three translation models are trained, i.e. $T_{1,n}^{x\to\text{PAS}}$, $T_{2,n}^{x\to\text{PAS}}$, $T_{3,n}^{x\to\text{PAS}}$, where $x \in \{\text{Jones H\&E, Sirius Red, CD68, CD34}\}$ and $n \in \{\text{Instance, Batch, Layer, Group}_8\text{, Group}_{16}\text{, Group}_{32}\text{, None}\}$. In this way, one pretrained segmentation model, e.g. S_1^{PAS}, is applied to the trans-

lations from all stains, $T_{i,n}^{x \to \text{PAS}}$ allowing analysis of the effects of a stain transfer model (Hypothesis 2.1) and target stain (Hypothesis 2.2). The three stain translation models that are obtained for each combination of target stain and normalisation layer allow the measurement of short-cut learning (Hypothesis 1.1). The standard deviation in MDS1 performance using different stain translation models obtained in the same experimental setting can be attributed to a pretrained model's bias.

- PAS to target (Figure 3b): Five segmentation models are trained for each of the target stains (S_1^x, S_2^x, ..., S_5^x), where $x \in \{\text{Jones H\&E, Sirius Red, CD68, CD34}\}$. These are evaluated on translations from PAS to each stain, using different stain translation models $T_{i,n}^{\text{PAS} \to x}$, where $n \in \{\text{Instance, Batch, Layer, Group}_8, \text{Group}_{16}, \text{Group}_{32}, \text{None}\}$ and $i \in \{1, 2, 3\}$. As such, the test images and the segmentation models within one target stain remain constant, and therefore the variation in results within stains can be attributed to translation quality (Hypothesis 2.1). Moreover, by comparing to the results from the previous experiment, the influence of translation direction can be investigated (Hypothesis 2.2). Similarly as previously stated, the standard deviation within several runs of the same experimental setting can be related to a short-cut learning (Hypothesis 1.1).

In the case of intra-stain variability, we measure the PAS pretrained models' sensitivity to the translation (stain normalisation) from the publicly available AIDPATH dataset [45]. From this perspective, Hypotheses 1.1 and 2.1 can be investigated.

3.1. CycleGAN models

CycleGAN has been shown to be an effective way to obtain plausible stain transfer (see definition in Introduction, page 4) [14, 47, 9, 41, 28, 12]. The overall architecture contains two generators: $G_{AB} : A \to B$ and $G_{BA} : B \to A$; and two discriminators D_A and D_B. The aim of generator G_{AB} (G_{BA}) is to translate an image originally stained with stain A (B) to appear as though it had been stained with stain B (A). In an adversarial manner, the aim of D_A (D_B) is to distinguish between real patches stained with stain A (B) and those translated from stain B (A) to A (B).

Originally, the CycleGAN architecture contains Instance normalisation layers in both the generator and the discriminator. Since there is plenty of variation in the literature regarding the loss function, architecture, and normalisation strategy, the following experiments use the originally proposed architecture (ResNet-9 generator, PatchCNN discriminator) and mean-squared error as the loss function. The variation in each model is introduced by changing the normalisation layer, by omitting Instance normalisation completely or by replacing it with Batch, Layer, or Group normalisation in both the generator and discriminator.

Since the CycleGAN model is able to translate between two stains, separate CycleGAN models are trained for each pair of stainings. More training details are given in Appendix A.1.

3.2. Normalisation techniques

Several normalisation strategies have been introduced in order to improve learning in specific tasks. In this study, however, we focus on the most popular strategies—Batch [48], Instance [40], Layer [49], and Group [50] normalisation.

In the case of 2D images a feature computed by a model's layer, x, is a 4D tensor $x = (N, C, H, W)$ where N denotes batch size, C is the number of

channels and H and W are spatial height and width. A normalisation layer performs normalisation of x such that

$$\hat{x} = \frac{x - \mu_{norm}}{\sigma_{norm}}, \tag{1}$$

where μ_{norm} and σ_{norm} are the mean and standard deviation computed over different axes depending on the normalisation technique used.

In the case of Batch Normalisation (BN) [48], μ_{norm} and σ_{norm} are computed channel-wise, along the (N, H, W) axes, thus normalising all feature elements that share the same channel across a batch. Layer Normalisation (LN) [49], calculates μ_{norm} and σ_{norm} over the (C, H, W) axes, normalising features for each sample in a batch separately. Instance Normalisation (IN) [40] computes μ_{norm} and σ_{norm} across the (H, W) axes, thus normalising features for each sample and each channel separately. Similarly to Layer Normalisation, Group Normalisation [50] computes μ_{norm} and σ_{norm} over the (H, W) axes, but instead of normalisation over all channels, a specific number of groups of adjacent channels is chosen. Thus, when the number of groups is equal to 1, GN becomes LN, and it reduces to IN when the number of groups is equal to the number of channels. Thus, the number of groups is a hyperparameter of this layer. In the literature, it is usually chosen to be a factor of 2, and herein groups of 8, 16 and 32 are tested (32 being the maximum possible due to the minimal number of filters used in the CycleGAN convolutional layers).

3.3. Data

3.3.1. Inter-stain

Tissue samples were collected from a cohort of 10 patients who underwent tumor nephrectomy due to renal carcinoma. The kidney tissue was selected as distant as possible from the tumors to display largely normal renal glomeruli,

some samples included variable degrees of pathological changes such as full or partial replacement of the functional tissue by fibrotic changes ("sclerosis") reflecting normal age-related changes or the renal consequences of general cardiovascular comorbidity (e.g. cardial arrhythmia, hypertension, arteriosclerosis). The paraffin-embedded samples were cut into $3\mu m$ thick sections and stained with either Jones H&E basement membrane stain (Jones), PAS or Sirius Red, in addition to two immunohistochemistry markers (CD34, CD68), using an automated staining instrument (Ventana Benchmark Ultra). Whole slide images were acquired using an Aperio AT2 scanner at $40\times$ magnification (a resolution of $0.253\mu m$ / pixel). All the glomeruli in each WSI were annotated and validated by pathology experts by outlining them using Cytomine [51]. The dataset was divided into 4 training, 2 validation, and 4 test patients. The number of glomeruli in each staining dataset is given in Table 1.

Staining	**Training**	**Validation**	**Test**
PAS	662	588	1092
Jones H&E	624	593	1043
Sirius Red	654	579	1049
CD34	568	598	1019
CD68	529	524	1046

Table 1: Number of glomeruli in each staining dataset.

Glomeruli segmentation is framed as a two class problem: glomeruli (pixels that belong to glomerulus), and tissue (pixels outside a glomerulus). The training set comprised all glomeruli from a given staining's training patients plus seven times more tissue (i.e. non-glomeruli) patches (to account for the variance observed in non-glomeruli tissue). During the inter-stain experiments, this dataset is referred to as 'Hanover dataset'.

3.3.2. Intra-stain

For intra-stain analysis, the publicly available AIDPATH dataset [45] is used. AIDPATH is a collection of five different datasets of human kidney tissue cohorts acquired and digitised from three European institutions: Castilla-La Mancha's Healthcare services (Spain), The Andalusian Health Service (Spain) and The Vilnius University Hospital Santaros Klinikos (Lithuania). Tissue samples were collected with a biopsy needle having an outer diameter between $100\mu m$ and $300\mu m$ and paraffin blocks were prepared using tissue sections 4µm thick, then stained using PAS [45, 52]. In total, the dataset contains 47 WSIs. The data from Castilla-La Mancha's Healthcare services (SESCAM) were used in this study[2] since it represents the greatest difference to the PAS staining present in the intra-stain dataset. All slides were manually annotated and the same data extraction strategy was applied as previously mentioned.

In all experiments, patches of size 512×512 pixels (at 20x magnification) are used since glomeruli and part of the surrounding fit within this size of patch at the level-of-detail used. Glomeruli segmentation is framed as a two class problem: glomeruli (pixels that belong to glomerulus), and tissue (pixels outside a glomerulus). Separate models are trained for each of available stainings.

4. Results

4.1. Inter-Stain Variability

The translations obtained by many of the stain transfer models are plausible (see definition in Introduction, page 4), as will be discussed in more details in Section 5.1.1. Nevertheless, the quantitative analysis performed using pre-trained models shows that there are significant differences between their ability to reduce domain shift. Here, two directions are taken: by evaluating the PAS

[2]Specifically images 1, 3 and 7.

