HistoStarGAN: A Unified Approach to Stain Normalisation, Stain Transfer and Stain Invariant Segmentation in Renal Histopathology

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

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

Abstract

Virtual stain transfer is a promising area of research in Computational Pathology, which has a great potential to alleviate important limitations when applying deep-learning-based solutions such as lack of annotations and sensitivity to a domain shift. However, in the literature, the majority of virtual staining approaches are trained for a specific staining or stain combination, and their extension to unseen stainings requires the acquisition of additional data and training. In this paper, we propose HistoStar-GAN, a unified framework that performs stain transfer between multiple stainings, stain normalisation and stain invariant segmentation, all in one inference of the model. We demonstrate the generalisation abilities of the proposed solution to perform diverse stain transfer and accurate stain invariant segmentation over numerous unseen stainings, which is the first such demonstration in the field. Moreover, the pre-trained HistoStarGAN model can serve as a synthetic data generator, which paves the way for the use of fully annotated synthetic image data to improve the training of deep learning-based algorithms. To illustrate the capabilities of our approach, as well as the potential risks in the microscopy domain, inspired by applications in natural images, we generated KidneyArtPathology, a fully annotated artificial image dataset for renal pathology.

---

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

1. Introduction

The deep learning revolution (Sejnowski, 2018) opens the door for remarkable applications in the medical domain. Plenty of routine clinical tasks have great potential to be fully automatised, which triggered a staggering amount of research (Piccialli et al., 2021; Varoquaux and Cheplygina, 2022; Liu et al., 2019). In such an environment, Computational Pathology is not an exception (Van der Laak et al., 2021; Srinidhi et al., 2021). However, currently many state-of-the-art deep learning methods are data-hungry approaches that require huge collections of annotated data to be trained. However, collecting medical data is strictly regulated, while obtaining high-quality annotations can only be effectively performed by trained experts (Grote et al., 2018). All this poses important constraints for the development of automated solutions. Moreover, existing data and annotations can only be reused with limited success, because the staining process is prone to high variability (Bancroft and Gamble, 2008), representing a source of a domain shift that negatively affects pre-trained models (Tellez et al., 2019; Gadermayr et al., 2018; Vasiljević et al., 2021a).

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), being able to generate samples from complex data distributions (Karras et al., 2020b), have great potential to overcome some of the limitations in applying deep learning-based solutions in Computational Pathology. Several class of applications, such as image synthesis and stain transfer have been identified in the literature (Tschuchnig et al., 2020). A particularly promising application area is stain transfer, which enables virtual re-staining of a histopathological image, i.e. changing its appearance to look as though it has been stained with another staining or staining variation. Many attempts in the field employ image-to-image translation-based approaches for stain normalisation (Shaban et al., 2019; Ke et al., 2021; de Bel et al., 2021), i.e. reducing the domain shift caused by intra-stain variation, which is a variation in the appearance of the same staining (e.g. due to different laboratories procedures). Other approaches try to reduce inter-stain variation, which is a variation in the appearance of different stainings (Liu et al., 2021; Gadermayr et al., 2018; Levy et al., 2020; Vasiljević et al., 2021a). Only a few attempts try to address the generalisation problem of deep-learning-based solutions by propos-

ing a model that is robust to intra-stain variation (Wagner et al., 2021; Xue et al., 2021; Levine et al., 2020) or inter-stain-variation (Vasiljević et al., 2021a).

However, considering numerous factors of variations which affect the appearance of one staining, in addition to the differences in tissue structures visible under different stainings, the task of virtual stain transfer is naturally non-deterministic and ill-posed. Although state-of-the-art virtual staining strategies, mainly based on GANs, result in visually plausible translations, assessing the results' quality is hard. Obtaining pixel-wise annotations across different staining modalities is limited by the requirement of either processing consecutive tissue sections separately, or re-staining identical sections several times. The latter has the disadvantage of tissue destruction, or processing arte-facts. Using consecutive sections that are only stained once minimises this problem, but the sequentially produced sections are never identical, and their co-registration can be limited by various factors (Merveille et al., 2021). Thus, a virtual staining model resulting in multiple translations can support the digital pathology domain to overcome the shortage of high-quality annotations and improve the training of networks. There-fore, it would be reasonable to expect from a virtual staining model to result in diverse translations. However, the approaches which enable such diverse translations are not very common in the field.

In this paper, we propose HistoStarGAN, a model that can perform diverse trans-lations for stain normalisation and stain transfer tasks, in addition to stain invariant segmentation. The main contributions of this article are:

- HistoStarGAN, the first model able to simultaneously perform diverse stain trans-fer, stain normalisation and stain invariant segmentation.

- The proposed model generalises stain transfer across many stainings, including unseen stainings.

- The model learns using limited annotations from one staining and generalises to a wide range of unseen stainings without requiring additional annotations.

- KindeyArtPathology - the first artificially created fully-annotated histological dataset, illustrating the potential and the risks of generating synthetic training

3

data targeting staining methods relevant in renal pathology.

- Pre-trained HistoStarGAN, which can be used for numerous offline applications, e.g. augmentation of small-size private datasets.

The rest of the paper is organised as follows. The HistoStarGAN model is explained in more detail in Section 3.1; visual and quantitative results are given in Section 4. The HistoStarGAN architecture and training parameters are analysed in an ablation study presented in Section 5. KindeyArtPathology dataset is explained in more detail Section 6. The KidneyArtPathology dataset and pre-trained models can be available upon request. Demonstrative examples are provided online[1]. The limitations and opportunities of the proposed solution are presented in Section 7; conclusions of the paper are presented in Section 8.

2. Related work

Virtual staining approaches are greatly inspired by style transfer literature (Jose et al., 2021) where GAN-based approaches achieve state-of-the-art results in many application areas. Therefore, some authors try to bypass the physical staining process, which introduces most of the variation (Li et al., 2020; Rana et al., 2020) while others try to normalise slide appearance after staining (Shaban et al., 2019). Nevertheless, the methods are usually based on image-to-image translations approaches, where one domain (stain) is translated to look like a reference staining. Some approaches (Salehi and Chalechale, 2020; Cho et al., 2017; Cong et al., 2021) simulate paired datasets by discarding colour information (e.g. by using a greyscale version of the image or extracting a haematoxylin channel from the image) and learning paired image-to-image translations models, such as pix2pix (Isola et al., 2017). However, discarding the majority of colour information can also eliminate relevant diagnostic details (Moghadam et al., 2022; BenTaieb and Hamarneh, 2017). Thus, many approaches build upon the idea of unpaired image-to-image translation methods, where CycleGAN-based (Zhu et al., 2017) approaches are dominant (Shaban et al., 2019; Kang et al., 2021; Shrivastava

---

[1]https://main.d33ezaxrmu3m4a.amplifyapp.com/

et al., 2021; de Bel et al., 2021; Lo et al., 2021; Liu et al., 2021). However, CycleGAN reduces virtual staining to a deterministic mapping, producing a single and fixed output for a given input. Although training repetitions of the same model, or different epochs during training, can result in different translations (Vasiljević et al., 2022), the common characteristic of CycleGAN-based methods is a deterministic translation, which limits output diversity. Moreover, the method is bi-directional, meaning that its application to more than two stainings requires multiple trainings of the model. Alternative multi-domain approaches, such as those based on StarGAN (Choi et al., 2018) are shown to be inferior to CycleGAN in specific applications (Vasiljević et al., 2021a). The quality of more complex models (Lin et al., 2022) to reduce stain variation in multi-domain translations remains to be explored. Alternative approaches try to overcome the stain variation problem by aiming for stain invariant solutions. Existing methods mainly exploit virtual staining to perform training time augmentation (Vasiljević et al., 2021a; Wagner et al., 2021), or test-time augmentation (Scalbert et al., 2022). Such a solution can greatly benefit from diverse virtual staining as a single model can perform more extensive augmentation. The current state-of-the-art methods for diverse multi-domain style transfer, such as StarGANv2 (Choi et al., 2020) and TUINT (Baek et al., 2021), could be used for diverse virtual staining. However, contrary to the domain of natural images where style transfer should alter source-specific image characteristics in the output (e.g. the size of ears when translating a cat to a dog, the amount of hair and hairstyle in female to male transfer etc.), in the medical domain, such extensive alterations are not desirable as they could result in removing/inventing specific cell population (like cancerous) or structures, for example glomeruli in renal pathology. Thus, the direct application of current state-of-the-art style transfer models is not straightforward. A recent attempt by Scalbert et al. (2022) employs StarGANv2 translations for test-time augmentation. However, the benefits of this approach are demonstrated in H&E staining only, and it requires annotated samples from several domains to learn the translation. In the absence of annotated samples, the training of such models becomes difficult, as will be demonstrated in this paper and discussed in more detail in Section 5.3. Alternatively, the application of unconditional GAN-based approaches for style transfer such as StyleGAN (Karras et al., 2020b) or BigGAN (Brock et al., 2018), al-

This paper proposes an extension of the StarGANv2 model, named HistoStarGAN. The presented approach can perform plausible and diverse virtual staining, preserving the structure of interest during translation. Moreover, the model obtains stain invariant segmentation of the selected structure. The proposed solution results in a single model which is, for the first time, able to perform simultaneous stain normalisation, stain transfer and stain invariant segmentation without any additional data manipulation during test time. The HistoStarGAN is the new state-of-the-art method for stain invariant segmentation, outperforming the current by a large margin. The benefits are demonstrated in several stainings, both histochemical and immunohistochemical, for the task of glomeruli segmentation.

3. Method

3.1. Model Description

HistoStarGAN architecture is presented in Figure 1. The model is composed of five modules: generator (G), discriminator (D), mapping network (F), style encoder (E) and segmentation network (S). The models G, D, F and E are elements of Star-GANv2's architecture.

