Towards Histopathological Stain Invariance by Unsupervised Domain Augmentation using Generative Adversarial Networks

Jelica Vasiljević
ICube, University of Strasbourg, CNRS (UMR 7357), France
University of Belgrade, Belgrade, Serbia
University of Kragujevac, Kragujevac, Serbia
jvasiljevic@unistra.fr

Friedrich Feuerhake
Institute of Pathology, Hannover Medical School, Germany
University Clinic, Freiburg, Germany
Feuerhake.Friedrich@mh-hannover.de

Cédric Wemmert
ICube, University of Strasbourg, CNRS (UMR 7357), France
wemmert@unistra.fr

Thomas Lampert
ICube, University of Strasbourg, CNRS (UMR 7357), France
lampert@unistra.fr

December 24, 2020

Abstract

The application of supervised deep learning methods in digital pathology is limited due to their sensitivity to domain shift. Digital Pathology is an area prone to high variability due to many sources, including the common practice of evaluating several consecutive tissue sections stained with different staining protocols. Obtaining labels for each stain is very expensive and time consuming as it requires a high level of domain knowledge. In this article, we propose an unsupervised augmentation approach based on adversarial image-to-image translation, which facilitates the training of stain invariant supervised convolutional neural networks. By training the network on one commonly used staining modality and applying it to images that include corresponding, but differently stained, tissue structures, the presented method demonstrates significant improvements over other approaches. These benefits are illustrated in the problem of glomeruli segmentation in seven different staining modalities (PAS, Jones H&E, CD68, Sirius Red, CD34, H&E and CD3) and analysis of the learned representations demonstrate their stain invariance.

1 Introduction

The introduction of Whole Slide Imaging (WSI) scanners enables the production of huge amounts of histological image data. A crucial step in the tissue preparation process is staining, which consists of dyeing the tissue slices with a specific stain in order to highlight internal structures when viewed under a microscope. Image datasets in digital pathology often consist of (nearly consecutive) tissue slides stained differently, each staining providing specific morphological information. As each staining highlights different components of the tissue, even consecutive sections

largely representing almost identical anatomical structures can take on very different appearances. On the other hand, the staining process itself is prone to high variability due to inter-subject differences, the staining procedure in a specific lab, or scanner characteristics, leading to different appearances of the tissues coloured with the same stain [1]. This high variability in the appearance of digital histological images affects the performance of machine learning algorithms used for automatic WSI analysis [2]. In specific applications such as the detection or segmentation of objects that maintain morphological consistency across stains (e.g. nuclei, glomeruli, etc), stain variability could be considered, from a computer vision viewpoint, as noise that dramatically hampers the performance of machine learning models. This motivates the development of image analysis methods that work across different stains in order to solve tasks related to morphologically consistent structures [3–6].

Most state-of-the-art computer vision algorithms are sensitive to domain shift [7]. This means that models trained for a specific task on histological images of stain A exhibit a significant drop in performance when applied to histological images of stain B (for the same task) [3,5,6] or variability of stain A (e.g. images from other laboratories) [2].

Typical solutions consider either fine tuning existing models or training a new model for each data variation, which requires the acquisition of additional labeled data. Nevertheless, medical image datasets are often characterised by their scarcity of annotations [8] and obtaining properly annotated images is time consuming and costly as expert knowledge is required for most complex annotations. Alternative approaches such as crowd sourcing are limited by the need of specific task design and intensive training [9]. While plenty of works address the problem of intra-stain variation [10–16] there are few that address the problem of inter-stain variation [3,4,6,17,18].

Common approaches to deal with inter-stain variability are stain colour augmentation, stain normalisation, and stain transfer [19]. A recent study on the effects of colour augmentation and stain normalisation for computational pathology shows that stain colour augmentation has greater impact on the robustness of machine learning models compared to stain normalisation [2]. On the other hand, stain transfer as a technique for virtual staining [17, 18] addresses the general problem of a lack of annotations in the medical domain and can be applied to various problem classes.

In this article we improve upon existing work [6] to introduce a general approach that combines stain transfer and stain augmentation to learn stain invariant representations[1]. Generally, it is assumed that annotations exist for one staining (called the source staining) and that other stainings (called the target stainings) are unannotated. The goal is to use the available annotations to learn a stain invariant representation by taking advantage of stain transfer techniques.

We focus on the segmentation of glomeruli, a highly relevant functional unit of the kidney, and consider a commonly used staining in renal pathology—Periodic acid-Schiff reaction (PAS)—as the source domain and two commonly applied histochemical (Sirius Red and Jones' basement membrane stain—Jones H&E) and two immunohistochemical stainings (CD34 highlighting blood vessel endothelium, CD68 for macrophages) as targets.

Specifically, we train GAN-based stain-translation models [20, 21] in order to achieve realistic translation between source and target stains.

In turn, these allow a segmentation model to be trained on source data augmented with random translations to the target stains. Compared to previous state-of-the-art approaches to stain-invariant segmentation [4, 6] the presented strategy demonstrates a significant improvement in segmentation performance.

The main contributions of this article are:

- to introduce a model that is capable of segmenting several stainings, i.e. is stain invariant, and outperforms all evaluated state-of-the-art approaches;

- an augmentation method that is general and unsupervised;

- to the best of our knowledge, this is the first approach to use image-to-image translation to force stain-invariant features to be learnt, rather than using the typical domain adaptation process that results in domain (staining) specific models, albeit different from the source staining;

- to increase flexibility regarding the direction of translation described in the literature [4];

- to present a discussion on the limitations of GAN-based image-to-image translations.

The remainder of this article is organised as follows: in Section 2, literature concerned with stain transfer and stain invariant approaches is reviewed. Section 3 gives a detailed description of the proposed method and dataset. Section 4 presents the experimental results with comparison to existing state-of-the-art methods. Finally, Section 5 analyses the approach in terms of stain invariance, quality of translations, and its limitations.

---

[1] 'Stain invariant' refers to the fact that the same model can be applied to multiple stains (possibly those not seen during training) without modification of the data at test time or adaptation of the model after training.
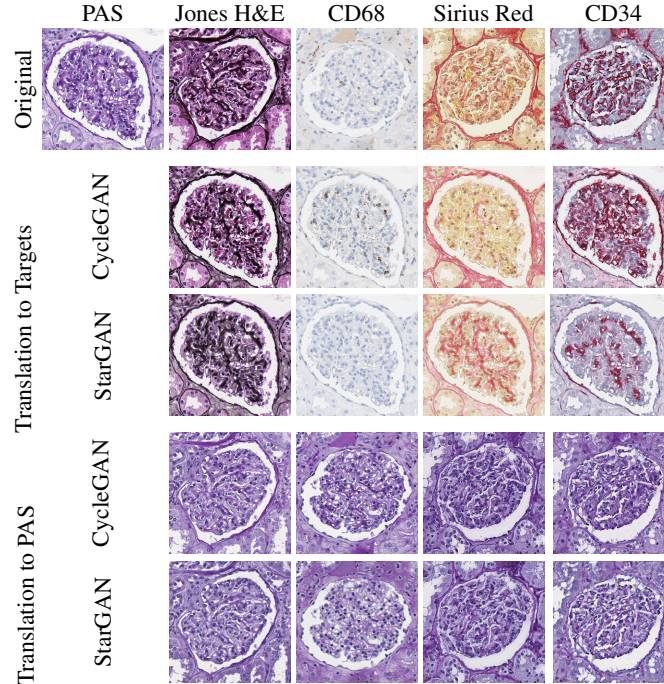
Figure 1: Examples of stain-to-stain translated images using a CycleGAN. The first row contains real patches from each staining. The second row represents the translation of a PAS patch to the target stainings. The third row represents translation of real target patches to the PAS staining.

2 Related work

The occurrence of stain variability can be grouped into intra-stain variability (variations in appearance of a particular stain) or inter-stain variability (the variability in appearance due to different stains). As has been shown in the literature, both types of variation hamper the performance of deep models [2, 6]. When the task at hand is related to the morphological structures in the tissue (such as glomeruli or nuclei detection) rather than stain-related markers (such as macrophages) approaches invariant to inter-stain variation are highly desirable.

Increasing a model's robustness to general stain variations can be considered a domain adaptation problem, which can be approached in either pixel-space or feature space. In feature space, the adaptation is performed by forcing the extraction of domain invariant features during model training. In pixel space, the adaptation can be performed by modifying the properties of a target image in order to match characteristics of the source images, so models trained on the annotated domain can be applied to the modified target images. Conversely the annotated domain can be translated to match the non-annotated (in a way that annotations are preserved), and train a model on the transformed source images such that it is directly applicable in the target domain. Adaptation in pixel space offers better interpretation of the model (from a end-user, i.e. clinician or pathologist, point of view), contrary to mapping them in a feature space, since the result can be viewed as a histopathological image. This is an important consideration in medical imaging. Moreover, once the pixel-space has been aligned, any task can be performed within it.

Inter-stain variability approached from the perspective of feature space alignment is addressed in recent works [22, 23].