Normalisation Layer	Test Staining				Average
	Jones H&E → PAS	Sirius Red → PAS	CD68 → PAS	CD34 → PAS	
Instance	0.849 (0.017)	0.870 (0.009)	0.684 (0.043)	0.754 (0.008)	0.789 (0.087)
Batch	0.339 (0.059)	0.508 (0.041)	0.002 (0.001)	0.400 (0.067)	0.312 (0.218)
Layer	0.816 (0.014)	0.832 (0.005)	0.167 (0.046)	0.754 (0.024)	0.642 (0.319)
Group_8	0.848 (0.011)	0.810 (0.006)	0.308 (0.101)	0.628 (0.040)	0.649 (0.246)
Group_{16}	0.849 (0.011)	0.800 (0.036)	0.486 (0.060)	0.650 (0.039)	0.696 (0.163)
Group_{32}	0.815 (0.007)	0.807 (0.017)	0.546 (0.049)	0.737 (0.015)	0.726 (0.125)
None	0.770 (0.003)	0.730 (0.035)	0.250 (0.028)	0.747 (0.047)	0.624 (0.250)
Average (excl. BN)	0.824 (0.031)	0.808 (0.046)	0.407 (0.197)	0.712 (0.057)	

Table 2: MDS1 F_1 scores with different CycleGAN normalisation layers (target stain translated to PAS). The values represent the average of 5 pretrained segmentation models (S_1^{PAS}, S_2^{PAS}, …, S_5^{PAS}), each applied to 3 repetitions of the translation model training ($T_{1,n}^{\text{y}\to\text{PAS}}$, $T_{2,n}^{\text{y}\to\text{PAS}}$, $T_{3,n}^{\text{y}\to\text{PAS}}$), therefore the average and standard deviations (in parenthesis) of 15 repetitions in total. The last row represents row-wise averages, excluding Batch normalisation results, since translations obtained by these models are often not plausible.

model's performance on translations from the target stains to PAS (see Table 2); and by testing the models pretrained on each target stain to translations of PAS images (see Table 3). The results presented in each table are the averages over three separate CycleGAN models, each applied to a five pretrained baseline models. The performance of the baseline models are given in Table 4, which serves as proof that the problem is solvable with high accuracy in all stainings.

Since all the results in Table 2 are calculated using the same PAS pretrained models, they can be used to determine the sensitivity of such models to: (column-wise) different types of normalisation (in which the translated stain, and therefore test images, in addition to the pretrained models are fixed); and (row-wise) different translation models having the same normalisation strategies. As is established in the style-transfer literature, Instance normalisation achieves the best overall performance, although in some cases other normalisation strategies achieve similar performance. For example, with CD34, Instance, Layer, Group_{32} and None (without a normalisation layer) all achieve similar results, whereas in CD68 Instance norm is the clear winner. This indicates that the choice of architecture is dependent on the stain, and most likely, therefore,

Normalisation Layer	**Test Staining**				**Average**
	PAS → Jones H&E	PAS → Sirius Red	PAS → CD68	PAS → CD34	
Instance	0.891↑ (0.001)	0.744↓ (0.079)	0.630↓ (0.019)	0.641↓ (0.087)	0.726↓ (0.121)
Batch	0.134↓ (0.022)	0.002↓ (0.001)	0.133↑ (0.087)	0.049↓ (0.008)	0.079↓ (0.066)
Layer	0.879↑ (0.002)	0.172↓ (0.080)	0.459↑ (0.111)	0.524↓ (0.106)	0.509↓ (0.291)
Group_8	0.873↑ (0.008)	0.470↓ (0.387)	0.444↑ (0.053)	0.373↓ (0.078)	0.540↓ (0.226)
Group_{16}	0.876↑ (0.002)	0.118↓ (0.025)	0.423↓ (0.121)	0.503↓ (0.106)	0.480↓ (0.312)
Group_{32}	0.883↑ (0.006)	0.320↓ (0.198)	0.577↑ (0.068)	0.377↓ (0.269)	0.539↓ (0.255)
None	0.862↑ (0.009)	0.075↓ (0.055)	0.568↑ (0.078)	0.483↓ (0.115)	0.497↓ (0.325)
Average (excl. BN)	0.877↑ (0.010)	0.316↓ (0.255)	0.517↑ (0.085)	0.483↓ (0.100)	

Table 3: MDS1 F_1 scores with different CycleGAN normalisation layers (PAS translated to target stains). The values represent the average of 5 pretrained segmentation models ($(S_1^x, S_2^x, \ldots, S_5^x)$, where $x \in \{\text{Jones H\&E}, \text{Sirius Red}, \text{CD68}, \text{CD34}\}$), each applied to 3 repetitions of the translation model training ($T_{1,n}^{\text{PAS}\to x}$, $T_{2,n}^{\text{PAS}\to x}$, $T_{3,n}^{\text{PAS}\to x}$), therefore the average and standard deviations (in parenthesis) of 15 repetitions in total. ↑ indicates improved performance compared to the reverse translation, see Table 2, and a ↓ a decrease in performance.

PAS	**Jones H&E**	**Sirius Red**	**CD68**	**CD34**	**Overall**
0.907 (0.009)	0.864 (0.011)	0.867 (0.016)	0.853 (0.018)	0.888 (0.015)	**0.876 (0.022)**

Table 4: F_1 scores for the baseline results (standard deviations are in parentheses).

the complexity of the translation required. However, the fact that none of the pretrained models applied to CD34 and CD68 translations are able to achieve baseline results indicates that either the pretrained PAS models are sensitive to some features not captured by the translation models, and/or the translation models induce a domain-shift.

This can be explained, to some extent, by the difference between IHC and HC stainings since PAS, Jones H&E and Sirius Red use chemicals that interact with several tissue components and multiple normalisation strategies are able to approach baseline performance. On the other hand, CD34 and CD68 are designed to detect specific proteins and here, performance varies greatly.

The results in each column of Table 3 are calculated using the same pretrained segmentation model but now on the target stains, therefore each column represents a different model tested on the same PAS data translated to each target stain. As such they complement the conclusions from Table 2, that is from the target stain perspective, by representing the sensitivity of the pretrained target models to different normalisation strategies. For example, it becomes clear that the normalisation strategy has very little effect when applying the Jones H&E segmentation models to the PAS translations (except with Batch normalisation). The row-wise results are calculated, again using the same PAS images but now translated to different stains, and therefore different pretrained models are used.

As previously discussed, it seems that in this particular application differences in staining type (e.g. HC vs IHC) can play an import role regarding a sensitivity of pretrained model to translations obtained by different stain transfer models.

Comparing the performances between Tables 2 and 3 represents the two directions of the same translation (PAS $\rightarrow$ Target and Target $\rightarrow$ PAS). Overall,

better results are obtained when translating in the Target → PAS direction, which could be related to the fact that the translation difficulty is not symmetrical. Even when accounting for the fact that segmentation is more difficult in non-PAS stains (see Table 4), more significant drops in performance are observed between Tables 2 and 3. The differences between performance are indicated by an up or down arrow in each cell of Table 3, representing an increase or decrease compared to Table 2. When translating from a general staining such as PAS to more specific stainings, the translation model must 'invent' stain-specific markers since they are not specifically marked in the general-purpose stain. Thus, this direction of translation can be harder than the other way-around and the translation model may fail to reconstruct the finer details that the pretrained segmentation model relies on. Moreover, these pretrained segmentation models could be biased towards stain-specific markers (e.g. due to short-cut learning) and thus its performance can be highly dependent on translation quality. Evidence for this is given by the large standard deviations observed when applying the same translation architecture to the same pretrained segmentation models (e.g. Table 3, Sirius Red, Group_8 and Group_{32}). When the translations contain the specific features focused on by the pretrained models, they perform well (e.g. the best performing translation model in Group_8 achieves an average segmentation score of 0.776), otherwise the translated images can be seen as out-of distribution examples in which the segmentation model fails (the worst translation model in results in Group_8 achieves an average segmentation score of 0.034), even though the translations appear plausible, see Figure 4.

Additional evidence for this will be given in Section 5.1.2 when the segmentation model's variance will be considered from the perspective of stain translation model training.

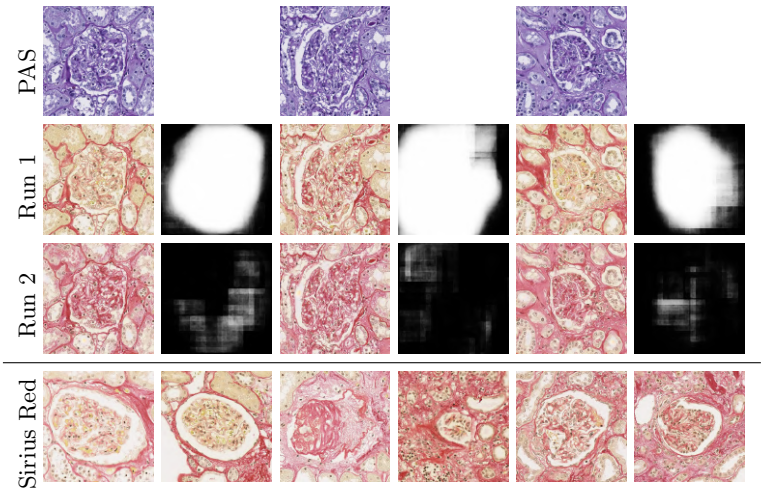


Figure 4: PAS patches translated to Sirius Red with two repetitions of the CycleGAN (Group_8) model alongside corresponding segmentations from a pretrained Sirius Red model. The last row represents real patches from the Sirius Red domain.