The mapping network F generates a stain-specific style by transforming the random latent code z into the target stain's style. The style is injected into the generator G during translation, which enables diverse generations as different latent codes result in different stain-specific styles. In order to ensure that the generator uses the injected style information, the model is constrained with a style reconstruction loss, i.e. the style encoder E extracts the style from the generated image, and the difference between that style and the style provided to the generator during translation is minimised (blue arrows in the Figure 1). In order to explicitly allow style diversification, the model is trained to produce different outputs for different styles in the given target domain by the style diversification loss (orange arrows in the Figure 1). Moreover, the model is constrained using a cycle-consistency loss (green brackets in Figure 1), e.g. the dif-

6

ference between the original and reconstructed images is minimised. Reconstruction is performed using the same generator, but the style information is extracted by the mapping network by taking the original image as input.
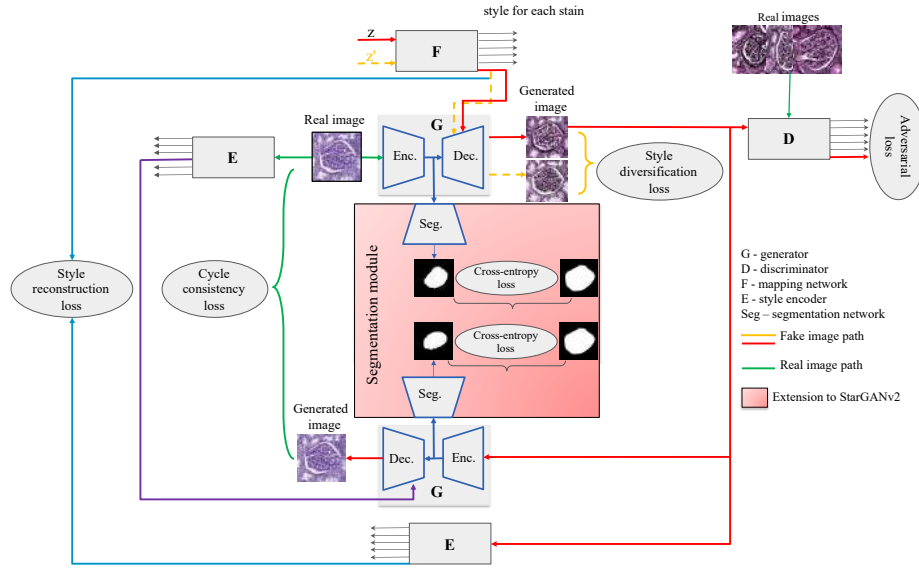

Figure 1: HistoStarGAN – an end-to-end trainable model for simultaneous stain transfer and stain invariant segmentation. The red block denotes the difference compared to the StarGANv2 model (Choi et al., 2020). The model is composed of five modules – generator (G), discriminator (D), mapping network (F), style encoder (E) and segmentation network (S) (the generator and style encoder are illustrated twice for clarity of the flow of information, however, these are the same networks). The mapping network F generates a stain-specific style by transforming the random latent code z into the target stain's style. The style is injected into the generator G during translation, which enables diverse generations, which is also explicitly enforced by the style diversification loss. This is illustrated by the red and orange lines that indicate different styles when performing stain transfer of real PAS image to the Jones H&E staining. In order to ensure that the generator uses the injected style information, the model is constrained with a style reconstruction loss (blue arrows). Additionally, the model is constrained by cycle-consistency loss (green brackets) which is imposed as the difference between the real image and the translation of a fake image given the style of the real image (violet line). Given that generator G has an encoder-decoder structure, to ensure that the features extracted by the encoder are stain invariant, the training is constrained by a segmentation loss (red module). Finally, the training of the model is also guided by an adversarial loss.

The generator G is an encoder-decoder network, with an instance normalisation layer in the encoder and an adaptive instance normalisation layer in the decoder. In

7

this way, the encoder removes stain-specific characteristics from the image while the decoder injects target-stain characteristics during the generation process. Thus, the features extracted by the generator's encoder should be stain invariant. Under the assumption that a structure of interest is visible in all stainings, i.e. the stain invariant solution is feasible, the representation extracted in the bottleneck should be sufficient to perform the considered object-related task. Therefore, a segmentation module is attached to the bottleneck. Trained end-to-end with the other modules, this extension forces the preservation of structures of interest during the translation process.

More formally, let $\mathcal{X}$ be the set of histopathological images, $\mathcal{S}$ the set of available segmentation masks for the structure of interest, and $\mathcal{Y}$ the set of stainings found in $\mathcal{X}$. Given an image $x \in \mathcal{X}$, its original staining $y \in \mathcal{Y}$ and corresponding segmentation mask m_{seg}, the model is trained using the following objectives.

Adversarial objective:. The latent code $z \in \mathcal{Z}$, source domain $y \in \mathcal{Y}$ and target domain $\tilde{y} \in \mathcal{Y}$ are randomly sampled. The target style code $\tilde{s} = F_{\tilde{y}}(z)$ is obtained via the mapping network. Thus the generator G, with input image x from domain y and style $\tilde{s}$, generates and output image $G(x, \tilde{s})$. This image is evaluated by the discriminator's output, which corresponds to a domain $\tilde{y}$, $D_{\tilde{y}}$. These are trained using the adversarial loss (Goodfellow et al., 2014) , such that

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{x,y}[\log D_y(x)] + \mathbb{E}_{x,\tilde{y},z}[\log(1 - D_{\tilde{y}}(G(x, \tilde{s})))], \tag{1}$$

where $\mathbb{E}$ denotes mathematical expectation.

Style reconstruction:. In order to ensure that generator G uses the provided style code $\tilde{s}$ when producing output $G(x, \tilde{s})$, the mapping network E is used to extract the style of the generated image, $E_{\tilde{y}}(G(x, \tilde{s}))$, and the following style reconstruction loss (Choi et al., 2020) is used:

$$\mathcal{L}_{\text{sty}} = \mathbb{E}_{x,\tilde{y},z}[\|\tilde{s} - E_{\tilde{y}}(G(x, \tilde{s}))\|_1]. \tag{2}$$

Style diversification:. The generator is forced to produce different outputs for different styles produced by the mapping network F. Given different latent codes z_1 and z_2, the

following diversity loss (Choi et al., 2020) is maximised:

$$\mathcal{L}_{\text{ds}} = \mathbb{E}_{x,\tilde{y},z_1,z_2}[\|G(x, \tilde{s}_1) - G(x, \tilde{s}_1)\|_1]. \tag{3}$$

Cycle-consistency:. The following cycle-consistency loss Zhu et al. (2017) forces the generator to reconstruct the original image given the source style code $\hat{s} = E_y(x)$:

$$\mathcal{L}_{\text{cyc}} = \mathbb{E}_{x,y,\tilde{y},z}[\|x - G(G(x, \tilde{s}), \hat{s})\|_1]. \tag{4}$$

Segmentation objective:. The cross-entropy loss Jadon (2020) is used to train the segmentation branch on both real data and their translation to a random stain, such that

$$\mathcal{L}_{\text{seg}} = \mathbb{E}_{x,m_{seg}}[m_{seg} \log S eg(x)] + \mathbb{E}_{x,m_{seg},z}[m_{seg} \log S eg(G(x, \tilde{s}))]. \tag{5}$$

Full objective:. The importance of each of these losses is controlled by hyperparameters, and are combined in the following full objective:

$$\min_{G,F,E} \max_{D} \mathcal{L}_{\text{adv}} + \lambda_{\text{sty}}\mathcal{L}_{\text{sty}} - \lambda_{\text{ds}}\mathcal{L}_{\text{ds}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}. \tag{6}$$

3.2. Training Setup

Training the HistoStarGAN model's segmentation branch is supervised, and requires the segmentation masks for all images in the dataset. However, in Computational Pathology this assumption is unrealistic. It is more reasonable to assume that annotations exist for limited examples, e.g. as is common in the field, only for one staining (Gadermayr et al., 2018; Vasiljević et al., 2021a). Thus, the PAS staining is considered to be annotated (the source stain), while the other stainings (Jones H&E, Sirius Red, CD34, CD68) are considered to be unannotated.

To overcome the lack of annotations in target stainings, the CycleGAN's deterministic nature and limited capacity to perform geometrical changes are used to artificially generate a fully-annotated dataset. The CycleGAN model is trained in an unsupervised manner, using randomly extracted patches from a given pair of stainings. Separate CycleGAN models are trained for each pair of PAS-target stains. When trained, the CycleGAN models are applied to the source (annotated) dataset in order to generate annotated samples in the target stainings.

9

3.2.1. Dataset

The dataset contains tissue samples collected from a cohort of 10 patients who underwent tumour nephrectomy due to renal carcinoma. The kidney tissue was selected as distant as possible from the tumours 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. cardiac 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 (WSIs) were acquired using an Aperio AT2 scanner at 40× magnification (a resolution of 0.253 μm / pixel). All the glomeruli in each WSI were annotated and validated by pathology experts by outlining them using Cytomine (Marée et al., 2016). The dataset was divided into 4 training, 2 validation and 4 test patients for each staining, the number of annotated glomeruli in each staining is given in Table 1. The sets of patches of size 512×512 pixels (at 40× magnification) are extracted from each staining in the following manner:

- For CycleGAN training, 5000 random patches are extracted from training WSIs for each staining (no supervision required for any staining).

- For HistoStarGAN training, patches are extracted from PAS staining only and translated to all other stainings by pre-trained CycleGAN models (no supervision in target domains). The segmentation performances of HistoStarGAN model in Section 4 are reported based on WSIs from test patients in each staining.

- Fully supervised training of baseline segmentation networks (UNet models (Ronneberger et al., 2015), only used for evaluation purposes in Section 5.3) are trained for each staining separately. All glomeruli patches and seven times more tissue patches are extracted from each staining. The best models are determined based on performances on the validation set (patches extracted from validation

10

Staining	Training	Validation	Test
PAS	662	588	1092
Jones H&E	621	590	1043
Sirius Red	651	576	1049
CD34	565	595	1019
CD68	526	521	1046

Table 1: The number of annotated glomeruli in each staining. Tissue patches are extracted in proportion 1 : 7 to account for to account for tissue variability. All patches are of size 512×512 pixels (at 40× magnification)

patients from each staining) and the final performances are reported on the WSIs of the test patients.

3.2.2. Training Details

The CycleGAN models have a 9 ResNet blocks architecture, and they are trained for 50 epochs, with patches of size 512×512 pixels extracted from the training patients in each staining. Upon training, a balanced dataset is formed by translating 500 glomeruli and 500 random negative patches from PAS to all target stains. When training the HistoStarGAN model, extensive data augmentation is performed, following the conclusions by Karras et al. (2020a) that data augmentation is a crucial factor when training GANs with limited data.

The following augmentations are applied 50% of time with an independent probability of 0.5 (batches are augmented 'on the fly') for each method; elastic deformation ($\sigma = 10$); affine transformations – random rotation in the range $[0°, 180°]$, random shift sampled from $[-5, 5]$ pixels, random magnification sampled from $[0.95, 1]$, and horizontal/vertical flip; brightness and contrast enhancements with factors sampled from $[0.0, 0.2]$ and $[0.8, 1.2]$ respectfully; additive Gaussian noise with $\sigma \in [0, 0.01]$.

