HoechstGAN: Virtual Lymphocyte Staining Using Generative Adversarial Networks

Georg Wölflein[1,*] In Hwa Um[2] David J. Harrison[2,3] Ognjen Arandjelović[1]

[1]School of Computer Science, University of St Andrews
[2]School of Medicine, University of St Andrews
[3]Division of Laboratory Medicine, Lothian NHS University Hospitals, Edinburgh
[*]georg@woelflein.de

Abstract

The presence and density of specific types of immune cells are important to understand a patient's immune response to cancer. However, immunofluorescence staining required to identify T cell subtypes is expensive, time-consuming, and rarely performed in clinical settings. We present a framework to virtually stain Hoechst images (which are cheap and widespread) with both CD3 and CD8 to identify T cell subtypes in clear cell renal cell carcinoma using generative adversarial networks. Our proposed method jointly learns both staining tasks, incentivising the network to incorporate mutually beneficial information from each task. We devise a novel metric to quantify the virtual staining quality, and use it to evaluate our method.

1. Introduction

The UK incidence of kidney cancer is projected to rise by 26% to 32 cases per 100,000 by 2035 [46]. Furthermore, in a 2018 survey, the Royal College of Pathologists reported a severe shortage of histopathologists and expected this problem to grow due to increased demand in pathology services in conjunction with an impending retirement crisis of the workforce [48]. The advent of deep learning alongside the availability of high-resolution imaging data provides a promising avenue for the development of new diagnostic tools to aid pathologists in their work [8].

Clear cell renal cell carcinoma (ccRCC) is the most common type of kidney cancer [45, 32] and is characterised by a highly heterogeneous and vascularised tumour microenvironment (TME) [40, 38]. Pathologists grade ccRCC based on the nuclear morphology of cancer cells: the Fuhrman grade considers morphological features such as nuclear shape, size and prominence of the nucleolus [17], but has largely been replaced by the WHO/ISUP scale in 2012 which concerns itself with the nucleolar morphology instead [32, 31]. This grade is one of the main components in ccRCC metastasis risk stratification systems such as the Leibovich score [27], together with the TNM (tumour, node, metastasis) stage [1]. The nuclear grade is obtained from a mere hematoxylin and eosin (H&E) stained slide without relying on costly procedures such as multiplex immunofluorescence (mIF) and immunohistochemistry (IHC). However, even though this nuclear grading system is deployed worldwide, it has often encountered inter- and intra-observer variability [49, 5, 24] due to the high intra-tumour heterogeneity (ITH). The use of machine learning provides a promising avenue to overcome both of these problems [8, 13].

Tumour progression is understood to be the outcome of two competing mechanisms: the invasive tumour process and the opposing defense system which is effectively the host's immune response [18]. The Leibovich score, which has been widely used as a prognostic tool for ccRCC, assesses factors related to the invasive tumour process such as tumour size and tumour cell nuclear morphology, but neglects to take into account the TME, wherein the host immune system plays a role of self-defence. Cluster of differentiation 3 and 8 (CD3 and CD8) are the antigens expressed in T cells and cytotoxic T cells, respectively. Recently, evidence has underscored the prognostic value of immune cell density such as the ratio of $CD8^+$ to $CD3^+$ cells[1] in various types of tumours such as colorectal cancer [16, 18, 35], skin cancer [30] and ccRCC [22].

Contributions This paper makes two main contributions. First, we present a framework that extends the pix2pix architecture [25] for translating a single cheap stain (Hoechst) to multiple more expensive ones (CD3 and CD8). The framework, summarised in Figure 1, is designed in a manner that enables both target domains to mutually benefit each other in the learning process. Second, we introduce a novel metric for the task of virtual staining and use it to evaluate our model. Our code and data are available at `https://georg.woelflein.eu/hoechstgan`.

[1]$CD3^+$ and $CD8^+$ refer to cells that express CD3 and CD8, respectively.

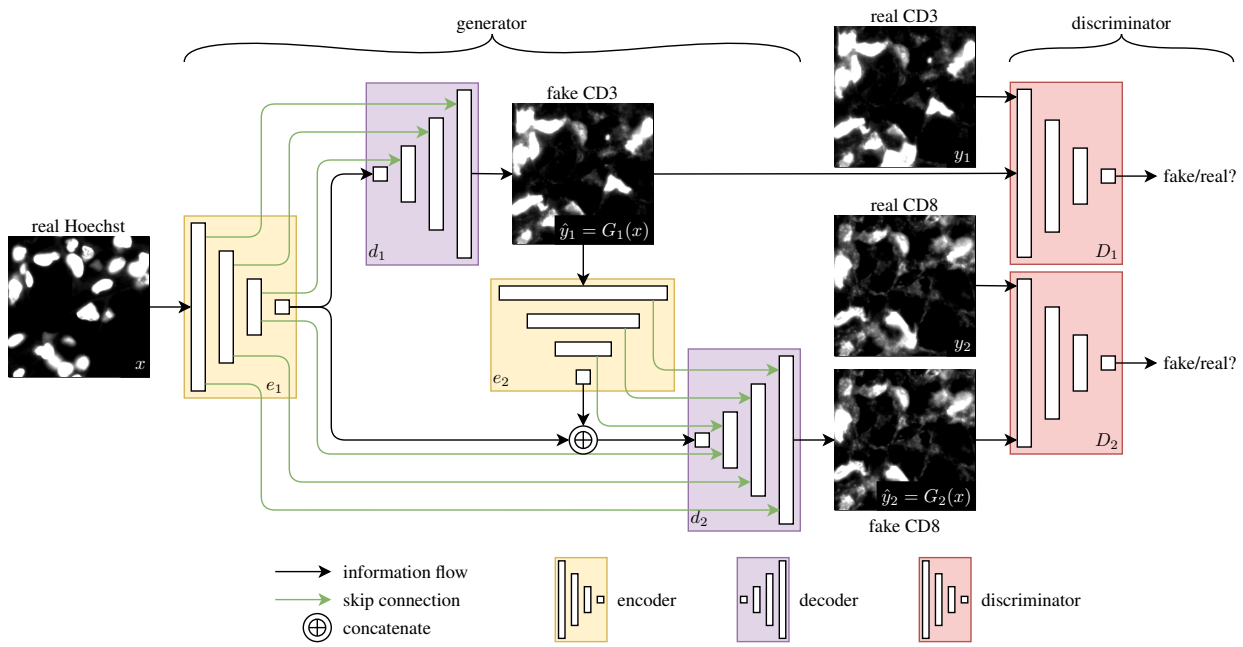


Figure 1: Data flow in the HoechstGAN model for translating Hoechst patches to CD3 and CD8. Generation of the CD3 patch (top part of the figure) can be understood to be a simple pix2pix model [25] (see Figure 4). However, for the CD8 patch, we re-use the encoder part e_1 of the generator alongside a second encoder e_2 from the fake CD3 image. We concatenate the latent codes of both encoders and feed the combined code into the decoder d_2 which is ultimately responsible for generating the fake CD8 image. Skip connections in U-Net style [43] between the encoders and decoders (shown in green) allow the model to directly pass low-level information across the network without needing to compress it through the latent code bottlenecks. Finally, we have a separate discriminator for both target stains.