4.2. Intra-stain variability

In the case of intra-stain variability, the same pretrained PAS segmentation models used previously are evaluated on the AIDPATH dataset containing PAS-stained WSIs from the Servicio de Salud de Castilla-La Mancha (SESCAM) (see Figure 5 for a visual comparison between the two datasets). Direct application of the segmentation models to this variation of PAS is not successful, missing the majority of glomeruli (see the vPAS column in Table 5), which confirms the need of a stain normalisation procedure. As in Section 4.1, CycleGAN models were trained to translate the AIDPATH dataset to the source PAS dataset, using different normalisation strategies. Full training details are given in Appendix Appendix A.1.

Table 5 presents these results, in which it can be observed that the normalisation strategy also has an important role when performing stain normalisation and segmentation performance does not correlate with visual quality, see Figure

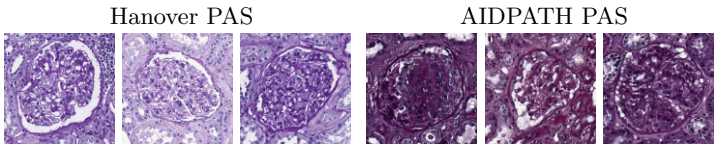


Figure 5: Glomeruli PAS variation between Servicio de Salud de Castilla-La Mancha (SESCAM) and Hanover.

6.

Score	vPAS	Instance	Layer	Batch	Group_8	Group_{16}	Group_{32}
F_1	0.183 (0.091)	0.351 (0.042)	0.504 (0.029)	0.532 (0.034)	0.223 (0.053)	0.236 (0.046)	0.282 (0.019)
Precision	0.229 (0.175)	0.819 (0.028)	0.806 (0.024)	0.434 (0.047)	0.680 (0.119)	0.633 (0.105)	0.775 (0.044)
Recall	0.385 (0.256)	0.226 (0.035)	0.370 (0.033)	0.738 (0.012)	0.135 (0.034)	0.148 (0.031)	0.174 (0.015)

Table 5: Stain normalisation, the effects of different CycleGAN normalisation layers on the F_1 scores of pretrained PAS models.

The results presented in Table 5 should be interpreted with caution. Since the AIDPATH dataset is composed of biopsies, the number of glomeruli in each image is an order of magnitude smaller then in the Hanover dataset. Thus, a small number of false positives (or negatives) has a big effect on the overall score. Also, the images contain a significant portion of sclerotic glomeruli which is not the case in the Hanover dataset and therefore lower segmentation performance should be expected due to dataset bias. For example, translation models with Batch normalisation obtain the best overall recall, i.e. lowest rate of false negatives, which means that the segmentation masks predicted by pretrained models cover the majority of glomeruli. However, its low precision indicates that there are more false positives, i.e. more structures are wrongly classified as glomeruli. Contrarily, Instance normalisation has the best overall precision, meaning that the pretrained models produce fewer false positives, but the detection is less robust, i.e. not all of the glomeruli structures are detected.

Nevertheless, this study is concerned with performance relative to each nor-

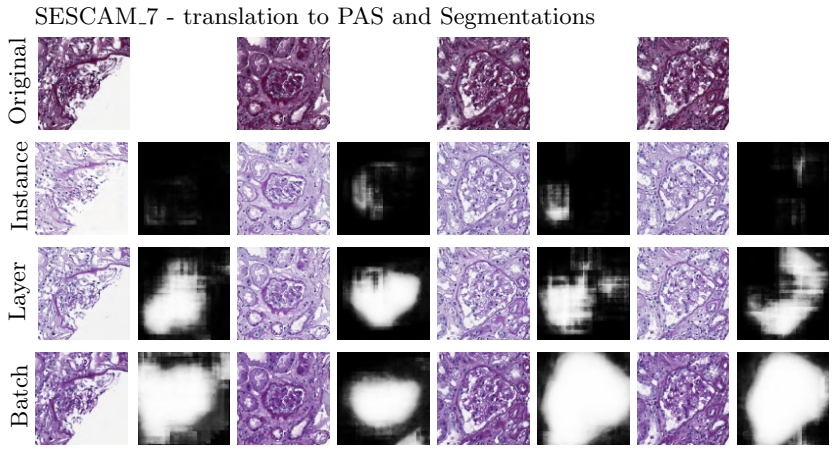


Figure 6: Glomeruli patches extracted from SESCAM_7 image (first row) and their translations to our dataset PAS using different CycleGAN models trained on SESCAM_1 and SESCAM_3 images, with corresponding segmentations from a pretrained segmentation model on our dataset.

malisation strategy and since the same pretrained models are used for these evaluations, the effect of the translation model is evident. Taken together with the results of inter-stain variability (Section 4.1), there is no 'golden' rule for the best choice of normalisation strategy, and it is rather dependent on the problem at hand.

5. Discussion

In this section, qualitative and quantitative assessments of the stain transfer models will be presented. The qualitative analysis includes visual assessment, which is presented in Section 5.1.1. However, the findings in Section 4 give strong evidence that this cannot be relied upon. Section 5.1.2 will further demonstrate this by highlighting the model's instability during different training stages. Moreover, Section 5.1.3 presents some failure cases that can be easily overlooked by non-experts. The quantitative analysis includes assessment via evaluation approaches found in the literature [35, 30], which are given in Section

5.2.1, and a comparison of image distributions is presented in Section 5.2.2. Furthermore, some guidelines about the clinical usage of artificially stained images is presented in Section 5.3.

5.1. Qualitative analysis

5.1.1. Visual quality

Figure 7 illustrates the visual quality of the obtained translations, in which each staining has been translated to PAS using different CycleGAN models. Furthermore, Figure 8 presents the translations of a PAS patch to each of the target stainings. Visually, all translations (except Batch normalisation) look plausible (see definition in Introduction, page 4). Of course, not every normalisation type produces the same output since the translation between stains is not a one-to-one mapping. This is more noticeable in stains CD34 or CD68. Nevertheless, these variations fall within the range of those that can occur naturally. Furthermore, the same variations can be observed for one translation model in different epochs, or different training repetition, as shown on Figure 9.

However, it is significant to note that the translation process can greatly affect the appearance of stain-specific markers in CD68 and CD34. There is an important difference between IHC staining methods highlighting only one specific protein with a chromogenic label, as opposed to HC staining methods that are less specific but nevertheless may result in color enrichment for certain anatomical structures. In the current example, brown IHC staining (CD68) reflects expression of a specific protein during macrophage differentiation and activation, whereas gradual enrichment of purple staining as result of the chemical PAS reaction reflects the presence of carbohydrate macromolecules that are not specific for macrophages, but enriched in their phagocytic subset and thus associated with protein degradation. Thus, both methods highlight slightly different populations of macrophages, illustrating the important caveat of translating HC

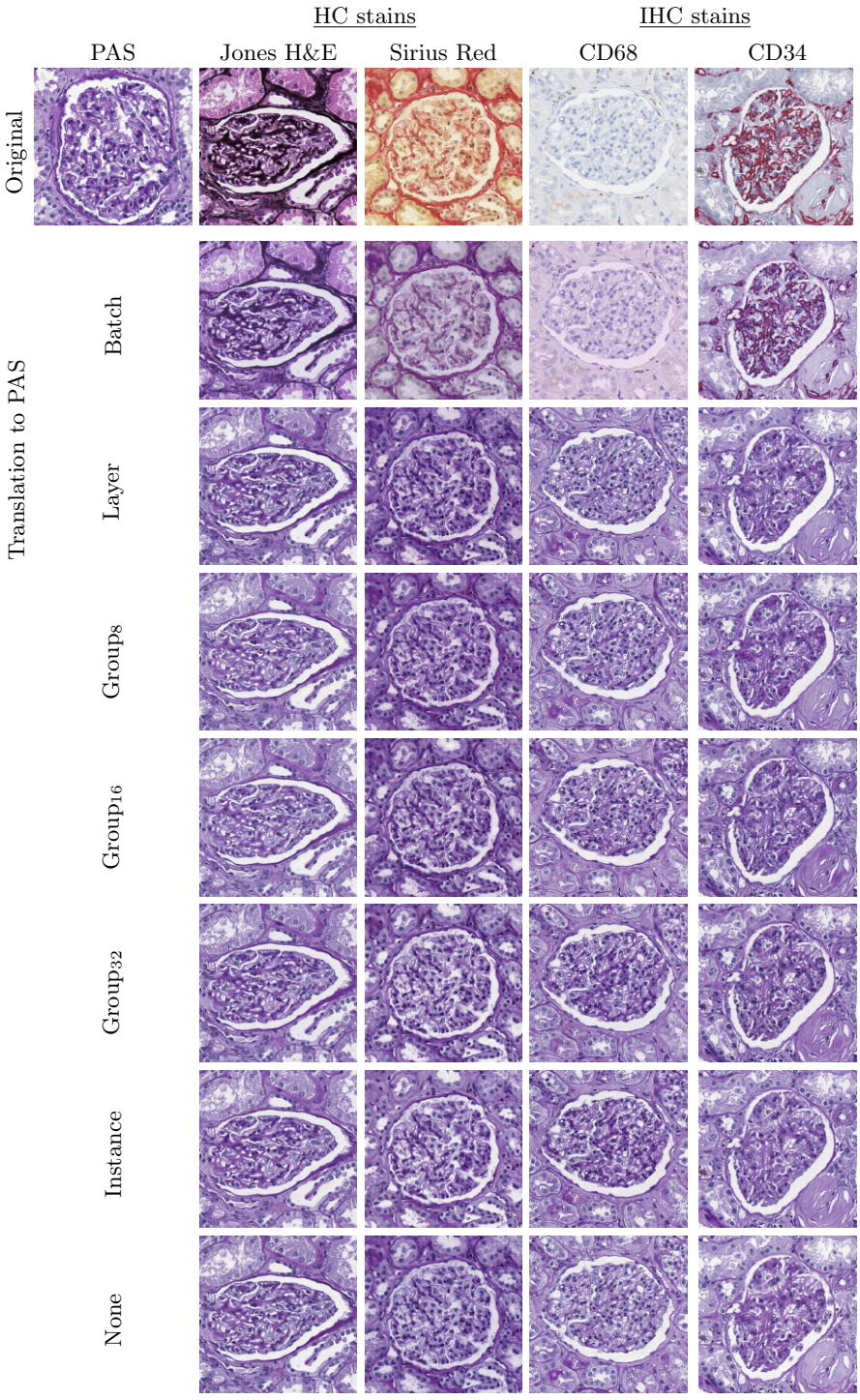


Figure 7: Target stain patch translated to PAS using CycleGAN models trained with different normalisation layers.

