Contents lists available at ScienceDirect

Journal of Pathology Informatics

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

Unsupervised many-to-many stain translation for histological image augmentation to improve classification accuracy

Maryam Berijanian [a,g], Nadine S. Schaadt [b], Boqiang Huang [c], Johannes Lotz [d], Friedrich Feuerhake [b,e], Dorit Merhof [c,f,*]

[a] *Department of Computational Mathematics, Science and Engineering (CMSE), Michigan State University, East Lansing, USA*
[b] *Institute for Pathology, Hannover Medical School, Hannover, Germany*
[c] *Institute of Image Analysis and Computer Vision, Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany*
[d] *Fraunhofer Institute for Digital Medicine MEVIS, Lübeck, Germany*
[e] *Institute for Neuropathology, University Clinic Freiburg, Freiburg, Germany*
[f] *Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany*
[g] *Institute of Imaging and Computer Vision, RWTH Aachen University, Aachen, Germany*

ARTICLE INFO

Keywords::
Deep learning
Stain translation
Unsupervised
Digital pathology
Breast cancer classifier
Adversarial networks

ABSTRACT

Background: Deep learning tasks, which require large numbers of images, are widely applied in digital pathology. This poses challenges especially for supervised tasks since manual image annotation is an expensive and laborious process. This situation deteriorates even more in the case of a large variability of images. Coping with this problem requires methods such as image augmentation and synthetic image generation. In this regard, unsupervised stain translation via GANs has gained much attention recently, but a separate network must be trained for each pair of source and target domains. This work enables unsupervised many-to-many translation of histopathological stains with a single network while seeking to maintain the shape and structure of the tissues.

Methods: StarGAN-v2 is adapted for unsupervised many-to-many stain translation of histopathology images of breast tissues. An edge detector is incorporated to motivate the network to maintain the shape and structure of the tissues and to have an edge-preserving translation. Additionally, a subjective test is conducted on medical and technical experts in the field of digital pathology to evaluate the quality of generated images and to verify that they are indistinguishable from real images. As a proof of concept, breast cancer classifiers are trained with and without the generated images to quantify the effect of image augmentation using the synthetized images on classification accuracy.

Results: The results show that adding an edge detector helps to improve the quality of translated images and to preserve the general structure of tissues. Quality control and subjective tests on our medical and technical experts show that the real and artificial images cannot be distinguished, thereby confirming that the synthetic images are technically plausible. Moreover, this research shows that, by augmenting the training dataset with the outputs of the proposed stain translation method, the accuracy of breast cancer classifier with ResNet-50 and VGG-16 improves by 8.0% and 9.3%, respectively.

Conclusions: This research indicates that a translation from an arbitrary source stain to other stains can be performed effectively within the proposed framework. The generated images are realistic and could be employed to train deep neural networks to improve their performance and cope with the problem of insufficient numbers of annotated images.

Background

Histological tissues are extensively examined for research, education, and diagnostic purposes. Due to the rising applicability of digital whole slide scanners, the field of digital pathology has been growing recently and moving toward automatic workflows. Large numbers of high-resolution whole slide images (WSIs) are stored for computational analyses to extract patterns, features, and quantitative information in general. Deep learning and neural networks are widely adopted in digital pathology for segmentation, detection, annotation, registration, and classification of WSIs.[1]

A major challenge in the field of digital pathology is the large variability of WSIs due to the differences in cutting thickness, staining process, intra- and inter-subject variabilities, and scanner characteristics. It is not feasible to train neural networks separately for each type of variation, since a large

[*] Corresponding author at: University of Regensburg, 93040 Regensburg, Germany.
E-mail addresses: Berijani@msu.edu (M. Berijanian), Schaadt.Nadine@mh-hannover.de (N.S. Schaadt), Boqiang.Huang@informatik.uni-regensburg.de (B. Huang), Johannes.Lotz@mevis.fraunhofer.de (J. Lotz), Feuerhake.Friedrich@mh-hannover.de (F. Feuerhake), Dorit.Merhof@informatik.uni-regensburg.de (D. Merhof).

http://dx.doi.org/10.1016/j.jpi.2023.100195
Received 15 September 2022; Received in revised form 19 January 2023; Accepted 20 January 2023
Available online 25 January 2023

number of annotated images would be required for each supervised training, which incurs high manual efforts and costs.[2]

Solutions that deal with the variability of WSIs, especially regarding staining variabilities, are image augmentation, stain normalization, and stain translation. Recent research also focuses on increasing the level of automation in the workflow and proposes semi- or unsupervised learning approaches to exploit un-annotated WSIs which are available in large numbers in medical units.[2,3]

