Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Original Research Article

Tackling stain variability using CycleGAN-based stain augmentation

Nassim Bouteldja [a,*,1], David L. Hölscher [a,1], Roman D. Bülow [a], Ian S.D. Roberts [c], Rosanna Coppo [d,e], Peter Boor [a,b]

[a] Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany
[b] Department of Nephrology and Immunology, RWTH Aachen University Hospital, Aachen, Germany
[c] Department of Cellular Pathology, Oxford University Hospitals National Health Service Foundation Trust, Oxford, United Kingdom
[d] Fondazione Ricerca Molinette, Torino, Italy
[e] Regina Margherita Children's University Hospital, Torino, Italy

ARTICLE INFO

Keywords:
Stain augmentation
Stain normalization
Digital pathology
Segmentation
Deep learning
Kidney

ABSTRACT

Background: Considerable inter- and intra-laboratory stain variability exists in pathology, representing a challenge in development and application of deep learning (DL) approaches. Since tackling all sources of stain variability with manual annotation is not feasible, we here investigated and compared unsupervised DL approaches to reduce the consequences of stain variability in kidney pathology.

Methods: We aimed to improve the applicability of a pretrained DL segmentation model to 3 external multi-centric cohorts with large stain variability. In contrast to the traditional approach of training generative adversarial networks (GAN) for stain normalization, we here propose to tackle stain variability by data augmentation. We augment the training data of the pretrained model by the stain variability using CycleGANs and then retrain the model on the stain-augmented dataset. We compared the performance of i/ the unmodified pretrained segmentation model with ii/ CycleGAN-based stain normalization, iii/ a feature-preserving modification to ii/ for improved normalization, and iv/ the proposed stain-augmented model.

Results: The proposed stain-augmented model showed highest mean segmentation accuracy in all external cohorts and maintained comparable performance on the training cohort. However, the increase in performance was only marginal compared to the pretrained model. CycleGAN-based stain normalization suffered from encoded imperceptible information into the normalizations that confused the pretrained model and thus resulted in slightly worse performance.

Conclusions: Our findings suggest that stain variability can be tackled more effectively by augmenting data by it than by following the commonly used approach of normalizing the stain. However, the applicability of this approach providing only a rather slight performance increase has to be weighted against an additional carbon footprint.

Introduction

Histological analysis represents the gold-standard for diagnosing many diseases including almost all types of cancer and majority of kidney diseases.[1] The widespread use of digital whole-slide scanners in pathology has paved the way for digital pathology to generate large amounts of digitized highly resolved histological data, i.e. the so-called whole-slide images (WSIs). This allows the use of computational approaches, in particular deep learning (DL) techniques, for automated and efficient image analysis. However, due to the lack of standardization, the process of tissue preparation and digitization involves multiple degrees of variability, e.g. in cutting thicknesses, staining protocols, dye compositions, scanner characteristics, and modalities. This results in various appearances and considerable variability of the same histological stain between laboratories and even within a laboratory. In turn, this impedes computational image analysis resulting in lower performance of computer-aided diagnosis systems (Fig. 1).[2]

Generative adversarial networks (GANs) have recently been proposed to tackle stain variability.[3] Such networks generate synthetic image data. In this work, they were employed to synthesize images to augment the training data. GANs are successfully used in 3 major applications in digital pathology. First, stain normalization,[2,4–9] i.e. reducing color variations within a specific stain. Second, stain translation,[10–14] i.e. converting between different stains. And third, the conversion of different modalities, e.g. histology and fluorescence.[15,16] Most of those approaches employ the cycle-consistent generative adversarial network (CycleGAN).[17] It represents the state-of-the-art for unsupervised domain adaptation based on image-to-image translation, i.e. converting between 2 image domains, e.g. horse and zebra images, by transferring the image style. This is performed

* Corresponding author at: University Hospital Aachen, Pauwelsstr. 16, 52074 Aachen, Germany.
E-mail addresses: nbouteldja@ukaachen.de (N. Bouteldja), dhoelscher@ukaachen.de (D.L. Hölscher), rbuelow@ukaachen.de (R.D. Bülow), ian.roberts@ouh.nhs.uk (I.S.D. Roberts), rosanna.coppo@unito.it (R. Coppo), pboor@ukaachen.de (P. Boor).

[1] Authors contributed equally.

http://dx.doi.org/10.1016/j.jpi.2022.100140
Received 1 July 2022; Received in revised form 4 September 2022; Accepted 7 September 2022
Available online 13 September 2022