The HistoStarGAN model is trained using the following loss weights: $\lambda_{sty} = 1$, $\lambda_{ds} = 1$, $\lambda_{cyc} = 1$, and $\lambda_{seg} = 5$. To stabilise the training, the weight of the style diversification loss, λ_{ds}, is linearly decreased to zero over $100\,000$ iterations (Choi et al., 2020). Also, the weight of the segmentation loss, λ_{seg}, is 0 for the first $10\,000$ iterations

11

until the model starts to generate recognisable images from each stain. Although their fidelity is not high enough, at this moment the segmentation model receives enough meaningful information to start learning. HistoStarGAN is trained for 100 000 iterations. The architectures for the generator, discriminator, style encoder, and mapping network are the same as in the original StarGANv2 architecture (Choi et al., 2020) as well as optimisers and learning rates. The segmentation branch's architecture is the same as the generator's decoder, without the adaptive instance normalisation layer. The segmentation branch is trained using the Adam optimiser with a learning rate 10^{-5}. As for the other networks, exponential moving averages over parameters (Karras et al., 2018) is applied during training to obtain the final segmentation network, as experimentally, it gives better results than the best model saved based on validation performance.

The HistoStarGAN model's segmentation branch is trained using a balanced dataset, mainly containing artificial histological images produced by both CycleGAN and HistoStarGAN. However, tasks in Computational Pathology are usually concerned with sparse structures, and using an imbalanced dataset to account for the tissue diversity is beneficial for learning (Lampert et al., 2019; Vasiljević et al., 2021a). Moreover, CycleGAN-based translations can be noisy (Vasiljević et al., 2022), which could affect the stain invariant properties of the segmentation module. Thus, after training HistoStarGAN as a whole, the segmentation module is fine-tuned for one epoch using real unbalanced PAS-stained images while the rest of the model is fixed (the choice of the number of fine-tuning epochs is discussed in Section 5). This dataset contains all PAS glomeruli (662 extracted from the training patients) and seven times more negative patches (4634) to account for tissue variability. The Adam optimiser is used with a batch size of 8 and a learning rate of 0.0001. The same augmentation as in Lampert et al. (2019) is applied during fine-tuning.

4. Results

This section will demonstrate that HistoStarGAN results in a single model able to perform diverse stain transfer, stain normalisation and stain invariant segmentation. Moreover, having a stain-invariant encoder, the HistoStarGAN model can, for the first

12

time, generalise stain transfer to unseen stainings.

4.1. Diverse Multi-Domain Stain Transfer

A trained model is able to perform diverse stain transfer between any pair of stainings seen during training. The diverse transfer is obtained by sampling different random codes, which are transformed by the mapping network into target-stain specific styles. Some examples of PAS image translations, alongside corresponding segmentations, are provided in Figure 2. The obtained translations are plausible histopathological images, where the structures of interest, glomeruli in this case, are preserved during translation. The difference between translations are at the level of microscopic structures (e.g. appearance of nuclei, the thickness of a membrane, etc.), see the Figure 3. Since the HistoStarGAN is multi-domain, the same model is also able to perform translations between other staining pairs, examples of which are provided in Figure 4.

4.2. Generalisation of stain transfer

Since HistoStarGAN is trained on a variety of stains, the generator's encoder is stain invariant, which for the first time enables virtual staining of unseen stains. Examples of which are presented in Figure 5, in which a new stain modality named H&E in addition to three double-stainings CD3-CD68, CD3-CD163 and CD3-CD206 are translated to stainings seen during training. Moreover, in Figure 6 the model is applied to the AIDPATH dataset (Bueno et al., 2020) composed of images which are publicly available variations of the PAS stain. This demonstrates that HistoStarGAN can generalise and obtain stain normalisation (column PAS) and stain transfer (other columns) simultaneously, alongside stain-invariant segmentation. Videos representing the exploration of the latent space during translation, are provided online[2].

4.3. Stain invariance

A model composed of the generator's encoder and the segmentation branch, can perform stain-invariant segmentation of WSIs across various stainings. Table 2 presents

---

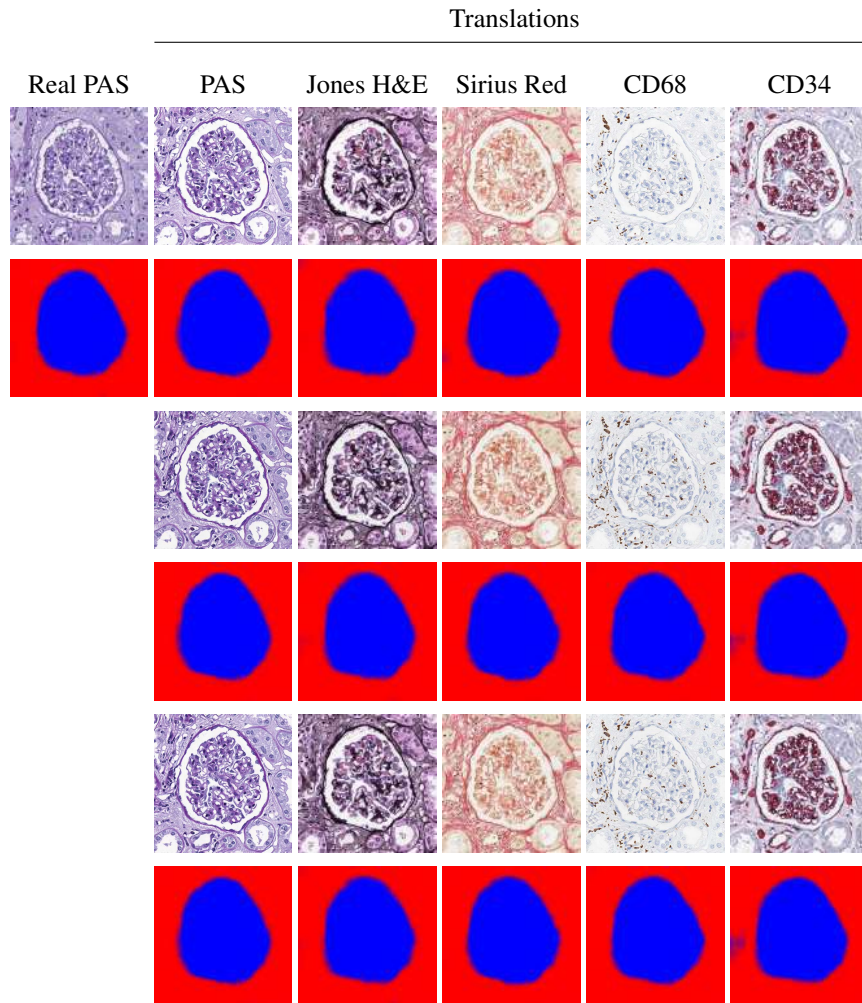[2]https://main.d33ezaxrmu3m4a.amplifyapp.com/