Generative adversarial networks (GANs) have been recently employed for image augmentation and translation. An interesting point about GANs is that image pairs are not required for training. Highly realistic images, although artificial, can be generated without manual effort to improve the performance of neural networks or to facilitate semi-supervised or unsupervised pipelines.[2,3]

GANs provide state-of-the-art methods for unsupervised stain translation with promising results. In Gadermayr et al.,[3] stain translation via Cycle-GANs enabled, in addition to domain adaptation, a fully unsupervised and stain-independent segmentation approach. In another study,[4] multi-channel images were created by a concatenation of differently stained images as a result of stain translation. It proved that such “image enrichment” improves the accuracy of a segmentation network. In a related work,[2] classifiers were integrated into the generator and discriminator networks of a Cycle-GAN to identify the domain (staining) of the input image and to learn domain-specific attention maps. As a result, image-to-image translation was improved in a self-supervised way. Experiments were performed by applying a previously trained segmentation network from Bouteldja et al.[5] to synthetic images translated from other stains. An ablation study proved that this method has a superior performance compared to a simple Cycle-GAN model. Additionally, 3 extra feature channels were added to both the input and output of each generator in the Cycle-GAN. They can be used to store the necessary information of the input image for an improved reconstruction. The extra channels ensure that the image generator can avoid the unintended encoding of domain-specific features from the input image into the translated image. This approach has proved to have a positive effect for some stains on performance compared to an unmodified Cycle-GAN.[2]

Generally, an image-to-image translation via standard GANs is only possible between 2 domains. Two different GANs must be trained for each pair of domains, since 2 different translations can be defined within 2 domains. StarGAN[6] is a novel and scalable GAN that can be trained to perform translations between multiple domains with a single generator and a discriminator. This results in an increase in flexibility since any input image can be translated to any other domain. In addition, not only a single model has to be trained instead of many pairs of generators and discriminators, but also, all training images from different domains can be used for training, which enhances the quality of results. An improved version is StarGAN-v2,[7] which can generate diverse images, meaning that output images are generated in each desired domain in different styles and variabilities. In this network, a style encoder extracts the style code of images of different domains and the generator translates the input images to the desired output domains using the corresponding style codes. It should be noted that it does not need annotations or image pairs for training, and it can learn the mappings in an unsupervised way.

Contribution

The focus of this paper is on the augmentation of WSIs with many-to-many stain translations to improve the performance of classification tasks. The original StarGAN was implemented for the translation of human and animal faces. However, the focus of this work is on histological WSIs, which have a higher resolution and special structures that make the translation task more difficult. Also, the morphology should be kept constant during stain translation, which is not the case in face translation.

To perform a stain translation (or a many-to-many translation) between any 2 arbitrary domains with the same network, the StarGAN-v2 was implemented. Due to unavoidable morphological changes in the histological tissues, an edge detector in the form of a Canny filter was incorporated into the network and an additional term was added to the loss function to minimize the difference between the edges of the source image and the translated image. It was shown that these adjustments help to improve the results by keeping the main structure of the tissues during translation. Since having diverse translations from the same input image is not part of the goal, the loss term related to diversity was omitted.

Breast cancer was the most common cancer in 2020 and is one of the deadliest cancers in general.[8] Thus, researches are dedicated to its detection, prevention, and treatment worldwide. Convolutional Neural Networks (CNNs) and classifiers have been recently adopted for binary and multi-class classification tasks. Binary classification is used for identifying whether a tissue is malignant or non-malignant. Multi-class classification, however, also distinguishes the type of cancer.[9]

In this work, we performed a binary classification into samples that included malignant cells (“malignant”), or were tumor-free (“non-malignant”) as a clinically relevant proof-of-concept to quantify the effect of data augmentation with translated images on classification accuracy. Breast cancer classifiers were trained with and without our generated synthetic images and reveal a substantial improvement in the accuracy of the classifiers. Additionally, using a subjective test on medical and digital pathology experts, it is proved that the translated images are realistic and indistinguishable from real images.

Methods

The goal is to train a single generator that can generate translated images from a given input image to any arbitrary domain using the same network. For this purpose, the original StarGAN-v2[7] was applied to histological images. Different domain-specific style codes were learned for each domain and the generator reflects any arbitrary style in the output image. To realize a many-to-many stain translation, at least 3 domains (stains) must be present, but no labels or image pairs are required. However, due to the large extent of morphological and structural changes of the input images during translation, the original network has been improved in this work.