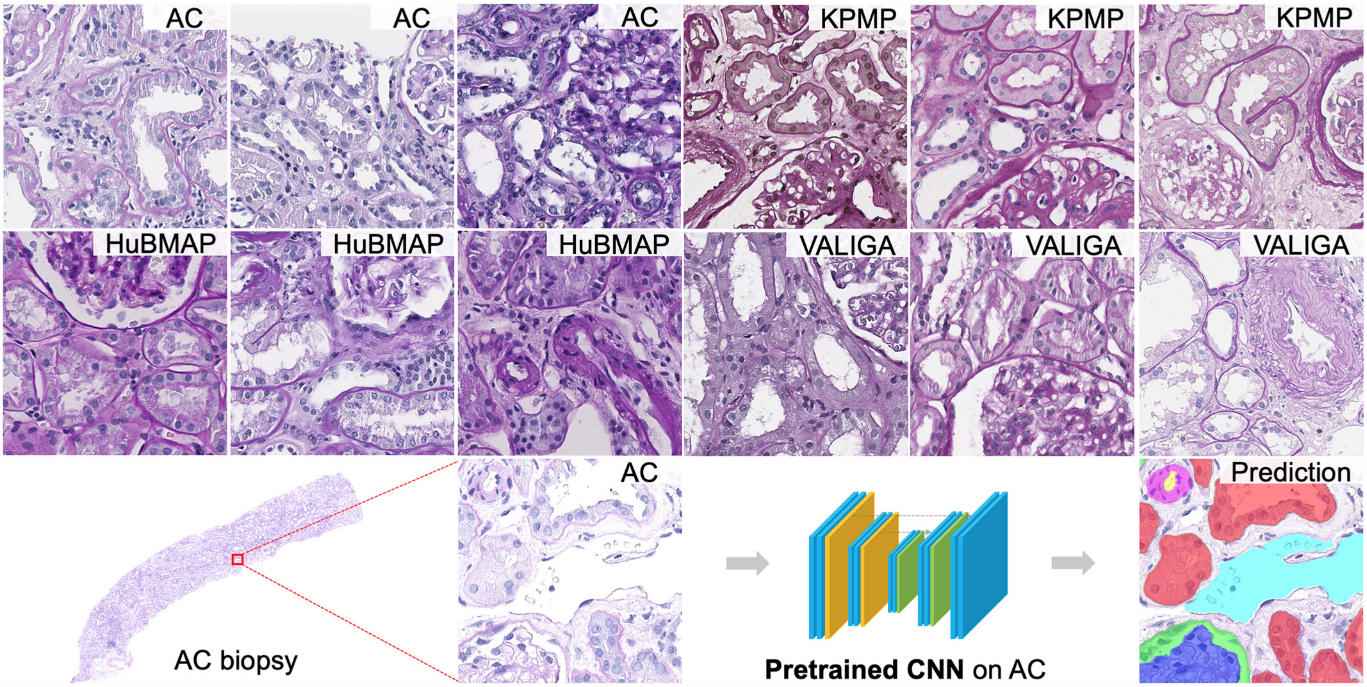


Fig. 1. Overview of stain variations in all cohorts.

in an unsupervised manner, i.e. without the need for matching image pairs to reduce manual overhead. In this study, CycleGANs were used to convert between different stain variations. More specifically, we tackle intra-stain variation to improve the applicability of DL models to external data of the same stain showing distinct color distributions. This represents the major aim of stain normalization, addressing the problem by transferring the color distribution of the training images to the external data.

Related work

Traditional non-DL approaches for stain normalization perform color matrix projections,[18] color deconvolution,[19,20] or match chromatic and density distributions using color and contextual information[21] to finally align the color profile of target WSIs to reference images. However, an exhaustive study[2] examined the effects of stain normalization and conventional (non-DL) data augmentation on classification performance of convolutional neural networks (CNN) in various histopathology tasks. They found that extensive color augmentation during training already outperformed traditional stain normalization approaches on external cohorts, making them redundant.

DL approaches for stain normalization have often shown its superiority over the aforementioned traditional methods in terms of structure similarity or classification performance.[2,4–6,8,22] For instance, hematoxylin-eosin (HE)-stained images from different centers were converted into grayscale and then mapped to the stain style of a particular center using conditional GAN (cGAN)-based frameworks.[5,6] Whereas Salehi et al.[5] trained the cGAN on a single center and conditioned the mapping on the original source image, Cho et al.[6] trained it jointly on all centers and conditioned on the selected center for stain style transfer.

Many studies have employed CycleGANs for stain normalization.[4,8,9] Shaban et al.[4] used them to translate HE data between Aperio and Hamamatsu scanners. Mahapatra et al.[8] added a feature-preserving loss to CycleGAN training for semantic guidance. They penalized the discrepancy of intermediate feature maps of input images and their translations by a pretrained segmentation CNN to preserve structural information. Several works have also included a feature-preserving loss on feature maps or outputs by a prior CNN to improve the image translation.[6,7,10] De Bel et al.[9] trained a CNN for the segmentation of glomerular tufts on one data center (Amsterdam) and applied it to another center (Radboud) using extensive color augmentation during training and CycleGANs for stain normalization. They found improvements in segmentation performance when additionally employing CycleGAN-based stain normalization on top of extensive color augmentations.

However, very little research has been performed to examine the effects of stain normalization on segmentation performance of strongly augmented CNNs.[7,9] Also, both studies[7,9] only analyzed a single structure, i.e. the glomerular tuft, although CycleGAN-based image translation can show highly varying effects on the segmentation of different structures.[10]

Contributions

Our aim is to improve the applicability of an already existing DL segmentation model[23] to external cohorts of the same stain. In contrast to previous studies, we do not follow the classical approach of synthesizing stain-normalized images using GANs, but instead we propose to tackle this task from a data augmentation perspective. We employ CycleGANs to translate images from the training cohort (of the segmentation model) to the external cohorts and use those for data augmentation during training of a new segmentation model (Fig. 2 iv/). To the best of our knowledge, this is the first study to use DL-based data augmentation for improved generalization to external cohorts in digital pathology. We evaluate our proposed approach for the segmentation of 6 major kidney structures on biopsy slides from 3 external centers. We also compare it with 3 approaches including traditional CycleGAN-based stain normalization (Fig. 2 ii/), adding a feature-preserving loss into CycleGAN training (Fig. 2 iii/), and finally using only the already existing segmentation model without stain normalization or the proposed augmentation (Fig. 2 i/).

Methods

In our application scenario, we presume a pretrained segmentation CNN including its underlying annotated training data. We selected a previously developed DL model for the segmentation of kidney structures from human tissue that has been trained on a single-centric cohort *A*.[23] We aim at improving its applicability to external cohorts *E* by retraining it on its stain-augmented training data using CycleGANs.