In general, feature space alignment has proven effective in transferring knowledge from single (or multiple) source domains to a single target domain, i.e. producing a task and stain specific model. Recently, advances move towards single-source multi-target approaches [24], however they still rely on task specific features. Most of the work on unsupervised domain adaption (i.e. no labels in the target domain) are single-source single-target approaches, which fail to leverage knowledge shared across multiple domains.

From a pixel-space alignment perspective, stain colour augmentation, stain normalisation and stain transfer are commonly used to prevent models from being biased towards the source stain [19]. Stain normalisation refers to the standardisation of a stain's visual appearance and is commonly used as a pre-processing technique to reduce intra-
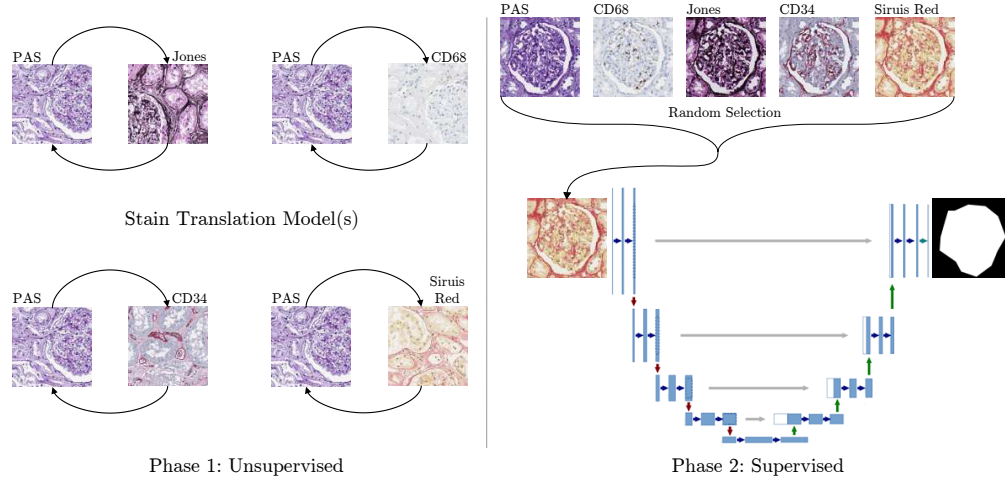
Figure 2: Overall diagram of the proposed approach. Phase 1, translation models are learnt to translate images from the source domain to the target domains; Phase 2, patches of the source domain are randomly translated to the target domains during training (U-Net image taken from [34]).

stain variability [10, 11, 16]. Stain colour augmentation and stain transfer are broader techniques that can address inter-stain variability.

Stain augmentation approaches aim to synthesise new images to enforce colour variation robustness, assuming that objects of interests are invariant to colour intensity and illumination. A typical approach is to perturb image pixels in a random or weighted manner [6, 25, 26]. Alternatively, the stain variation problem can be bypassed by using grayscale images but several studies show this to be inferior [2, 6].

Stain transfer as a virtual staining technique aims to translate an image of tissue stained with a particular stain (or even unstained tissue [27, 28]) into another stain as realistically as possible [17, 18, 29]. Earlier methods [14, 30, 31] rely on decomposing the image into concentration and colour matrices in order to stain new images using equivalent matrices form a reference image. Despite the use of a reference image, these approaches suffer from poor translational realism [6]. Recently, GANs [32] have been used for this purpose due to their ability to synthesize realistic outputs. These are applied in two settings: unpaired translations [4, 29] and paired translation [18]. The CycleGAN architecture [21], as a widely used technique for unpaired image-to-image translation, has recently been applied to stain transfer [3, 4, 18, 29]. And StarGAN [20], a technique of multi domain image-to-image translation, has recently been applied to virtual histological staining of bright-field microscopic images [33].

Gadermayr et al. [4] showed that, when applied to the task of histological stain transfer, CycleGAN is able to produce highly realistic translations, increasing a deep model's robustness in both supervised and unsupervised settings. They proposed three approaches to overcome a lack of target domain annotations [4]: 1) train a segmentation model on source data and apply it to target data translated to the source domain, referred to as MultiDomain Supervised 1 (MDS1); 2) train on the target stain translated to the source, and apply the model directly on the target images, MDS2. Both MDS1 and MDS2 approaches build stain-specific models that can be applied only to a specific staining. The third approach directly trains the segmentation model by optimising unpaired image-to-segmentation translations in an adversarial manner [4]. However, due to the specificity of cycle-consistency, this approach requires that additional information be artificially generated and inserted into the annotation (e.g. cell nuclei). Furthermore, the performance of such an approach was consistently surpassed by the MDS approaches.

The work proposed herein represents a fourth approach that uses the principles of pixel-space alignment to achieve stain transfer and to form a stain-invariant model capable of direct application to several stains. The presented method is general, unsupervised, and applicable to any problem that concerns training a segmentation/classification model for problems that tackle morphologically consistent structures in digital pathology. The only existing similar approach is UDA-SD [6], which we build upon by using GANs to obtain high-quality translations between annotated and unannotated stains (see Figure 1). These are used to augment a training set by random translations of the source domain, resulting in one stain-invariant model capable of segmenting various (unlabeled) domains. As such, this work proposes an single-source, multi-target unsupervised domain adaptation approach.
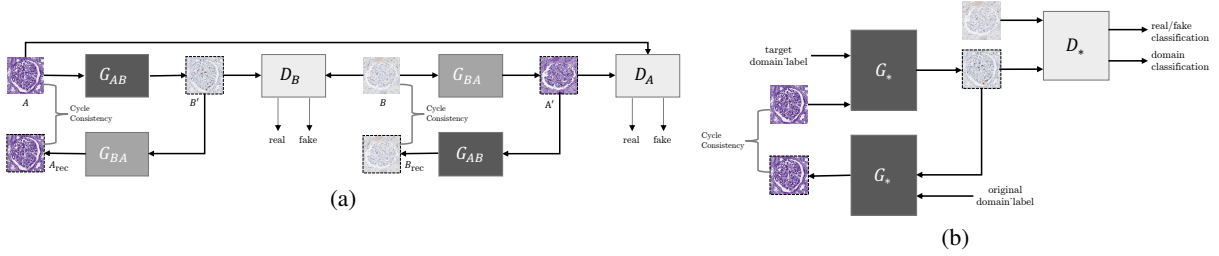
4

Figure 3: CycleGAN (a) and StarGAN (b) diagrams.

3 Method

We propose an approach for training stain invariant Convolutional Neural Networks (CNNs). It is assumed that annotated WSIs are available for a stain A while WSIs of other stains $B_1, B_2, \ldots, B_N$ are unannotated. The aim is to increase the variability of the (annotated) training set through augmentation by randomly translating it to the non-annotated domains (including the original, annotated domain). The overall architecture of the proposed method is presented in Figure 2.

The method consists of the following steps.

a) Stain Translation - in order to obtain realistic translations of the annotated stain A to unannotated stains $B_1, B_2, \ldots, B_N$, a GAN-based unsupervised, unpaired image-to-image translation model is employed. In this study we evaluate both the CycleGAN [21] and StarGAN [20] models (see Figure 3). Once these translation models are obtained, any supervised model (for which labels exist in the source domain and tackles morphologically consistent structures) can be trained. This study focuses on glomeruli segmentation.

b) Segmentation Model - this model is trained on the annotated data after being translated to a random non-annotated stain. Since translation does not change the overall structure of the image (Figure 1), the original domain's label/ground truth is still valid. Thus, during training, various annotated samples of all available stainings are presented to the model, forcing it to learn stain invariant features. Once the segmentation model has been trained, it can be directly applied to the unannotated stains, without any further translations.

3.1 Stain Translation Model

The problem of stain transfer is posed as an unpaired image-to-image translation, which is approached using GAN-based models. The overall structure of these approaches assume the existence of one (or more) generators that translate an image from one domain to another, and one (or more) discriminators that distinguish between translated and real images from corresponding domains. The generator(s) and discriminator(s) play an adversarial game in which the goal of the generator(s) is to prevent the discriminator(s) from distinguishing between the translated samples and those belonging to the real data distribution, while the goal of the discriminator(s) is to distinguish them. Eventually, this game leads to an equilibrium in which the translated and real samples are indistinguishable [20, 21, 32]. Since the task of image-to-image translation is usually applied to natural images, the visual quality and variation of translated samples are very important. Thus, various architectures have been proposed in order to obtain better quality translations and greater style transfer capabilities (e.g. StarGAN v2 [35]).

Nevertheless, since the purpose of stain transfer is to translate an image of stain A to have the appearance of stain B, it is necessary for the translation model to preserve the image's global structure (e.g. glomeruli should remain in the same location, regardless of the target stain). Thus, image-to-image translation architectures which are less constrained and allow more diverse translation (including shape change) such as StarGAN v2 [35] are not suitable. The specific architecture used in the CycleGAN and StarGAN models, combined with the loss functions (particularly cycle-consistency) in practice prevents geometrical changes [21], as is attested by the number of prior works using CycleGANs in digital pathology [3, 4, 10, 16–18, 36].