StarGAN with edge detector

The original StarGAN-v2 consists of 4 modules. One of them is the style encoder, which extracts the style of a reference image when given its domain. The style encoder can produce diverse style codes based on multi-task learning, in which a Multilayer Perceptron (MLP) with multiple outputs provides different styles for different domains, but during training, only 1 branch is randomly selected. Another part is the mapping network, which generates different style codes when given a latent vector and an arbitrary domain. It works based on a multi-task architecture (similar to the style encoder) and can generate diverse style codes by sampling the latent vector. Another module is the generator, which translates a given input image into the output image to reflect a given style code. The style code is given either by the style encoder or the mapping network. The final part is the discriminator, which consists of different branches for different domains. Each branch is basically a binary classifier for each domain, determining whether a given image is real or artificial.[7]

The full objective function for training the original StarGAN consists of adversarial loss L_{adv}, style reconstruction loss L_{sty}, style diversification loss L_{ds}, and a term for preserving source characteristics (i.e., cycle consistency loss L_{cyc}).[7] The original objective function is therefore:

$$min_{G,F,E}\ max_D \left[L_{adv} + \lambda_{sty} L_{sty} - \lambda_{ds} L_{ds} + \lambda_{cyc} L_{cyc}\right], \tag{1}$$

where λ_{sty}, λ_{ds}, and λ_{cyc} are hyperparameters that control the contribution of each term, and G, F, E, and D are, respectively, the generator, mapping network, style encoder, and the discriminator.[7]

The style diversification term in the loss function encourages the generator to create diverse output images. However, in the application of this work, diverse translation results are not required, and the focus is to maintain the morphology while translating the stain. Therefore, the style diversification term was omitted from the training loss function. In addition, a Canny edge detector[10] was incorporated into the network which acts on the input and translated images. An additional term was added to the training loss function to minimize the difference between the edges extracted from the source image and the translated image, and to encourage the network to maintain the morphology and structure.

Considering the mentioned alterations, the full objective function becomes:

$$min_{G,F,E}\ max_D\left[L_{adv}+\lambda_{sty}L_{sty}+\lambda_{cyc}L_{cyc}+\lambda_{edg}L_{edg}\right]. \quad (2)$$

The adversarial, style reconstruction, and cycle consistency losses are the same as defined in Choi et al.[7] The new part, L_{edg}, is the edge detector's loss as follows, and the coefficient λ_{edg} is an additional hyperparameter to control the effect of edge detection.

$$L_{edg}=\frac{1}{n}\sum_{i=1}^{n}\left|f\left(X_{fake_i},\theta_1,\theta_2\right)-f\left(X_{real_i},\theta_1,\theta_2\right)\right|, \quad (3)$$

where n is the number of images in each training batch, X_{fake_i} and X_{real_i} are, respectively, an artificial (translated) and its corresponding real (input) image in a batch, f is the Canny filter, and θ_1 and θ_2 are 2 hyperparameters.

If the Canny filter f is applied on a 3-channel input image X with 2 thresholds θ_1 and θ_2, the output will be the edges of the input image in the form of a binary image (mask) according to the algorithm proposed by John Canny,[10] which is implemented in OpenCV (Open Source Computer Vision Library).[11]

Evaluation methods

The goal is to perform stain translation via the upgraded StarGAN to create synthetic images. Then, as a proof of concept, the existing real dataset is augmented with artificial data and the improvement of the classification task is evaluated. The artificial images are deliberately created to be realistic and in good quality. Thus, 2 evaluation methods are introduced here: (1) the subjective or qualitative evaluation of the generated artificial images, and (2) the objective or quantitative evaluation of the effect on classification accuracy.