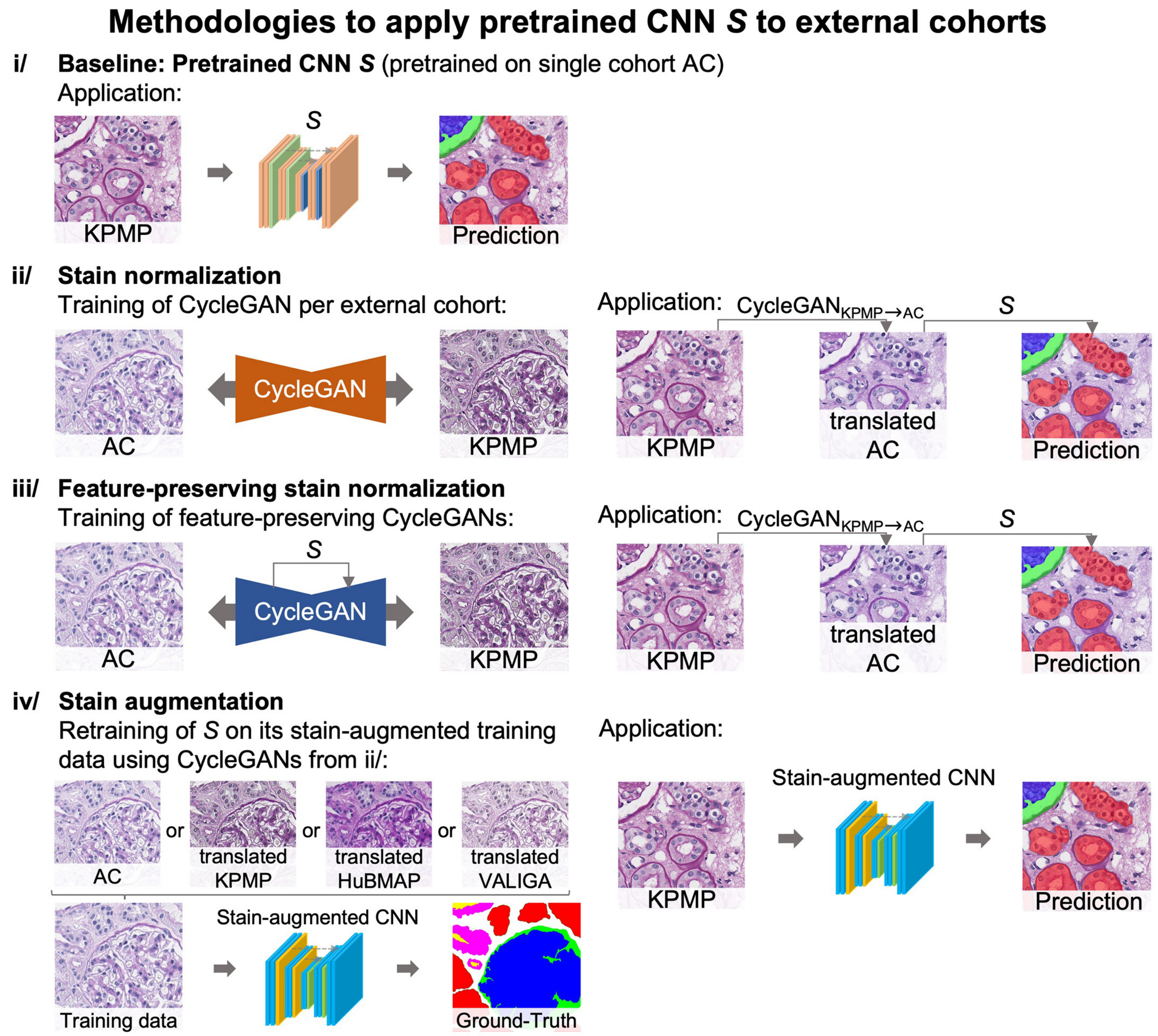

Methodologies to apply pretrained CNN S to external cohorts
i/ Baseline: Pretrained CNN S (pretrained on single cohort AC)
Application:
S
KPMP
Prediction
ii/ Stain normalization
Training of CycleGAN per external cohort:
CycleGAN
AC
KPMP
Application: CycleGAN KPMP→AC
S
KPMP
translated AC
Prediction
iii/ Feature-preserving stain normalization
Training of feature-preserving CycleGANs:
S
CycleGAN
AC
KPMP
Application: CycleGAN KPMP→AC
S
KPMP
translated AC
Prediction
iv/ Stain augmentation
Retraining of S on its stain-augmented training data using CycleGANs from ii/:
AC
or
translated KPMP
or
translated HuBMAP
or
translated VALIGA
Training data
Stain-augmented CNN
Ground-Truth
Application:
KPMP
Stain-augmented CNN
Prediction

Fig. 2. Approaches for improved generalization of the pretrained segmentation model to external cohorts. To improve the pretrained model S (i/), CycleGANs are trained for translation between its training cohort AC and the external cohorts, respectively. Stain normalization then uses the CycleGANs to translate the external domains into the training cohort AC for improved application (ii/ + iii/). In contrast, the proposed stain augmentation augments the annotated training data of S by the external stain variations using the CycleGANs (iv/). Then, a new and cohort-independent segmentation model is trained on the stain-augmented annotated training data.

CycleGANs

The CycleGAN[17] is a widely used approach to train CNNs for unsupervised image style transfer. It consists of 2 generators $g_{A\to E}$, $g_{E\to A}$ and 2 discriminators d_A, d_E that are trained in an unsupervised fashion, i.e. using unpaired images, on the following 3 loss functions: The adversarial loss

$$\begin{aligned}\ell_{adv} = {} & \mathbb{E}_{x\sim p_A(x)}[\log(d_A(x)) + \log(1 - d_E(g_{A\to E}(x)))] \\ & + \mathbb{E}_{y\sim p_E(y)}[\log(d_E(y)) + \log(1 - d_A(g_{E\to A}(y)))]\end{aligned}$$

forces the generators to synthesize realistic image translations between the stain styles A and E by fooling the discriminators, since they are trained to distinguish the translations from real images $x \in A$, $y \in P$. The cycle-consistency loss

$$\ell_{cyc} = \mathbb{E}_{x\sim p_A(x)}[\|g_{E\to A}(g_{A\to E}(x)) - x\|_1] + \mathbb{E}_{y\sim p_E(y)}[\|G_{A\to E}(G_{E\to A}(y)) - y\|_1]$$

incentivizes the generators to synthesize translations with spatially consistent and aligned structures, since mapping back such translations facilitate the input reconstruction. This cycle-based reconstruction loss represents the gist behind CycleGANs. Finally, the identity loss

$$\ell_{idt} = \mathbb{E}_{y\sim p_A(y)}[\|g_{A\to E}(y) - y\|_1] + \mathbb{E}_{x\sim p_A(x)}[\|g_{E\to A}(x) - x\|_1]$$

is used to stabilize training in particular at the beginning as it enforces the generators to learn an identity mapping for images from the target domain to prevent bad early optimizations. It was also shown to improve color preservation.[24]

By adding weights to the loss terms, the overall loss function ℓ is represented by:

$$\ell = \lambda_{adv}\ell_{adv} + \lambda_{cyc}\ell_{cyc} + \lambda_{idt}\ell_{idt}$$

CycleGAN-based stain augmentation

We propose to stain-augment the annotated training data of the pretrained segmentation CNN by translating it to the external cohorts using CycleGANs, and then retrain a segmentation model on the stain-augmented dataset (Fig. 2 iv/). Since this approach brings the various external stain variations into the training dataset, it facilitates learning respective color distributions and thus the generalization across external centers. Since we included 3 external cohorts, 3 CycleGANs were trained for the translation of the trained stain style *A* to the respective external cohort *E*. The segmentation model is then retrained from scratch on the 4 times enlarged annotated training and validation dataset using the same training routine.[23] We compare this approach with CycleGAN-based stain normalization that works the other way around. Here, *E* is now translated into the training domain *A* of the existing segmentation model for application (Fig. 2 ii/ + iii/). Briefly summarized, to tackle data variability, the proposed approach augments data by the variability, while stain normalization normalizes the variability.