13

Figure 2: Diverse HistoStarGAN translations of a PAS glomeruli patch to target stains (including PAS) with corresponding segmentations. Fake images in each row are generated using the same random vector transformed into a stain-specific style by the Mapping network. The differences between translations are in microscopic structures (e.g. membrane weight or nucleus appearing), which is barely visible in these figures. Full resolution images, in which these differences are more visible, are available online.
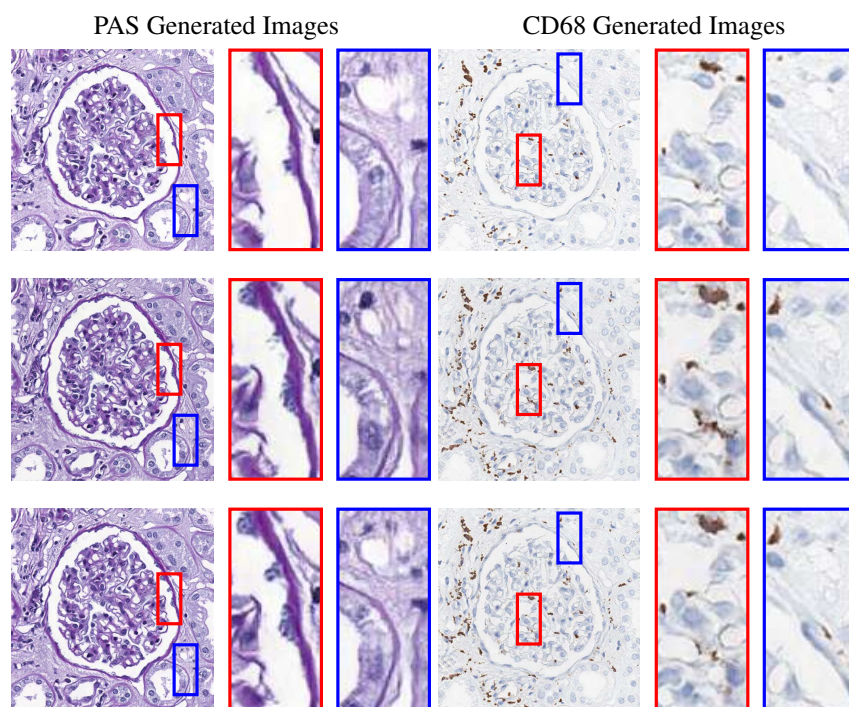
Figure 3: A closer look at the differences between translations.
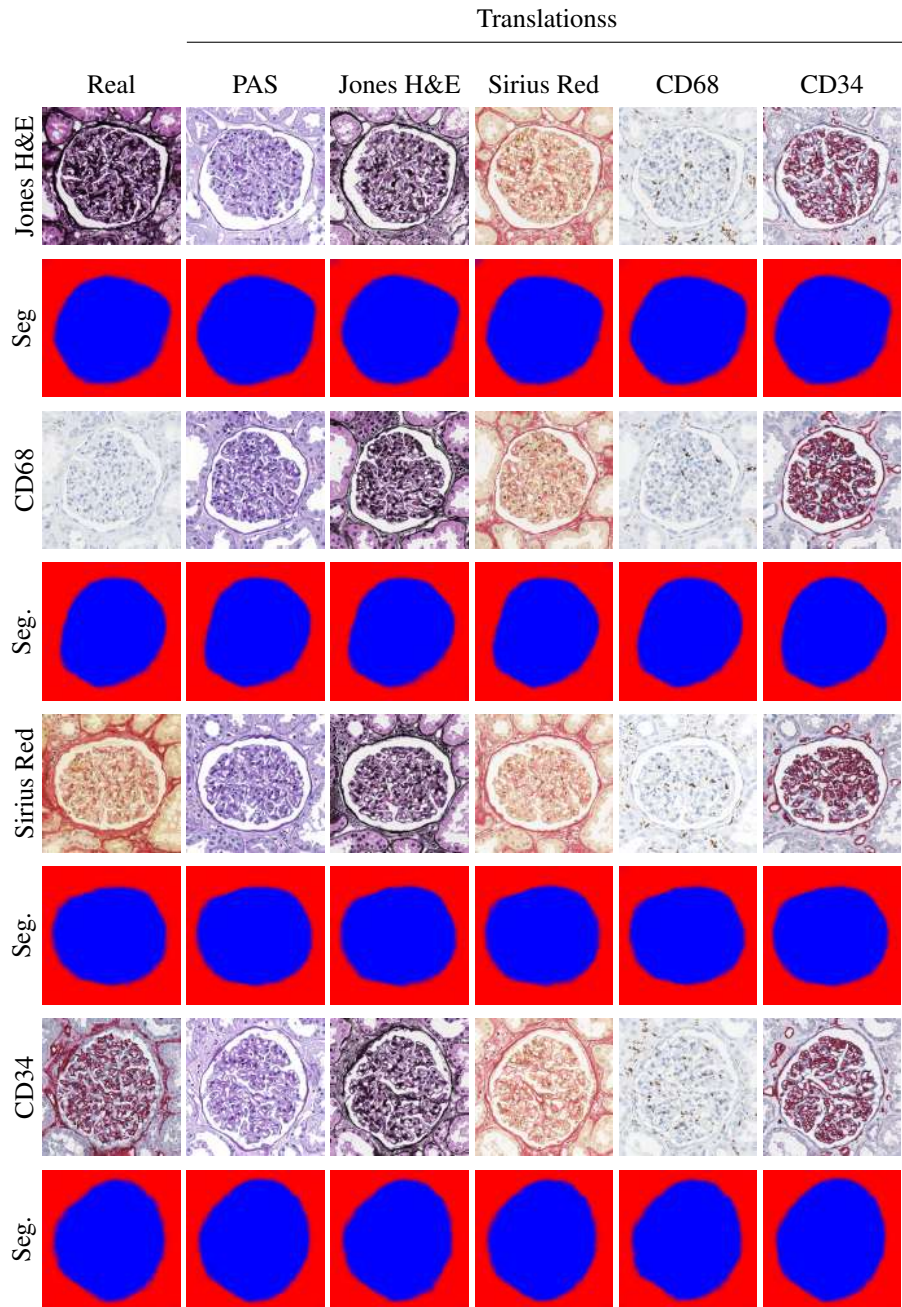
Figure 4: HistoStarGAN translations between different stains with corresponding segmentation. Each translation is obtained by using different latent codes.
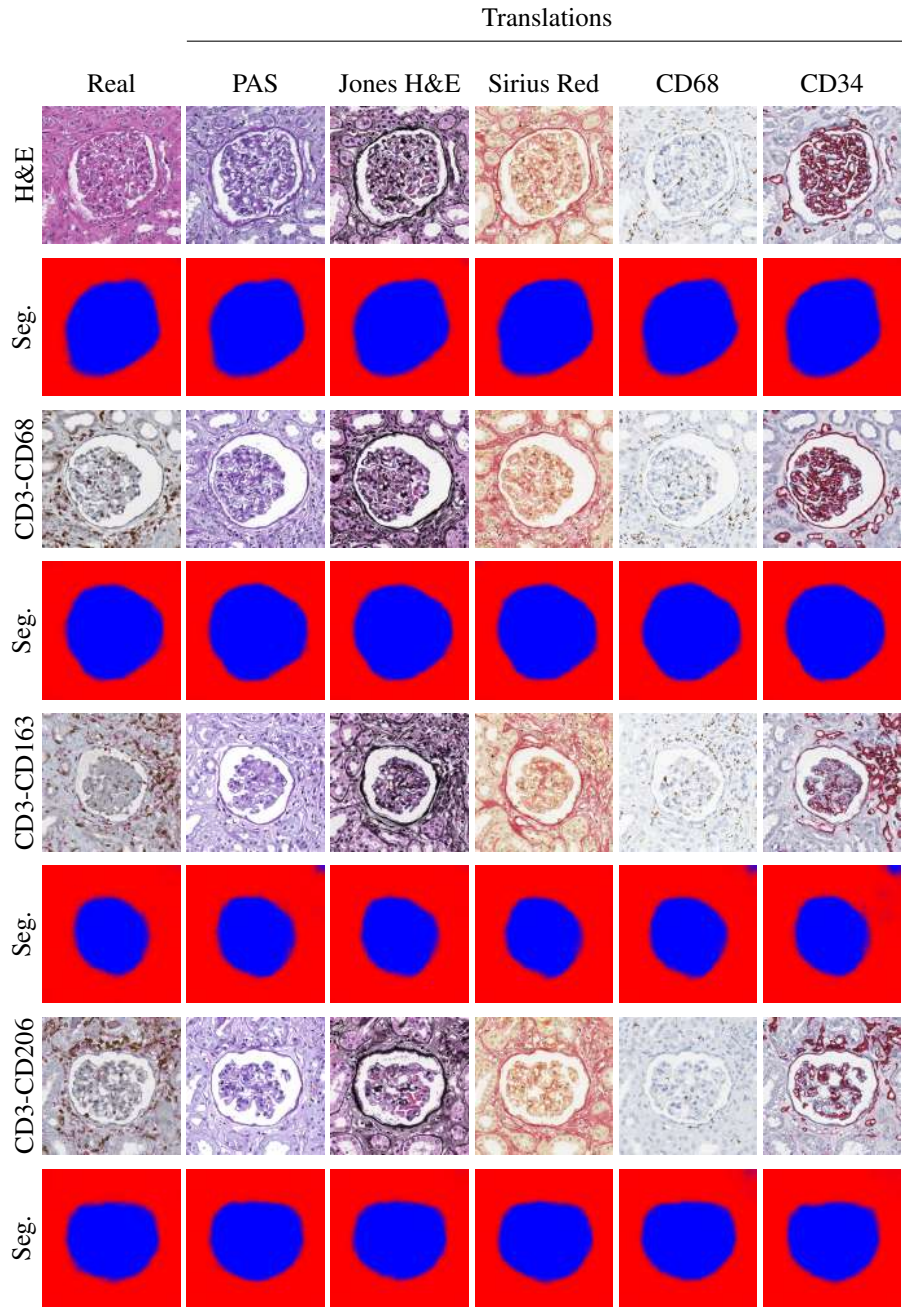
Figure 5: HistoStarGAN – generalisation of stain transfer and segmentation to unseen stain modalities.
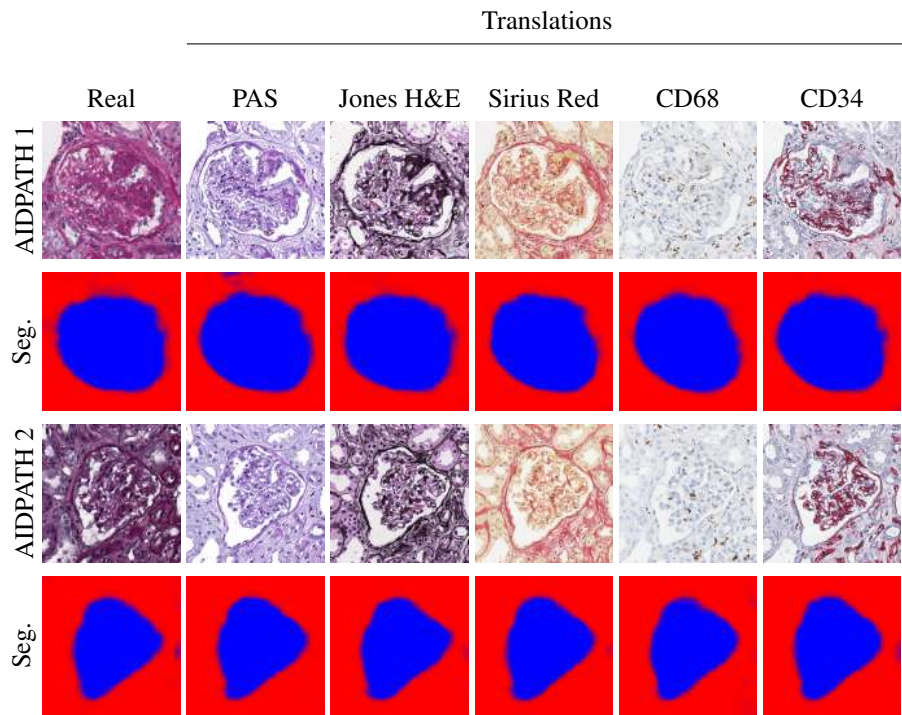
Figure 6: HistoStarGAN applied for stain normalisation, stain transfer and glomeruli segmentation of the publicly available AIDPATH (PAS-based) dataset.

Model	Score	Test Staining					
		PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
UDA-GAN	F_1	**0.903** (0.003)	0.849 (0.031)	0.720 (0.016)	**0.875** (0.016)	0.800 (0.033)	0.829 (0.072)
	Precision	0.878 (0.018)	0.787 (0.060)	0.688 (0.110)	0.835 (0.035)	0.720 (0.064)	0.782 (0.079)
	Recall	0.930 (0.014)	0.923 (0.010)	0.777 (0.095)	0.921 (0.007)	0.903 (0.016)	0.891 (0.064)
HistoStar-GAN	F_1	0.871 (0.009)	**0.870** (0.007)	**0.755** (0.006)	0.859 (0.004)	**0.840** (0.004)	**0.839** (0.048)
	Precision	0.845 (0.029)	0.864 (0.019)	0.845 (0.039)	0.883 (0.018)	0.839 (0.032)	0.855 (0.018)
	Recall	0.899 (0.016)	0.877 (0.007)	0.684 (0.024)	0.836 (0.017)	0.842 (0.024)	0.828 (0.084)

Table 2: Quantitative results for HistoStarGAN compared to UDA-GAN. Each model is trained on annotated PAS (source staining) and tested on different (target) stainings. Standard deviations are in parentheses, the highest F_1 scores for each staining are in bold.

the segmentation results for test WSIs from all stainings (virtually) seen during training. The model's performance is compared to UDA-GAN, which uses the same Cycle-GAN models for data augmentation. Since the patch size is 512×512, each patch is cropped to 508×508 during UDA-GAN training. The presented results are the averages of three independent training repetitions with corresponding standard deviations.