The subjective or qualitative evaluation

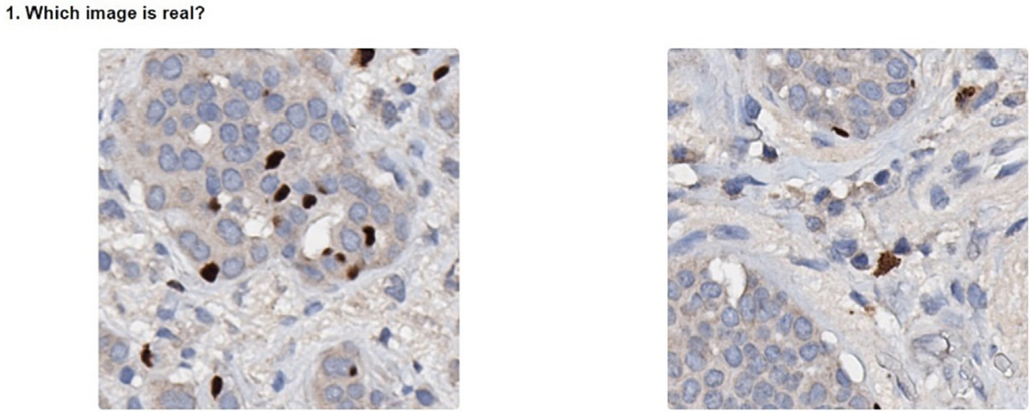
The subjective evaluation consists of quality control by 7 digital pathology experts to verify that the artificial images are realistic. Real and synthetic image patches from 3 stains H&E, p63, and FoxP3/CD3 were randomly selected to create the subjective test. Considering the 3 aforementioned domains, 6 different translations exist. For each of the 6 translations, 10 questions were considered in order to compare the translations when analyzing the test results. So, in total, 60 questions were designed in the form of real–artificial pairs and each participant was asked to choose the real image. The participants did not know the underlying translation behind the artificial images. If the real and artificial images are, indeed, indistinguishable, the participants will answer randomly, and therefore, 50% of their answers will be correct and the other half will be incorrect. The participants were medical and technical experts in the field of digital pathology. The order of all the 60 questions and of the real or artificial images was random, but each question contained exactly 1 real and 1 artificial image. An example question of the subjective test is represented in Fig. 1. Before the actual test began, 6 additional sample questions with real–artificial pairs with the correct labels were shown to the participants to familiarize themselves with the problem.

The objective or quantitative evaluation

As a quantitative (objective) evaluation and, simultaneously, as a proof of concept, the generated synthetic images were employed to train a breast cancer classifier to quantify the improvement of performance in the classification task. The classifier was trained to determine if each patch contains a tumor or not with a binary label of "malignant" or "non-malignant". The binary labels were defined as "malignant" for patches containing any malignant cells, including invasive breast cancer of any histopathological subtype and intraductal malignancies termed ductal carcinonma in situ (DCIS), as opposed to tumor-free "non-malignant" patches that included pre-existing normal tissues (glands, ducts, connective/fat tissues, or occasionally benign lesions such as apocrine metaplasia). For the evaluation, 2 classifiers were trained: the first was trained with real images only, and the second with real and artificial images augmented together. For the test phase, the test images only contained real patches and the accuracy values of the 2 classifiers were compared to prove that if the data is augmented with synthetic images, the classification accuracy will increase.

Data preparation

Cancerous human breast tissues obtained from anonymized surplus archival material processed in the context of previous retrospective studies,[12,13] were extracted and embedded in paraffin blocks (the use of anonymized surplus archival materials for research purposes is covered by the institutional review board (ethics committee of Hannover Medical School), approval number 2063–2013). Then, they were cut into 3 μm slices and placed on a glass slide. They were stained with several agents, of which the following ones were selected for this work: forkhead box P3 and cluster of differentiation 3 (FoxP3/CD3), hematoxylin and eosin (H&E), p63 immunohistochemical staining, and Cytokeratin 5/14 (CK5/14). De-paraffinized sections were stained with hematoxylin and eosin (H&E) and chromogenic immunohistochemistry to detect p63 Cytokeratin 5/14 (CK5/14) by 3,3 diaminobenzidine (DAB) staining, and forkhead box P3 and cluster of



1. Which image is real?

Fig. 1. A sample question of the subjective test, in which 1 real and 2 artificial image are presented in random order. The participant is asked to identify the real image by clicking on it.

differentiation 3 (FoxP3/CD3) in a chromogenic double staining with a red chromogen in addition to the brown DAB signal. Then, they were scanned and digitized by the whole slide scanner Aperio AT2 by Leica with a 40× objective lens. The dataset consists of 1 whole-slide image (WSI) without annotations in FoxP3/CD3 staining, 1 WSI in H&E, 1 in p63, 2 annotated WSIs in CK5/14, and 1 annotated WSI in H&E. In total, there are 6 WSIs, 3 of which were annotated for breast tumors.

Data pre-processing includes the automatic detection of the tissue contents from the background using Otsu's thresholding[14] and the extraction of the patches that fulfil the requirement of a minimum content threshold. The patch extraction was performed with OpenSlide, a widely used open-source library for reading and manipulating WSIs.[15] The threshold was chosen to be 0.6, meaning that each patch could only be selected if at least 60% of the content is tissue with a maximum of 40% background. Patches were extracted from the highest resolution (40× magnification or 0.253 μm / pixel), and they were randomly selected from each WSI with a resolution of 512 × 512 pixels. The patches were then downsampled by factor of 2 to the final patch size of 256 × 256 pixels to have a size and resolution comparable to similar works.[7] The fat tissues were regarded as background since they mostly dissolve and appear as empty space after the staining process.