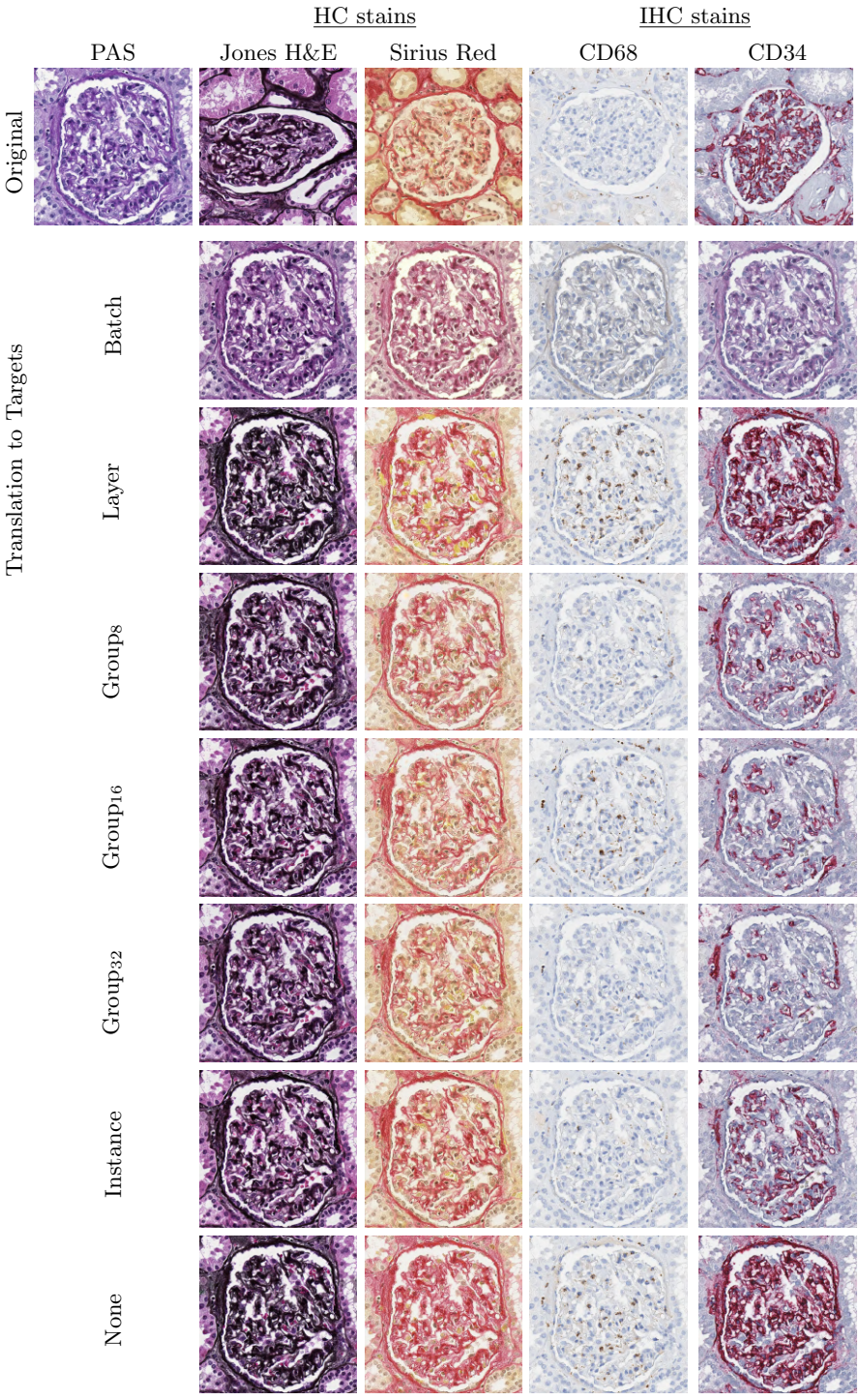


Figure 8: PAS patch translated to target stain using CycleGAN models trained with different normalisation layers.

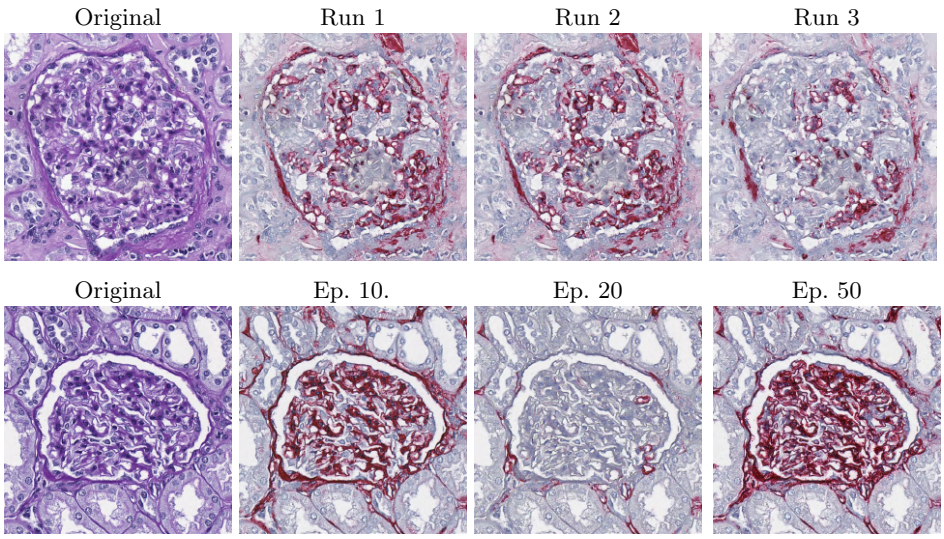


Figure 9: An illustration of inter-stain variance, PAS patches translated to the CD34 target stain using (first row) CycleGAN with instance norm from the 50th epochs in three separate training repetitions; (second row) CycleGAN with layer norm from different epochs of the same training run.

in IHC and vice versa: the translation result looks "plausible" (see definition in Introduction, page 4). In this context this means that their visual appearance is consistent with the target staining and reflects regular morphological features of macrophages such as size and shape. However, this visual plausibility may not be accurate for biomedical evaluation. Examples include biological features such as "macrophage activation" (reflected by expression of the CD68 protein on the cell surface), or "protein digestion by macrophages" (reflected by PAS-positive granular substance within the macrophage cytoplasm). The same holds true for translations between CD34 and PAS: both methods can highlight blood vessels, CD34 (IHC) specifically the inner layer (endothelium) and PAS (HC) less specifically components of the entire vessel wall. Again, translations between stains in any direction are likely to look "plausible" and may even be useful for general visual detection of blood vessels, but they are clearly misleading for

other purposes (e.g. specific evaluation of endothelial pathology).

Moreover, a general biological aspect makes visual comparison between different stain transfer models, such as the one done in [35], more unreliable due to technical consideration. Therefore, drawing general conclusions about model capacity may lead to incorrect findings. For example, Liu et al. [35] observe such behaviour when CycleGAN models are trained on different datasets (e.g. Figure 6 and 9 of [35]), which could be explained by this observation.

5.1.2. Training Stability

As previously noted by several authors, CycleGAN-based stain transfer is able to reach plausible translations early during training [14, 10]. Since there is no explicit stopping criteria in the training process, one can stop training at any moment when no obvious artifacts are produced and the translations are plausible. Taking into account that there is no ground truth for stain translation (the staining process is irreversible), and that the stain process itself is prone to high variation (particularly between labs), many possible translations are valid. Thus, it is possible that for the same patch, a stain translation model produces different valid translations during training (as shown in Figure 9).

In order to investigate how the quality of translation varies during training, the test set (4 WSI images) is evaluated using CycleGAN models from five different epochs—10th, 20th, 30th, 40th and 50th using stains CD34 and CD68, since they are (biologically) the most different to PAS and perform the worst in the previous section (see Tables 2 and 3). It is assumed that translations between them and PAS are hard.