Data

Four cohorts showing formalin-fixed, paraffin-embedded, and Periodic Acid-Schiff (PAS)-stained kidney tissue from humans were used. The local cohort from Aachen (AC) consisted of 1009 whole-slide images (WSIs) extracted from the archive of the Institute of Pathology of the RWTH Aachen university clinic. The Kidney Precision Medicine Project (KPMP, accessed on March 15, 2021)[25] cohort, the Human BioMolecular Atlas Program (HuBMAP)[26] cohort, and the international VALIGA cohort[27] each included data from multiple centers comprising 85, 9, and 648 WSIs, respectively. All cohorts showed large stain variability and a large morphological spectrum of kidney diseases (Fig. 1). For the AC cohort, tissue was cut into 1–3 μm thick sections. Both the AC and VALIGA cohorts were digitized by the Aperio AT2 whole-slide scanner (Leica Biosystems, Wetzlar, Germany) with a 40x objective lens. Further detailed information on all cohorts and their underlying trials is provided in [23,25,26]. The KPMP and HuBMAP cohorts are publicly available and accessible (via atlas.kpmp.org/repository and portal.hubmapconsortium.org, respectively). The 2 other cohorts, i.e. AC and VALIGA, are not publicly available due to legal and ethical issues. Access can be provided on a reasonable request and only for non-commercial research purposes. In total, 1751 WSIs were used.

For performance assessment, we used the manually annotated test dataset from Boor et al.[23] It consisted of 240, 198, and 214 image patches of size 174 x 174 μm^2 from 5 randomly selected WSIs from the external KPMP, HuBMAP, and VALIGA cohorts, respectively. The annotations were considered the ground-truth for evaluation and were exhaustively performed using the publicly available and widely used software QuPath[28].

Pretrained segmentation model

As pretrained model, we selected the previously developed segmentation model.[23] Detailed information on this model is provided in Boor et al.[23] Briefly described, this pretrained DL segmentation model has solely been trained on the AC cohort. 2821, 108, and 664 annotated image patches of size 174 × 174 μm^2 from 64, 4, and 17 WSIs were used for training, validation, and testing, respectively. A U-Net-like[29] architecture was trained for the segmentation of 6 kidney structures including tubules, full glomeruli, glomerular tufts, veins & non-tissue background, arteries, and arterial lumina (Fig. 1) from PAS-stained human tissue. RAdam[30] was used as optimizer and the initial learning rate of 0.001 was divided by 3 on validation loss plateus until it fell below 4E-6 for training termination. During training, extensive data augmentation was performed, i.e. spatial transformations including flipping, 90° rotation, elastic, affine, and piecewise affine transformations as well as color transformations including hue and saturation shifting, gamma contrast, and intensity normalization.

Experimental setting

We evaluated 4 approaches in this study for improving the applicability of the pretrained DL segmentation model[23] to 3 external cohorts. First, we only applied the unmodified pretrained model as baseline (Fig. 2 i/), second, we performed the proposed CycleGAN-based stain augmentation (Fig. 2 iv/), third, we used the same CycleGANs for stain normalization (Fig. 2 ii/), and fourth, included a feature-aware loss to the latter for feature-aware stain normalization (Fig. 2 iii/).

For CycleGAN training, we took the whole dataset of 1751 WSIs and excluded all slides from patients of the annotated test data in all 4 cohorts to ensure a patient-level data split. Three CycleGANs were then trained on the remaining slides for the translation between AC and each external cohort. The data preprocessing pipeline started with the application of an already existing segmentation model[23] for the automated detection of kidney tissue in WSIs. This removed interfering, non-kidney structures such as muscles, fat, or connective tissue from the analysis. Detailed information on this model are provided in Boor et al.[23] Briefly summarized, an nn-Unet[31] was trained on the AC cohort and showed high segmentation performance in AC, KPMP, and HuBMAP. We then resampled the detected tissue into 0.337 μm pixel spacing and tessellated it into images of size 640 x 640 pixels to comply with the input requirements of the pretrained segmentation model. The CycleGANs were then trained on the extracted tiles using the same architecture and training routine from Bouteldja et al,[10] which showed promising performance for domain adaptation across stains. In short, U-Net-like generators with a depth of seven and PatchGAN[32] discriminators with a depth of 4 were trained for 300 000 iterations using RAdam on random minibatches of size 3. After 150 000 iterations, the initial learning rate of 10^{-4} began to decrease to zero linearly. The loss terms were equally weighted ($\lambda_{adv} = \lambda_{cyc} = \lambda_{idt} = 1$), and flipping, 90° rotation, and gamma correction were employed for data augmentation. Detailed information on this model are provided in Bouteldja et al.[10]

Regarding the feature-preserving loss for CycleGAN-based stain normalization, we applied the same pipeline from Bouteldja et al[10] that showed improved segmentation performance on CycleGAN-transformed images by a prior CNN. For this, the pretrained CNN was integrated into the CycleGAN by penalizing the discrepancy between its predictions on AC inputs and their reconstructions during training. This modification aims to leverage semantic guidance to preserve relevant features in stain normalized images for an improved applicability of the pretrained CNN.

All experiments were implemented using PyTorch and run on an NVIDIA A100 GPU. CycleGAN training and its feature-preserving variant consumed about 7 and 11 gigabytes of VRAM and took 8 and 19 hours, respectively. Training of the pre-existing segmentation model and its stain-augmented variant required 10 gigabytes of VRAM each and took 19 and 50 hours, respectively. Our source code is publicly available at https://github.com/NBouteldja/KidneyStainAugmentation.

Evaluation

Since the selected segmentation model for this study segmented multiple kidney structures on instance level, we measured quantitative segmentation performance on the external cohorts using instance-level Dice Scores (IDSC)[33] for each class. For a a set of test images $t \in T$ with n_{p_t} binary prediction instances $p_{t,x}$ and n_{g_t} ground-truth instances $g_{t,y}$ indexed by $x = 0, \ldots, n_{p_t}$

and $y = 0, \ldots, n_{g_t}$ for an arbitrary image t, the IDSC is computed over the whole test set T as follows:

$$IDSC = \frac{1}{\sum_{t \in T} n_{p_t} + n_{g_t}} \sum_{t \in T} \left(\sum_{x}^{n_{p_t}} DSC\left(p_{t,i}, g_{t,*}\right) + \sum_{y}^{n_{g_t}} DSC\left(g_{t,j}, p_{t,*}\right) \right)$$

For prediction instance $p_{t,i}$, $g_{t,*}$ represents its maximally overlapping ground-truth instance (0 for false positives), and vice versa for $p_{t,*}$. In contrast to the image-level dice similarity coefficient (DSC), the IDSC weights each prediction and ground-truth instance equally across the whole test set. It ranges between 0 (no single overlap in any test image) and 1 (perfect overlaps in all test images). By averaging the DSC for all prediction and ground-truth instances in T, the IDSC measures the mean area overlap per instance.