Experimental settings for stain translation

This work focuses on unsupervised stain translation, i.e., different WSIs from different patients with various stains were utilized for training the translation between any stains. The improved StarGAN, as described in Section 2.1, was trained once for a translation between FoxP3/CD3, H&E, and p63 stains for the qualitative evaluation. Further, another StarGAN was trained for a translation between H&E and CK5/14 stains for the quantitative evaluation.

For the stain translation, 3 WSIs without any breast tumor annotations, each with one of the stains FoxP3/CD3, H&E, and p63, were employed for training the network for a many-to-many stain translation. One thousand patches were extracted from each WSI, resulting in 3000 patches in total. For each stain, 90% of the patches (900) were used for training. After training, the remaining 100 patches per domain were used for a qualitative evaluation (ref. the subjective test in Section 3.2).

The proposed upgraded StarGAN with Canny edge detector was trained based on the objective function in Eq. ((2), and the style codes were extracted from reference images instead of being generated by the mapping network from latent vectors. The generator was updated once after 5 updates of the discriminator.

The training was performed for 100 000 iterations in a batch size of 4, with an Adam optimizer. The parameters for the optimizer are the learning rate, weight decay, β_1, and β_2 (coefficients for computing running averages of gradient and its square), which were respectively 1e-4, 1e-4, 0.0, and 0.99, and the hyper-parameters λ_{cyc} and λ_{sty} which were set equal ($\lambda_{cyc} = \lambda_{sty} = 1$). In addition, different coefficients for the edge loss were tested to show the effect of L_{edg}. The network was trained with different λ_{edg} values of integers between 0 and 5 for comparison, and also with $\lambda_{edg} = 0$ for ablation study. The hyperparameters of the Canny edge detector θ_1 and θ_2 are 0.3 and 0.5, respectively. Classic data augmentation methods were also applied, which included random horizontal and vertical flipping and random resized cropping with a scale of between 0.8 and 1.0 (the lower and upper bounds for the random area of the crop) and a ratio of between 0.9 and 1.1 (lower and upper bounds for the random aspect ratio of the crop). All of the aforementioned transformations were performed with the probability of 0.5. The experiments are implemented in PyTorch[16] and conducted on NVIDIA GeForce RTX 2080 Ti GPU with 11.3 GB VRAM. The training of the StarGAN for translations between 3 stains (i.e., 6 translations) required approximately 1 week, and between 2 stains (i.e., 2 translations) 2 days and 7 h.

Experimental settings for classification

The classification pipeline consists of stain translation, augmentation, and classification, as depicted in Fig. 2. Patches were extracted from 3 breast cancer WSIs, 2 stained with CK5/14 and the third stained with H&E, and each patch was labeled as “malignant” or “non-malignant”. The WSIs were originally annotated manually by our medical expert for segmentation purposes and the binary labels were inferred automatically based on the annotations. In doing so, even if only a very small part of a tumor is contained within a patch, it is labeled “malignant”.

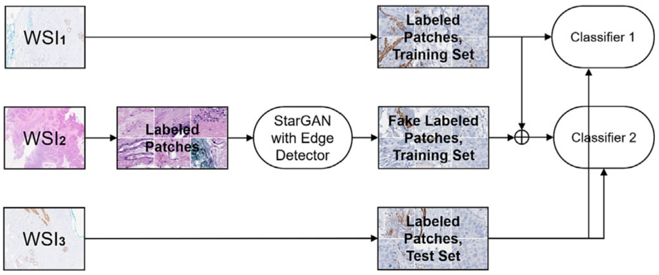

WSI₁
Labeled Patches, Training Set
Classifier 1
WSI₂
Labeled Patches
StarGAN with Edge Detector
Fake Labeled Patches, Training Set
Classifier 2
WSI₃
Labeled Patches, Test Set

Fig. 2. A schematic picture of the classification workflow. Two different WSIs of stain CK5/14 are considered for training and testing and 1 WSI of stain H&E for image augmentation. Patches are extracted from each WSI and are labeled as “malignant” or “non-malignant”. The patches with H&E stain are translated to CK5/14 using the upgraded StarGAN. The labels do not change during translation. Then, two classifiers are trained: one with the real CK5/14 patches extracted directly from a WSI as the training set, and the other with the same and augmented with the translation outputs. The levels of accuracy by the classifiers are then compared based on the real test patches.

