



**HAL**  
open science

# Towards Measuring Domain Shift in Histopathological Stain Translation in an Unsupervised Manner

Zeeshan Nisar, Jelica Vasiljevic, Pierre Gançarski, Thomas Lampert

► **To cite this version:**

Zeeshan Nisar, Jelica Vasiljevic, Pierre Gançarski, Thomas Lampert. Towards Measuring Domain Shift in Histopathological Stain Translation in an Unsupervised Manner. IEEE International Symposium on Biomedical Imaging, from March 28, 2022 to March 31, 2022, Kolkata, India, Mar 2022, Kolkata, India. 10.1109/ISBI52829.2022.9761411 . hal-03799541

**HAL Id: hal-03799541**

**<https://hal.science/hal-03799541>**

Submitted on 5 Oct 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# TOWARDS MEASURING DOMAIN SHIFT IN HISTOPATHOLOGICAL STAIN TRANSLATION IN AN UNSUPERVISED MANNER

Zeeshan Nisar\* Jelica Vasiljević<sup>\*,†,‡</sup> Pierre Gançarski\* Thomas Lampert\*

\*ICube, University of Strasbourg, France †University of Belgrade, Serbia

‡Faculty of Science, University of Kragujevac, Serbia

## ABSTRACT

Domain shift in digital histopathology can occur when different stains or scanners are used, during stain translation, etc. A deep neural network trained on source data may not generalise well to data that has undergone some domain shift. An important step towards being robust to domain shift is the ability to detect and measure it. This article demonstrates that the PixelCNN and domain shift metric can be used to detect and quantify domain shift in digital histopathology, and they demonstrate a strong correlation with generalisation performance. These findings pave the way for a mechanism to infer the average performance of a model (trained on source data) on unseen and unlabelled target data.

*Index Terms*— digital pathology, image-to-image translation, domain shift, generative models, segmentation

## 1. INTRODUCTION

Image datasets in digital pathology often consist of consecutive tissue slides stained differently [1], with each stain providing different information on the same region of interest. Since each stain highlights different tissue structures, even consecutive slides (representing identical anatomical structures e.g. glomeruli) can appear very different, see Fig. 1 (1<sup>st</sup> row). Furthermore, the staining procedure is vulnerable to high variability due to inter-subject variations, lab specific techniques, capturing pipeline changes [2], scanner characteristics, and staining protocols, and these can introduce further variation of a tissue’s appearance [3].

Although large-scale biological structures retain morphological structure across each stain as with the glomeruli in Fig. 1, state-of-art deep learning (DL) methods trained for some task (i.e. glomeruli segmentation) on one (source) stain (e.g. PAS) do not generalise well to histological images of the target stain (e.g. Jones H&E, CD68, CD34, Sirius Red), see Table 1 (1<sup>st</sup> row). This degradation in performance is caused by the inter-stain variation, see Fig. 1 (1<sup>st</sup> row), or intra-stain variation (e.g. one stain collected from different laboratories). Even small domain shifts may cause significant drops in performance. Many DL algorithms are vulnerable to this shift

[4], meaning that proper care needs to be taken when deploying them for clinical aid. As such, it is of great importance to handle this variance or at least to estimate when it is likely to significantly effect an algorithm’s performance.

Therefore an important step towards handling domain shift in digital histopathology is the ability to detect it. To the best of our knowledge, no such work exists for digital pathology, particularly for segmentation. This article concentrates on detecting domain shift during stain style transfer between a source stain (PAS) and translated Target→PAS, nevertheless the presented solution is general and unsupervised, and is therefore applicable to other types of domain shift.

Two approaches are investigated: the PixelCNN [5], which is used to model the distribution of the source data in an unsupervised manner, and the Domain Shift Metric (DSM) [2], which uses a pre-trained feature representation to measure domain shift and therefore integrates some knowledge of the task to be performed, in this study the segmentation representation trained on the source stain is used. It is shown that both of these measures have high correlation with whole slide image (WSI) segmentation scores, even though the domain shift is calculated on a small subset of the data. Except for pre-training the segmentation model in DSM, both of these approaches measure domain shift in an unsupervised manner.