The architectures with Instance normalisation, Group_{16} and without any normalisation (None) are used since they obtained respectively the best, average, and worst overall scores of the models producing plausible translations in the previous section (i.e. Batch normalisation is excluded). To visualise this

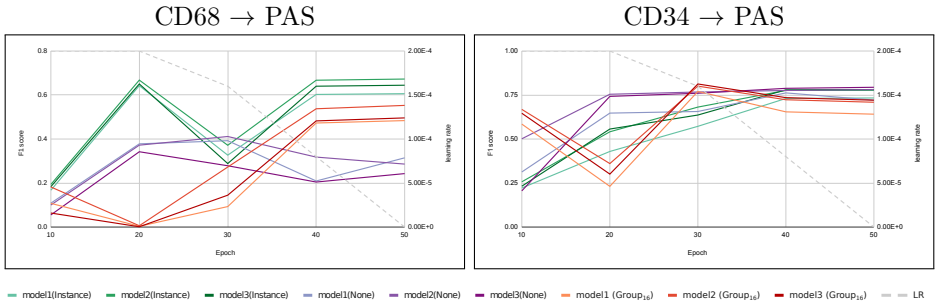


Figure 10: (PAS) MDS1 performance in different CycleGAN epochs.

effect, three pretrained PAS models were randomly selected and the (MDS1) segmentation output is shown in Figure 10. Concerning CD34, better performance is generally obtained in later epoch, however, this is not the case with CD68. In both cases, longer training does not necessarily correlate with better translations. Note that the learning rate (also included on the figures) decreases during training (see the training policy presented in Appendix A), explaining the stability obtained at later epochs.

Moreover, the ranking of the normalisation strategies is not constant in each epoch, for example in case of CD34, the translations obtained using Group_{16} in the 30th epoch are better segmented then those obtained by Instance norm in the final epoch. As such, the ranking presented in Tables 2 and 3 may vary depending on the experimental setup (training duration, etc). Apart from visual differences, an additional cause of the variance of pretrained model performance could be different levels or types of noise being injected into the translations at different epochs due to self-adversarial attack to which CycleGAN-based architectures are prone [53]. Additional evidence for this is the that all segmentation models are affected similarly at the same epoch (e.g. in the case of CD68 and Group_{16}, all models have almost 0 F-score), indicating that the problem originates in the translation, rather than short-cut learning. This also goes inline

with the well-known phenomenon of transferability of adversarial examples [54].

To confirm that visual quality is not related to segmentation performance, Figure 11 presents translations to PAS at different epochs during training using Instance norm (since it is found to be the best strategy overall), along with their corresponding segmentations (using PAS model 2 from Figure 10). As can be seen, they are all plausible, however, the segmentations vary greatly.

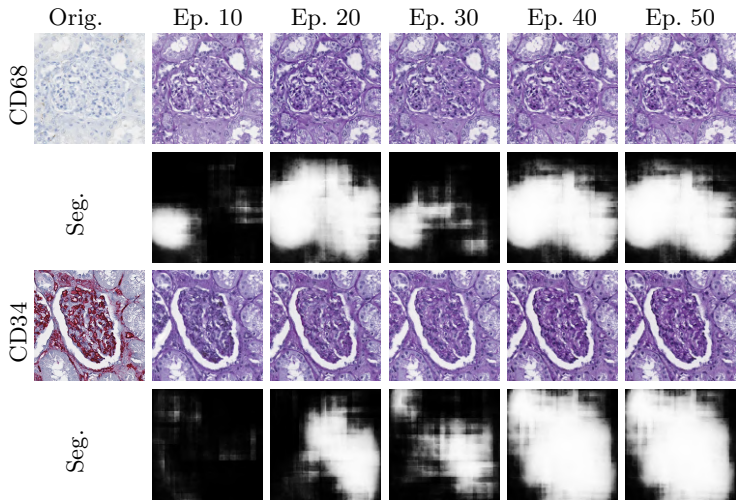


Figure 11: Glomeruli patches translated to PAS using CycleGAN (Instance) models from different training epochs and their segmentation using pretrained PAS model 2 from Figure 10).

5.1.3. CycleGAN Failure Cases

In addition to replacing Instance normalisation with other type of normalisation, the normalisation layer was removed entirely from the CycleGAN architecture. Although this modification can sometimes lead to more unstable training (in our case, the translation between PAS and CD68 or Sirius Red were more frequently unstable), the obtained results are still visually appealing and even better then with some normalisation strategies (e.g. Batch normalisation). Nevertheless, in this setting it was found that the CycleGAN models

are more likely to produce artifacts. The model is prone to hallucinate features, such as those presented in Figure 12. This behaviour was observed particularly often when CD34 and CD68 were the target stains. Since the produced artifacts are visually in accordance with the overall image texture, these cases could be easily unnoticed by the untrained eye, highlighting the importance of including pathologists in the stain translation development process.

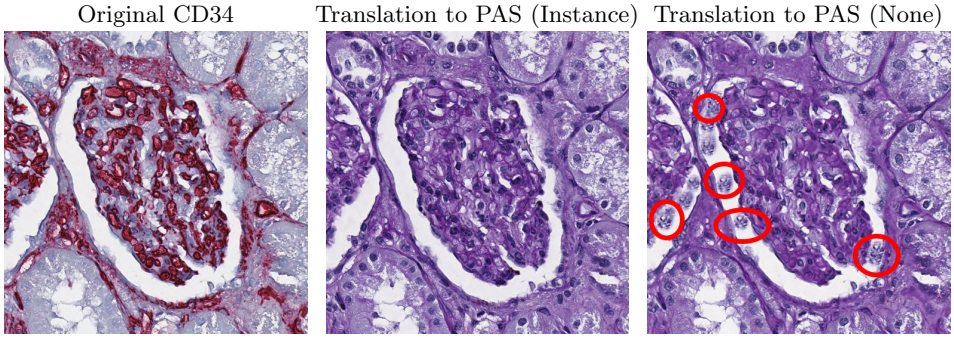


Figure 12: Hallucination effect of CycleGAN without normalisation.

5.2. Quantitative evaluation metrics

5.2.1. Reconstruction assessment

Modifications to CycleGAN architecture can include specific modules or loss functions in order to ensure that the model preserves important structural information [35]. However, it has been shown that changing the normalisation layers of even the most basic CycleGAN architecture can cause differences in the preservation of structural information during translation. Figure 13 presents the SSIM and PSNR of PAS images reconstructed via translation to different target stains with different CycleGAN normalisation layers. These are calculated over 200 random patches (100 glomeruli and 100 negative). As it can be observed from these figures that significant variation in both metrics is present in all target stains. More importantly, the order of the metrics does not cor-

relate with the ability of the architecture to reduce domain shift (see Table 2 and Table 3). This indicates that using these metrics in this setting may not accurately reflect the benefits of modifications to CycleGAN-based models.

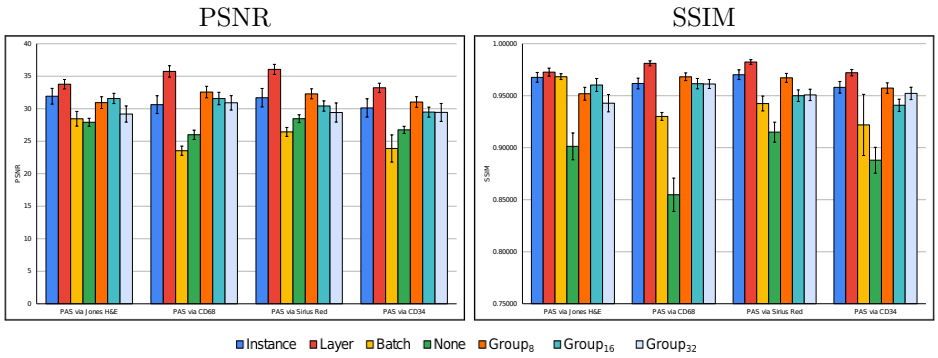


Figure 13: (PAS) MDS1 performance in different CycleGAN epochs.

5.2.2. Translation Distributions

Despite the success of CycleGANs, they are prone to self-adversarial attacks [55, 53, 21]. The cycle-consistency constraint forces the generator to hide information necessary to reconstruct the input image as imperceptible noise and since it has been shown that the results appear plausible (see definition in Introduction, page 4), one possible hypothesis is that this imperceptible noise causes the domain shift observed in Section 4 [56]. Song et al. [57] show that a PixelCNN++ generative model can be used to detect adversarial attacks in images and it is therefore used here to detect the presence (or not) of adversarial noise in the obtained translations.