To avoid bias, the test data patches were extracted from a WSI different from the training data (i.e., from different patients). This means that the trained classifier will not be patient-specific but, rather, will be general. Around 2000 patches with stain CK5/14 extracted from one WSI were considered as the training set, and 500 patches from the other WSI with the same stain were taken as the test set. Then, approximately 2000 patches stained with H&E were translated to CK5/14 via the proposed network. The network here is trained with the same settings as in Section 2.4 for the translation between H&E and CK5/14 stains using the patches extracted from H&E and the training dataset from the classifiers (i.e., patches from CK5/14). Note that the binary labels of the patches do not change during translation, because whether a patch contains a tumor, or not, does not change with a different staining.

Then, 2 classifiers are trained with the goal of quantifying the effect of data augmentation with artificial images on classification accuracy: one with the real training set of CK5/14 stain only (2000 patches), and the other with the same training set as well as synthetic CK5/14 images (4000 patches together) that are obtained from the translation of the H&E stain. Then, the performance of the 2 classifiers is compared on the real test data from the CK5/14 stain.

Two networks, VGG-16 and ResNet-50, were employed for classification. The VGG network won first place in the ImageNet challenge (ILSVRC) 2014, and ResNet won the ILSVRC and MS COCO 2015 competitions.[17,18] Exactly, 50% of the training images had “malignant” labels and the other half were “non-malignant ” and, therefore, the classes are balanced. The dataset was pre-processed, as described in Section 2.3. The pre-trained networks on the ImageNet dataset were adapted for breast cancer classification using a transfer-learning approach. More specifically, the weights of the pre-trained networks were fixed and used for feature extraction and the last layer was replaced by a trainable dense layer with Softmax activation for the intended binary classification problem.

All settings for training the classifiers in all cases with, and without, the augmented synthetic images are the same to enable a fair comparison of results. The pre-trained networks were fine-tuned on the training data for 100 epochs, in a batch size of 8, with an Adam optimizer and a categorical cross-entropy loss function. The parameters for the optimizer are the learning rate, β_1, and β_2, which were equal to 0.001, 0.9, and 0.999, respectively. Standard data augmentation methods were also applied, which were random horizontal and vertical flipping with the probability of 0.5. The pre-trained networks were implemented and fine-tuned in Keras, an open-source deep learning API.[19]

Results

Many-to-many stain translation

The addition of an edge detector, as described previously, improved the many-to-many stain translation in terms of the reduction of morphological changes and in having an edge-preserving translation. Initial qualitative results for the many-to-many translation with a single network without an edge detector (for $\lambda_{edg}= 0$) are shown in Fig. 3. The first row shows sample input images in the experiment, and the first column presents a few of the reference images. Each input image is translated in such a way as to reflect the staining of the reference image, which results in 6 synthetic images in this example. All translated images are technically plausible and cannot be distinguished from real patches. Fig. 4 and Fig. 5 present similar results but with an incorporated edge detector with $\lambda_{edg} = 1$ and $\lambda_{edg} = 4$, respectively. The general structure and main edges of the input images are easily observable in the artificial images with a visible improvement with higher values of λ_{edg}. However, even with the highest λ_{edg}, the structure is not 100% preserved. With higher values for $\lambda_{edg} = 5$ or higher, the results appear the same as in $\lambda_{edg} = 4$ to human eyes. Therefore, $\lambda_{edg} = 4$ produces the highest quality of translation results.

Note that since the images are from cancerous tissues, which are inherently chaotic, often there are no clear edges and structures, even in real images. Thus, adding an edge detector does not fully preserve the structure, but it still improves the images.

Subjective test

There are 6 possible translations between the 3 stains FoxP3/CD3, H&E, and p63. The participants were asked 10 questions per translation (60 questions in total) in the form of real–artificial image pairs. The coefficient for the edge detector during the stain translation which produces the artificial images was $\lambda_{edg} = 4$, as it produces the seemingly best translation results compared to lower values, as depicted in Fig. 5.