One such example is given by [36], who uses the CycleGAN for realistic stain-transfer. StarGAN, however, has thus-far not been used for this purpose and we observe the same geometrical consistency when applied to stain transfer.

3.1.1 CycleGAN Based Stain Transfer

Separate CycleGAN models, see Figure 3a, are trained to obtain translators between each target stain (B) and the source stain (A). Each model consists of two generators: $G_{AB} : A \to B$ and $G_{BA} : B \to A$; and two discriminators D_A and D_B. The aim of D_A is to distinguish between real source stain patches and those translated from the target stain to the source; while D_B aims to distinguish between real target stain patches and those translated from the source stain to the target. These are trained in an adversarial manner [32].

3.1.2 StarGAN Based Stain Transfer

Rather than training a translation model for each pair of stains, StarGAN [20] results in one, multi-stain translation model, see Figure 3b. This model contains one generator G_*, conditioned on the domain label (stain), which translates an input image x_i from stain i to image x_j, having the characteristics of stain j, i.e. $G_*(x_i, j) \to x_j$; and one multi-task discriminator D_* that simultaneously distinguishes between real and generated samples (D_{adv}) and classifies each patch's stain label (D_{stain}), i.e. $D_*(x) \to (D_{stain}(x), D_{adv}(x))$. As such, a single discriminator controls the translation into multiple stains.

3.2 Segmentation Model

The U-Net [34] is adopted as it has been proven successful in biomedical imaging [37] and, in particular, glomeruli detection [38]. Glomeruli segmentation is framed as a two class problem: glomeruli (pixels that belong to glomerulus), and tissue (pixels outside a glomerulus).

During training, a training patch with a probability of $\frac{N-1}{N}$, where N is the number of stains, is converted into randomly selected stain using the pre-trained translator(s). Thus, all available stains (including the annotated one) are presented to the network with equal probability $\frac{1}{N}$, forcing the network to learn more stain invariant features

3.3 Training Details

Throughout the study, patches of size 508×508 pixels are used since glomeruli and part of the surrounding fit within this size of patch at the level-of-detail used, see Section 3.4.

3.3.1 Stain translation

Gadermayr et al. [36] showed that different sampling strategies for the annotated and unannotated domains could negatively impact a stain transfer model's performance, and patches are therefore randomly extracted using a uniform sampling strategy (in an unsupervised manner).

CycleGAN: In order to stabilise adversarial training, CycleGAN [21] uses least-squares loss instead of the original negative log-likelihood, such that

$$\mathcal{L}_{\text{CycleGAN}}(G_{AB}, G_{BA}, D_A, D_B) = \mathcal{L}_{\text{adv}}(G_{AB}, D_B)$$
$$+ \mathcal{L}_{\text{adv}}(G_{BA}, D_A) + w_{\text{cyc}}\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA})$$
$$+ w_{\text{id}}\mathcal{L}_{\text{identity}}(G_{AB}, G_{BA}), \tag{1}$$

where

$$\mathcal{L}_{\text{adv}}(G_{AB}, D_B) = \mathbb{E}_{t \sim P_{T_i}(t)}[(D_B(t) - 1)^2]$$
$$+ \mathbb{E}_{s \sim P_S(s)}[(D_B(G_{AB}(s)) - 1)^2],$$

$$\mathcal{L}_{\text{adv}}(G_{BA}, D_A) = \mathbb{E}_{s \sim P_S(s)}[(D_A(s) - 1)^2]$$
$$+ \mathbb{E}_{t \sim P_{T_i}(t)}[(D_A(G_{BA}(t)) - 1)^2],$$

$$\mathcal{L}_{\text{cyc}}(G_{AB}, G_{BA}) = \mathbb{E}_{s \sim P_S(s)}[\|G_{BA}(G_{AB}(s)) - s\|_1]$$
$$+ \mathbb{E}_{t \sim P_{T_i}(t)}[\|G_{AB}(G_{BA}(t)) - t\|_1],$$

$$\mathcal{L}_{\text{identity}}(G_{AB}, G_{BA}) = \mathbb{E}_{s \sim P_S(s)}[\|G_{BA}(s) - s\|_1]$$
$$+ \mathbb{E}_{t \sim P_{T_i}(t)}[\|G_{AB}(t) - t\|_1].$$

The loss weights and architecture are taken from the original paper ($w_{\mathrm{cyc}} = 10$, $w_{\mathrm{id}} = 5$) [21] since they produced realistic output, see Fig. 1. A translation network with nine ResNet blocks was used, as suggested by the authors for high dimensional data (above 256×256 pixels) [21].

The models are trained for 50 epochs, with a learning rate of 0.0002, using the Adam optimiser, and a batch size of 1. From the 25$^{\mathrm{th}}$ epoch, the learning rate linearly decayed to 0, and the cycle-consistency and identity weights halved. In all experiments, the translation model from last (50$^{\mathrm{th}}$) epoch is used. Moreover, to reduce model oscillation, Shrivastava et al.'s strategy [39] of updating the discriminator using the 50 previously generated samples is adopted.

StarGAN:

In order to promote adversarial training stability, StarGAN [20] uses Wasserstein loss with gradient penalty instead of the original negative log-likelihood, such that

$$\mathcal{L}_{\mathrm{StarGAN}}(G_*, D_*) = \mathcal{L}_{\mathrm{adv}}(G_*, D_*) + w_{\mathrm{cyc}}\mathcal{L}_{cyc}(G_*)$$
$$+ w_{\mathrm{cls}}\mathcal{L}_{cls}(G_*, D_*),$$

where

$$\mathcal{L}_{\mathrm{adv}}(G_*, D_*) = \mathbb{E}_{x \sim P_j(x)}[D_{\mathrm{adv}}(x)]$$
$$- \mathbb{E}_{(x \sim P_i(x), j)}[D_{\mathrm{adv}}(G_*(x, j)))]$$
$$- \lambda_{gp}\mathbb{E}_{\hat{x}}[(\|\nabla_{\hat{x}} D_{adv}(\hat{x})\|_2 - 1)^2],$$
$$\mathcal{L}_{\mathrm{cyc}}(G_*) = \mathbb{E}_{x \sim P_i(x), i, j}[\|G_*(G_*(x, j), i) - x\|_1],$$
$$\mathcal{L}_{\mathrm{cls}}(G_*, D_*) = \mathbb{E}_{(x \sim P_i(x), j)}[-\log D_{\mathrm{stain}}(j|G_*(x_i, j))]$$
$$+ \mathbb{E}_{(x \sim P_i(x))}[-\log D_{\mathrm{stain}}(i|x)],$$

where $\hat{x}$ is sampled uniformly between the real and generated images [20].

The generator's and discriminator's architecture and training settings were the same as the CycleGAN model described previously (without instance normalisation in the discriminator). The loss weights are taken from the original paper ($w_{\mathrm{cyc}} = 10$, $w_{\mathrm{cls}} = 1$, $\lambda_{gp} = 10$) [20] since they produced realistic output, as can be seen in Figure 1.

3.3.2 U-Net

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.

After translation (or not), 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 [25], α sampled from $[-0.25, 0.25]$ and β from $[-0.05, 0.05]$.

3.4 Data

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 ("scerosis") 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, Sirius Red or H&E, in addition to three immunohistochemistry markers (CD34, CD68 and CD3), using an automated staining instrument (Ventana Benchmark Ultra). H&E and CD3 were reserved as unseen stain, therefore the total number of augmentation stains is $N = 5$. Whole slide images were acquired using an Aperio AT2 scanner at $40\times$ 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 [40]. The dataset was divided into 4 training, 2 validation, and 4 test patients. The number of glomeruli in each staining dataset was: PAS - 662 (train.), 588 (valid.), 1092 (test); Jones H&E - 1043 (test);H&E - 1151; Sirius Red - 1049 (test); CD34 - 1019 (test); CD68 - 1046 (test);CD3 - 1083. The training set comprised all glomeruli from the source staining training patients (662) plus 4634 tissue (i.e. non-glomeruli) patches (to account for the variance observed in non-glomeruli tissue).

4 Results

In the following, the proposed approach is referred to as Unsupervised Domain Augmentation using GANs (UDA-GAN), with two variants: with CycleGANs (UDA-CGAN) and StarGAN (UDA-∗GAN). These are presented and compared to the MDS1 [4], MDS2 [4], Unsupervised Domain Augmentation via Stain Decomposition (UDA-SD) [6] and vanilla PAS (vPAS), which is the direct application of a model trained on PAS to the target domains without any adaptation (i.e. the control).