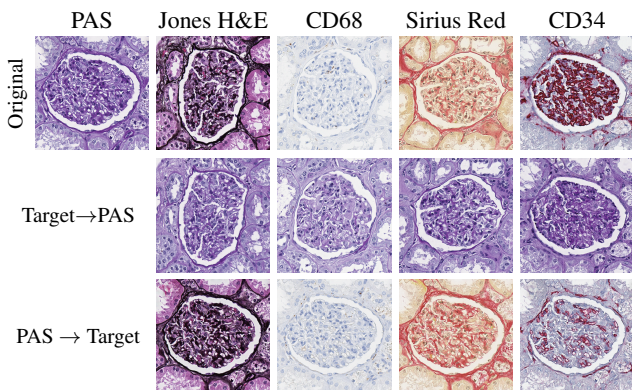
The rest of this paper is organised as follows: Section 2 reviews the literature on stain translation and how it can introduce domain shift, Section 3 details the methods used. Section 4 presents the dataset, experiments, and results. Finally, Section 5 concludes the paper with possible future directions.

## 2. LITERATURE REVIEW

Though several approaches [6]–[9] deal with the problem of intra-stain variation, few address the problem of inter-stain variation [1], [10], [11]. Stain colour augmentation, stain normalisation, and stain transfer [12] are the current standard approaches to learn stain invariant representations. The term stain invariant indicates the ability of the same model to be applied across multiple stains (possibly to those that are not used during training). In general, it is assumed that annotations for the same task are available for the source stain but

Training Strategy	Test Stain				
	PAS	Jones H&E	CD68	Sirius Red	CD34
Baseline PAS (Full slide)	0.907 (0.009)	0.084 (0.033)	0.001 (0.001)	0.016 (0.018)	0.070 (0.063)
MDS1 Target→PAS (Full slide)	- (-)	0.849 (0.017)	0.683 (0.043)	0.870 (0.009)	0.754 (0.008)

**Table 1:** Segmentation scores ( $F_1$ ) of the U-Net (trained on PAS) applied to full test slides of different stains (1<sup>st</sup> row) and translated (Target→PAS) slides (2<sup>nd</sup> row). Averages of 5 U-Net repetitions applied to 3 Cycle-GAN repetitions, i.e. 15 repetitions in total, standard deviations are in parentheses.



**Fig. 1:** Stain translations using CycleGAN.

not for the target stains since acquiring labels for each staining is expensive and laborious. Tellez *et al.* [13] state that stain colour augmentation has a greater influence on the robustness of DL methods than stain normalisation. While stain transfer (a technique for virtual staining) tackles the generic problem of lack of annotations in the medical domain [14], [15] and can be applicable to various related scenarios. Recently Gadermayr *et al.* [11] propose to use an unpaired adversarial image-to-image translation approach called CycleGAN [16] to overcome the lack of annotations for the target stain by: 1) training a segmentation model on the source stain and apply it to the target stain translated to the source stain, named as MultiDomain Supervised 1 (MDS1); 2) translate the source to target stain and directly train the segmentation model for the target stain, named as MultiDomain Supervised 2 (MDS2). Vasiljević *et al.* [1] extend this to create UDA-GAN, a stain augmentation procedure that uses multiple target stains to learn a stain invariant representation for segmentation, which can even be applied to out-of-sample stains.

Although visually these unpaired translations look very realistic, see Fig. 1 (2<sup>nd</sup> and 3<sup>rd</sup> row), in accordance with recent advances in the theoretical understanding of CycleGANs

[17], when translating from an information rich domain to an information poor domain some hidden information is embedded within them as imperceptible noise [1], [18]. This can cause domain shift in the translated stains that can affect the final predictions, see Table 1 (2<sup>nd</sup> row: CD68 → PAS and CD34 → PAS), since a majority of state-of-the-art computer vision algorithms are vulnerable to domain shift [4].

### 3. METHODS

#### 3.1. PixelCNN

Song *et al.* [19] have shown that PixelCNNs can be used to detect adversarial attacks in images, and we hypothesise that the hidden information may be detectable in a similar manner.

The PixelCNN [5] is a generative model built specifically for images and to have tractable likelihood calculation. The model quantifies the pixels of an image  $x$  over all its sub-pixels as a product of conditional distributions, such that

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

These conditional distributions are parameterised by a CNN and hence shared across all pixels in the image. PixelCNN is used to model the underlying data distribution of the source (i.e. PAS) and the translated Target→PAS stains.