The HistoStarGAN can generalise across all virtually seen stainings during training, outperforming UDA-GAN trained using the same translation models. Apart from being an end-to-end model that simultaneously performs virtual staining and segmentation, HistoStarGAN also results in an increase in precision. This can be attributed to the fact that HistoStarGAN better recognises the negative tissue (less false positives). However, recall is lower, which indicates that more glomeruli (or parts of them) are missed compared to UDA-GAN. This could be the consequence of predicting a segmentation mask from the feature space directly, without skip-connections. However,

		Test Staining						
Model	**Score**	H&E	CD3	CD3-CD68	CD3-CD163	CD3-CD206	CD3-MS4A4A	Overall
UDA-GAN	F_1	0.681	0.648	0.258	0.260	0.330	0.330	0.418
		(0.031)	(0.111)	(0.062	(0.050)	(0.058)	(0.071)	(0.194)
	Precision	0.865	0.550	0.171	0.168	0.230	0.240	0.371
		(0.079)	(0.161)	(0.047)	(0.039)	(0.054)	(0.070)	(0.281)
	Recall	0.563	0.824	0.538	0.586	0.598	0.542	0.608
		(0.028)	(0.032)	(0.070)	(0.039)	(0.029)	(0.029)	(0.108)
HistoStar-GAN	F_1	**0.813**	**0.741**	**0.597**	**0.611**	**0.593**	**0.570**	**0.654**
		(0.022)	(0.009)	(0.011)	(0.015)	(0.014)	(0.012)	(0.099)
	Precision	0.855	0.850	0.835	0.891	0.881	0.882	0.866
		(0.018)	(0.007)	(0.022)	(0.021)	(0.026)	(0.025)	(0.022)
	Recall	0.777	0.656	0.465	0.465	0.447	0.422	0.539
		(0.056)	(0.011)	(0.017)	(0.021)	(0.020)	(0.016)	(0.144)

Table 3: Quantitative results for HistoStarGAN compared to UDA-GAN on unseen stains. Each model is trained on annotated PAS (source staining). Standard deviations are in parentheses, the highest F_1 scores for each staining are in bold.

extending the HistoStarGAN model with skip-connections between the encoder and segmentation branch, experimentally showed to negatively affect training stability.

4.3.1. Generalisation of stain invariance

To test the generalisation of HistoStarGAN, the model is applied to new stainings not seen during training. Using the annotations of four whole-slide images from each of the new stainings (H&E, CD3, CD3-CD68, CD3-CD163 and CD3-CD206), the averages of three independent training repetitions with corresponding standard deviations are presented in Table 3. Overall, HistoStarGAN generalises better and is more robust compared to UDA-GAN. A potential cause of UDA-GAN failure in some stains can be its learning process where stain invariance is forced only on the pixel level. Although the HistoStarGAN uses the identical CyleGAN translations, the framework extracts stain-invariant features, which are better transferred to unseen stains. Moreover, the

segmentation branch of HistoStarGAN is trained on a dataset with more variety since HistoStarGAN translations are also included (see Figure 1).

5. Ablation Studies

HistoStarGAN model builds upon StarGANv2 in several aspects – the first extension is attaching a segmentation module to the generator; the second is using pre-trained CycleGANs to create the training dataset, and the third is fine-tuning the segmentation module using only source data. In the following, each of these aspects will be discussed via an ablation study, in a bottom-to-top direction as illustrated in Figure 7.
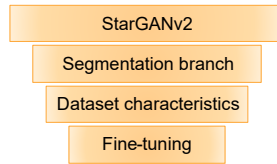
Figure 7: Illustration of the differences between the StarGANv2 and HistoStarGAN. Each of these differences is justified via an ablation study.

| Model | Score | Test Staining | | | | | |
		PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
HistoStar-GAN	F_1	0.820	0.816	0.705	0.792	0.777	0.782
	Precision	0.779	0.791	0.764	0.792	0.802	0.761
	Recall	0.865	0.843	0.655	0.782	0.795	0.779
HistoStar-GAN (fine-tuned)	F_1	**0.876**	**0.875**	**0.762**	**0.855**	**0.842**	**0.842**
	Precision	**0.851**	**0.871**	**0.861**	**0.886**	**0.842**	**0.862**
	Recall	**0.902**	**0.879**	**0.683**	**0.825**	**0.843**	**0.826**

Table 4: Fine-tuning effects on the segmentation model's performance, in which the model is fine-tuned for one epoch.

5.1. Fine-Tuning

Table 4 demonstrates that fine-tuning the segmentation branch using the fixed generator's encoder increases segmentation performance. In practice, fine-tuning for just a single epoch on an imbalanced source dataset increases overall performance on virtually seen stains (test patient WSIs) by around 6% in overall F_1-score. However, longer fine-tuning, although potentially beneficial for particular stainings, does not offer any benefits. Also, the performance on the source's validation set does not correlate with the obtained improvements, i.e. the fine-tuned model with the lowest validation loss does not bring the best overall results on unseen stains. Thus, fine-tuning for one epoch is chosen.

5.2. Dataset Characteristics

A balanced and (virtually) fully annotated dataset is used to train HistoStarGAN. A labelled dataset is required since the segmentation branch is trained in a supervised manner, and thus alternative data sampling strategies such as random data sampling (Gadermayr et al., 2018; Vasiljević et al., 2021a) are not suitable. However, the fully annotated dataset does not need to be balanced. Thus, using the findings of Lampert et al. (2019), an imbalanced source (PAS) dataset is formed where all glomeruli in the training patient images are extracted (662) and seven times as many negative patches (4634). This dataset is translated using CycleGAN to all target stains (CycleGAN models are always trained in an unsupervised way, on random patches). Thus, a fully annotated and highly imbalanced dataset is created on top of which the HistoStarGAN model is trained.

The results obtained in this setting are presented in Figure 8, in which one glomeruli patch from the PAS stain is translated into multiple stainings using different latent codes. The obtained translations for one stain pair differ from each other in terms of glomeruli texture and surrounding structures. The model is able to change the size of glomeruli by changing the size of the Bowman's capsule (white space), in addition to varying the appearance of stain-specific markers (i.e. macrophages in CD68 stain). Nevertheless, the segmentation branch successfully recognises all of glomeruli variations. Compared to training with a balanced dataset, this setting offers more translation

variability related to the surrounding tissue and structure of the glomeruli themselves. However, although globally these translations look realistic, the internal structure inside the glomeruli are not. Since in HistoStarGAN these translations must also be successfully segmented, this leads to an increase in the false positive rate. For example, it is evident that the produced translations often contain 'artificial looking patterns' such as a tendency to group microscopic structures (e.g. nuclei) into diagonal, horizontal or vertical stripes (in Figure 8 this is mostly visible in translations to Sirius Red and Jones H&E). Therefore, the proposed model uses a balanced dataset since it greatly reduces such possibilities.

5.3. Segmentation Branch

Without the segmentation branch, HistoStarGAN is reduced to StarGANv2 (see Figure 7). Moreover, removing the segmentation module removes the requirement for the dataset to be fully annotated. Thus, such a model can be trained using a dataset composed of random patches (uniformly sampled in each stain), as well as created by CycleGAN translations of the annotated stain (PAS), both balanced and imbalanced. Each of these will be separately analysed.