As mentioned, the CycleGAN models obtained after 50 epochs are used, however, Brieu et al. [3] show that realistic translations are obtained even in early epochs, which is confirmed in Figure 4. They therefore train a segmentation model using translators taken from multiple epochs to increased augmentation variability. Therefore we include the Multi UDA-CGAN model in which translation models from each 5th epoch are used, resulting in 40 translation models.
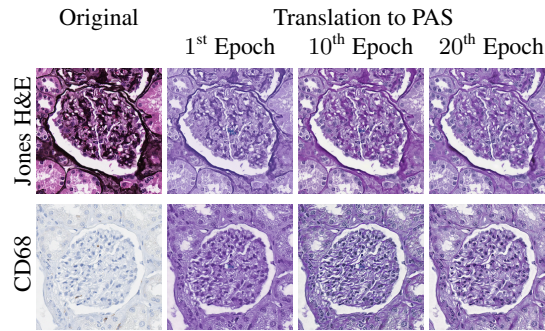

Figure 4: Jones H&E and CD68 glomeruli patches translated to PAS using CycleGAN models from different training epochs.

For MDS1 and MDS2, the translated models are trained according to Section 3.1. Variants of MDS1 and MDS2 using the StarGAN translators were evaluated, and are referred to as MDS∗1 and MDS∗2. The F_1-score, along with precision and recall, are used to measure the performance. The presented results are the averages of five independent training repetitions, with corresponding standard deviations.

The experimental results are presented in Table 1. Baseline performances (U-Net models, as described above, trained and tested on the same stain using each stain's ground truth) were determined for each staining and presented in Table 2.

Despite the fact that the translations obtained using both CycleGAN and StarGAN look realistic (see Figure 1), it can be observed that the direction of translation (MDS1 vs MDS2) and translation model (StarGAN vs CycleGAN) influence the results. This is best illustrated with MDS1, in which a model trained on the original PAS data is applied to target data translated to PAS. It can be observed that in each target stain, the difference between CycleGAN and StarGAN translations is significant, although there appears to be no significant difference in the quality of the translations. The proposed UDA-GAN approaches show more stable performance while the best results are obtained using UDA-CGAN.

Furthermore, UDA-CGAN reaches almost baseline performance in three out of five test stainings. Despite the fact that the model has seen data from PAS stain only 20% of the time during training, the model has baseline performance on this (source) domain. The model also approaches baseline performance in target stains Jones and Sirius Red. For stains CD68 and CD34, the model reaches an F_1-score of 0.705 and 0.799, meaning that it gives an improvement of 11.9% and 6% respectively over the next best CycleGAN method (MDS2). The average performance over the five different stainings show that UDA-CGAN reaches an average F_1-score of 0.827 (0.808 without including the PAS staining, in order to be fairly compared to the MDS approaches), while MDS2, as the next best method, reaches an F_1-score of 0.748. The biggest relative difference is observed in staining CD68 where the overall improvement is 55.8% compared to the original approach [6] and 11.9% compared to MDS2. Other than the baseline, UDA-CGAN

Training Strategy	Score	Test Staining					
		PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
vPAS	F_1	0.907 (0.009)	0.085 (0.034)	0.001 (0.002)	0.016 (0.018)	0.071 (0.063)	0.043 (0.041)
	Precision	0.885 (0.023)	0.055 (0.021)	0.097 (0.129)	0.034 (0.034)	0.257 (0.243)	0.111 (0.101)
	Recall	0.932 (0.014)	0.418 (0.316)	0.001 (0.001)	0.073 (0.101)	0.058 (0.039)	0.137 (0.190)
UDA-SD [6]	F_1	0.891 (0.007)	0.791 (0.079)	0.147 (0.048)	0.828 (0.046)	0.739 (0.026)	0.679 (0.303)
	Precision	0.840 (0.018)	0.699 (0.116)	0.365 (0.238)	0.778 (0.088)	0.695 (0.054)	0.675 (0.183)
	Recall	0.950 (0.007)	0.926 (0.015)	0.099 (0.032)	0.892 (0.024)	0.795 (0.059)	0.732 (0.359)
MDS1 [4]	F_1	-	**0.872** (0.016)	0.395 (0.057)	0.828 (0.040)	0.673 (0.033)	0.692 (0.215)
	Precision	-	0.843 (0.036)	0.447 (0.092)	0.787 (0.071)	0.857 (0.033)	0.734 (0.193)
	Recall	-	0.904 (0.018)	0.364 (0.071)	0.877 (0.020)	0.556 (0.047)	0.675 (0.261)
MDS2 [4]	F_1	-	0.869 (0.020)	0.586 (0.059)	0.797 (0.040)	0.739 (0.044)	0.748 (0.121)
	Precision	-	0.833 (0.049)	0.519 (0.108)	0.699 (0.061)	0.723 (0.051)	0.695 (0.132)
	Recall	-	0.909 (0.013)	0.697 (0.059)	0.929 (0.004)	0.765 (0.106)	0.825 (0.112)
UDA-CGAN	F_1	**0.901** (0.011)	0.856 (0.036)	**0.705** (0.031)	**0.873** (0.025)	0.799 (0.034)	**0.827** (0.078)
	Precision	0.869 (0.034)	0.800 (0.069)	0.690 (0.059)	0.830 (0.051)	0.754 (0.076)	0.789 (0.069)
	Recall	0.936 (0.014)	0.924 (0.012)	0.723 (0.034)	0.922 (0.009)	0.856 (0.036)	0.872 (0.089)
Multi UDA-CGAN	F_1	0.897 (0.010)	0.863 (0.030)	0.684 (0.046)	0.861 (0.021)	**0.808** (0.023)	0.822 (0.084)
	Precision	0.860 (0.021)	0.812 (0.057)	0.648 (0.098)	0.813 (0.043)	0.764 (0.061)	0.779 (0.081)
	Recall	0.937 (0.007)	0.922 (0.009)	0.736 (0.038)	0.917 (0.010)	0.862 (0.032)	0.875 (0.083)
MDS*1	F_1	-	0.756 (0.086)	0.092 (0.055)	0.599 (0.108)	0.751 (0.033)	0.550 (0.314)
	Precision	-	0.675 (0.136)	0.242 (0.116)	0.496 (0.123)	0.742 (0.092)	0.539 (0.223)
	Recall	-	0.881 (0.029)	0.061 (0.044)	0.780 (0.099)	0.774 (0.070)	0.624 (0.379)
MDS*2	F_1	-	0.816 (0.060)	0.525 (0.048)	0.837 (0.032)	0.766 (0.030)	0.736 (0.144)
	Precision	-	0.740 (0.096)	0.874 (0.037)	0.785 (0.059)	0.752 (0.030)	0.787 (0.061)
	Recall	-	0.918 (0.008)	0.376 (0.046)	0.901 (0.014)	0.785 (0.073)	0.745 (0.253)
UDA-*GAN	F_1	0.890 (0.022)	0.807 (0.031)	0.549 (0.081)	0.792 (0.052)	0.758 (0.076)	0.759 (0.127)
	Precision	0.853 (0.043)	0.717 (0.050)	0.794 (0.044)	0.703 (0.085)	0.738 (0.082)	0.761 (0.062)
	Recall	0.933 (0.008)	0.926 (0.010)	0.426 (0.090)	0.913 (0.013)	0.796 (0.135)	0.799 (0.216)

Table 1: Quantitative results for each strategy trained on 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 (PAS is not included in the vPAS average since it is the training staining.)

is the only to achieve acceptable results in this staining. Multi UDA-CGAN does not improve upon this, possibly because it introduces too much variability.

	PAS	Jones H&E	CD68	Sirius Red	CD34	Overall
F_1	0.907 (0.009)	0.864 (0.011)	0.853 (0.018)	0.867 (0.016)	0.888 (0.015)	0.876 (0.022)
Precision	0.885 (0.023)	0.824 (0.020)	0.846 (0.027)	0.801 (0.042)	0.862 (0.015)	0.844 (0.032)
Recall	0.932 (0.014)	0.911 (0.005)	0.856 (0.022)	0.957 (0.018)	0.929 (0.011)	0.917 (0.038)

Table 2: Quantitative baseline results (standard deviations are in parentheses).

4.1 Unseen Stains

In order to further evaluate the stain invariance of the UDA-GAN approaches, they are tested on two unseen stains (unseen to both the CycleGAN/StarGAN and UDA-GAN models), see Figure 5: histological stain H&E (a general overview staining not specific for a protein) and immunohistochemical stain CD3 (T cell marker). Even though they highlight similar structures to the 'virtually' seen stains, they are visually very different in appearance. For each stain, images are taken from 3 patients containing 1151 (H&E) and 1083 (CD3) glomeruli. Table 3 presents the results, averaged over the previously trained UDA-GAN models.
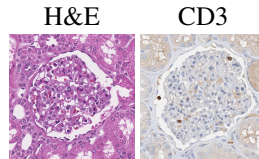
H&E   CD3

Figure 5: Examples of glomeruli from the unseen stains.