#### 3.2. Domain Shift Metric

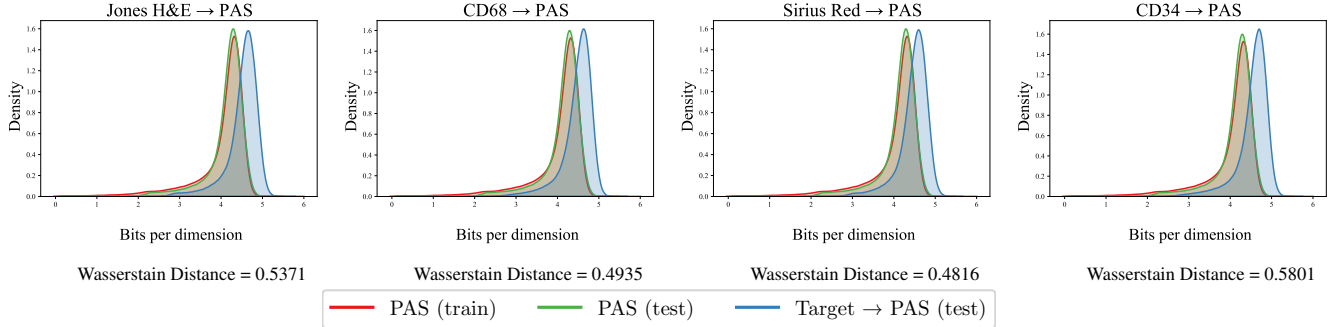
The Domain Shift Metric (DSM) [6] measures the difference between two domains' distributions using the feature representations of a pre-trained model (referred to herein as Domain Shift Scores or DSS). Consider a CNN with layers  $\{l_1, \dots, l_L\}$ . Let  $\Phi(x) = \{\phi_{l1}(x), \dots, \phi_{lk}(x)\}$  such that  $\Phi_{lk}(x) \in \{\mathbb{R}^{h \times w}\}$  denote the filter activations at layer  $l$  and filter  $k$ . The mean value of each  $\Phi_{lk}(x)$  is calculated as

$$c_{lk}(x) = \frac{1}{hw} \sum_{i,j} \Phi_{lk}(x)_{i,j}. \quad (2)$$

Let  $p_{c_{lk}}^S(x)$  denotes continuous distribution of  $c_{lk}(x)$  over the source stain  $\mathcal{S}$  and  $p_{c_{lk}}^T(x)$  denotes the same over the translated Target→PAS stain  $\mathcal{T}$ , then the DSM  $R_l$  is defined as

$$R_l(p^S, p^T) = \frac{1}{k} \sum_{i=1}^k \mathcal{D}(p_{c_{lk}}^S, p_{c_{lk}}^T), \quad (3)$$

where  $\mathcal{D}$  is the Wasserstein distance [20] between  $p_{c_{lk}}^S(x)$  and  $p_{c_{lk}}^T(x)$ , which tends towards zero when  $\mathcal{S}$  and  $\mathcal{T}$  are similar.



**Fig. 2:** PixelCNN based domain shift measures of the translated Target→PAS stains w.r.t. real PAS train and test sets.

## 4. EXPERIMENTS AND RESULTS

### 4.1. Data

Tissue samples were obtained from a group of 10 patients who had tumor nephrectomy for renal carcinoma. Renal tissue was chosen to be as far from the tumors as possible to represent largely normal renal glomeruli; however, certain samples contained varying degrees of pathological modifications such as complete or partial displacement of functional tissue by fibrotic changes (“sclerosis”) indicating normal age-related changes or the renal effects of general cardiovascular comorbidity (e.g. cardiac arrhythmia, hypertension, arteriosclerosis). Using an automated staining tool (Ventana Benchmark Ultra), the paraffin-embedded samples were sliced into 3 $\mu$ m thick sections and stained with either Jones H&E basement membrane stain (Jones), PAS, Sirius Red, as well as two immunohistochemistry markers (CD34, and CD68). An Aperio AT2 scanner was used to capture whole slide pictures at 40 magnification (a resolution of 0.253 m / pixel). Pathology specialists annotated and verified all of the glomeruli in each WSI by labeling them with Cytomine [21]. The whole dataset was split 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 - 590 (valid), 1043 (test); Sirius Red - 576 (valid), 1049 (test); CD34 - 595 (valid), 1019 (test); CD68 - 521 (valid), 1046 (test).