For comparison of segmentation performance of all 4 approaches on the external cohorts, we performed Kruskal–Wallis tests followed by pairwise Dunn's post-hoc tests with Bonferroni correction against the proposed stain-augmented model.

We also compared the performance of the pretrained segmentation model and its stain-augmented variant on the internal AC test data comprising 664 annotated patches (Section Pretrained segmentation model) to assess whether the stain-augmented model was able to maintain performance on the training cohort. For this 2-group comparison, we performed Mann–Whitney U tests.

Results

The proposed stain-augmented model (StainAugm) showed high segmentation accuracies in all external cohorts and structures except for the more challenging arteries that were detected with considerably worse performance (Table 1). Performance ranged between 88.8% and 94.6% instance-level Dice Scores for non-arterial structures. It also outperformed all other baseline approaches in most cases. Significant improvement was only found in a few cohorts and only in tubules and arterial structures. Averaged over all structures, the stain augmentation model showed the highest mean performance in all external cohorts.

Interestingly, the pretrained baseline model (Baseline) provided higher mean accuracies in all cohorts when being applied on the raw data rather than its stain-normalized version. It outperformed the stain normalization (StainNorm) in most cases and showed only 2 cases of significantly inferior accuracy compared to the best performing stain-augmented model, i.e. tubules and arteries in HuBMAP. Besides, the inclusion of the feature-preserving loss into CycleGANs (w/ SegNet) improved segmentation performance on stain-normalized images in the majority of cases compared to their unmodified version.

Regarding performance on AC, i.e. the cohort used for trained the pretrained segmentation model, the stain-augmented model demonstrated somewhat comparable accuracy compared to the baseline model. Only the slightly worse predicted tubules showed statistically significant differences. Despite the stain variability across all cohorts, the pretrained baseline model still showed somewhat comparable performances in the external cohorts compared to AC. Except for HuBMAP, only arterial structures were predicted considerably worse. All other structures even showed improved accuracies in most cases. In contrast, the stain-augmented model demonstrated comparable mean performance in HuBMAP and considerably higher mean accuracies in KPMP and VALIGA compared to AC.

Qualitative segmentation results of the stain-augmented model demonstrated high accuracies in selected test images from the external cohorts across the stain variations (Fig. 3). Prediction results of all approaches in selected images from HuBMAP and VALIGA are depicted in Fig. 4. Both the baseline model and its stain-augmented variant showed high accuracy. Regarding stain normalization, realistic and fitting image transformations were demonstrated. However, the application of the pretrained baseline model on those normalized images showed shortcomings. In the HuBMAP image, the upper left artery was confused with a tubule (red arrow). The inclusion of the feature-preserving loss normalized that artery in a hardly distinguishable way that could be, in turn, detected by the baseline model. Similar results were observed for the VALIGA image. Also here, tubules and the glomerulus were normalized in a realistic way, but could only be partly or hardly detected. The feature-preserving loss improved the normalization in this regard but still showed shortcomings.