		H&E	CD3	Average
vPAS	F_1	0.126 (0.066)	0.000 (0.000)	0.063
	Precision	0.070 (0.041)	0.018 (0.018)	0.044
	Recall	0.854 (0.109)	0.000 (0.000)	0.427
UDA-CGAN	F_1	0.731 (0.100)	0.658 (0.075)	0.694
	Precision	0.781 (0.122)	0.569 (0.124)	0.675
	Recall	0.697 (0.123)	0.802 (0.042)	0.750
UDA-∗GAN	F_1	0.752 (0.087)	0.650 (0.030)	0.701
	Precision	0.824 (0.057)	0.853 (0.065)	0.839
	Recall	0.706 (0.147)	0.531 (0.058)	0.618

Table 3: Quantitative results of UDA-GAN models on unseen stains (standard deviations are in parentheses).

Although the results are lower than those obtained using stains virtually seen during training, i.e. stain translation targets during augmentation, they confirm the network's capacity for stain invariant segmentation.

On average, both UDA-CGAN and UDA-∗GAN perform equally well on unseen stains. And when taken in context of Table 1, although they generally achieve lower results, UDA-∗GAN remains within the range previously seen and UDA-CGAN exhibits more variance.

A high F_1 score is achieved when faced with a completely new stain colour profile (H&E). When faced with a similar stain profile to one virtually seen (CD3, similar to CD68), the corresponding UDA-CGAN F_1 score is similar and UDA-∗GAN improves, likely because CD3 has more contrast (due to an unspecific reaction of the primary antibody)

when compared to CD68, with which it struggles. It is worth noting that the models with the best PAS performance (which can be determined since annotations exist) are also those that perform best on the unseen stains.

5 Further Analysis and Discussion

In this section the results are analysed in more detail, including calculating additional statistics and, and related to those found in the literature. Further analysis of the trained models is presented by visualising their feature distributions and their attention. Finally, the topic of CycleGAN training is discussed.

5.1 Translation Direction

Gadermayr et al. suggest that translation "should always be performed from the difficult-to-segment to the easy-to-segment domain" and that MDS1 is the preferred method [4]. Table 2 shows that the difficult-to-segment stain is CD68 and the easy is PAS as it results in a more accurate baseline segmentation. It is also observed in Table 1 that MDS1 with PAS as the source domain is in fact surpassed by MDS2 in all but one target stainings (in which it is equal). These results suggest that the characteristics of "difficult-to-segment" stainings may vary, and the method/translation direction should be adjusted to the specific requirements of a given biological question, i.e. the panel of necessary staining methods.

5.2 Representation Comparison

In the case of staining CD68, neither MDS1, MDS∗1, MDS2, nor, MDS∗2 perform well. Poor MDS1 (MDS∗1) performance could indicate that the CycleGAN (StarGAN) translation between the CD68 and PAS domains do not capture the features the PAS model uses for segmentation. On the other hand, poor MDS2 and MDS∗2 performance could indicate that the translation between PAS and CD68 contains features that are not present in real CD68. From a biological viewpoint, this most likely represents the fact that immunohistochemistry for CD68 highlights just one specific, migratory cell population (macrophages) that is not part of the pre-existing tissue architecture, with a brown chromogen, while the anatomical structures are only faintly stained (blue "counterstain" using hemalaun). Strikingly, immunohistochemistry for CD34, a marker for vascular endothelial cells, labeled with a red chromogen, performs much better. This can probably be explained by a specific immunohistochemical labeling of anatomical structures (blood vessels) in addition to the blueish counterstain, containing more features that are also covered in the other staining methods (PAS, Jones H&E). This is particularly evident with StarGAN, which exploits common inter-stain characteristics.

Between MDS1 and MDS2 (both CycleGAN and StarGAN translations) the largest difference is seen in stain CD68, which marks a protein exclusively produced by macrophages. PAS, as a chemical reaction staining glycolysated proteins in general, highlights a part of macrophages (co-located, but not overlapping, with CD68). Thus the translation from PAS to CD68 (MDS2, MDS∗2) is easier than the reverse (MDS1, MDS∗1), since PAS contains some (but not all) of the information exposed by CD68.

The fact that UDA-CGAN outperforms both MDS1 and MDS2 using the same translation functions indicates that it is capable of extracting more general (stain invariant) features, i.e. it avoids learning stain specific and 'false' features introduced by the CycleGAN. This is also the case for UDA-∗GAN but to a lesser extent. UDA-∗GAN uses the same generator for all translations so it is likely to extract similar features between the source stain and all target stains. This reduces the impact of the multi-stain augmentation, which becomes evident when comparing UDA-∗GAN to MDS∗2.

5.2.1 Feature Distributions

In order to visualise the distributions of each model's extracted features, Figure 6 presents UMAP (Uniform Manifold Approximation and Projection) embeddings [41] of the penultimate convolutional layer's activations in the best performing model (over all stainings) for two hundred random glomeruli and two hundred random tissue patches of each staining using MDS1, MDS∗1, UDA-CGAN, and UDA-∗GAN (MDS2 and MDS∗2 models are stain specific, therefore cannot be applied to all stainings). For MDS1 and MDS∗1, the translations to the PAS stain is achieved using CycleGAN and StarGAN respectively, and UDA-CGAN and UDA-∗GAN are applied to original data without any modification.

In order to quantitatively measure these distributions, silhouette scores [42] have been calculated between:

- each stain's glomeruli class and the union of the glomeruli samples from all other stains;
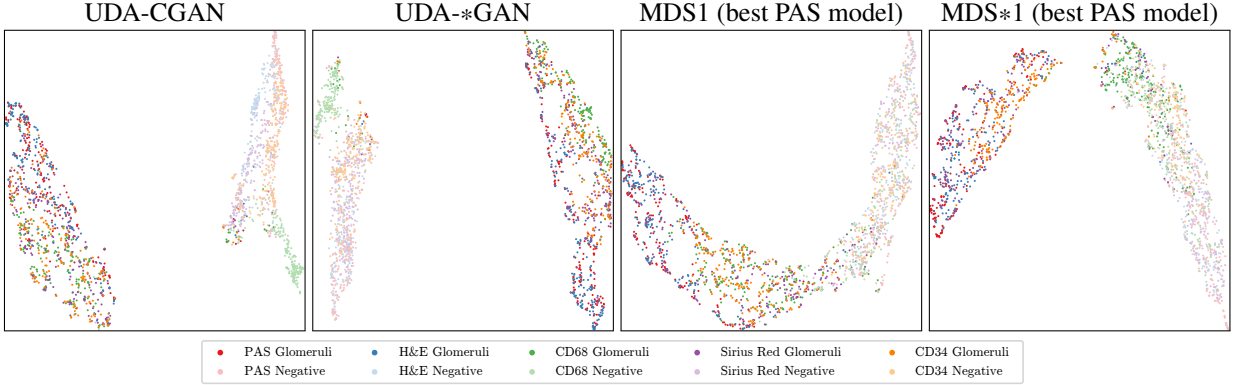- PAS glomeruli and each target stain's glomeruli;

Figure 6: Two-dimensional UMAP embeddings of the representation learnt, sampled from the penultimate convolutional layer using 200 patches per stain per class from the overall best performing PAS and UDA-CGAN models. Each point represents a patch from the respective class and staining (glomeruli patches are centred on a glomeruli).

	Training Strategy	PAS	Jones H&E	CD68	Sirius Red	CD34
Glomeruli-Combined	UDA-CGAN	0.069 (0.004)	0.106 (0.008)	**0.000** (0.007)	-0.044 (0.003)	**-0.019** (0.003)
	UDA-∗GAN	**0.051** (0.008)	**0.094** (0.005)	0.069 (0.010)	-0.041 (0.003)	-0.024 (0.003)
	MDS1	0.198 (0.011)	0.176 (0.014)	-0.057 (0.002)	-0.038 (0.004)	-0.070 (0.005)
	MDS∗1	0.219 (0.006)	0.048 (0.011)	0.166 (0.004)	**-0.027** (0.003)	-0.063 (0.004)
Glomeruli-PAS Glomeruli	UDA-CGAN	-	0.004 (0.002)	**0.123** (0.008)	0.078 (0.003)	**0.071** (0.006)
	UDA-∗GAN	-	0.004 (0.001)	0.179 (0.003)	**0.070** (0.002)	0.090 (0.003)
	MDS1	-	**0.002** (0.001)	0.253 (0.007)	0.175 (0.009)	0.255 (0.003)
	MDS∗1	-	0.037 (0.005)	0.477 (0.007)	0.098 (0.003)	0.186 (0.010)
Glomeruli-Tissue	UDA-CGAN	0.594 (0.009)	0.567 (0.017)	**0.533** (0.009)	0.554 (0.012)	**0.551** (0.010)
	UDA-∗GAN	**0.625** (0.016)	**0.584** (0.015)	0.475 (0.014)	**0.581** (0.012)	0.530 (0.005)
	MDS1	0.489 (0.007)	0.481 (0.017)	0.300 (0.009)	0.300 (0.016)	0.424 (0.006)
	MDS∗1	0.489 (0.007)	0.456 (0.019)	0.051 (0.004)	0.354 (0.023)	0.449 (0.005)

Table 4: Silhouette scores measuring (averaged over 3 different random samplings): the separation between the each stain's glomeruli class and the glomeruli class formed from all other stains (Combined), the separation between the each stain's glomeruli class and the PAS glomeruli class (PAS), and the separation between each stain's glomeruli class and its negative class. Calculated using the features of 200 random patches per stain per class, extracted from the penultimate convolutional layer: 0 means total overlap, 1 means total separation, -1 means that samples are more similar to the other class than their own, therefore values closer to 0 are better (in bold) for Glomeruli-Combined and Glomeruli-PAS Glomeruli, and 1 is better (in bold) for Glomeruli-Tissue.