CycleGAN-Generated Balanced Dataset. A StarGANv2 model trained on a balanced dataset produced by CycleGAN translations of the annotated stain, can also result in diverse multi-domain translations between multiple stainings, as illustrated in Figure 9. Usually, the glomeruli structures are visually preserved. However, when measuring the quality of obtained translation using the performances of pre-trained segmentation models for each staining (baseline models trained in a fully-supervised setting), the conclusion can be different. Figure 10 presents a visual comparisons of HistoStarGAN and StarGANv2, where translations obtained by StarGANv2 are segmented using pre-trained models from each staining. Since both models can obtain diverse translations, Figure 10 represent the average translation and average segmentation over 50 random latent codes. It can be seen that HistoStarGAN gives a more accurate segmentations, especially in difficult cases, e.g. sclerotic glomeruli, rows 5 and 7. Nevertheless, in the absence of a segmentation branch, the dataset composition itself cannot be a strong
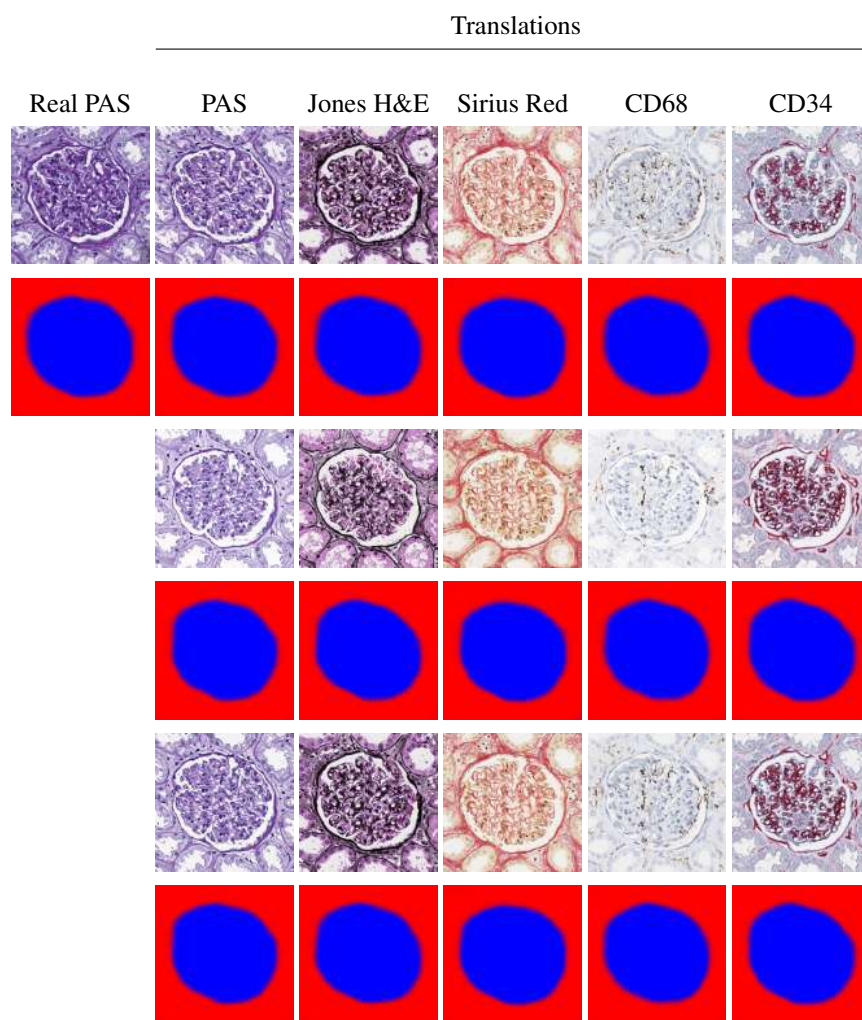
Figure 8: HistoStarGAN translations using a model trained using a CycleGAN-generated imbalanced annotated dataset. The images in each column are generated using the same latent codes.
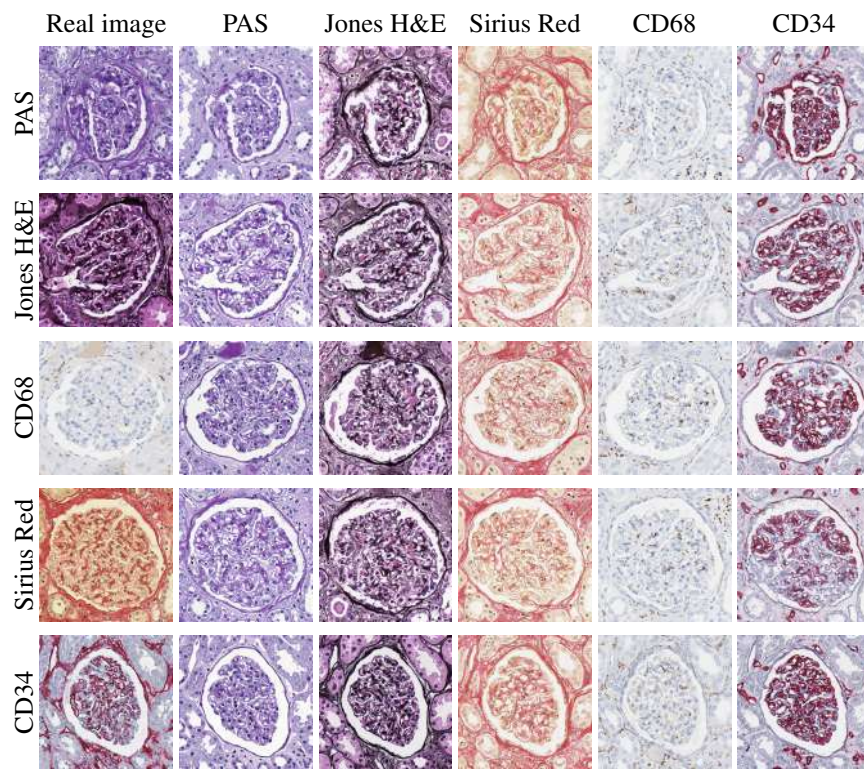
Figure 9: StarGANv2 trained on CycleGAN-generated translations of a balanced PAS dataset to other stains. Each image in generated using different latent code.
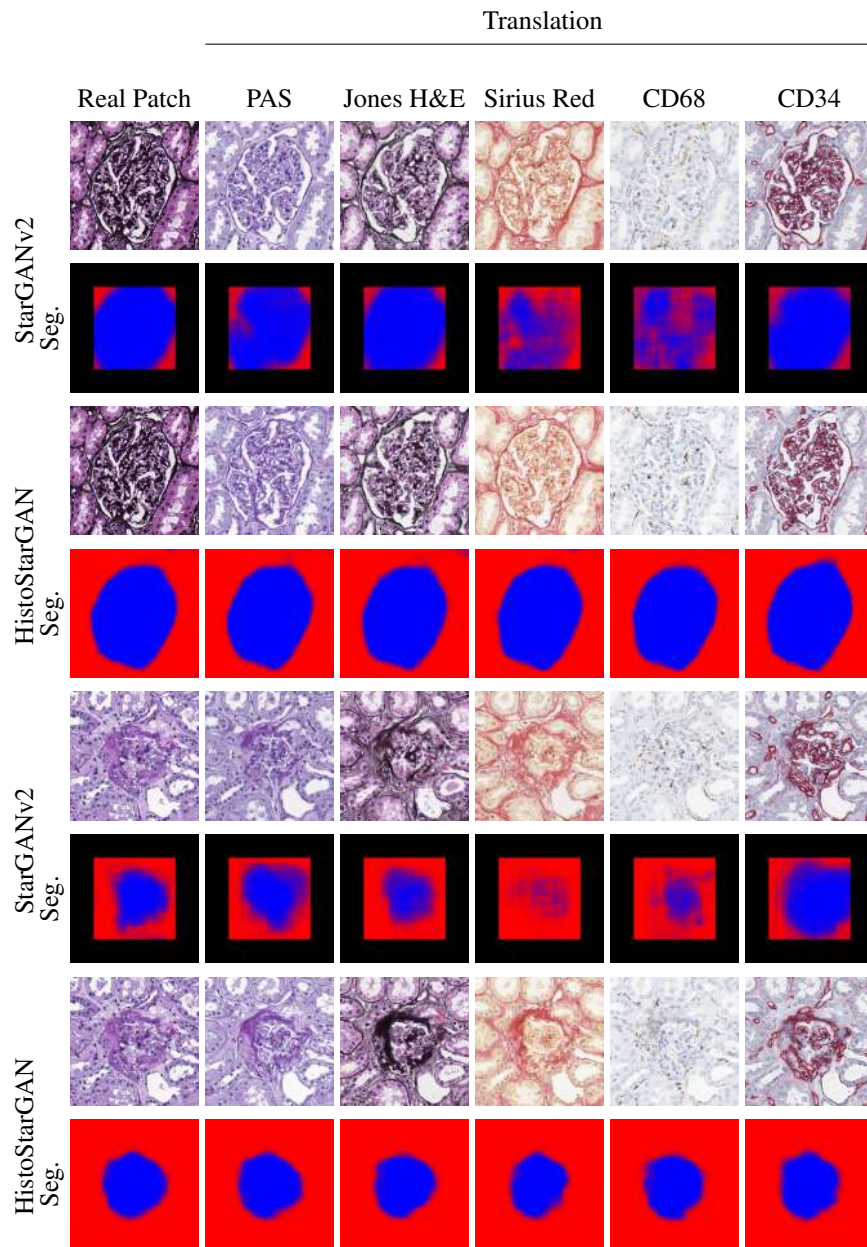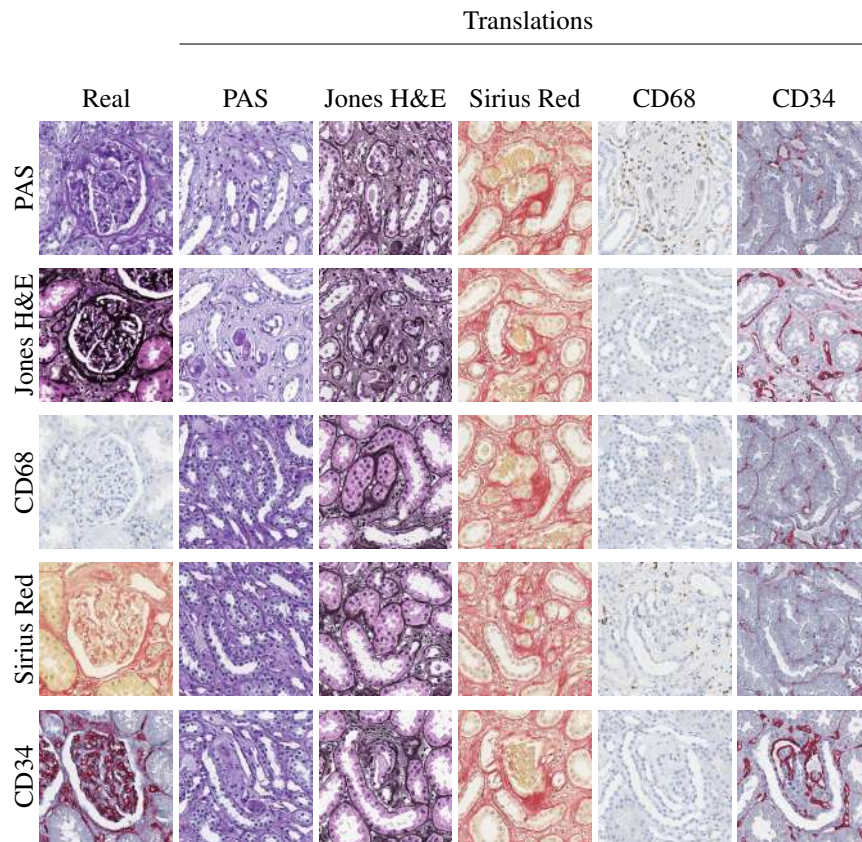
Figure 10: StarGANv2 trained on a balanced CycleGAN-generated dataset compared to HistoStarGAN. The segmentations for StarGANv2 are obtained by stain-specific, pre-trained baseline models. Each translation and segmentation is averaged over 50 random latent codes. N.B: HistoStarGAN yields more accurate segmentations. Pre-trained models result in 324×324 pixel images which are placed in the centre of a 512×512 pixel black squares.

guarantee that glomeruli structures will be preserved. Thus, an explicit requirement, such as attaching the proposed segmentation branch, is beneficial to ensure the correctness of the translation and robust segmentation.

Figure 11: StarGANv2 model trained on CycleGAN-generated translations of an imbalanced PAS dataset to other stains. Each image is generated using different latent codes.