### 4.2. Experiments

Throughout the experiments, patches of size 508 $\times$ 508 pixels are used since glomeruli and part of the surrounding area fit within this patch size at the level-of-detail used. To account for random variations, all experiments are repeated with three different CycleGAN models and each is used with five different U-Net models, i.e. 15 total repetitions. When training (PixelCNN & CycleGAN) 5000 training and 500 validation patches were used, and when evaluating (PixelCNN & DSM) 5 sets of 1000 test patches were used. These were extracted in a random, uniform manner from the corresponding patients.

#### 4.2.1. U-Net

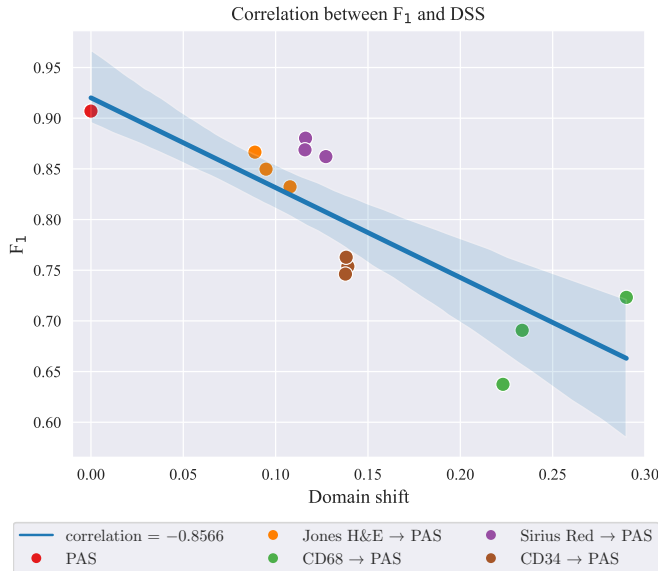
The U-Net [22] architecture is used to segment the glomerulus region in the source staining (PAS), since it has been proven successful in biomedical imaging [23] and, in particular, glomeruli detection [24]. Glomeruli segmentation is framed as a two class problem: glomeruli (pixels that belong to glomerulus), and tissue (pixels outside a glomerulus). The training set comprised all glomeruli from the source stain (PAS) training patients (662) plus 4634 tissue (i.e. non-glomeruli) patches (accounting for the variance in non-glomeruli tissue). The network was trained using the same parameters and procedures as used by Lampert *et al.* [10]. The segmentation scores for each stain are presented in the 1<sup>st</sup> row of Table 1.

#### 4.2.2. CycleGAN

CycleGAN [16] is an unpaired image-to-image translation network widely used for style transfer in digital pathology [1], [11], [18]. Given images of a source stain  $s \sim \mathcal{S}$  and a target stain  $t \sim \mathcal{T}$ , the goal is to learn a two way mapping between  $t$  and  $s$ . This network is used to translate all of the target stains (i.e. Jones H&E, CD68, Sirius Red, and CD34) to the source stain (PAS). Gadermayr *et al.* [11] briefly explain how different sampling strategies for the annotated and unannotated stains can negatively impact a stain transfer model’s performance, and therefore patches are randomly extracted in an unsupervised manner using a uniform sampling strategy. The same training strategies as used by Vasiljević *et al.* [1] was employed for training the CycleGAN networks to translate the target stains to PAS. Fig. 1 (2<sup>nd</sup>, and 3<sup>rd</sup> row) present the results for each of these translations and the second row in Table 2 presents the U-Net segmentation scores when the target stains are translated to PAS and the PAS trained network used to segment them.

#### 4.2.3. PixelCNN

As was shown in Table 1 (2<sup>nd</sup> row), a segmentation model trained on PAS experiences a degradation in performance



**Fig. 3:** Correlation between segmentation scores of the whole test slides translated to PAS and the average DSS of 5 sets of 1000 randomly sampled test patches. Each point is the average of 5 U-Net repetitions for each CycleGAN model.

when applied to translated images. It is our hypothesis that this is caused by the imperceptible noise added during stain style transfer. This noise is imperceptible to humans but causes a domain/covariate shift. To test this, the probability density of the underlying source (PAS) training distribution is estimated using a PixelCNN model.