- each stain's glomeruli and negative samples.

These are presented in Table 4. The first should favour the UDA-GAN approaches since their goal is to learn a stain invariant representation, the second should favour MDS1 and MDS∗1, since their objective is to translate the target stainings to the PAS distribution, while the third should be the goal of all approaches.

In the first case, the UDA-GAN approaches exhibit larger (or equal) overlap between the glomeruli in all stainings, indicating greater clustering, which is reflected in the fact that these models are able to segment all stainings. Higher scores for the MDS1 approaches indicate less concentrated clustering (e.g. Jones H&E , which appears to be concentrated in one part of the glomeruli space, separate from the other stains).

Since the MDS1 approaches are trained on PAS, they rely on accurate translation models, which must result in a direct overlap with the PAS distribution. Whereas the UDA-GAN approaches can tolerate more translation variance (that does not result in glomeruli - tissue overlap) since they are trained on the translated data. Interestingly the scores for MDS1 and MDS∗1 for PAS are relatively high, indicating that this approach fails to completely overlap the target stains with PAS.

In the second case, it can be observed that UDA-GAN approaches have better (or equal) overlap with the PAS glomeruli in all stainings. The score for MDS∗1 in CD68 is much higher then for any other stain, and Figure 6 shows that the CD68 glomeruli class has been merged with the negative class, explaining the very low recall in Table 1.

In the third case, it can be observed that the UDA-GAN approaches better separate the glomeruli and tissue classes. This is especially illustrated in the case of CD68 in which UDA-∗GAN learns to separate the glomeruli and tissue classes, whereas MDS∗1 fails. As mentioned, MDS∗1 relies on accurate translations, whereas the UDA-GAN approaches are able to correct for weak translations during training.

5.2.2 Attention

Another approach to illustrate the representations learned is to use attention visualisations. Grad-CAM [43] is used to visualise the attention of the penultimate convolutional layer, see Figure 7. For fair comparison, the best performing MDS1, MDS∗1, MDS2, and MDS∗2 model for each staining, and the baseline approaches are used. For UDA-CGAN and UDA-∗GAN, the overall best-performing model is taken.
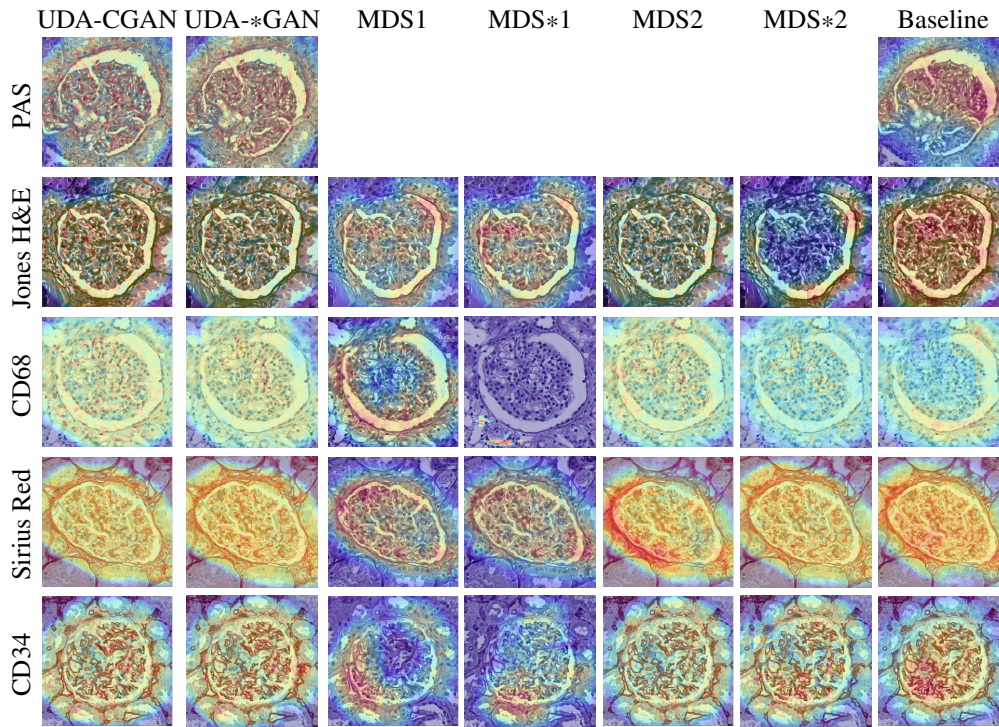

Figure 7: GradCAM visualisation of (column-wise) 1) overall best performing UDA-CGAN model; 2) best MDS1 model in each particular stain; 2) best MDS2 model in each particular stain; 3) best baseline model in each particular stain. N.B. the first column represents the attention of the same model over different stainings.

The attention of the best-performing MDS1 (and MDS∗1, except CD68 for which MDS∗1 does not work) models in all stainings is focused on border-like features, which the model uses in the original PAS domain (on which it was trained). Poor MDS1 and MDS∗1 performance can be explained by an absence of the specific features on which the trained model is focused on, and which are not necessarily present in the target domains nor relevant for the detection of the structures in general. When comparing attention in all the baseline models, it can be observed that in each staining, the models have a tendency to focus on stain specific features. The attention of the MDS2 models is more general and close to the stain-invariant model's attention. According to the results presented in Table 1, the stain-invariant approach (UDA-CGAN) gives an improvement in precision, while the recalls of both models are similar (in all stainings except CD34). Meaning that the stain-invariant model reduces false positives and, to some extent, false negatives (while both models detect true positives similarly). Thus, attention in the true positive class is expected to be similar. Both the MDS∗2 and UDA-∗GAN use a common translation model (StarGAN), which, as previously mentioned, is biased towards common features between the stains. Thus both models are likely to focus on these common features, explaining their similar attentions (this is also confirmed in Table 1). In addition to common

features, the MDS2 models can also use stain-specific features because they are only exposed to one type of image (as is the case with the presented Jones H&E example).

The main advantage of the UDA-GAN approaches is their ability to properly generalise across different stainings, as the presented attentions represent the features learnt by one model, while in all other cases a domain (stain) dependent model is obtained.

5.2.3 Multi vs Single Stain Translation

The fact that StarGAN has a single translator may force it to preserve common features between stains, features that UDA-∗GAN is likely to focus on. It is also likely that these features are general and present in unseen stains, therefore UDA-∗GAN's performance is similar within the virtually seen and unseen stains. Nevertheless, the single translator may force StarGAN to ignore hard-to-translate stains, e.g. CD68, and to minimise its loss upon the other stains. This would cause the general features to be extracted from the remaining stains. Low recall for CD68 and CD3 indicates that the model struggles to identify glomeruli, i.e. the general features do not exist in these stains. This is confirmed in Figure 6, in which the CD68 glomeruli class overlaps the negative class. UDA-CGAN, on the other hand, is trained using the translators and has more sources of variation since each augmentation translator is independent. It can therefore leverage the additional information present in the augmented translations to achieve higher accuracy during application to them, while still learning a stain invariant representation.

5.3 Translation Quality

5.3.1 Quality Measure

One way to validate the quality of such translations is to consult a pathologist, however this is expensive and time consuming. Preliminary validation by a pathologist confirms that the translations are indistinguishable from real images.

One of the most widely used measures for validating the quality of generated natural images is the Frećhet Inception Distance (FID) [44].

FID measures the distance between features of the generated and real images extracted from the Inception network. Since this network is trained on natural images, it is not directly applicable to the medical domain. Instead, a new network could be trained on histpathological data and used as a feature extractor. However, due to (relatively) limited available data (in comparison to the ImageNet database which contains ∼14 million training images), it is likely that the trained network will capture dataset specific features (as confirmed by the experiments presented in this discussion), and thus will not reflect the accuracy of translated patches.

Moreover, FID is a highly biased estimator [45], and has been shown to not be correlated with classification accuracy [46], as is the concern of this work.

5.3.2 Noise

Since both directions of translation (PAS to target and vice versa) obtain visually good results, it is unclear why the performance of MDS1 and MDS2 vary so much. Recent literature [47] indicates that cycle-consistency $\mathcal{L}_{\mathrm{cyc}}$, Eq. (1), forces image-to-image translation models to hide information necessary for proper reconstruction of A_{rec} and B_{rec} in the form of imperceptible low amplitude, high frequency noise in A' and B' (see Figure 3). It is reasonable to assume that such noise will have an affect on the model's performance.