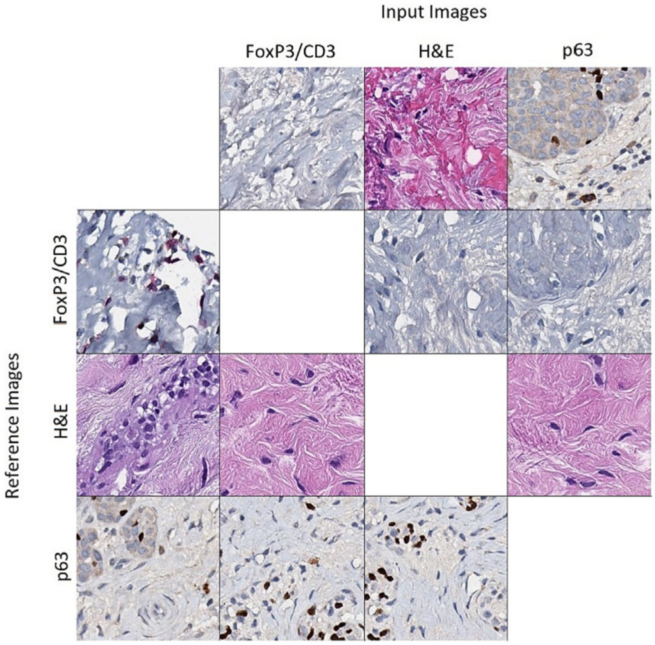


Fig. 3. Qualitative results of a many-to-many stain translation with a single network. The experiment was performed on patches extracted from breast cancer WSIs in 3 stains without an edge detector with $\lambda_{edg}= 0$. First row: selected source images, first column: selected reference images. The source images are translated with the same network to reflect the staining of the reference images. All other images (6) are synthetic.

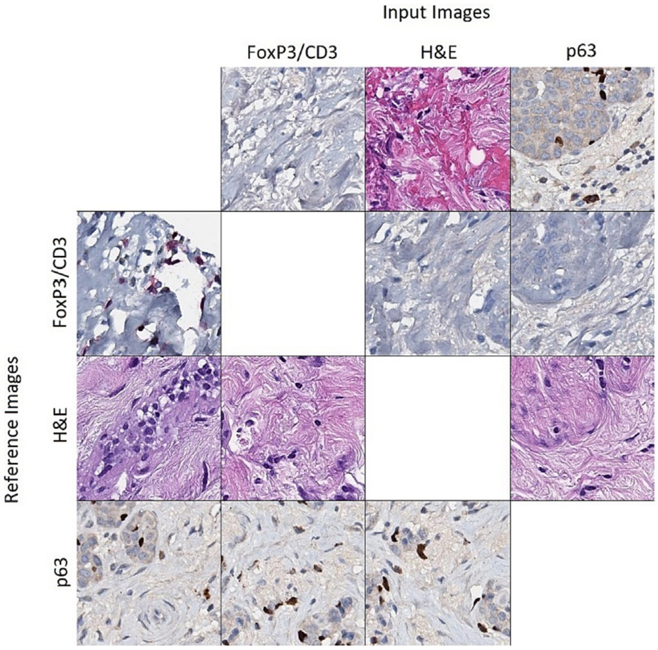


Fig. 4. Qualitative results similar to Fig. 3 but with an incorporated edge detector with $\lambda_{edg}= 1$.

The number of correctly distinguished real images is presented in Fig. 6. It can be observed that among different translations, FoxP3/CD3 to H&E has the lowest, and H&E to p63 has the highest average. Note that the average for the translation FoxP3/CD3 to H&E (32.9%) differs greatly from the ideal value of 50%. This shows that the participants mistook the artificial images for real ones. It does not, however, question the quality of artificial images. The average scores for all translations are sufficiently close to 50% and the variabilities are random and not meaningful.

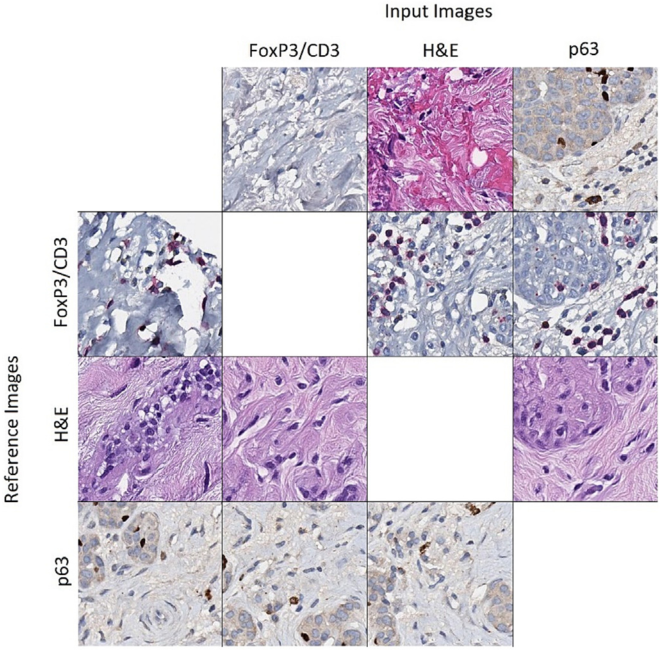


Fig. 5. Qualitative results similar to Fig. 3 but with an incorporated edge detector with $\lambda_{edg}= 4$.