As each pixel value is conditioned on the product of all previously generated pixels, the original architectures were trained and evaluated on patches of size  $32 \times 32$  due to GPU memory limitations. Therefore 5000 (training & test) and 500 (validation) patches were decomposed into 1,280,000 non-overlapping training and test and 128,000 validation patches. The same training parameters as used by Salimans *et al.* [5]<sup>1</sup> were employed. The training of one PixelCNN model took approximately 15 days using four V100 GPUs (in parallel).

The PixelCNN model is first validated using the PAS training data and an unseen PAS test set. It is found that their log-likelihoods follow the same order of magnitudes, see Fig. 2. The log-likelihood distributions of the Target→PAS stains are also included in this figure and they clearly show that there is a domain shift compared to the overlapping test distributions. The Wasserstein distance can be used to measure the similarity of two distributions, where smaller distances indicate more similar distributions. This is the case for the train (PAS) and test (PAS) distributions, giving a Wasserstein distance of 0.0879 (average over 5 sets of 1000 randomly sampled patches). In comparison, the distance between PAS train and Target→PAS for all stains is relatively large, see Fig.

<sup>1</sup><https://github.com/openai/pixel-cnn>

Jones H&E →PAS	CD68→PAS	Sirius Red→PAS	CD34→PAS
0.097 (0.008)	0.248 (0.002)	0.119 (0.003)	0.138 (0.002)

**Table 2:** Average Domain Shift Scores ( $R_l$ ) of 5 sets of 1000 randomly sampled patches for the Target→PAS translated stains. Averages of 5 U-Net and 3 CycleGAN repetitions, i.e. 15 repetitions in total; standard deviations are in parentheses.

2, highlighting a greater difference between the distributions.

There is a strong correlation,  $-0.7339$ , between full slide segmentation scores (Table 1: 2<sup>nd</sup> row) and distribution distance. This unsupervised approach can therefore be used for insight into the success of applying the segmentation model to unseen, unannotated data that has undergone domain shift by using a sample of the data. Furthermore, it is also able to detect imperceptible (even to domain experts) domain shifts.

#### 4.2.4. Domain Shift Metric

Using the pre-trained PAS source network as the feature representations, the domain shift can also be calculated using the domain shift metric, Eq. (3). The DSS are presented in Table 2. Since the model is supervised on the same task, the average segmentation score (of 5 models) has a stronger negative correlation than observed with the PixelCNN, as in Fig. 3.

This stronger correlation is observed when compared to the PixelCNN since DSM uses the same representation as the segmentation model, which has been trained in a supervised manner for a specific task. It is therefore sensitive to the type of domain shift that will affect the segmentation performance.

## 5. CONCLUSIONS

This article has investigated unsupervised approaches to propose a method to detect domain shift in histopathological images and shown that domain shift has a strong correlation with the segmentation performance of stain translated data. As such, the work focused on detecting imperceptible noise that is introduced by CycleGAN models, however the solution is general and can detect any kind of domain shift.

These measures offer a mechanism to estimate the average performance of pre-trained neural networks when applied to unseen target stains (for the same task) without having any expert opinion or ground-truth. This has been achieved using two approaches, one that uses an unsupervised, generative model of the data and another that uses a pre-trained (supervised representation). Since the purpose of this work is to predict how domain shift will affect a pre-trained model, this representation would be available, however, if this is not the case, then the completely unsupervised PixelCNN also offers strong correlation with segmentation score.



## 6. ACKNOWLEDGMENTS

This work was supported by: ArtIC project "Artificial Intelligence for Care" (grant ANR-20-THIA-0006-01) and co-funded by Région Grand Est, Inria Nancy - Grand Est, IHU of Strasbourg, University of Strasbourg and University of Haute-Alsace; ERACoSysMed and e:Med initiatives by BMBF; SysMIFTA (project management PTJ, FKZ 031L-0085A; ANR, project number ANR-15-CMED-0004); and the French Government for co-tutelle funding (Jelica Vasiljevic). We thank the Nvidia Corporation, the Centre de Calcul de l'Université de Strasbourg, and GENCI-IDRIS (Grant 2020-A0091011872) for access to the GPUs used for this research. We also thank the Medizinische Hochschule Hannover for providing high-quality images and annotations.