CycleGAN-Generated Imbalanced Dataset. If the dataset used to train the StarGANv2 is imbalanced, the model can no longer preserve the structure of interest. Some examples of stain transfers obtained with this model are presented in Figure 11. Since the HistoStarGAN model trained using the same imbalanced dataset does not have such a problem, this demonstrates the significance of the proposed segmentation branch.
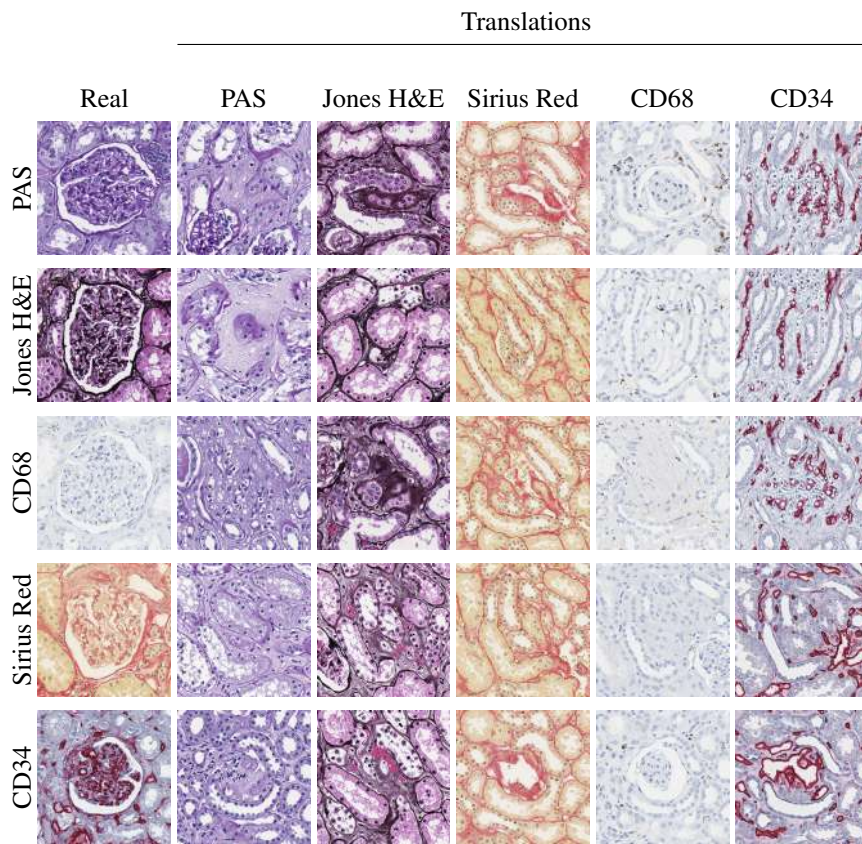
27

Figure 12: StarGANv2 model trained on random patches extracted from all stains, applied to translated glomeruli patches. The images in each column are generated using different latent codes.

Random Dataset:. StarGANv2 can also be trained on a dataset composed of random patches extracted from each stain (PAS, Jones H&E, Sirius Red, CD68 and CD34). Thus, 5000 random patches from the training patients are extracted uniformly, and the total dataset containing 25 000 images is used to train the StarGANv2 model. As in the CycleGAN-generated imbalanced data setting, this model cannot preserve the structure of interest, as demonstrated in Figure 12. Moreover, it is prone to bigger alterations of microscopic structures, which limits the usefulness of such translations and confirms the benefits of the proposed HistoStarGAN model.
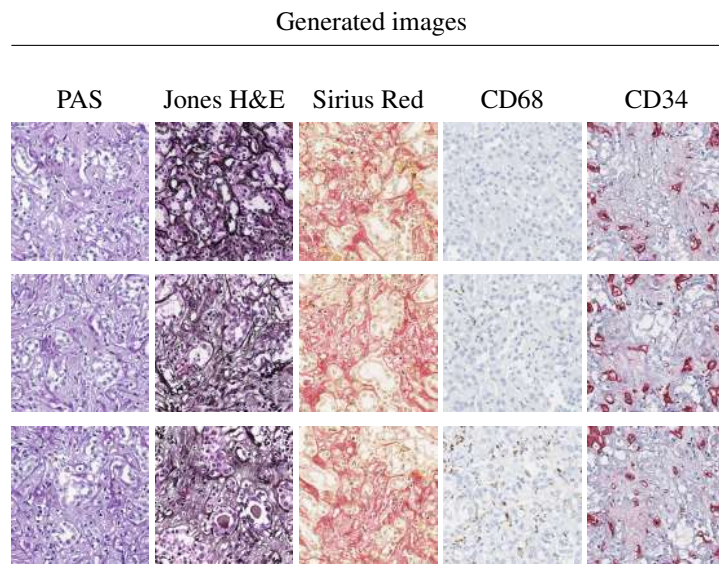
6. KidneyArtPathology Dataset

Figure 13: The HistoStarGAN model generates plausible histopathological images from random noise.
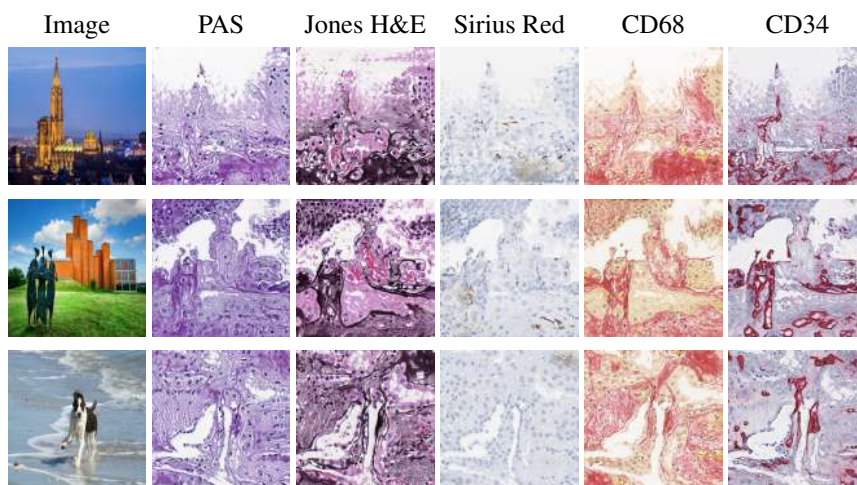
Figure 14: KidneyArtPathology - histopathological image generation from natural images[3].

Here we describe a new dataset released to encourage the progress of deep learning-based solutions in the field of renal pathology. The dataset contains 5000 images from five stainings, in a resolution of 512×512 pixels, fully annotated for the task of glomeruli segmentation. To achieve this, a HistoStarGAN model is trained using a dataset that is as representative as possible, i.e. two immunohistochemical stainings (CD68 and CD34) in addition to three histochemical stainings (PAS, Jones H&E and Sirius Red). Thus, the model can generate various translations in multiple stainings, which has allowed the creation of the KidneyArtPathology dataset [4]. The dataset is composed of HistoStarGAN translations of PAS stain images (the same as those used to generate the annotated dataset used for HistoStarGAN training) which are translated using 10 random latent vectors into each stainings, including PAS.

Moreover, the associated pre-trained HistoStarGAN model (also publicly available) can be further used to augment private datasets by augmenting the number of stainings

---

[4]KidneyArtPathology has been release online, https://main.d33ezaxrmu3m4a.amplifyapp.com/

to those used during HistStarGAN's training. There are several benefits of such a dataset, which fall under three categories:

- Pathological – Non-invasive Pathology Training, the diverse appearance of glomeruli can be helpful in the early stages of a pathologist's training.

- Benchmarks – the absence of publicly available real-world datasets poses huge challenges for rigorous comparison in the literature. Thus, such a large collection of annotated patches can serve as a benchmark.

- Domain adaptation – The data can be used in addition to a private dataset (which can contain limited data) to build more robust models, e.g. as an augmentation or domain adaptation strategy.

New histopathological images: Since the style of a stain is encoded by the mapping network, it is possible to generate new patches in each training staining by providing a random image, rather than a source histopathological patch, to be translated to a given stain. In addition to a fully-annotated glomeruli dataset, new histopathological images in different stains can be generated. Some generated examples are provided in Figure 13. Alternatively, by providing non-histopathological image, the HistoStarGAN is able to convert it to a histopathological image, examples of which are provided in Figure 14 (although medical use of this is admittedly most likely limited to non-existent, it is an interesting side property of HistoStarGAN) .

7. Limitations and Opportunities

HistoStarGAN is an end-to-end model that can segment and generate diverse plausible histopathological images alongside their segmentation masks. However, for some latent codes the model can produce specific artefacts in the translations. These are usually well-incorporated into the overall texture of the image, which makes them not obvious at first glance. A closer look at one such example is given in Figure 15. The primary hypothesis is that the discriminator does not have enough capacity to spot these artefacts, and thus a more sophisticated discriminator could be considered.
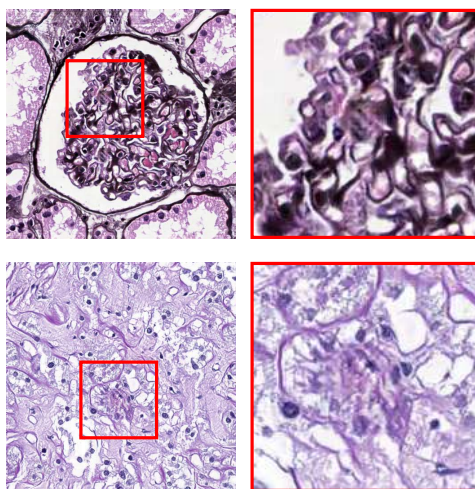
31

Figure 15: HistoStarGAN - common artefacts. N.B. the artefacts are well-incorporated into the image texture and not obviously noticeable at first glance.

Furthermore, the performance of the HistoStarGAN model can be affected by the choice of source and target stainings and the quality of CycleGAN translations, which remains to be explored in future work. HistoStarGAN is based on CycleGAN translations, which are proven to be noisy (Vasiljević et al., 2021b, 2022), and therefore some of these invisible artefacts may be recreated by HistoStarGAN. Thus, it is unclear what information from the source domain is preserved in translations and whether any of the applied augmentations can perturb them. It is possible that incorporating additional real examples from the target stainings can improve translation quality. Additionally, it has been confirmed experimentally that the choice of loss functions used for the diversity loss can significantly affect the HistoStarGAN translations. For example, the model will be forced to perform structural changes if diversity loss is set to maximise the structural similarity between images (this is illustrated on the project's website linked to above). Although from a style-transfer perspective, such a solution is interesting, changing internal structures in this way is not biologically justifiable. However, according to a recent study (Yamashita et al., 2021), medically-irrelevant augmentation

can increase model robustness; thus, this represents another exciting direction for the work ahead.

Nevertheless, the model has great potential to be used for tackling the lack of annotations in histopathological datasets. Depending on the availability of labels in a given target domain, diverse HistoStarGAN translations with corresponding segmentation masks can help training from unsupervised to fully supervised settings, using one or more source (annotated) domains.