PixelCNN++ [58] quantifies the pixels of an image x over all its sub-pixels as a product of conditional distributions, such that it learns to predict the next

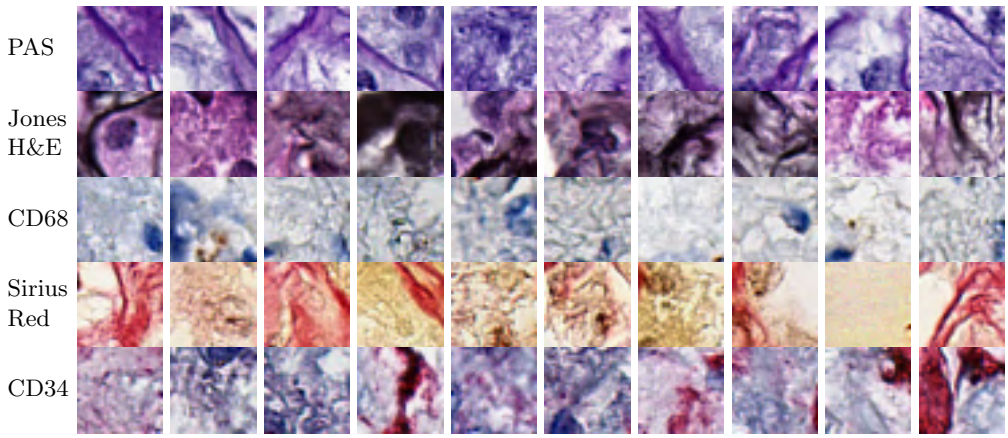


Figure 14: Samples generated from the trained PixelCNN++ for each stain (PAS, Jones H&E, CD68, Sirius Red, and CD34).

pixel value given all previously generated pixels, that is

$$p(x) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1}). \tag{2}$$

These conditional distributions are parameterised by a convolutional neural network (CNN) and hence shared across all pixel positions in the image. The PixelCNN++ [58] architecture is used to model the underlying distribution of each stain separately (training details are presented in Appendix A): PAS, Jones H&E, CD68, Sirius Red, and CD34. As such, the PixelCNN++ models are able to generate images that belong to the real data distribution. Figure 14 presents examples of several such patches. Due to memory limitations, the models are trained on 32×32 pixel patches (therefore each 512×512 patch is decomposed into non-overlapping patches), and therefore the models are able to generate only structures visible at this patch size. Visual evaluation can clearly identify cell nuclei, endothelial lining, a partially granular cytoplasmic texture, extracellular matrix components (such as collagen fibers), and even some cell borders are faintly outlined, recapitulating the cell membranes of some epithelial

cells.

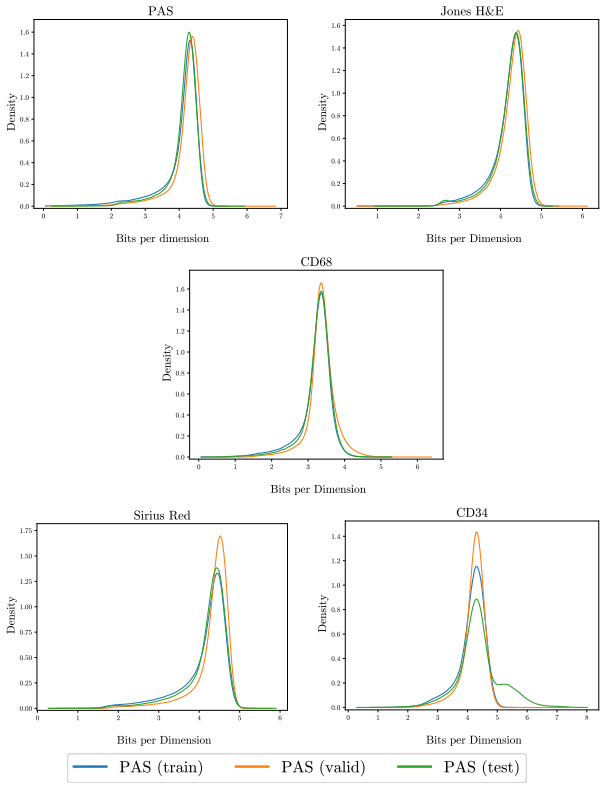


Figure 15: Visualisation of training, validation, and test data distributions for each stain under PixelCNN++.

To further validate the efficacy of the PixelCNN++ models, the distributions of the training, validation, and test sets are plotted for all stains, see Figure 15. This confirms that the PixelCNN++ model is able to accurately estimate real data distributions, since there is an overlap between all three distributions in all stains. To investigate whether the drop in performance of the pretrained models is caused by an imperceptible domain shift, all the test target stains are translated to PAS using the CycleGANs models with different normalisation layers. Figure 16 shows the distributions of the resulting images compared to the real PAS test set. It can be observed that the translated target-to-source

stains have a different data distribution, confirming the existence of a domain shift, which causes the pretrained models to fail. If the translation is performed in the opposite direction (from PAS to target), the same domain shift is found, see Figure 17.

It is important to note when interpreting these figures, that the relative distance between the graphs of real and translated distributions does not necessarily correlate to the performance of pretrained models [57].

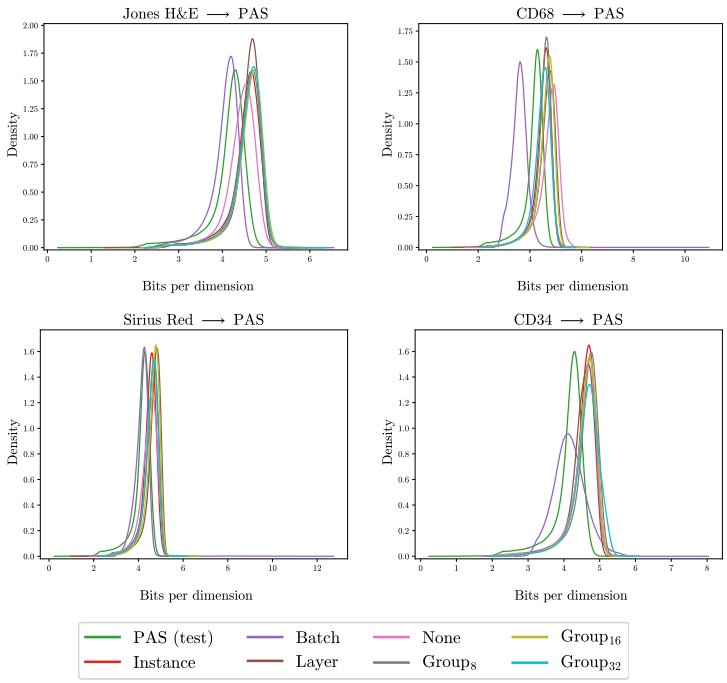


Figure 16: Qualitative comparison of the real PAS (test) distribution and translated target-to-PAS distributions using different CycleGAN normalisation layers. Each distribution is calculated using the test set for each stain.

These results confirm that, although plausible, the translations obtained with various stain translation models, actually generate data in a manner that slightly mismatches the real data distributions. Thus, the pretrained models can exhibit variation in performance when applied to such data even though

the output is visually plausible. This additionally confirms that stain transfer (Hypothesis 2.1 and 2.2) is the cause of MDS1 performance variability.

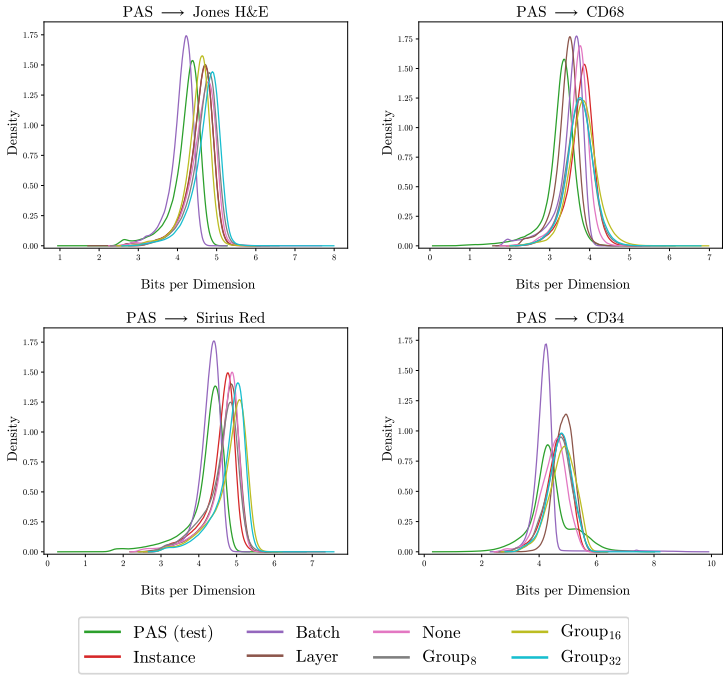


Figure 17: Qualitative comparison of real target stain distributions (test set) and translated PAS-to-target stain distribution (test set) for each type of CycleGAN normalisation layer.

5.3. Clinical usage of stain transfer

When it comes to the clinical usage of images obtained via a stain transfer model, stain translations can be useful and hold great potential for future development of digital pathology. The potential risk that their results may be misleading under certain circumstances can be mitigated by carefully considering the biological and image-related context and the intended use of their application. For example, the general detection of blood vessels (e.g., their quantification) is feasible for larger vessels and even down to the size of arterioles in PAS → CD34 translations and vice versa. However, the evaluation of microves-

sel density including very small vessels (capillaries) would be difficult, as they are strongly labelled by CD34 but not necessarily by PAS. Likewise, there are specifically PAS-positive structures; e.g., so-called granular osmiophilic material ("GOM", a diagnostic hallmark of a vascular disease abbreviated "CADASIL" [59, 60]) that would not be visible in a CD34 staining. Translations between stains would be misleading in those cases. Other examples include the detection of glomeruli in kidney tissue. This would be possible in both stainings and in translations thereof if the aim is solely glomeruli quantification, but misleading if particular substructures are evaluated for diagnostic reasons.