1.1. Physical immunostaining

One of the most important techniques in clinical oncology and cancer research is the process of *immunostaining* a number of different proteins in cancer tissue sections which is used to assist the diagnosis and decide treatment options using immunoglobulin-based dection systems via chromogen or fluorophores [26, 20, 15]. Immunostaining facilitates the visualisation of various proteins in cells using artificial colouration [11] to distinguish different cell types such as CD105 for enothelial cells representing blood vessels, CD3 for pan T cells, CD20 for B cells, and pan-cytokeratin for epithelial cells. This staining enables pathologists to explore spatial relationships and to assess differences in intensities. Analysing the spatial distribution of specific molecules can aid in understanding the TME and cancer mechanisms, laying the foundation to personalised therapy by predicting which patients are more likely to respond to immunotherapy [37]. For example, a recent study showed how a specific marker identified using IHC can be used to obtain better risk stratification in ccRCC [36].

The two main types of molecular staining assays are IHC and immunofluorescence (IF). While in brightfield IHC the antigens on cells are visualised by chromogen converting into a brown colour in the presence of horseradish peroxidase (HRP), IF is visualised by the enzymatic reaction between fluorescent coated tyramide and HRP [53, 7, 26]. Due to a higher signal to noise ratio and less interference between channels, IF is better suited for detecting multiple different molecules simultaneously (known as mIF), and hence we employ it in this study.

Hoechst 33342 is a widely used counterstaining fluorescent dye that binds to the adenine-thymine (A and T) boundary in DNA, thereby highlighting cell nuclei [9]. It is relatively cheap at around $1 per slide and takes approximately 30 minutes to prepare. On the other hand, it is much more expensive and time-consuming to identify specific subtypes of cells such as T cells. The antigens CD3 and CD8 are expressed in T cells and cytotoxic T cells, respectively, and thus used in CD3 or CD8 staining to highlight the respective cell subtypes. However, even single IF (required to stain just one of these two markers) costs upwards of $20 per slide and takes 3.5 hours to perform; this is much more expensive and time-consuming than Hoechst counterstaining. Nonetheless, features such as the densities of $CD3^+$ and $CD8^+$ cells are very useful in assessing the immune response [16, 18, 22]. Moreover, this information can improve the accuracy of currently available patient risk stratification systems such as the Leibovich score after nephrectomy, which could aid therapies tailored to individual pa-

tients. For example, patients with a very low risk of disease recurrence can be saved from undergoing expensive treatments with severe side-effects such as anti-cancer drugs, tyrosine kinase inhibitor (TKI) therapy or immunotherapy and could live with minimal fear of cancer returning.

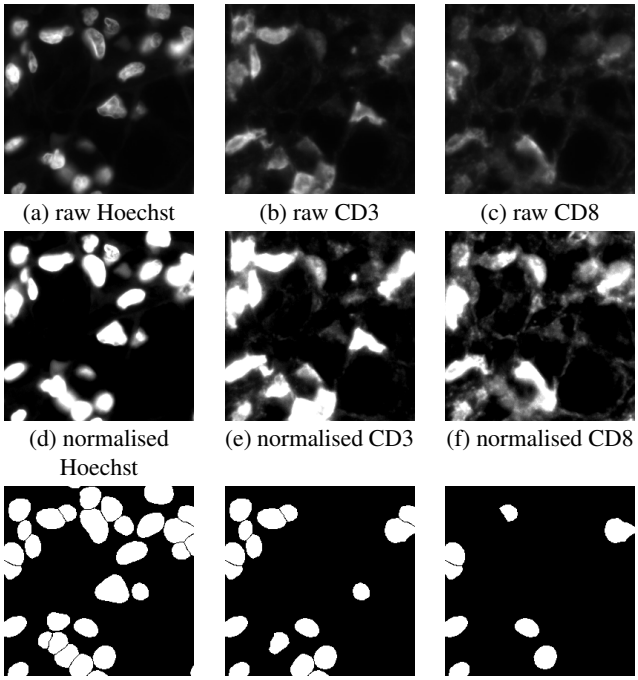
Given the benefit of the Immunoscore™ suggested by Galon *et al*. [18] in colorectal cancer, CD3 and CD8 immunostaining could improve metastasis risk stratification in ccRCC. Since immunostaining is costly and time-consuming compared to Hoechst counterstaining, a system that is able to artificially generate CD3 and CD8 stained output based on Hoechst counterstained input (*i.e*. the focus of this work) would provide practical value in the clinic.

1.2. Virtual staining

The process of artificially transforming digital images of physical stains is known as *virtual staining*. More specifically, given a whole slide image (WSI) of a tissue sample physically stained with some stain A, virtual staining means artificially generating another WSI of how that tissue would present itself had it been stained with another stain B. We distinguish between two types of virtual staining: *stain normalisation* and *stain-to-stain transformation* [41]. Stain normalisation attempts to standardise the input image to produce a uniform and consistent stain [41, 34], thereby providing an antidote to the high level of variations observed within whole slide imaging systems and laboratories that are due to different protocols, dyes and scanners [4, 2]. On the other hand, stain-to-stain transformation concerns itself with the situation where source and target stains are different [41], which is the focus of this work. Nonetheless, preserving tissue structure and morphology is an important consideration in both cases [41, 34].

Virtual staining is inherently posed as an image-to-image translation problem, and the underlying dataset may either be paired or unpaired. In the paired setting, each tissue sample is stained with both the source and target stains (*e.g*. by Hoechst and CD3/CD8 using mIF, as is the case in this work), whereas in the unpaired setting, the source and target images do not need to originate from the same tissue samples. Due to the nature of the problem, it is not surprising that generative deep learning models constitute the predominant approach. Generative adversarial networks (GANs) [21] have enjoyed immense popularity in this domain, especially variants of pix2pix [25] in the paired setting [19, 14, 54, 55, 3, 42], and of CycleGAN [56] or StarGAN [10] for unpaired datasets [14, 52, 28]. However, GANs are notoriously difficult to evaluate [51, 6], a problem that is exacerbated in the unpaired setting by the lack of a ground truth. Therefore, we formulate a novel evaluation metric for this type of paired virtual staining problem.

It is worth noting that the task of identifying lymphocytes from Hoechst stained images has also been posed as a segmentation problem in the past [12], circumventing the need for virtual staining. Rather than a CD3-stained image, we would obtain a segmentation mask of $CD3^+$ cells; in effect we would be translating between Figure 2d and Figure 2h. However, we found that the CD3/CD8-stained images contain information valuable in supervising the learning process that is lost when the target domain is a binary segmentation mask. In fact, training our model directly on the binary masks produced completely black images.