Finally, it is crucial to note that translations obtained by the HistoStarGAN model should not be used for diagnostic purposes. All images are artificially generated, and the model can perturb diagnostic markers. Thus, the dataset composed of HistoStarGAN translations should only be used for general-purpose analysis related to glomeruli (e.g. counting).

8. Conclusions

The HistoStarGAN model represents the first end-to-end trainable solution for simultaneous stain normalisation, stain transfer and stain invariant segmentation. For the first time, obtaining highly plausible stain transfer from unseen stainings is possible without any additional change to the model (e.g. fine-tuning). Moreover, the model achieves new state-of-the-art results in stain invariant segmentation, successfully generalising to six unseen stainings. The proposed solution is general and extensible to new stainings or use cases.

Being able to generate diverse translations for a given input, the proposed solution paved the way to generate KidneyArtPathology, an artificially created and fully annotated data set illustrating a broad spectrum of synthetic image data ranging from biologically usable results, to visually plausible but biologically useless output.

However, the obtained virtual staining can realistically change the appearance of microscopic structures, such as the number of visible nuclei, size of Bowman's capsule or membrane thickness. Thus, it is ill-advised to use these translations for applications that rely on the assumption that the translation process preserves all structures in the original image. Moreover, such virtually stained images must not be used for

diagnostic purposes since diagnostic markers can be altered. As such, this study goes towards saving resources and time for annotating the high-quality datasets required for developing deep learning based models, not to replace the process of physical staining.

Nevertheless, it remains an open question why the model's performance varies across different stainings. One hypothesis is that the model, being based on Cycle-GAN translations, replicates some of their limitations. However, the influence of these CycleGAN models and the choice of the stainings used in the dataset remains to be explored in future work.

Acknowledgements

References

Kyungjune Baek, Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Hyunjung Shim. Rethinking the truly unsupervised image-to-image translation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14154–14163, 2021.

John D Bancroft and Marilyn Gamble. *Theory and Practice of Histological Techniques (sixth Edition)*. Elsevier Health Sciences, sixth edition edition, 2008.

Aïcha BenTaieb and Ghassan Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE Transactions on Medical Imaging*, 37(3):792–802, 2017.

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

Gloria Bueno, Lucia Gonzalez-Lopez, Marcial Garcia-Rojo, Arvydas Laurinavicius, and Oscar Deniz. Data for glomeruli characterization in histopathological images. *Data in Brief*, 29:105314, 2020.

Hyungjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. Neural stain-style transfer learning using gan for histopathological images. *arXiv preprint arXiv:1710.08543*, 2017.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8789–8797, 2018.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8188–8197, 2020.

Cong Cong, Sidong Liu, Antonio Di Ieva, Maurice Pagnucco, Shlomo Berkovsky, and Yang Song. Semi-supervised adversarial learning for stain normalisation in histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 581–591, 2021.

Thomas de Bel, John Melle Bokhorst, Jeroen van der Laak, and Geert Litjens. Residual cyclegan for robust domain transformation of histopathological tissue slides. *Medical Image Analysis*, 70, 2021.

Michael Gadermayr, Vitus Appel, Mara Barbara 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 (MICCAI), volume 11071, pages 165–173, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, volume 27, pages 2672–2680, 2014.

Anne Grote, Nadine S Schaadt, Germain Forestier, Cédric Wemmert, and Friedrich Feuerhake. Crowdsourcing of histological image labeling and object delineation by medical students. *IEEE Transactions on Medical Imaging*, 38:1284–1294, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017.

Shruti Jadon. A survey of loss functions for semantic segmentation. In *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, pages 1–7, 2020.

Laya Jose, Sidong Liu, Carlo Russo, Annemarie Nadort, and Antonio Di Ieva. Generative adversarial networks in digital pathology and histopathological image processing: A review. *Journal of Pathology Informatics*, 12:43, November 2021.

Hongtao Kang, Die Luo, Weihua Feng, Shaoqun Zeng, Tingwei Quan, Junbo Hu, and Xiuli Liu. Stainnet: A fast and robust stain normalization network. *Frontiers in Medicine*, 8, 2021.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020b.

Jing Ke, Yiqing Shen, Xiaoyao Liang, and Dinggang Shen. Contrastive learning based stain normalization across multiple tumor in histopathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 571–580, 2021.

Thomas Lampert, Odyssée Merveille, Jessica Schmitz, Germain Forestier, Friedrich Feuerhake, and Cédric Wemmert. Strategies for training stain invariant CNNs. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 905–909, 2019.

Adrian B Levine, Jason Peng, David Farnell, Mitchell Nursey, Yiping Wang, Julia R Naso, Hezhen Ren, Hossein Farahani, Colin Chen, Derek Chiu, Aline Talhouk, Brandon Sheffield, Maziar Riazy, Philip P Ip, Carlos Parra-Herran, Anne Mills, Naveena Singh, Basile Tessier-Cloutier, Taylor Salisbury, Jonathan Lee, Tim Salcudean, Steven Jm Jones, David G Huntsman, C Blake Gilks, Stephen Yip, and Ali Bashashat. Synthesis of diagnostic quality cancer pathology images by generative adversarial networks. *The Journal of Pathology*, 252(2):178–188, 2020.

Joshua J Levy, Christopher R Jackson, Aravindhan Sriharan, Brock C Christensen, and Louis J Vaickus. Preliminary evaluation of the utility of deep generative histopathology image translation at a mid-sized nci cancer center. *bioRxiv*, 2020.

Dan Li, Hui Hui, Yingqian Zhang, Wei Tong, Feng Tian, Xin Yang, Jie Liu, Yundai Chen, and Jie Tian. Deep learning for virtual histological staining of bright-field microscopic images of unlabeled carotid artery tissue. *Molecular Imaging and Biology*, 22(5):1301–1309, 2020.

Yiyang Lin, Bowei Zeng, Yifeng Wang, Yang Chen, Zijie Fang, Jian Zhang, Xiangyang Ji, Haoqian Wang, and Yongbing Zhang. Unpaired multi-domain stain transfer for

kidney histopathological images. In *AAAI Conference on Artificial Intelligence*, 2022.

Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE Transactions on Medical Imaging*, 40:1977–1989, 2021.

Xiaoxuan Liu, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis. *The Lancet Digital Health*, 1(6):e271–e297, 2019.

Ying-Chih Lo, I-Fang Chung, Shin-Ning Guo, Mei-Chin Wen, and Chia-Feng Juang. Cycle-consistent gan-based stain translation of renal pathology images with glomerulus detection application. *Applied Soft Computing*, 98:106822, 2021.

Raphaël Marée, Loïc Rollus, Benjamin Stévens, Renaud Hoyoux, Gilles Louppe, Rémy Vandaele, Jean-Michel Begon, Philipp Kainz, Pierre Geurts, and Louis Wehenkel. Collaborative analysis of multi-gigapixel imaging data using cytomine. *Bioinformatics*, 32(9):1395–1401, 2016.

Odyssee Merveille, Thomas Lampert, Jessica Schmitz, Germain Forestier, Friedrich Feuerhake, and Cédric Wemmert. An automatic framework for fusing information from differently stained consecutive digital whole slide images: A case study in renal histology. *Computer Methods and Programs in Biomedicine*, 208:106157, 2021.

Atefeh Ziaei Moghadam, Hamed Azarnoush, Seyyed Ali Seyyedsalehi, and Mohammad Havaei. Stain transfer using generative adversarial networks and disentangled features. *Computers in Biology and Medicine*, 142:105219, 2022.

Francesco Piccialli, Vittorio Di Somma, Fabio Giampaolo, Salvatore Cuomo, and Giancarlo Fortino. A survey on deep learning in medicine: Why, how and when? *Information Fusion*, 66:111–137, 2021. doi: 10.1016/j.inffus.2020.09.006.

Aman Rana, Alarice Lowe, Marie Lithgow, Katharine Horback, Tyler Janovitz, Annacarolina Da Silva, Harrison Tsai, Vignesh Shanmugam, Akram Bayat, and Pratik Shah. Use of deep learning to develop and analyze computational hematoxylin and eosin staining of prostate core biopsy images for tumor diagnosis. *JAMA Network Open*, 3(5):e205111–e205111, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241, 2015.

Pegah Salehi and Abdolah Chalechale. Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. In *International Conference on Machine Vision and Image Processing (MVIP)*, pages 1–7, 2020.

Marin Scalbert, Maria Vakalopoulou, and Florent Couzinié-Devy. Test-time image-to-image translation ensembling improves out-of-distribution generalization in histopathology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2022.

Terrence J. Sejnowski. *The Deep Learning Revolution*. The MIT Press, 2018. ISBN 9780262038034.

M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 953–956, 2019.

Aman Shrivastava, Will Adorno, Yash Sharma, Lubaina Ehsan, S. Asad Ali, Sean R. Moore, Beatrice C. Amadi, Paul Kelly, Sana Syed, and Donald E. Brown. Self-

attentive adversarial stain normalization. In *International Conference on Pattern Recognition*, pages 120–140, 2021.

Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical Image Analysis*, 67, 2021.

David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen Van Der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019.

Maximilian E. Tschuchnig, Gertie J. Oostingh, and Michael Gadermayr. Generative adversarial networks in digital pathology: A survey on trends and future potential. *Patterns*, 1(6):100089, 2020.

Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: The path to the clinic. *Nature Medicine*, 27(5):775–784, May 2021.

Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: Methodological failures and recommendations for the future. *NPJ Digital Medicine*, 5(1):1–8, 2022.

Jelica Vasiljević, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks. *Neurocomputing*, 460:277–291, 2021a.

Jelica Vasiljević, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Self adversarial attack as an augmentation method for immunohistochemical stainings. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1939–1943, 2021b.

Jelica Vasiljević, Zeeshan Nisar, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Cycleggan for virtual stain transfer:is seeing really believing? *Artificial Intelligence in Medicine*, 2022.

Sophia Wagner, Nadieh Khalili, Raghav Sharma, Melanie Boxberg, Carsten Marr, Walter de Back, and Tingying Peng. Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 257–266, 2021.

Yuan Xue, Jiarong Ye, Qianying Zhou, L. Rodney Long, Sameer Antani, Zhiyun Xue, Carl Cornwell, Richard Zaino, Keith C. Cheng, and Xiaolei Huang. Selective synthetic augmentation with HistoGAN for improved histopathology image classification. *Medical Image Analysis*, 67:101816, 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101816.

Rikiya Yamashita, Jin Long, Snikitha Banda, Jeanne Shen, and Daniel L Rubin. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Transactions on Medical Imaging*, 40:3945–3954, 2021.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.