To illustrate this phenomena, we use the composition of translations PAS $\rightarrow$ Target$+\mathcal{N}(0,\sigma) \rightarrow$ PAS, where $\mathcal{N}(0,\sigma)$ is a zero mean Gaussian distribution with standard deviation σ. Figure 8 presents reconstructions of the same PAS image after translation to each target stain, with different standard deviation of additive noise. This shows that not all target stain translators encode information in the same way. For example, adding noise with a standard deviation of 0.5 to the CD68 intermediate stain results in a higher reconstruction error than adding the same noise in the Jones H&E intermediate stain. We suspect that the level of noise in each target staining correlates with the difficulty of translation i.e. harder translations require more noise. StarGAN appears to be more sensitive to the additive noise than CycleGAN. Potentially because it performs the much harder task of multi-stain translation. In order to properly reconstruct its input, more information may need to be hidden and thus it may be more sensitive to additive noise.

In order to quantify this sensitivity to noise, Bashkirova et al. [47] propose the following measure: the mean pixel difference $\bar{\delta}$ between the translations with and without noise is measured and is repeated for a range of standard deviations σ. The area under the resulting σ - $\bar{\delta}$ curve summarises the translator's sensitivity (lower is better, i.e. less
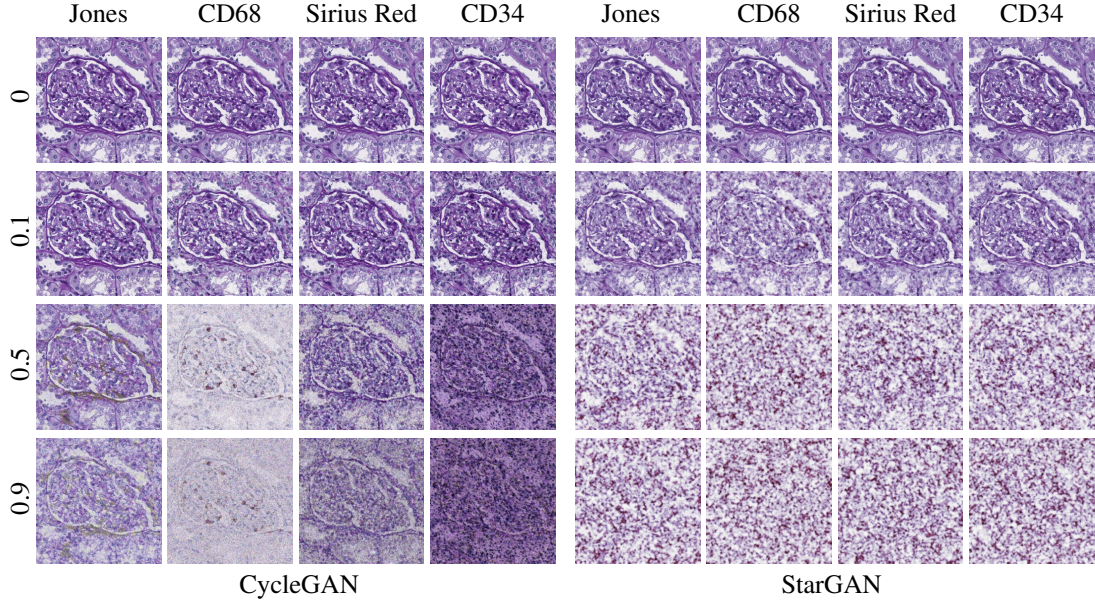
Figure 8: Effects of additive zero-mean Gaussian noise added to intermediate representation of PAS patch to reconstruction.

sensitive). This is calculated here using the target-to-source translations of 100 random PAS patches from each class and $\sigma \in [0, 0.025, 0.05, 0.075, 0.1]$, see Table 5.

	Jones H&E	CD68	Sirius Red	CD34
CycleGAN	5.359	14.778	7.423	11.841
StarGAN	11.672	43.602	17.28	16.288

Table 5: Sensitivity to noise calculated for each target stain based on 100 random patches from each class.

These were derived from the same PAS source patches but the generators were trained on different target data. Some caution should therefore be taken when comparing between stains, however, these are as close as possible to being comparable. The orders of sensitivity correlate with both the MDS1 and MDS∗1 F_1 scores presented in Table 1. It is impossible to do the same analysis for MDS2, i.e. Target → PAS → Target, as the target datasets are different.

The relatively high sensitivity to noise (therefore potentially high levels of noise present in the translation) offers an explanation for UDA-CGAN's relatively low precision in CD68 and CD34, and UDA-∗GAN's relatively low precision overall.

For further illustration, Figure 9 presents a Sirius Red and CD34 glomerulus patch translated to PAS using translators taken from different CycleGAN training epochs, along with its segmentation using the overall best performing PAS model (which, being trained on PAS, is sensitive to its distribution). Unintuitively, good visual appearance does not correlate with reliable segmentation (this holds for all stains), which suggests that there is a limitation on the direct application of cycle-constrained image-to-image translation methods in clinical practice. Brieu et al. [3] chose the set of training epochs by visual inspection, however, these findings indicate that this is not a good strategy and may lead to the model learning irrelevant features. Furthermore, weakly-supervised approaches based on the performance of an already trained model may not be a good alternative, as the learned features may differ between models and are not necessary task-related.

These analyses, and recent studies on the adversarial nature of CycleGANs [48, 49], lead to the hypothesis that the translations suffer from invisible artifacts produced by the CycleGAN. The extent and type of these artifacts could be related to stain differences. Stains with a greater difference require more complicated translation, forcing the translators to hallucinate specific features, as confirmed by Mercan et al. [18]. From this perspective, it can be understood why UDA-CGAN outperforms MDS1 and MDS2 as the model is forced to be robust to artifacts produced by different translations. In this sense, the lack of improvement offered by Multi UDA-CGAN model (see Table 1) over UDA-CGAN can be understood. In MDS1, these artifacts hamper the performances as translated images could act as
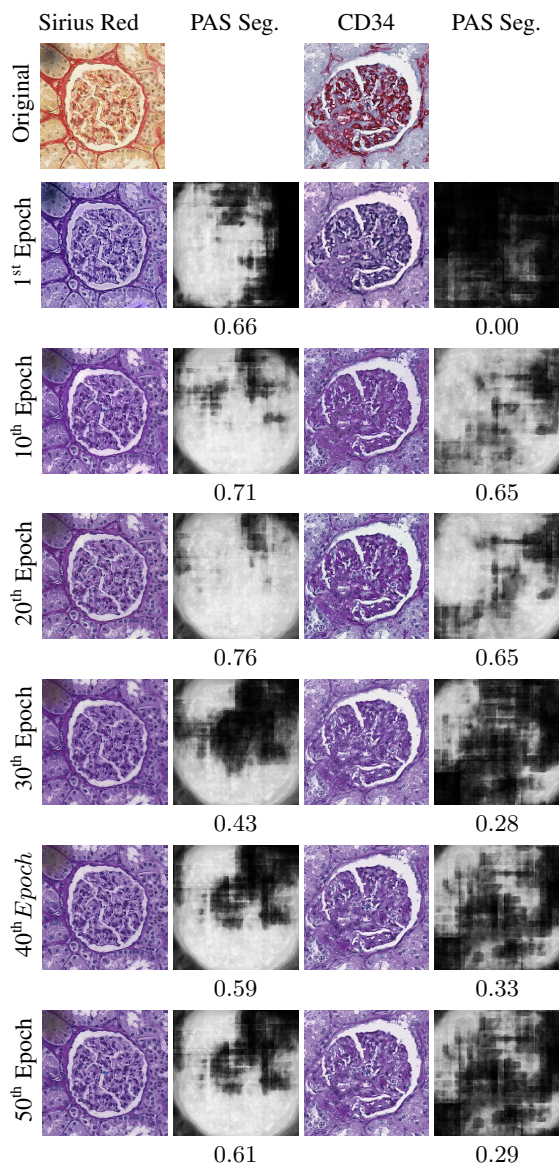
Figure 9: Glomeruli patches translated to PAS using CycleGAN models from different training epochs with PAS glomeruli segmentations (i.e. MDS1 approach) and their F_1 scores. N.B. images look realistic but segmentations vary widely.

adversarial examples, while in MDS2, these artifacts could be considered by the model as features. We hypothesise that a mechanism for assessing the quality of the translation, such as FID [44] in natural images, would offer further improvement.

This only goes to emphasise the care and consideration needed for the direct application of unpaired image-to-image translations in the medical domain, and when used in clinical practice the implications should be held up to even greater scrutiny. Since the proposed approach merely uses the translations to remove a model's bias, see Figure 6, we largely avoid such concerns.

6 Conclusions

To summarise, this article presents a state-of-the-art model, UDA-GAN, that outperforms all other existing pixel-space alignment approaches in five different stainings. The model is domain invariant (including to unseen stains), can be

easily extended to new stainings, and the training procedure is general, in that it can be used in different segmentation and classification tasks.