(a) raw Hoechst (b) raw CD3 (c) raw CD8

(d) normalised Hoechst (e) normalised CD3 (f) normalised CD8

(g) StarDist [44] cell mask (h) $CD3^+$ cells (i) $CD8^+$ cells

Figure 2: A 256×256 pixel patch (corresponding to $58\,\mu\mathrm{m} \times 58\,\mu\mathrm{m}$ in physical size) from a WSI (itself around $20\,\mathrm{mm} \times 20\,\mathrm{mm}$), showing raw and normalised intensities for Hoechst, CD3, and CD8, as well as masks for different cell types. Note that $CD8^+$ cells are a subset of $CD3^+$ cells.

Combining physical and virtual staining (*i.e*. performing virtual stain translation) has the potential to save time and money, while also providing standardised staining [41].

2. Materials and methods

Our dataset consists of whole slide images digitised from tumour tissue from ten patients with ccRCC. These were sourced from the Pathology Archive in Lothian NHS (ethics reference 10/S1402/33). Using mIF, the slides were stained with Hoechst, CD3 and CD8 before being scanned, resulting in a dataset of ten paired WSIs (the pairs being Hoechst $\rightarrow \{\mathrm{CD3}, \mathrm{CD8}\}$).

2.1. Multiplex immunofluorescence protocol

Leica BOND RX automated immunostainer (Leica Microsystems, Milton Keynes, UK) was utilised to perform mIF. The sections were dewaxed at $72\,^{\circ}\mathrm{C}$ using BOND

dewax solution (Leica, AR9222) and rehydrated in absolute alcohol and deionised water, respectively. The sections were treated with BOND epitope retrieval 1 (ER1) buffer (Leica, AR9961) for 20 min at 100 °C to unmask the epitopes. The endogeneous peroxidase was blocked with peroxide block (Leica, DS9800), followed by serum free protein block (Agilent, x090930-2). The sections were incubated with the first primary antibody (CD8, Agilent, M710301-2, 1:400 dilution) for 40 min at room temperature, followed by anti-mouse HRP conjugated secondary antibody (Agilent, K400111-2) for 40 min. Then CD8 antigen was visualised by Cy5 conjugated tyramide signal amplification (TSA) (Akoya Bioscience, NEL745001KT). Redundant antibodies, which were not covalently bound, were stripped off by ER1 buffer at 95 °C for 20 min. Then the second primary antibody (CD3, Agilent, A045229-2, 1:400 dilution) was visualised by TSA Cy3, taking the same steps from peroxide block to the ER1 buffer stripping of the first antibody visualisation. Cell nuclei were counterstained by Hoechst 33342 (Thermo Fisher, H3570, 1:100) and the sections were mounted with prolong gold antifade mountant (Thermo Fisher, P36930).

2.2. Image acquisition

The fluorescent images were captured using Zeiss Axio Scan Z1. We used three different fluorescent channels (Hoechst 33342, Cy3 and Cy5) simultaneously to capture individual channel images under 20× object magnification with respective exposure times of 10 ms, 20 ms and 30 ms.

The original dataset of ten WSIs was split into a training set of eight and a test set of two WSIs. Per slide, we generate approximately 60,000 non-overlapping patches of size 256 × 256 pixels, discarding empty patches. In total, that makes 475,334 patches in the training set and 152,185 patches in the test set. Table 1 lists the representation of each cell type in the dataset.

2.3. Preprocessing

2.3.1 Intensity normalisation

Digital images are commonly processed with 8-bit precision. However, the histograms in Figure 3 reveal that most of that pixel luminance is concentrated at the lower end of the range. Therefore, simply quantising the image to the range of $[0, 255]$ would lose important information. Instead, we apply a form of thresholding to normalise the intensities.

Notice that in Figure 3, each histogram exhibits one main peak (the leftmost maximum can be disregarded because it occurs at an intensity of almost zero corresponding to the background pixels). We found it sufficient to assume that the histogram follows a normal distribution $\mathcal{N}(\mu, \sigma^2)$, the parameters of which are obtained using maximum likelihood estimation. We found that the important information is concentrated between the peak and three standard deviations, *i.e.* in the range $[\mu, \mu + 3\sigma]$ which is indicated by the red lines in Figure 3. In particular, we determined that eliminating intensities to the left of that peak ($x < \mu$) effectively reduces background noise. Furthermore, pixels with very high intensities ($x > \mu + 3\sigma$) are rare and can be discarded because they do not add much information. Consequently, we transform the intensities x to a scale $[0, 1]$ by

$$f(x) = \min\left(1, \max\left(0, \frac{x - \mu}{3\sigma}\right)\right).$$

We estimate the parameters μ and σ based on the histograms of the WSIs and not on a patch level because the patches exhibit high variance. The described intensity normalisation procedure is applied to each stain separately, as illustrated by the sample patch in Figure 2.

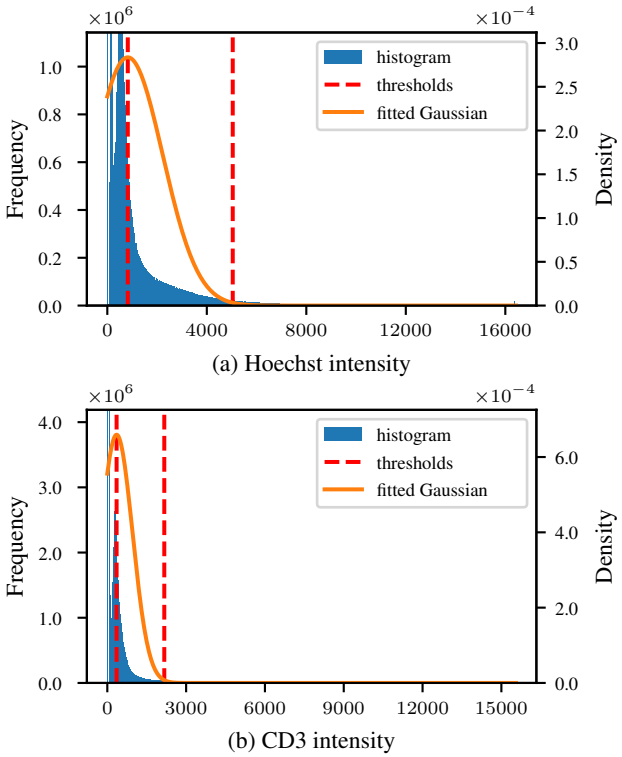


Figure 3: Intensity histograms (left axes) and fitted normal distributions (right axes) of the Hoechst and CD3 WSIs.