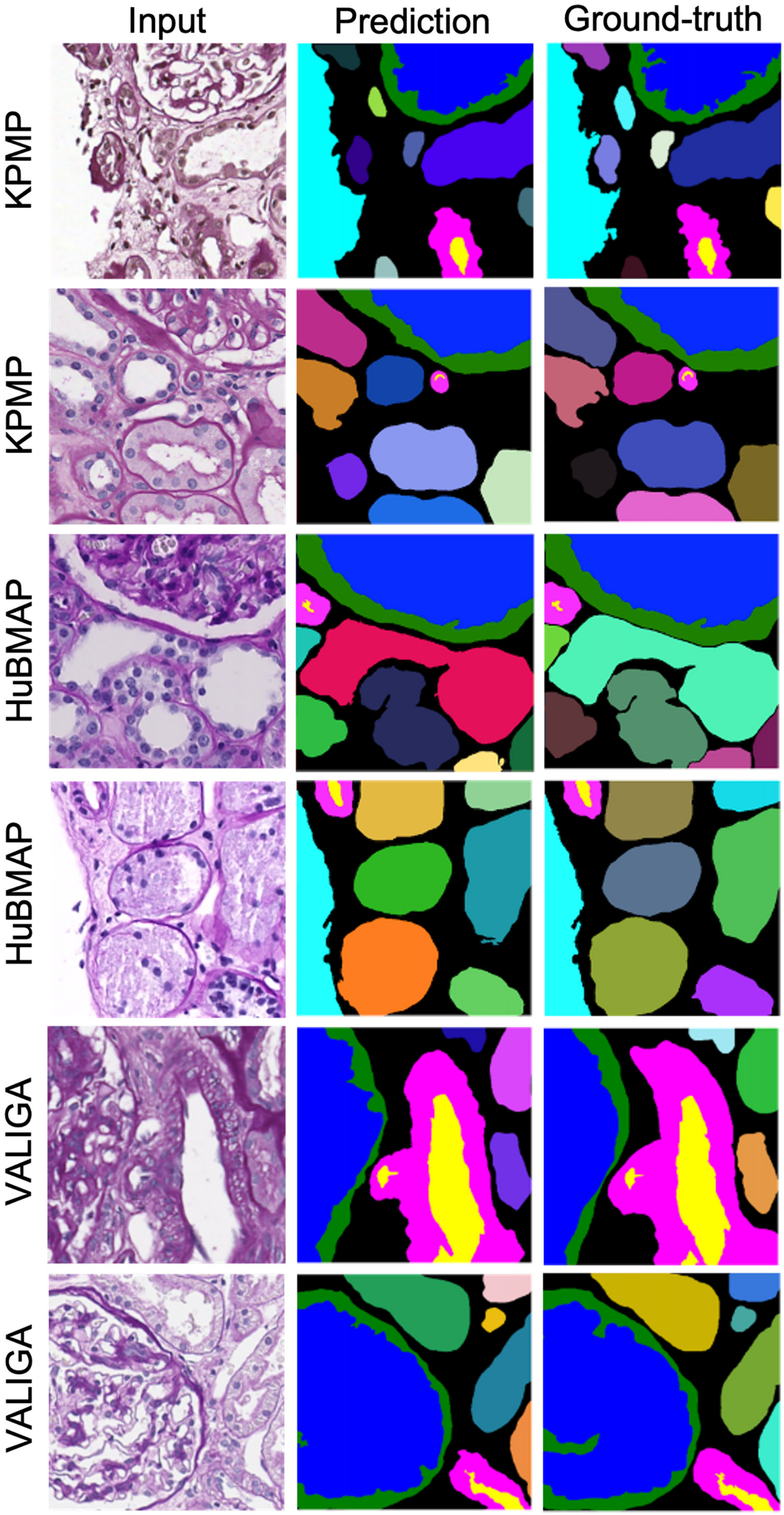
Discussion

In this study, we investigated different unsupervised approaches to improve the generalization of a pretrained segmentation CNN to external cohorts showing distinct stain variations. We proposed to augment the training data of the CNN by the stain variations using CycleGANs, and

Table 1
Segmentation performance in all cohorts.

KPMP	Classes						∅
	Full glomerulus	Glomerular tuft	Tubule	Artery	Arterial lumen	Vein + background	
Baseline	**94.1**	**94.1**	91.1	64.3	59.2	93.1	82.6
StainNorm	93.3	93.2	89.8*	59.5*	58.2	92.2	81.0
w/ SegNet	93.3	93.5	90.1*	58.7*	55.0	92.5	80.5
StainAugm	93.8	93.3	**91.7**	**65.2**	**61.4**	**93.8**	**83.2**
HuBMAP							
Baseline	92.2	93.5	89.2*	70.0*	70.8	89.5	84.2
StainNorm	93.9	93.4	88.8*	67.1*	67.4*	92.0	83.8
w/ SegNet	**94.6**	92.7	89.5	71.5	69.2	92.5	85.0
StainAugm	94.3	**94.3**	**90.0**	**73.2**	**73.8**	**93.2**	**86.5**
VALIGA							
Baseline	93.4	87.9	88.4	62.4	**63.6**	88.4	80.7
StainNorm	93.2	86.6	86.5*	61.3	62.9	**89.2**	80.0
w/ SegNet	**95.1**	88.6	87.5	62.4	60.3	87.1	80.2
StainAugm	94.6	**90.1**	**88.8**	**67.3**	61.8	**89.2**	**82.0**
AC				*Internal training cohort of pretrained segmentation model*			
Baseline	**92.6**	88.5	**88.5***	**70.6**	**66.8**	92.0	**83.2**
StainAugm	92.5	**88.7**	87.8	69.7	63.0	**92.4**	82.4

Segmentation performance was quantified using instance-level Dice scores (%). The pretrained segmentation model (Baseline), the stain normalization approach (StainNorm), and its feature-preserving modification (w/ SegNet) were compared against the proposed stain augmentation (StainAugm) using Bonferroni-Dunn's tests (statistical significance: *$p < 0.05$). The baseline was also compared to StainAugm on its training cohort AC using Mann–Whitney U tests (*$p < 0.05$). Highest performance per class and cohort is in bold, and mean accuracies per row are shown in the rightmost column.



(caption on next page)