The usefulness of "translating" stainings has therefore been demonstrated in this context but it should be emphasised that diagnostically relevant information can be lost on the way. We should also emphasise that the primary intention of this translation is to save on resources and time for annotation, not to replace the process of physical staining modalities.

UDA-GAN's stain invariance has been shown through quantitative results and by analysing its feature distribution and attention, which demonstrate that patches from all stainings are more aligned than with competing approaches. The approach uses pixel space alignment, which aids in visual interpretation and verification. The results have been discussed and related to those found in the literature. Namely that choosing translation model epochs by visual inspection is not the best approach (although no formal method for selecting the epochs to be used nor evaluating quality exists) and that the direction of translation of the data cannot be simply prescribed a priori. Finally, UDA-GANs limitations were presented; it is still unclear exactly why the model's performance degrades with certain stainings despite visually accurate translations. We hypothesise that some stainings are far from each other in terms of biological structures highlighted, in particular immunohistochemical staining methods highlighting migratory immune cells can be far from conventional histology staining methods making the translation task harder. Therefore, the translation model is forced to introduce hidden information.

Although UDA-GAN approaches diminish these effects to some extent, all these findings indicate that the application of image-to-image translation to medical imaging should be done with careful consideration.

Acknowledgements

References

[1] Patrick Leo, George Lee, Natalie N. C. Shih, Robin M. Elliott, Michael D. Feldman, and Anant Madabhushi. Evaluating stability of histomorphometric features across scanner and staining variations: prostate cancer diagnosis from whole slide images. *J Med Imaging*, 3(4), 2016.

[2] David Tellez, Geert J. S. 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. *Med Image Anal*, 58:101544, 2019.

[3] N. Brieu, Armin Meier, Ansh Kapil, Ralf Schoenmeyer, Christoc G Gavriel, Peter D. Caie, and Guenter Schmidt. Domain adaptation-based augmentation for weakly supervised nuclei detection. In *COMPAY@MICCAI*, 2019.

[4] Michael Gadermayr, Laxmi Gupta, Vitus Appel, Peter Boor, Barbara M. Klinkhammer, and Dorit Merhof. Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: A study on kidney histology. *IEEE Trans Med Imaging*, 38:2293–2302, 2019.

[5] Laxmi Gupta, Barbara Mara Klinkhammer, Peter Boor, Dorit Merhof, and Michael Gadermayr. Stain independent segmentation of whole slide images: A case study in renal histology. In *ISBI*, pages 1360–1364, 2018.

[6] Thomas Lampert, Odyssée Merveille, Jessica Schmitz, Germain Forestier, Friedrich Feuerhake, and Cédric Wemmert. Strategies for training stain invariant CNNs. In *ISBI*, pages 905–909, 2019.

[7] Gabriela Csurka. A comprehensive survey on domain adaptation for visual applications. In *Domain adaptation in computer vision applications*, chapter 1, pages 1–35. Springer, 2017.

[8] Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, and Jeffrey Chiang. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med Image Anal*, 63, 2020.

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

[10] Thomas de Bel, Meyke Hermsen, Jesper Kers, Jeroen van der Laak, and Geert J. S. Litjens. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In *MIDL*, volume 102, pages 151–163, 2019.

[11] Aïcha Bentaieb and Ghassan Hamarneh. Adversarial stain transfer for histopathology image analysis. *IEEE Trans Med Imaging*, 37(3):792–802, 2018.

[12] Maxime W. Lafarge, Josien P. W. Pluim, Koen A. J. Eppenhof, Pim Moeskops, and Mitko Veta. Domain-adversarial neural networks to address the appearance variability of histopathology images. In *DLMIA*, 2017.

[13] Maxime W. Lafarge, Josien P. W. Pluim, Koen A. J. Eppenhof, and Mitko Veta. Learning domain-invariant representations of histological images. *Frontiers in Medicine*, 6:162, 2019.

[14] M. Macenko, M. Niethammer, J. S. Marron, D. Borland, J. T. Woosley, Xiaojun Guan, C. Schmitt, and N. E. Thomas. A method for normalizing histology slides for quantitative analysis. In *ISBI*, pages 1107–1110, 2009.

[15] Sebastian Otálora, Manfredo Atzori, Vincent Andrearczyk, Amjad Khan, and Henning Müller. Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology. *Front Bioeng and Biotech*, 7:198, 2019.

[16] M. Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. StainGAN: Stain style transfer for digital histological images. In *ISBI*, pages 953–956, 2019.

[17] Amal Lahiani, Jacob Gildenblat, Irina Klaman, Shadi Albarqouni, Nassir Navab, and Eldad Klaiman. Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach. *ECDP*, 11435, 2019.

[18] Caner Mercan, Germonda Reijnen-Mooij, David T. Martin, Johannes Lotz, Nick Weiss, Marcel van Gerven, and Francesco Ciompi. Virtual staining for mitosis detection in breast histopathology. In *ISBI*, 2020.

[19] Chetan L. Srinidhi, Ozan Ciga, and Anne L. Martel. Deep neural network models for computational histopathology: A survey. *Med Image Anal*, 67, 2021.

[20] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018.

[21] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.

[22] Xianxu Hou, Jingxin Liu, Bolei Xu, Bozhi Liu, Xin Chen, Mohammad Ilyas, Ian O. Ellis, Jon Garibaldi, and Guoping Qiu. Dual adaptive pyramid network for cross-stain histopathology image segmentation. In *MICCAI*, 2019.

[23] K. Mei, C. Zhu, L. Jiang, J. Liu, and Y. Qiao. Cross-stained segmentation from renal biopsy images using multi-level adversarial learning. In *ICASSP*, pages 1424–1428, 2020.

[24] B. Gholami, P. Sahu, O. Rudovic, K. Bousmalis, and V. Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29:3993–4002, 2020.

[25] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, Geert Litjens, Jeroen van der Laak, and Francesco 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(9):2126–2136, 2018.

[26] Y. Xiao, E. Decencière, S. Velasco-Forero, H. Burdin, T. Bornschlögl, F. Bernerd, E. Warrick, and T. Baldeweck. A new color augmentation method for deep learning segmentation of histological images. In *ISBI*, pages 886–890, 2019.

[27] Neslihan Bayramoglu, Mika Kaakinen, Lauri Eklund, and Janne Heikkilä. Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks. In *ICCVW*, pages 64–71, 2017.

[28] Aman Rana, Gregory Yauney, Alarice Lowe, and Pratik Shah. Computational histological staining and destaining of prostate core biopsy RGB images with generative adversarial neural networks. In *ICMLA*, pages 828–834, 2018.

[29] Zhaoyang Xu, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. GAN-based virtual re-staining: A promising solution for whole slide image analysis. *arXiv*, 1901.04059, 2019.

[30] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Comput Graph*, 21(5):34–41, 2001.

[31] Abhishek Vahadane, Tingying Peng, Shadi Albarqouni, Maximilian Baust, Katja Steiger, Anna Schlitter, Amit Sethi, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans Med Imaging*, 35(8):1962–1971, 2016.

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[33] 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.

[34] Olaf Ronneberger, Philipp Fischer, and Thomas Broxs. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[35] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse image synthesis for multiple domains. In *CVPR*, June 2020.

[36] 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. *MICCAI 2018*, 11071:165–173, 2018.

[37] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A.W.M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Med Image Anal*, 42:60–88, 2017.

[38] Thomas de Bel, Meyke Hermsen, Bart Smeets, Luuk Hilbrands, Jeroen van der Laak, and Geert Litjens. Automatic segmentation of histopathological slides of renal tissue using deep learning. In *Medical Imaging 2018: Digital Pathology*, volume 10581, page 1058112, 2018.

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

[40] Raphael Marée, Loic Rollus, Benjamin Stevens, Renaud Hoyoux, Gilles Louppe, R. Vandaele, Jean-Michel Begon, Philipp Kainz, P. Geurts, and L. Wehenkel. Collaborative analysis of multi-gigapixel imaging data using cytomine. *Bioinformatics*, 32(9):1395–1401, 2016.

[41] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *J. Open Source Softw*, 3(29):861, 2018.

[42] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math*, 20:53 – 65, 1987.

[43] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[44] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, page 6629–6640. 2017.

[45] Mikołaj Bińkowski, Dougal J. Sutherl, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *International Conference on Learning Representations*, 2018.

[46] Suman V. Ravuri and Oriol Vinyals. Seeing is not necessarily believing: Limitations of BigGANs for data augmentation. In *ICLR Learning from Limited Labeled Data Workshop*, 2019.

[47] Dina Bashkirova, Ben Usman, and Kate Saenko. Adversarial self-defense for cycle-consistent GANs. In *NeurIPS*, pages 635–645, 2019.

[48] Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN, a master of steganography. In *NIPS 2017, workshop on Machine Deception*, 2017.

[49] Jelmer M. Wolterink, Konstantinos Kamnitsas, Christian Ledig, and Ivana Ivsgum. Deep learning: Generative adversarial networks and adversarial methods. In *Handbook of Medical Image Computing and Computer Assisted Intervention*, chapter 23, pages 547–574. Elsevier, 2020.