2.3.2 Cell segmentation

To facilitate objective assessment of the performance of the evaluated methods, we pre-compute cell masks for each patch. Since the Hoechst channel is responsible for highlighting cell nuclei, we first segment all cells in this channel using the StarDist algorithm [44] (see Figure 2g). Then, we apply thresholding to identify $CD3^+$ (Figure 2h) and $CD8^+$ (Figure 2i) cells using their respective channels. We do not attempt to segment cells on the CD3 or CD8 channels directly because the Hoechst masks are more reliable.

Two main factors negatively influence the quality of segmentation. First, while Hoechst stains the cell nuclei, CD3

Table 1: Representation of cell subtypes across the dataset. Presence refers to the percentage of patches that contain at least one cell of the respective subtypes. Area coverage means the percentage of pixels that are occupied by each cell subtype. In total, the dataset consists of 10 WSIs which generate 627,519 non-overlapping patches.

	Overall			Train			Test		
	Hoechst	CD3	CD8	Hoechst	CD3	CD8	Hoechst	CD3	CD8
Total cells	15,956,049	3,390,533	1,894,016	12,922,586	2,733,198	1,513,501	3,033,463	657,335	380,515
Cells per patch	25.42	5.40	3.02	27.19	5.75	3.18	19.93	4.32	2.50
Presence	99.95 %	93.08 %	71.61 %	99.96 %	94.40 %	7.191 %	99.94 %	8.90 %	7.07 %
Area coverage	26.48 %	5.01 %	3.02 %	28.16 %	5.34 %	3.15 %	21.22 %	3.98 %	2.61 %

is expressed only in a tiny part of a T cell's cytoplasm. Therefore, Hoechst and CD3 stains often do not perfectly align, which is evident by the fact that high intensity blobs in Figure 2e do not exactly match with Figure 2d. Furthermore, some cells in the WSIs are out of focus due to the slide thickness of $4\,\mu\text{m}$ which becomes evident by the varying intensity levels in Figures 2a and 2b. As a result of both of these factors, some CD3$^+$ cells (and analogously also CD8$^+$ cells) may erroneously not be classified as such.

2.4. HoechstGAN

Pix2pix [25] is a popular GAN model for the task of image-to-image translation. However, we are interested in translating a *single* Hoechst image to *two* different stains (CD3 and CD8). Instead of training two separate pix2pix GANs (one for Hoechst $\rightarrow$ CD3 and another for Hoechst $\rightarrow$ CD8, see Figure 4), we propose a combined model that is able to leverage the fact that information learnt in the first task may be beneficial in the second, and vice-versa.

Our proposed method extends the pix2pix architecture [25] in two main ways. First, we reuse the encoder part of the generator for both tasks (see Figure 1) which means that the latent representation learnt by the encoder is used as input to two separate decoders that generate the two stains. Therefore, the U-Net skip connections (which lead from the respective encoder layers to both decoders), as well as the latent representation itself, ensure that the encoder is able to capture relevant information for both tasks. Secondly, we employ an additional encoder that creates a latent code for the generated fake CD3 image which is concatenated with the latent code of the original Hoechst image to form the input of the CD8 decoder. We hypothesise that this additional step allows the network to learn a transformation from CD3 to CD8 which is useful to our task because CD8$^+$ cells are a subset of CD3$^+$ cells. Section 3 tests this claim's validity.

2.4.1 Training

Given a Hoechst image x, we generate the fake CD3 stain $\hat{y}_1 = G_1(x)$ using encoder e_1 and decoder d_1, like in pix2pix [25]:

$$G_1(x) = d_1(e_1(x)). \tag{1}$$

However, we depart from pix2pix [25] in generating the fake CD8 image $\hat{y}_2 = G_2(x)$ by fusing the latent representations of the original Hoechst image with the fake CD3 image from (1), as shown in Figure 1. Another decoder d_2 then converts the merged latent code to an image, given by

$$G_2(x) = d_2\left(e_1(x) \oplus e_2\left(G_1(x)\right)\right), \tag{2}$$

where $\oplus$ denotes concatenation of latent representations. In fact, this implicitly entails concatenating the skip connections of both encoders e_1 and e_2 to the decoder d_2 as well.

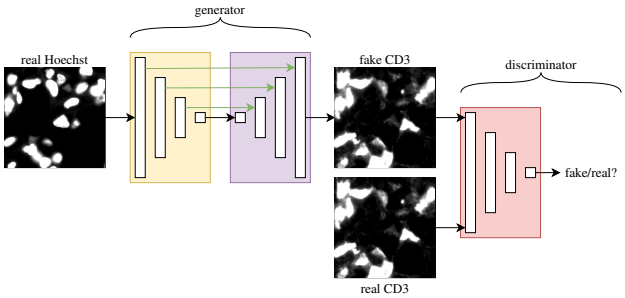


Figure 4: Data flow in the pix2pix model [25] applied to Hoechst $\rightarrow$ CD3 translation (notation is explained in Figure 1). Hoechst $\rightarrow$ CD8 translation follows the same idea.

Traditionally, noise is introduced in the image generation process of vanilla GANs directly by providing a Gaussian noise vector z as auxiliary input [21]. However, this turns out to be of little use for GANs conditioned on images because they simply learn to ignore the stochasticity provided by z [25]. Consequently, we infuse noise only in the form of dropout in the decoders, both at train and test time.

Our conditional GAN objective can be expressed as

$$\begin{aligned}\mathcal{L}_{\text{cGAN}}(G_1, G_2, D_1, D_2) = &\ \mathbb{E}_x\left[\log\left(1 - D_1(x, G_1(x))\right)\right] + \\ &\ \mathbb{E}_x\left[\log\left(1 - D_2(x, G_2(x))\right)\right] + \\ &\ \mathbb{E}_{x,y_1}\left[\log D_1(x, y_1)\right] + \\ &\ \mathbb{E}_{x,y_2}\left[\log D_2(x, y_2)\right]\end{aligned} \tag{3}$$

which the generators attempt to minimise against the adversarial discriminators (D_1, D_2) in turn trying to maximise it.