## 7. COMPLIANCE WITH ETHICAL STANDARDS

Study performed in line with the principles of the Declaration of Helsinki. Approval granted by the Ethics Committee of Hannover Medical School (Date 12/07/2015, No. 2968-2015).

## References

- [1] J. Vasiljević *et al.*, "Towards histopathological stain invariance by unsupervised domain augmentation using generative adversarial networks," *Neurocomputing*, vol. 460, pp. 277–291, 2021.
- [2] K. Stacke *et al.*, "Measuring domain shift for deep learning in histopathology," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 2, pp. 325–336, 2020.
- [3] P. Leo *et al.*, "Evaluating stability of histomorphometric features across scanner and staining variations: Prostate cancer diagnosis from whole slide images," *J. Med. Imaging*, vol. 3, no. 4, p. 047 502, 2016.
- [4] G. Csurka, "A comprehensive survey on domain adaptation for visual applications," *Domain adaptation in computer vision applications*, pp. 1–35, 2017.
- [5] T. Salimans *et al.*, "PixelCNN++: Improving the PixelCNN with discretized logistic mixture likelihood and other modifications," *ICLR*, 2017.
- [6] T. de Bel *et al.*, "Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology," in *MIDL*, 2019, pp. 151–163.
- [7] M. W. Lafarge *et al.*, "Learning domain-invariant representations of histological images," *Frontiers in medicine*, vol. 6, p. 162, 2019.
- [8] S. Otálora *et al.*, "Staining invariant features for improving generalization of deep convolutional neural networks in computational pathology," *Front. Bioeng. Biotechnol.*, vol. 7, p. 198, 2019.
- [9] M. T. Shaban *et al.*, "StainGAN: Stain style transfer for digital histological images," in *ISBI*, 2019, pp. 953–956.
- [10] T. Lampert *et al.*, "Strategies for training stain invariant CNNs," in *ISBI*, 2019, pp. 905–909.
- [11] M. Gadermayr *et al.*, "Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: A study on kidney histology," *IEEE Trans Med Imaging*, vol. 38, no. 10, pp. 2293–2302, 2019.
- [12] C. L. Srinidhi *et al.*, "Deep neural network models for computational histopathology: A survey," *Med Image Anal*, p. 101 813, 2020.
- [13] D. Tellez *et al.*, "Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology," *Med Image Anal*, vol. 58, p. 101 544, 2019.
- [14] A. Lahiani *et al.*, "Virtualization of tissue staining in digital pathology using an unsupervised deep learning approach," in *ECDP*, 2019, pp. 47–55.
- [15] C. Mercan *et al.*, "Virtual staining for mitosis detection in breast histopathology," in *ISBI*, 2020, pp. 1770–1774.
- [16] J.-Y. Zhu *et al.*, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232.
- [17] D. Bashkirova *et al.*, "Adversarial self-defense for cycle-consistent GANs," *NeurIPS*, 2019.
- [18] J. Vasiljević *et al.*, "Self adversarial attack as an augmentation method for immunohistochemical stainings," in *ISBI*, 2021, pp. 1939–1943.
- [19] Y. Song *et al.*, "Pixeldefend: Leveraging generative models to understand and defend against adversarial examples," *arXiv preprint arXiv:1710.10766*, 2017.
- [20] A. Ramdas *et al.*, "On Wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, p. 47, 2017.
- [21] R. Marée *et al.*, "Collaborative analysis of multi-gigapixel imaging data using Cytomine," *Bioinformatics*, vol. 32, no. 9, pp. 1395–1401, 2016.
- [22] O. Ronneberger *et al.*, "U-Net: Convolutional networks for biomedical image segmentation," in *MIC-CAI*, 2015, pp. 234–241.
- [23] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med Image Anal*, vol. 42, pp. 60–88, 2017.
- [24] T. de Bel *et al.*, "Automatic segmentation of histopathological slides of renal tissue using deep learning," in *Medical Imaging 2018: Digital Pathology*, vol. 10581, 2018, p. 1 058 112.