6. Conclusions

To summarise, this article presents a study on the sensitivity of virtual stain transfer obtained by the most commonly used technique, CycleGAN, when used to reduce the domain shift introduced by both inter- and intra- stain variation (commonly referred to as stain translation, and stain normalisation). In order to control the architectural differences between stain translation models the experiments focused on different normalisation layers in the CycleGAN architecture.

Surprisingly, the majority of architectures tested (including no normalisation at all) lead to visually appealing translations. However, by extensive experiments it was shown that those models generate data that belongs to different distributions, leading to unpredictable performance for pretrained segmentation models. Thus, the conclusion can be made that visual inspection is not sufficient in all situations and should be complemented by additional criteria for comparing and choosing stain transfer models. Specifically, we show that in both stain translation and stain normalisation, pretrained models exhibit huge performance differences even when there is no visual difference between the translations. We attribute this phenomena to self-adversarial attack, a con-

sequence of the natural many-to-many mapping which exists between different stains (or even the same stain in different labs). We also confirm this by showing that there is a difference between the distributions of real and translated images using PixelCNN++ generative models.

The architectural choice affects the appearance of important diagnostic evaluation criteria (such as markers for macrophages) and thus artificially generated images cannot be relied upon for diagnosis purposes. If one normalisation strategy was to be chosen for virtual staining, Instance normalisation would be recommended based on these results, although in some specific cases other normalisation strategies can give better results.

Acknowledgements

This work was supported by: ERACoSysMed and e:Med initiatives by the German Ministry of Research and Education (BMBF); SysMIFTA (project management PTJ, FKZ 031L-0085A; Agence National de la Recherche, ANR, project number ANR-15—CMED-0004); SYSIMIT (project management DLR, FKZ 01ZX1608A); ArtIC project "Artificial Intelligence for Care" (grant ANR-20-THIA-0006-01) and co funded by Région Grand Est, Inria Nancy - Grand Est, IHU of Strasbourg, University of Strasbourg and University of Haute-Alsace; and the French Government through co-tutelle PhD funding. We thank the Nvidia Corporation, the *Centre de Calcul de l'Université de Strasbourg*, and HPC resources of IDRIS under the allocation 2020-A0091011872 made by GENCI. We also thank the MHH team for providing high-quality images and annotations, specifically Nicole Kroenke for excellent technical assistance, Nadine Schaadt for image management and quality control, and Valery Volk and Jessica Schmitz for annotations under the supervision of domain experts.

Appendix A. Training details

Appendix A.1. CycleGAN models

Gadermayr et al. [22] showed that different sampling strategies have an impact on a stain transfer model's performance, therefore patches are randomly extracted using a uniform sampling strategy (i.e. in an unsupervised manner) from all training patients in all stainings. The training parameters for CycleGAN models are taken form the original paper ($w_{\text{cyc}} = 10$, $w_{\text{id}} = 5$) [13]. The models are trained for 50 epochs using the Adam optimiser, with a learning rate of 0.0002 and a batch size of 1. From the 25^{th} epoch, the learning rate linearly decayed to 0, and the cycle-consistency and identity weights halved. In all experiments, the translation model from the last (50^{th}) epoch is used. Moreover, to reduce model oscillation, Shrivastava et al.'s strategy [61] of updating the discriminator using the 50 previously generated samples is adopted.

In the case of intra-stain variation, patches were randomly extracted from patients 1 and 7 using a uniform sampling strategy for CycleGAN training. During test time, pretrained models were applied to patients 1, 3 and 7 (patient 3 is kept as an out-of-training distribution sample since it contains sufficient glomeruli - 49).

Appendix A.2. Segmentation U-Net Models

The same training parameters are used for all experiments: batch size of 8, learning rate of 0.0001, 250 epochs, and the network with the lowest validation loss is kept. The slide background (non-tissue) is removed by thresholding each image by its mean value then removing small objects and closing holes. All patches are standardised to $[0, 1]$ and normalised by the mean and standard deviation of the (labeled) training set. The following augmentations are applied with an independent probability of 0.5 (batches are augmented 'on the fly'), in

order to further force the network to learn general features: elastic deformation ($\sigma = 10$, $\alpha = 100$); random rotation in the range $[0°, 180°]$, random shift sampled from $[-205, 205]$ pixels, random magnification sampled from $[0.8, 1.2]$, and horizontal/vertical flip; additive Gaussian noise with $\sigma \in [0, 2.55]$; Gaussian filtering with $\sigma \in [0, 1]$; brightness, colour, and contrast enhancements with factors sampled from $[0.9, 1.1]$; stain variation by colour deconvolution [62], α sampled from $[-0.25, 0.25]$ and β from $[-0.05, 0.05]$. Due to specificity of the U-Net architecture with valid convolutions, the central part of each is used (resulting in a segmentation patch size of 508×508). The predicted segmentation has a size of 324×324 pixels.

The best model is saved based on performance on validation set which is composed of patches extracted from validation patients. The performances of best models are calculated over test patients in each of the experiment.

Appendix A.3. PixelCNN++ Model

The PixelCNN++ [58] architecture is used to model the underlying distribution of each stain: PAS, Jones H&E, CD68, Sirius Red, and CD34. The architectural configurations are formalised as: the model employs 3 Resnet [63] blocks consisting of 5 residual layers in the encoding phase, with 2×2 downsampling between the ResNet blocks. In the decoding phase, the same architecture is employed, but with upsampling layers instead of downsampling. All residual layers utilise 160 filter maps in their convolutional layers and have a dropout of 0.5. The overall training for one PixelCNN++ model took approximately 15 days on an HPC with 4 V100 GPUs (in parallel).

Since each pixel value is conditioned on the product of all previously generated pixels, the models were trained and evaluated on patches of size 32×32 due to GPU memory limitations. For each stain, we extracted 1280000 train, validation, and test patches from the corresponding patents. The model is trained

for 60 epochs with a learning rate of 0.001 and a decay rate of 0.999. The best model is saved with the lowest bits-per-dimension score [64] on the validation set. We use 128000 patches as the validation set, extracted randomly from the validation patients. We employed the original publicly available implementation [3].

References

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

[2] F. Ciompi, O. G. F. Geessink, B. E. Bejnordi, G. S. de Souza, A. Baidoshvili, G. Litjens, B. Ginneken, I. Nagtegaal, J. V. D. Laak, The importance of stain normalization in colorectal tissue classification with convolutional networks, ISBI (2017) 160–163.

[3] G. Csurka, A comprehensive survey on domain adaptation for visual applications, in: Domain adaptation in computer vision applications, Springer, 2017, pp. 1–35.

[4] M. Salvi, N. Michielli, F. Molinari, Stain color adaptive normalization (scan) algorithm: Separation and standardization of histological stains in digital pathology, Computer Methods and Programs in Biomedicine (2020).

[5] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, Xiaojun Guan, C. Schmitt, N. E. Thomas, A method for normalizing histology slides for quantitative analysis, in: ISBI, 2009, pp. 1107–1110.

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

[3] https://github.com/openai/pixel-cnn

[7] A. M. Khan, N. Rajpoot, D. Treanor, D. Magee, A nonlinear mapping approach to stain normalization in digital histopathology images using image-specific color deconvolution, IEEE Transactions on Biomedical Engineering 61 (2014) 1729–1738.

[8] A. Janowczyk, A. Basavanhally, A. Madabhushi, Stain normalization using sparse autoencoders (stanosa): Application to digital pathology, Computerized Medical Imaging and Graphics 57 (2017) 50–61.

[9] M. Gadermayr, L. Gupta, V. Appel, P. Boor, B. M. Klinkhammer, D. Merhof, Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: A study on kidney histology, IEEE Trans Med Imaging 38 (2019) 2293–2302.

[10] J. Vasiljević, F. Feuerhake, C. Wemmert, T. Lampert, Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks, Neurocomputing (2021).

[11] A. Rana, G. Yauney, A. Lowe, P. Shah, Computational histological staining and destaining of prostate core biopsy rgb images with generative adversarial neural networks, in: ICMLA, 2018.

[12] M. T. Shaban, C. Baur, N. Navab, S. Albarqouni, StainGAN: Stain style transfer for digital histological images, in: ISBI, 2019, pp. 953–956.

[13] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: ICCV, 2017, pp. 2223–2232.

[14] N. Brieu, A. Meier, A. Kapil, R. Schoenmeyer, C. G. Gavriel, P. D. Caie, G. Schmidt, Domain adaptation-based augmentation for weakly supervised nuclei detection, in: COMPAY@MICCAI, 2019.

[15] A. Shrivastava, W. Adorno, Y. Sharma, L. Ehsan, S. A. Ali, S. R. Moore, B. C. Amadi, P. Kelly, S. Syed, D. E. Brown, Self-Attentive Adversarial Stain Normalization, in: ICPR, 2021.