Note that each discriminator is a function of two images: the original Hoechst patch and the real/fake output stain (CD3 or CD8). In our implementation, we simply concatenate the two images before feeding them to the discriminator. However, in Section 3, we evaluate a variant of our model that employs a single shared discriminator D instead of the two separate discriminators D_1 and D_2. Here, the

new joint discriminator D obtains three (concatenated) inputs: the original Hoechst, a CD3 and a CD8 patch. It is tasked with deciding whether the CD3 and CD8 patches are both real or fake. As such, the conditional GAN loss from (3) reduces to only two terms:

$$\mathcal{L}_{\text{cGAN}}(G_1, G_2, D) = \mathbb{E}_x \left[\log\left(1 - D(x, G_1(x), G_2(x))\right)\right] + \mathbb{E}_{x,y_1,y_2} \left[\log D(x, y_1, y_2)\right]. \quad (4)$$

Our objective further incentivises the fake and real images to be similar using an L1 loss of the form

$$\mathcal{L}_{\text{L1}}(G_i) = \mathbb{E}_{x,y_i} \left[\|y_i - G_i(x)\|_1\right]. \quad (5)$$

The overall objective combines the GAN loss with the L1 penalties of both generators balanced by a hyperparameter λ to manage the trade-off, which we set to 100:

$$G^* = \arg \min_{G_1,G_2} \max_{D_1,D_2} \mathcal{L}_{\text{cGAN}}(G_1, G_2, D_1, D_2) + \lambda\mathcal{L}_{\text{L1}}(G_1) + \lambda\mathcal{L}_{\text{L1}}(G_2). \quad (6)$$

The training dataset consists of 256×256 pixel patches of normalised Hoechst images (Figure 2d) and corresponding normalised CD3 and CD8 images (Figures 2e and 2f). We train with a batch size of 64 for a total of 30 epochs. For the first 20 epochs, we use a learning rate of 0.0002 which is made to decay linearly to zero in the final ten epochs.

2.4.2 Compositing real and fake CD3 input

At the beginning of the training process, the generated CD3 image $G_1(x)$ is of very poor quality, and thus the encoder e_2 in Figure 1 is unable to learn useful representations. Therefore, we begin training by feeding the real CD3 image into the encoder e_2, and, as training progresses, gradually substitute the real image with the fake one generated by $G_1(x)$.

In other words, we substitute $G_1(x)$ in (2) with a composite image of $G_1(x)$ and the ground truth y_1 given by $\beta G_1(x) + (1 - \beta)y_1$ where the parameter β gradually increases from 0 to 1 during training. As such, (2) becomes

$$G_2(x) = d_2\left(e_1(x) \oplus e_2\left(\beta G_1(x) + (1 - \beta)y_1\right)\right), \quad (7)$$

where β depends on the epoch[2]. We achieve good results using $\beta = 0$ until epoch 8, and then employing a scaled sigmoid function between epochs 8 and 12 that transitions β to from 0 to 1 so that $\beta = 0.5$ at epoch 10.

2.4.3 Network architectures

The generator consists of encoders and decoders which are made up of convolution-BatchNorm-ReLU blocks [33, 23], as is common in deep convolutional GANs [39, 25]. In the encoder, we have eight such blocks with 64, 128, 256, 512, 512, 512, 512, and 512 convolutional filters, respectively. Each block halves the spatial dimensions of the input using 4×4 convolution kernels with a stride of 2. On the other hand, the corresponding blocks in the decoder achieve a doubling of spatial size by employing deconvolutions instead of convolutions. In decoder d_1, the number of filters per block is the same as in the encoders, but in reverse order. For decoder d_2, the number of filters is doubled because of the additional skip connections (*i.e.* 1024, 1024, 1024, 1024, 1024, 512, 256, and 128). We replace the ReLUs with 0.2-sloped leaky ReLUs [29, 50] in the deconvolution-BatchNorm-ReLU blocks of both decoders. Further, we apply 50% dropout [47] in the first four blocks of the decoders at train and test time which not only has a regularising effect during training, but also acts as a source of stochasticity in the image generation process. After the eight aforementioned blocks, the decoders have an additional convolution layer followed by a tanh nonlinearity in order to map to just one output channel (the patches are greyscale).

In our problem setting, it is desirable to retain some low-level information between input and output, such as the exact location and shape of specific cells. For this reason, we add the aforementioned U-Net-style skip connections [43] between the encoders and decoders, depicted by the green arrows in Figure 1. These skip connections, which are also used in the original pix2pix architecture [25], allow this type of information to be passed directly across the network.

We employ separate discriminators for each output staining (CD3 and CD8). Our discriminators are applied convolutionally across the generated and real images, averaging the fake vs. real classification per image, which is known as a PatchGAN [25]. This reduces the number of parameters in the discriminator while still being able to maintain (or even improve) image quality [25]. Both discriminators are identical in architecture: they consist of the aforementioned convolution-BatchNorm-ReLU blocks with 64, 128, 256, and 512 filters, respectively, using leaky ReLUs with a slope of 0.2, although the BatchNorm layer is omitted in the first block. This results in a receptive field of size 70×70.

3. Results

Quantitatively evaluating and comparing GANs is an infamously challenging task [51, 6] because metrics that objectively assess the quality of generated images are hard to come by. However, it turns out that virtual staining actually lends itself to objective evaluation by considering the signal to noise ratios of the real and generated stains. To this end, we introduce a novel evaluation metric for measuring the quality of virtual staining models which we refer to as the *masked intensity ratio (MIR)*. This metric scores highly if the generated CD3 patch exhibits high intensities in the areas of $CD3^+$ segmented cells and low intensities

[2]For compositing, we do not consider the epoch as an integer, but rather a real number that is updated after each batch. For example, if half of the batches in epoch 8 are processed, we consider that to be epoch 8.5.

Table 2: Results. We compare several variants of our HoechstGAN method which are denoted by capital letters in the suffix of the method column. 'M' indicates mutual learning, *i.e*. the generated CD8 stain depends on the generated CD3 stain. If 'M' is absent, encoder e_2 is removed from the model in Figure 1. If 'M' is present, 'C' indicates that compositing was used in the training process, *i.e*. the CD3 input to e_2 is transitioned from real to fake according to the schedule described in Section 2.4.2. Finally, 'D' indicates the use of a joint discriminator with the loss function given in (4). When 'D' is absent, we employ separate discriminators for CD3 and CD8, as depicted in Figure 1, with the loss function in (3).

Method	Parameters ↓	Train CD3 MIR_{rel} ↑	Test CD3 MIR_{rel} ↑	Train CD8 MIR_{rel} ↑	Test CD8 MIR_{rel} ↑
HoechstGAN-MCD	2.16×10^8	**1.23 ± 1.21**	**1.48 ± 1.27**	1.07 ± 0.97	1.43 ± 1.04
HoechstGAN-MC	2.19×10^8	0.87 ± 1.07	1.12 ± 1.23	**1.22 ± 0.88**	1.39 ± 1.02
HoechstGAN-MD	2.16×10^8	0.88 ± 1.07	1.15 ± 1.21	1.01 ± 0.93	**1.45 ± 1.03**
HoechstGAN-M	2.19×10^8	0.89 ± 1.26	1.24 ± 1.30	0.84 ± 0.97	1.26 ± 1.06
HoechstGAN-D	$\mathbf{9.20 \times 10^7}$	0.90 ± 1.01	1.14 ± 1.17	0.87 ± 0.94	1.33 ± 1.05
pix2pix	1.14×10^8	0.89 ± 1.04	1.15 ± 1.23	0.83 ± 0.87	1.16 ± 1.00
Regression-MC	2.13×10^8	0.62 ± 0.74	0.92 ± 0.75	0.57 ± 0.58	1.09 ± 0.73

everywhere else. As such, the MIR is simply defined as the ratio of the mean pixel intensity within the $CD3^+$ cell masks (Figure 2h) to the mean pixel intensity outwith that mask:

$$MIR = \frac{\textit{mean pixel intensity within mask}}{\textit{mean pixel intensity outwith mask}}. \quad (8)$$