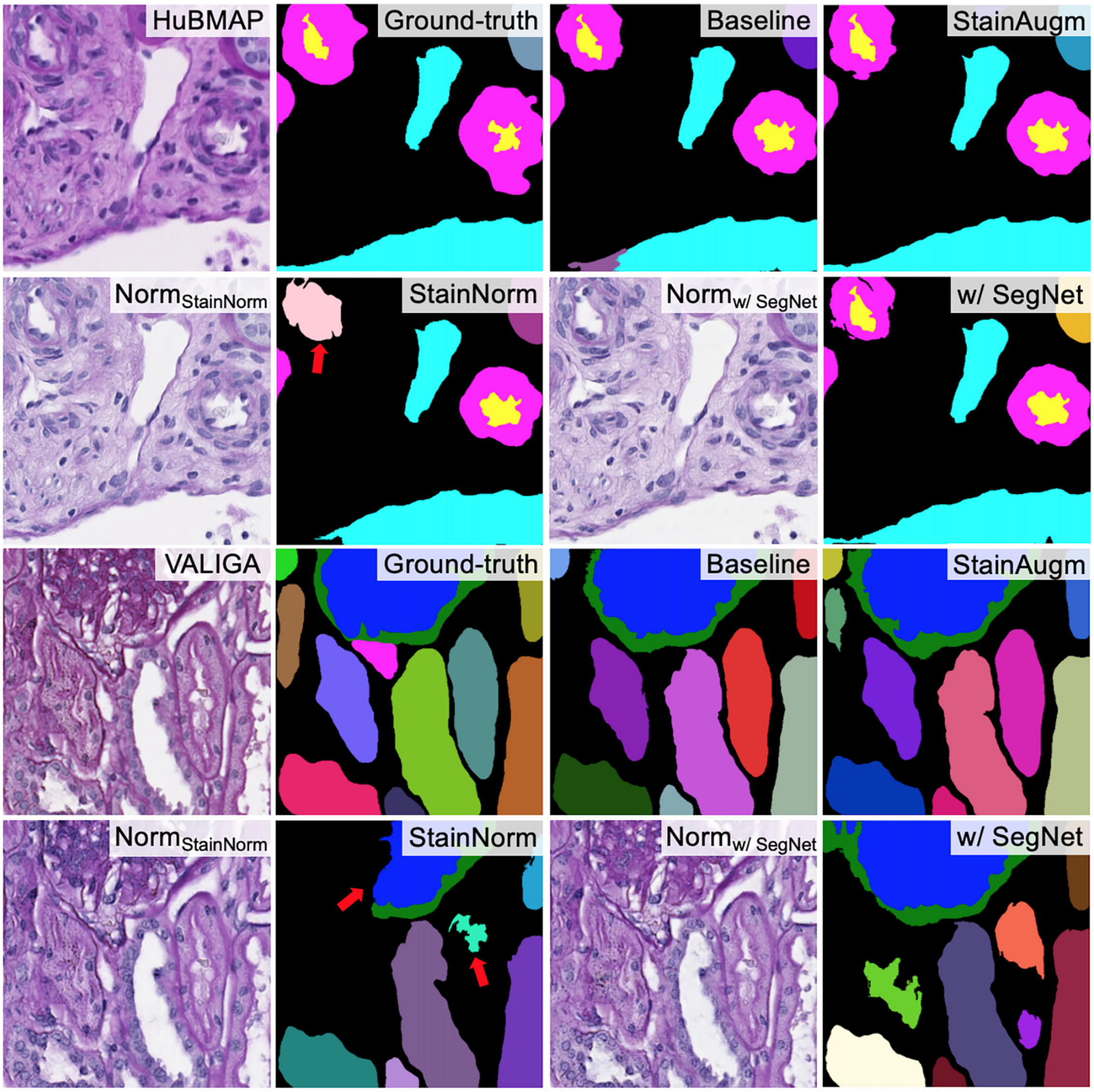

HuBMAP
Ground-truth
Baseline
StainAugm
$\text{Norm}_{\text{StainNorm}}$
StainNorm
$\text{Norm}_{\text{w/ SegNet}}$
w/ SegNet
VALIGA
Ground-truth
Baseline
StainAugm
$\text{Norm}_{\text{StainNorm}}$
StainNorm
$\text{Norm}_{\text{w/ SegNet}}$
w/ SegNet

Fig. 4. Qualitative normalization and segmentation results of all approaches. The pretrained segmentation model is denoted by “Baseline”, the proposed stain augmentation by “StainAugm”, CycleGAN-based stain normalization by “StainNorm”, and its feature-preserving modification by “w/ SegNet”. The stain-normalized translations by the latter 2 approaches are denoted by “$\text{Norm}_{\text{StainNorm}}$” and “$\text{Norm}_{\text{w/ SegNet}}$”, respectively. Interfering information got encoded into the stain-normalized images that confused the pretrained segmentation model and led to erroneous predictions (a few marked by red arrows).

compared this with traditional CycleGAN-based stain normalization, a feature-preserving variant, and simply the pretrained CNN itself.

By stain-augmenting the pretrained CNN, highest mean accuracies in all external cohorts were achieved while maintaining performance on the internal training cohort. The proposed model showed comparable mean performance once and even higher mean accuracies twice in the external cohorts compared to the internal training cohort. In contrast, the pretrained baseline model demonstrated slightly improved mean performance once and decreased accuracies twice in the external cohorts. This contrast reveals that the proposed stain-augmentation better prevents overfitting on the training cohort and improves generalization to external cohorts. The result that the stain-augmented model outperformed the pretrained baseline CNN in all external cohorts, demonstrated that the external stain variations affected performance of the baseline model and that its conventional data augmentation pipeline was not sufficient to fully generalize to the stain variations. For HuBMAP, mean performance difference was largest indicating that certain stain variabilities were less well captured by the conventional data augmentation pipeline.

Performance comparison between the cohorts is challenging, since it is not possible to assess whether performance changes resulted from the stain variability or from different cohort-specific data characteristics. It should be noted that all cohorts show different pathologies with varying occurences and degrees of severity that likewise affect specific structures and thus make them hard to detect. E.g., the pretrained CNN demonstrated improved performance in the external HuBMAP compared to its training cohort AC. This likely means that the model generalized well to the stain variations in HuBMAP and showed higher accuracies due to easier-to-segment structures, e.g. showing less severe pathologies. Hence, for HuBMAP, pathological alterations of structures were the determining factor for performance.

← **Fig. 3.** Qualitative segmentation results in all external cohorts. For each external, multi-centric cohort, images with 2 representative stain variations are selected (column 1). Their predictions by the proposed stain-augmented model are shown (column 2). They are colored in accordance with Fig. 1, however tubules are colored randomly here to assess the feasibility of their instance separation.

Stain normalization was outperformed by the pretrained model in almost all cases. This suggested, that stain normalization is in fact counterproductive, at least for the particular use case of DL-based kidney tissue segmentation. It encoded visually imperceptible information into the stain-normalized images that confused or prevented the detection of structures. Along with adversarial examples,[34] this illustrates how vulnerable CNNs can be to even imperceptible image changes. The inclusion of a feature-preserving loss for stain normalization helped alleviate those predictability issues and showed improved accuracies in most cases compared to the unmodified CycleGAN. However, performance was still slightly inferior to the baseline model in the majority of cases.

Unexpectedly, the baseline model showed high segmentation accuracies in all external cohorts depite their different stain variations compared to AC. It outperformed the stain normalization approaches and provided only slightly inferior accuracy compared to the best performing stain augmentation in all external cohorts. In contrast to the latter, it did not require additional ressources such as training of CycleGANs and another segmentation model on an enlarged dataset. Considering the growing attention to climate change issues in the artificial intelligence community and the vast emissions of CO_2 generated by DL models,[35] it is to be questioned whether the reported performance gains justify the larger carbon footprint. Thus, and in line with related work,[2] this study also recommends the application of extensive conventional data augmentation to DL applications in computational pathology.

The methodology used in this study can be more generally applied to making DL models applicable to external target domains in an uninformed fashion,[36] i.e. annotations are only available for the source domain. Our study supports training a domain-augmented model on real source and simulated target data and applying it to real target data, rather than training a model only on the real source domain and applying it to simulated target-to-source domain translations.

Conclusions

This work represents the first investigation of DL-based data augmentation to improve the generalization of a CNN to external cohorts in computational pathology. In contrast to traditional approaches that normalize external stain variability to the training cohort, we proposed augmentation of the training cohort with the external stain variability using CycleGANs, then retraining the CNN on the stain variability-enriched data.

The stain-augmented model outperformed all baseline approaches in all external cohorts, but showed only slightly improved performance compared to the pretrained and strongly augmented CNN. Both evaluated stain normalization approaches were counter-productive as the pretrained CNN showed slightly higher accuracy in the raw external cohorts rather than their normalized versions. For improved and cost-effective generalization to external image domains, our study generally suggests to train models on domain-augmented training data rather than translate those domains into the training cohort for application. This study also underlines that, in computational pathology, extensive data augmentation can already provide highly generalizable models, reducing the need for the aforementioned approaches, and thus saving computational ressources and CO_2 emissions.

Authors' contributions

NB, DLH and PB planned and oversaw the study. NB planned and conducted experiments. NB, DLH and RDB performed annotations. NB performed statistical analyses. DLH, RDB, ISR, RC and PB aquired data. NB wrote the first draft of the manuscript and arranged figures. PB critically reviewed the manuscript and figures. All authors read and approved the final version of the article.