[16] S. Cai, Y. Xue, Q. Gao, M. Du, G. Chen, H. Zhang, T. Tong, Stain style transfer using transitive adversarial networks, in: MLMIR, 2019.

[17] D. Mahapatra, B. Bozorgtabar, J.-P. Thiran, L. Shao, Structure preserving stain normalization of histopathology images using self-supervised semantic guidance, in: MICCAI, 2021.

[18] H. Kang, D. Luo, W. Feng, S. Zeng, T. Quan, J. Hu, X. Liu, Stainnet: a fast and robust stain normalization network, Frontiers in Medicine 8 (2021).

[19] A. Lahiani, N. Navab, S. Albarqouni, E. Klaiman, Perceptual embedding consistency for seamless reconstruction of tilewise style transfer, in: MICCAI, 2019, pp. 568–576.

[20] J. P. Cohen, M. Luck, S. Honari, Distribution Matching Losses Can Hallucinate Features in Medical Image Translation, in: MICCAI, 2018.

[21] J. Vasiljević, F. Feuerhake, C. Wemmert, T. Lampert, Self Adversarial Attack as an Augmentation Method for Immunohistochemical Stainings, in: ISBI, 2021.

[22] M. Gadermayr, V. Appel, M. B. Klinkhammer, P. Boor, D. Merhof, Which way round? A study on the performance of stain-translation for segmenting arbitrarily dyed histological images, in: MICCAI, volume 11071, 2018, pp. 165–173.

[23] S. Wagner, N. Khalili, R. Sharma, M. Boxberg, C. Marr, W. de Back, T. Peng, Structure-Preserving Multi-domain Stain Color Augmentation

Using Style-Transfer with Disentangled Representations, in: MICCAI, 2021, pp. 257–266.

[24] Z. Xu, X. Li, X. Zhu, L. Chen, Y. He, Y. Chen, Effective immunohistochemistry pathology microscopy image generation using cyclegan, Frontiers in Molecular Biosciences 7 (2020) 243.

[25] A. Vahadane, T. Peng, S. Albarqouni, M. Baust, K. Steiger, A. Schlitter, A. Sethi, I. Esposito, N. Navab, Structure-preserving color normalization and sparse stain separation for histological images, IEEE Trans Med Imaging 35 (2016) 1962–1971.

[26] T. Lampert, O. Merveille, J. Schmitz, G. Forestier, F. Feuerhake, C. Wemmert, Strategies for training stain invariant CNNs, in: ISBI, 2019, pp. 905–909.

[27] P. Shamsolmoali, M. Zareapoor, E. Granger, H. Zhou, R. Wang, M. E. Celebi, J. Yang, Image synthesis with adversarial networks: A comprehensive survey and case studies, Information Fusion 72 (2021) 126–146.

[28] C. Mercan, G. Reijnen-Mooij, D. T. Martin, J. Lotz, N. Weiss, M. van Gerven, F. Ciompi, Virtual staining for mitosis detection in breast histopathology, in: ISBI, 2020.

[29] Z. Xu, C. F. Moro, B. Bozóky, Q. Zhang, GAN-based virtual re-staining: A promising solution for whole slide image analysis, arXiv 1901.04059 (2019).

[30] Y.-C. Lo, I.-F. Chung, S.-N. Guo, M.-C. Wen, C.-F. Juang, Cycle-consistent GAN-based stain translation of renal pathology images with glomerulus detection application, Applied Soft Computing 98 (2021) 106822.

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

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: NIPS, 2014, pp. 2672–2680.

[33] M. Runz, D. Rusche, S. Schmidt, M. R. Weihrauch, J. Hesser, C.-A. Weis, Normalization of HE-stained histological images using cycle consistent generative adversarial networks, Diagnostic Pathology 16 (2021) 71.

[34] S. J. Shin, S. C. You, H. Jeon, J. W. Jung, M. H. An, R. W. Park, J. Roh, Style transfer strategy for developing a generalizable deep learning application in digital pathology, Computer Methods and Programs in Biomedicine 198 (2021) 105815.

[35] S. Liu, B. Zhang, Y. Liu, A. Han, H. Shi, T. Guan, Y. He, Unpaired stain transfer using pathology-consistent constrained generative adversarial networks, IEEE Transactions on Medical Imaging 40 (2021) 1977–1989.

[36] N. Bouteldja, B. M. Klinkhammer, T. Schlaich, P. Boor, D. Merhof, Improving unsupervised stain-to-stain translation using self-supervision and meta-learning, arXiv preprint arXiv:2112.08837 (2021).

[37] J. Ke, Y. Shen, X. Liang, D. Shen, Contrastive Learning Based Stain Normalization Across Multiple Tumor in Histopathology, in: MICCAI, 2021, pp. 571–580.

[38] A. Z. Moghadam, H. Azarnoush, S. A. Seyyedsalehi, M. Havaei, Stain transfer using Generative Adversarial Networks and disentangled features, Computers in Biology and Medicine (2022) 105219.

[39] Y. Choi, Y. Uh, J. Yoo, J.-W. Ha, StarGAN v2: Diverse image synthesis for multiple domains, in: CVPR, 2020.

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

[41] A. Lahiani, et al., Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach, in: ECDP, 2019, pp. 47–55.

[42] H. Liang, K. Plataniotis, X. Li, Stain style transfer of histopathology images via structure-preserved generative learning, in: MLMIR, 2020.

[43] Y. Zhang, K. de Haan, Y. Rivenson, J. Li, A. Delis, A. Ozcan, Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue, Light, Science & Applications 9 (2020).

[44] O. Ronneberger, P. Fischer, T. Broxs, U-Net: Convolutional networks for biomedical image segmentation, in: MICCAI, 2015, pp. 234–241.

[45] G. Bueno, L. Gonzalez-Lopez, M. Garcia-Rojo, A. Laurinavicius, O. Deniz, Data for glomeruli characterization in histopathological images, Data in Brief 29 (2020) 105314.

[46] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks, Nature Machine Intelligence 2 (2020) 665–673.

[47] T. de Bel, M. Hermsen, J. Kers, J. van der Laak, G. J. S. Litjens, Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology, in: MIDL, volume 102, 2019, pp. 151–163.

[48] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: ICML, 2015, p. 448–456.

[49] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450 (2016).

[50] Y. Wu, K. He, Group normalization, in: ECCV, 2018, pp. 3–19.

[51] R. Marée, L. Rollus, B. Stevens, R. Hoyoux, G. Louppe, R. Vandaele, J.-M. Begon, P. Kainz, P. Geurts, L. Wehenkel, Collaborative analysis of multi-gigapixel imaging data using cytomine, Bioinformatics 32 (2016) 1395–1401.

[52] G. Bueno, M. M. Fernandez-Carrobles, L. Gonzalez-Lopez, O. Deniz, Glomerulosclerosis identification in whole slide images using semantic segmentation, Computer Methods and Programs in Biomedicine (2020).

[53] D. Bashkirova, B. Usman, K. Saenko, Adversarial self-defense for cycle-consistent gans, in: NeurIPS, volume 32, 2019.

[54] Y. Liu, X. Chen, C. Liu, D. Song, Delving into transferable adversarial examples and black-box attacks, in: ICLR, 2017.

[55] C. Chu, A. Zhmoginov, M. Sandler, CycleGAN, a master of steganography, in: NIPS, workshop on Machine Deception, 2017.

[56] Z. Nisar, J. Vasiljevic, P. Gançarski, T. Lampert, Towards measuring domain shift in histopathological stain translation in an unsupervised manner, in: ISBI, 2022.

[57] Y. Song, T. Kim, S. Nowozin, S. Ermon, N. Kushman, PixelDefend: Leveraging generative models to understand and defend against adversarial examples, in: ICLR, 2018.

[58] T. Salimans, A. Karpathy, X. Chen, D. P. Kingma, PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications, in: ICLR, 2017.

[59] F. F., B. Volk, C. Ostertag, F. Juengling, J. Kassubek, O. M., M. Dichgans, Reversible coma with raised intracranial pressure: An unusual clinical manifestation of cadasil, Acta neuropathologica 103 (2002) 188–92.

[60] J. Pettersen, J. Keith, F. Gao, J. D. Spence, S. Black, Cadasil accelerated by acute hypotension: Arterial and venous contribution to leukoaraiosis, Neurology 88 (2017).

[61] A. Shrivastava, T. Pfister, O. Tuzel, J. Susskind, W. Wang, R. Webb, Learning from simulated and unsupervised images through adversarial training, in: CVPR, 2017.

[62] D. Tellez, M. Balkenhol, I. Otte-Höller, R. Loo, R. Vogels, P. Bult, C. Wauters, W. Vreuls, S. Mol, N. Karssemeijer, G. Litjens, J. van der Laak, F. Ciompi, 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 (2018) 2126–2136.

[63] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

[64] A. Van Oord, N. Kalchbrenner, K. Kavukcuoglu, Pixel recurrent neural networks, in: ICML, PMLR, 2016, pp. 1747–1756.