Further, we consider the *relative MIR* as the MIR of the generated stain divided by the MIR of the ground truth stain,

$$MIR_{rel} = \frac{MIR_{fake}}{MIR_{real}}, \quad (9)$$

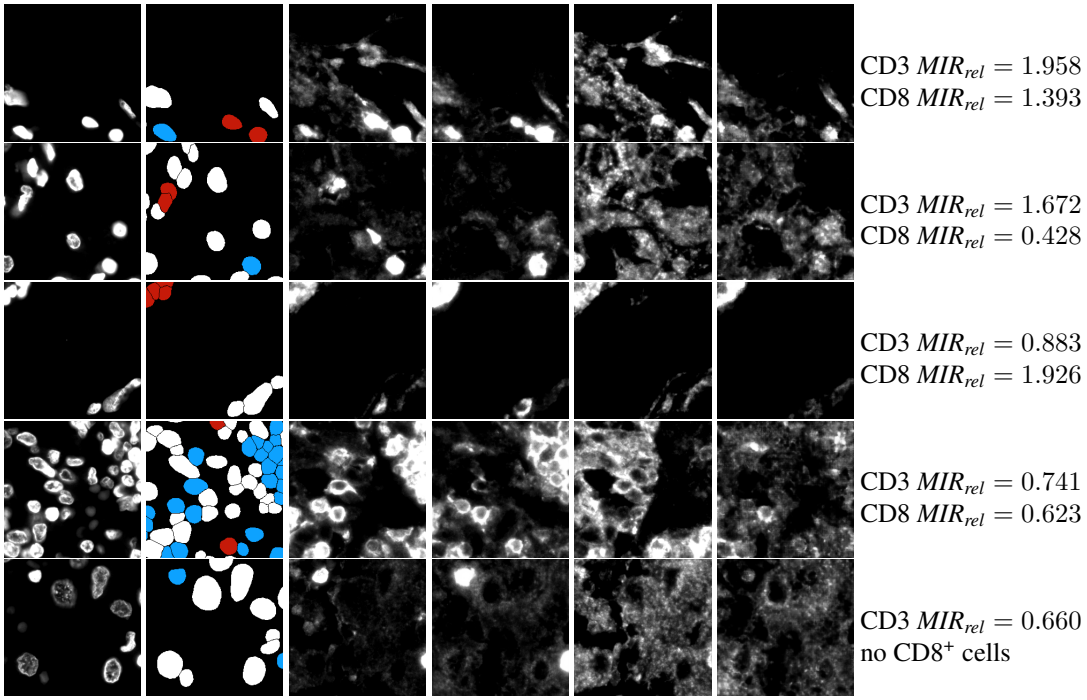
which essentially measures how much better the generated stain is than the ground truth stain. Therefore, the relative MIR is suitable for objectively comparing virtual staining models. A relative MIR greater than one means the fake stain has better signal, or less noise, than the ground truth.

We train several variants of the HoechstGAN architecture described in the previous section, reporting relative MIRs in Table 2 (for train/test sets and CD3/CD8). For comparative purposes, we establish a pix2pix [25] baseline which actually consists of two separate models (one Hoechst $\rightarrow$ CD3 as in Figure 4, and another Hoechst $\rightarrow$ CD8) because the original pix2pix architecture cannot handle multiple output modalities. Furthermore, we trained the best-performing model configuration, HoechstGAN-MCD, using just an L1 loss (without discriminator and GAN loss) as a regression baseline (bottom row in Table 2) to justify the need for the adversarial loss in the first place.

Contrary to what one would initially expect, it is evident in Table 2 that the MIRs are consistently higher in the test set compared to the training set across all models and output tasks. However, it becomes apparent in Table 1 that on average, there are fewer $CD3^+$ and $CD8^+$ cells in the test set patches than the training set patches, so one may argue that the difficulty of the test set patches is lower. This imbalance in train versus test set statistics occurred because we randomly selected eight WSIs for training and left the other two for testing, and the latter two just happened to have a lower density of cells. Had we split the dataset on the level of patches as opposed to WSIs, the cell densities would be more similar across the train and test sets, but, in return, we would (i) run the risk of data leakage, as there would be many incidences of two patches adjacent in one WSI being split into train and test sets, and (ii) reduce the clinical utility of our solution because it was never tested on unseen WSIs. Therefore, our analysis below focuses on comparing the relative test set MIRs across the evaluated models.

The results in Table 2 show that all GANs achieve better test set MIRs than the ground truth (relative MIRs on the test set are all greater than one), meaning that the generated patches have a better signal to noise ratio for identifying both CD3 and CD8 cells. On the test set, some HoechstGAN variants outperform the pix2pix baseline in the task of virtual CD3 generation, but for CD8, all variants improve upon the baseline. The smallest model, HoechstGAN-D, achieves similar performance on CD3 and better performance on CD8 than the pix2pix baseline [25] while requiring 22 million fewer parameters (because some parameters in the generator and all parameters in the discriminator are shared). HoechstGAN-MCD achieves the best performance overall, some samples of which are depicted in Figure 5.

Considering the fact that $CD8^+$ cells are a subset of $CD3^+$ cells, it comes at no surprise that most models struggle more with the former than the latter. However, HoechstGAN variants with joint discriminators (suffixed 'D' in Table 2) are able to close the gap between relative CD3 and CD8 MIRs. Why is this the case? In the HoechstGAN variants (except HoechstGAN-D), CD8 generation involves more parameters than CD3 generation, causing training to strike a trade-off biased towards improving the former at the cost of the latter in the separate discriminator setup. In other words, the two terms of the loss function in (3) that are concerned with D_2 will have greater magnitudes than the other two terms. Using a joint discriminator causes the two pairs of terms to be combined in the form of (4), where the discriminator sees the CD3 and CD8 patches simultaneously, thereby mitigating the balancing problem in the loss function.



(a) Hoechst (b) cell mask (c) real CD3 (d) fake CD3 (e) real CD8 (f) fake CD8

Figure 5: Results of HoechstGAN-MCD applied to five randomly selected samples from the test set. In the cell masks, the blue cells are CD3$^+$, the red ones are CD3$^+$CD8$^+$, and the white ones are neither.