Disclosure

The authors declare that there is nothing to disclose.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was funded by the German Research Foundation (DFG) (project IDs: 322900939, 454024652, 432698239, 445703531), the European Research Council (ERC) (Consolidator Grant AIM.imaging.CKD, No. 101001791), the German Federal Ministries of Education and Research (BMBF: STOPFSGS-01GM1901A), Health (DEEP LIVER, ZMVI1-2520DAT111) and Economic Affairs and Energy (EMPAIA) and the START-Program of the Faculty of Medicine of the RWTH Aachen University (Grant No. 148/21 to RDB). We are thankful to the NIH Human BioMolecular Atlas Program (HuBMAP) and the Kidney Precision Medicine Project (KPMP) for data sharing. We also thank the International IgA Nephropathy Network for access to the VALIGA dataset.

References

1. Gurcan MN, Boucheron LE, Can A, Madabhushi A, Rajpoot NM, Yener B. Histopathological image analysis: a review. IEEE Rev Biomed Eng 2009;2:147–171.
2. Tellez D, Litjens G, Bandi P, et al. *Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology*. 2019.
3. Tschuchnig ME, Oostingh GJ, Gadermayr M. Generative adversarial networks in digital pathology: a survey on trends and future potential. Patterns (N Y) 2020;1, 100089.
4. Shaban MT, Baur C, Navab N, Albarqouni S. *Staingan: Stain Style Transfer for Digital Histological Images*. 2019.
5. Salehi P, Chalechale A. *Pix2Pix-based Stain-to-Stain Translation: A Solution for Robust Stain Normalization in Histopathology Images Analysis*. 2020.
6. Cho H, Lim S, Choi G, Min H. *Neural Stain-Style Transfer Learning using GAN for Histopathological Images*. 2017.
7. Nishar H, Chavanke N, Singhal N. *Histopathological Stain Transfer Using Style Transfer Network with Adversarial Loss*. Springer International Publishing. 2020.
8. Mahapatra D, Bozorgtabar B, Thiran J-P, Shao L. *Structure Preserving Stain Normalization of Histopathology Images Using Self Supervised Semantic Guidance*. Springer International Publishing. 2020.
9. Td Bel, Hermsen M, Kers J, Laak Jvd, Litjens G. Stain-transforming cycle-consistent generative adversarial networks for improved segmentation of renal histopathology. In: *Cardoso MJ, Aasa F, Ben G, et al, eds.* Proceedings of The 2nd International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, PMLR; 2019. p. 151–163.
10. Bouteldja N, Klinkhammer BM, Schlaich T, Boor P, Merhof D. Improving unsupervised stain-to-stain translation using self-supervision and meta-learning. J Pathol Inform 2022;13:100–107.
11. Gadermayr M, Appel V, Klinkhammer B, Boor P, Merhof D. *Which Way Round? A Study on the Performance of Stain-Translation for Segmenting Arbitrarily Dyed Histological Images*. 2018.
12. Gadermayr M, Gupta L, Appel V, Boor P, Klinkhammer BM, Merhof D. Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: a study on kidney histology. IEEE Trans Med Imaging 2019;38:2293–2302.
13. Lo Y-C, Chung IF, Guo S-N, Wen M-C, Juang C-F. Cycle-consistent GAN-based stain translation of renal pathology images with glomerulus detection application. Appl Soft Comput 2021;98, 106822.
14. Xu Z, Fernández Moro C, Bozóky B, Zhang Q. *GAN-based Virtual Re-Staining: A Promising Solution for Whole Slide Image Analysis*. 2019.
15. Burlingame EA, McDonnell M, Schau GF, et al. SHIFT: speedy histological-to-immunofluorescent translation of a tumor signature enabled by deep learning. Sci Rep 2020;10: 17507.
16. Rivenson Y, Wang H, Wei Z, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. Nat Biomed Eng 2019;3:466–477.
17. Zhu J, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. 2017 IEEE International Conference on Computer Vision (ICCV) 2017:2242–2251.
18. Reinhard E, Adhikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Comput Graphics Appl 2001;21:34–41.
19. Macenko M, Niethammer M, Marron JS, et al. *A Method for Normalizing Histology Slides for Quantitative Analysis*. 2009.

20. Vahadane A, Peng T, Sethi A, et al. Structure-preserving color normalization and sparse stain separation for histological images. IEEE Trans Med Imaging 2016;35:1962–1971.
21. Bejnordi BE, Litjens G, Timofeeva N, et al. Stain specific standardization of whole-slide histopathological images. IEEE Trans Med Imaging 2016;35:404–415.
22. Bug D, Schneider S, Grote A, et al. *Context-Based Normalization of Histological Stains Using Deep Convolutional Features*. 2017.
23. Boor P, Hoelscher D, Bouteldja N, et al. *Next-Generation Morphometry for pathomics-data mining in histopathology*. PREPRINT (Version 1) available at Research Square. 2022. https://doi.org/10.21203/rs.3.rs-1609168/v1.
24. Taigman Y, Polyak A, Wolf L. *Unsupervised Cross-Domain Image Generation*. 2016.
25. de Boer IH, Alpers CE, Azeloglu EU, et al. Rationale and design of the Kidney Precision Medicine Project. Kidney Int 2021;99:498–510.
26. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. Nature 2019;574:187–192.
27. Coppo R, Troyanov S, Bellur S, et al. Validation of the Oxford classification of IgA nephropathy in cohorts with different presentations and treatments. Kidney Int 2014;86:828–836.
28. Bankhead P, Loughrey MB, Fernández JA, et al. QuPath: open source software for digital pathology image analysis. Sci Rep 2017;7:16878.
29. Ronneberger O, Fischer P, Brox T. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Springer International Publishing. 2015.
30. Liu L, Jiang H, He P, et al. *On the Variance of the Adaptive Learning Rate and Beyond*. 2019.
31. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203–211.
32. Isola P, Zhu J-Y, Zhou T, Efros A. *Image-to-Image Translation with Conditional Adversarial Networks*. 2017.
33. Bouteldja N, Klinkhammer BM, Bülow RD, et al. Deep learning–based segmentation and quantification in experimental kidney histopathology. J Am Soc Nephrol 2021;32:52–68.
34. Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv 14126572 2014:11.
35. Strubell E, Ganesh A, McCallum A. Energy and policy considerations for deep learning in NLP. arXiv preprint arXiv:190602243 2019:6.
36. Cook D, Feuz KD, Krishnan NC. Transfer learning for activity recognition: a survey. Knowled Inform Syst 2013;36:537–556.