Compositing (Section 2.4.2) has a net positive effect on performance. When using separate discriminators, compositing is able to increase CD3 performance with just a minor drop in relative CD8 MIR. This is in line with the originally intended purpose of ensuring that encoder e_2 learns a useful representation of the fake CD3 patch that benefits CD8 generation. Interestingly however, when deployed with a joint discriminator, compositing actually achieves the opposite effect: CD3 improves significantly while CD8 performance takes a slight hit. We speculate that this may be related to the aforementioned balancing problem in the loss function, but leave further investigation to future work.

Finally, to what extent does CD8 stain generation learn from CD3? We hypothesised in Section 2.4 that encoder e_2 helps CD8 stain generation by learning a useful mapping from CD3 to CD8. To validate this claim, we conduct an experiment on HoechstGAN-MCD, where we replace the input to encoder e_2 with either of the following: the corresponding real CD3 patch ($MIR_{rel} = 1.143$), another real CD3 patch randomly selected from the test set ($MIR_{rel} = 0.610$) or Gaussian noise ($MIR_{rel} = 0.431$). The MIR scores indicate that substituting the real CD3 patch achieves significantly better results than substituting a different randomly selected real CD3 patch or just random noise. We thus conclude that e_2 did, in fact, learn a useful representation of the CD3 input. Better yet, the actual HoechstGAN-MCD model (where e_2 is supplied with the generated CD3 patch) outperforms any of the three substitutions above ($MIR_{rel} = 1.426$ in Table 2), even the real CD3 patch. This means that e_2 has actually adapted itself to the generated CD3 image rather than the real one (in fact, the generated patches exhibit better signal to noise ratios than the real patches since the relative MIR is greater than one), so the compositing process achieved the desired effect.

4. Conclusion

We propose a framework that is able to translate from a cheap stain to multiple more expensive ones, alongside an evaluation metric to assess its performance. Our results suggest that GANs are well-suited for virtual staining, achieving better signal to noise ratios than the ground truth.

An avenue for future investigation could be making the HoechstGAN architecture symmetric between CD3 and CD8, *i.e.* allowing information to flow back from CD8 to CD3. Furthermore, different variations of the joint discriminator loss in (4) could be explored, where the discriminator is supplied combinations of real and fake patches (as opposed to either two real or two fake images).

Acknowledgements: GW is supported by Lothian NHS. This project received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 101017453 as part of the KATY project. This work is supported in part by the Industrial Centre for AI Research in Digital Diagnostics (iCAIRD) which is funded by Innovate UK on behalf of UK Research and Innovation (UKRI) (project number 104690).

References

[1] Mahul B Amin, Frederick L Greene, Stephen B Edge, Carolyn C Compton, Jeffrey E Gershenwald, Robert K Brookland, Laura Meyer, Donna M Gress, David R Byrd, and David P Winchester. The eighth edition AJCC cancer staging manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. *CA: a cancer journal for clinicians*, 67(2):93–99, 2017.

[2] Aldo Badano, Craig Revie, Andrew Casertano, Wei-Chung Cheng, Phil Green, Tom Kimpe, Elizabeth Krupinski, Christye Sisson, Stein Skrøvseth, Darren Treanor, et al. Consistency and standardization of color in medical imaging: A consensus report. *Journal of digital imaging*, 28(1):41–52, 2015.

[3] Bijie Bai, Hongda Wang, Yuzhu Li, Kevin de Haan, Francesco Colonnese, Yujie Wan, Jingyi Zuo, Ngan B Doan, Xiaoran Zhang, Yijie Zhang, et al. Label-free virtual HER2 immunohistochemical staining of breast tissue using deep learning. *arXiv preprint arXiv:2112.05240*, 2021.

[4] Babak Ehteshami Bejnordi, Nadya Timofeeva, Irene Otte-Höller, Nico Karssemeijer, and Jeroen AWM van der Laak. Quantitative analysis of stain variability in histology slides and an algorithm for standardization. In *Medical Imaging 2014: Digital Pathology*, volume 9041, page 904108. International Society for Optics and Photonics, 2014.

[5] Sibel Bektas, Burak Bahadir, Nilufer Onak Kandemir, Figen Barut, Aylin Ege Gul, and Sukru Oguz Ozdamar. Intraobserver and interobserver variability of Fuhrman and modified Fuhrman grading systems for conventional renal cell carcinoma. *The Kaohsiung Journal of Medical Sciences*, 25(11):596–600, 2009.

[6] Ali Borji. Pros and cons of gan evaluation measures. *Computer Vision and Image Understanding*, 179:41–65, 2019.

[7] Igor B. Buchwalow and Werner Böcker. *Immunohistochemistry: Basics and Methods* -. Springer Science & Business Media, Berlin Heidelberg, 2010.

[8] Peter D Caie, Neofytos Dimitriou, and Ognjen Arandjelović. Precision medicine in digital pathology via image analysis and machine learning. In *Artificial Intelligence and Deep Learning in Pathology*, pages 149–173. Elsevier, 2021.

[9] Brad Chazotte. Labeling nuclear DNA with hoechst 33342. *Cold Spring Harbor Protocols*, 2011(1):pdb–prot5557, 2011.

[10] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018.

[11] Albert H Coons, Hugh J Creech, R Norman Jones, and Ernst Berliner. The demonstration of pneumococcal antigen in tissues by the use of fluorescent antibody. *The Journal of Immunology*, 45(3):159–170, 1942.

[12] Jessica Cooper, In Hwa Um, Ognjen Arandjelović, and David J Harrison. Hoechst is all you need: Lymphocyte classification with deep learning. *arXiv preprint arXiv:2107.04388*, 2021.

[13] Raffaele De Filippis, Georg Wölflein, In Hwa Um, Peter D Caie, Sarah Warren, Andrew White, Elizabeth Suen, Emily To, Ognjen Arandjelović, and David J Harrison. Use of high-plex data reveals novel insights into the tumour microenvironment of clear cell renal cell carcinoma. *medRxiv*, 2022.

[14] Kevin de Haan, Yijie Zhang, Jonathan E Zuckerman, Tairan Liu, Anthony E Sisk, Miguel FP Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, et al. Deep learning-based transformation of H&E stained tissues into special stains. *Nature communications*, 12(1):1–13, 2021.

[15] Julie G Donaldson. Immunofluorescence staining. *Current protocols in cell biology*, 69(1):4–3, 2015.

[16] Wolf Herman Fridman, Catherine Sautès-Fridman, Jérôme Galon, et al. The immune contexture in human tumours: Impact on clinical outcome. *Nature Reviews Cancer*, 12(4):298–306, 2012.

[17] Susan A Fuhrman, Larry C Lasky, and Catherine Limas. Prognostic significance of morphologic parameters in renal cell carcinoma. *The American journal of surgical pathology*, 6(7):655–663, 1982.

[18] Jérôme Galon, Bernhard Mlecnik, Gabriela Bindea, Helen K Angell, Anne Berger, Christine Lagorce, Alessandro Lugli, Inti Zlobec, Arndt Hartmann, Carlo Bifulco, et al. Towards the introduction of the 'Immunoscore'in the classification of malignant tumours. *The Journal of pathology*, 232(2):199–209, 2014.

[19] Parmida Ghahremani, Yanyun Li, Arie Kaufman, Rami Vanguri, Noah Greenwald, Michael Angelo, Travis J Hollmann, and Saad Nadeem. Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. *Nature Machine Intelligence*, 4(4):401–412, 2022.

[20] Neal S Goldstein, Stephen M Hewitt, Clive R Taylor, Hadi Yaziji, David G Hicks, Members of Ad-Hoc Committee On Immunohistochemistry Standardization, et al. Recommendations for improved standardization of immunohistochemistry. *Applied Immunohistochemistry & Molecular Morphology*, 15(2):124–133, 2007.

[21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[22] Congfang Guo, Hua Zhao, Yu Wang, Shuai Bai, Zizhong Yang, Feng Wei, and Xiubao Ren. Prognostic value of the neo-immunoscore in renal cell carcinoma. *Frontiers in Oncology*, 9:439, 2019.

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456. PMLR, 2015.

[24] Humayun Irshad, Laleh Montaser-Kouhsari, Gail Waltz, Octavian Bucur, JA Nowak, Fei Dong, Nicholas W Knoblauch, and Andrew H Beck. Crowdsourcing image annotation for nucleus detection and segmentation in computational pathology: Evaluating experts, automated methods, and the crowd. In *Pacific Symposium on Biocomputing Co-chairs*, pages 294–305. World Scientific, 2014.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adver-

sarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[26] Alexander E. Kalyuzhny. *Immunohistochemistry - Essential Elements and Beyond*. Springer, Berlin, Heidelberg, 2016.

[27] Bradley C Leibovich, Michael L Blute, John C Cheville, Christine M Lohse, Igor Frank, Eugene D Kwon, Amy L Weaver, Alexander S Parker, and Horst Zincke. Prediction of progression after radical nephrectomy for patients with clear cell renal cell carcinoma: A stratification tool for prospective clinical trials. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 97(7):1663–1671, 2003.

[28] 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(8):1977–1989, 2021.

[29] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. Icml*, volume 30, page 3. Citeseer, 2013.

[30] Martin C Mihm Jr and James J Mulé. Reflections on the histopathology of tumor-infiltrating lymphocytes in melanoma and the host immune response. *Cancer immunology research*, 3(8):827–835, 2015.

[31] Holger Moch. The WHO/ISUP grading system for renal carcinoma. *Der Pathologe*, 37(4):355–360, 2016.

[32] Holger Moch, Antonio L Cubilla, Peter A Humphrey, Victor E Reuter, and Thomas M Ulbright. The 2016 WHO classification of tumours of the urinary system and male genital organs—part A: Renal, penile, and testicular tumours. *European urology*, 70(1):93–105, 2016.

[33] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Icml*, 2010.

[34] Haseeb Nazki, Ognjen Arandjelović, In Hwa Um, and David Harrison. MultiPathGAN: Structure preserving stain normalization using unsupervised multi-domain adversarial network with perception loss. *arXiv preprint arXiv:2204.09782*, 2022.

[35] Ines Panicou Nearchou, Daniel Alexander Soutar, Hideki Ueno, David James Harrison, Ognjen Arandjelović, and Peter David Caie. A comparison of methods for studying the tumor microenvironment's spatial heterogeneity in digital pathology specimens. *Journal of Pathology Informatics*, 12(1):6, 2021.

[36] Haruyuki Ohsugi, Takashi Yoshida, Chisato Ohe, Junichi Ikeda, Motohiko Sugi, Hidefumi Kinoshita, Koji Tsuta, and Tadashi Matsuda. The SSPN Score, a novel scoring system incorporating PBRM1 expression, predicts postoperative recurrence for patients with non-metastatic clear cell renal cell carcinoma. *Annals of Surgical Oncology*, 28(4):2359–2366, 2021.

[37] Edwin Roger Parra. Methods to determine and analyze the cellular spatial distribution extracted from multiplex immunofluorescence data to understand the tumor microenvironment. *Frontiers in Molecular Biosciences*, 8:575, 2021.

[38] Chao-Nan Qian, Dan Huang, Bill Wondergem, and Bin Tean Teh. Complexity of tumor vasculature in clear cell renal cell carcinoma. *Cancer*, 115(S10):2282–2289, 2009.

[39] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[40] Christopher J Ricketts and W Marston Linehan. Intratumoral heterogeneity in kidney cancer. *Nature genetics*, 46(3):214–215, 2014.

[41] Yair Rivenson, Kevin de Haan, W Dean Wallace, and Aydogan Ozcan. Emerging advances to transform histopathology using virtual staining. *BME Frontiers*, 2020, 2020.

[42] Yair Rivenson, Tairan Liu, Zhensong Wei, Yibo Zhang, Kevin de Haan, and Aydogan Ozcan. PhaseStain: The digital staining of label-free quantitative phase microscopy images using deep learning. *Light: Science & Applications*, 8(1):1–11, 2019.

[43] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[44] Uwe Schmidt, Martin Weigert, Coleman Broaddus, and Gene Myers. Cell detection with star-convex polygons. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2018 - 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II*, pages 265–273, 2018.

[45] Rebecca L Siegel, Kimberly D Miller, Hannah E Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. *CA: a cancer journal for clinicians*, 71(1):7–33, 2021.

[46] C R Smittenaar, K A Petersen, K Stewart, and N Moitt. Cancer incidence and mortality projections in the UK until 2035. *British Journal of Cancer*, 115(9):1147–1155, Oct. 2016.

[47] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[48] The Royal College of Pathologists. Meeting pathology demand: Histopathology workplace census. Technical report, London, UK, 2018.

[49] Anne Y Warren and David Harrison. WHO/ISUP classification, grading and pathological staging of renal cell carcinoma: Standards and controversies. *World journal of urology*, 36(12):1913–1926, 2018.

[50] Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[51] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.

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

[53] Aliya U Zaidi, Hideki Enomoto, Jeffrey Milbrandt, and Kevin A Roth. Dual fluorescent in situ hybridization and immunohistochemical detection with tyramide signal amplification. *Journal of Histochemistry & Cytochemistry*, 48(10):1369–1375, 2000.

[54] Guanghao Zhang, Hui Hui, Bin Ning, Di Dong, Jie Tian, and Wen He. Self-attention based virtual staining for bright-field images of label-free human carotid atherosclerotic plaque tissue section. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 3492–3495. IEEE, 2021.

[55] Yijie Zhang, Kevin de Haan, Yair Rivenson, Jingxi Li, Apostolos Delis, and Aydogan Ozcan. Digital synthesis of histological stains using micro-structured and multiplexed virtual staining of label-free tissue. *Light: Science & Applications*, 9(1):1–13, 2020.

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