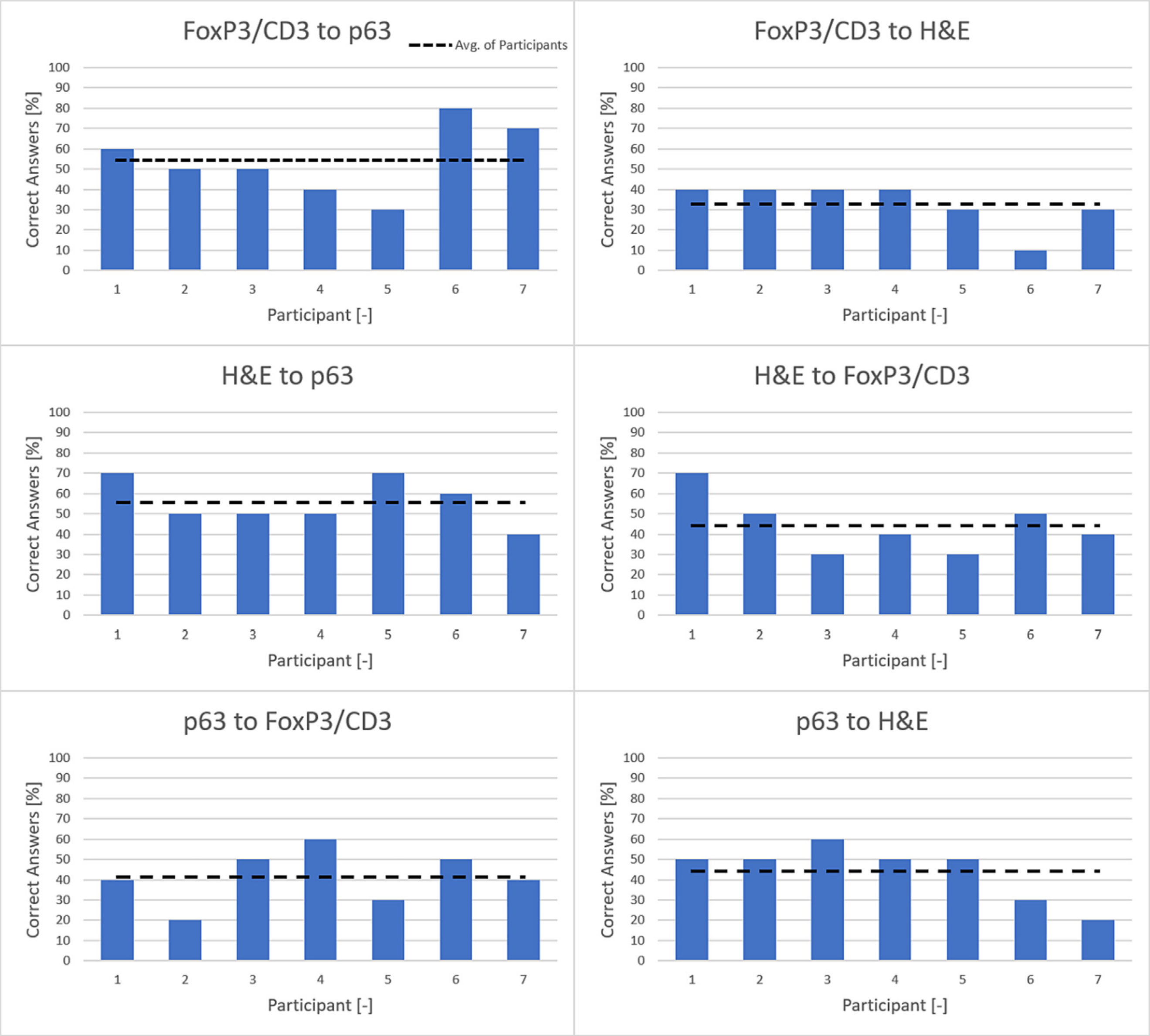


Fig. 6. Histogram of the percentage of correctly identified real images for the participants of the subjective test for each of 6 possible stain translations. The average score of all participants is indicated with a dashed horizontal line, which approaches 50% in most cases.

Fig. 7 indicates the total number of correct answers from all participants. The percentage of correctly identified real images was between 40% and 55% in total, and the average was 45.5%. The fact that the results approach 50% means that the synthetic images are, indeed, technically plausible and indistinguishable from the real ones.

Classification and data augmentation

The levels of accuracy of the classifiers were calculated based on the test set according to the pipeline in Fig. 2 and the results are presented in Table 1. By augmenting the training dataset with the artificial images and thus doubling its size, the test-set accuracy increased by 8.0% (from 82.1% to 90.1%) and by 9.3% (from 79.6% to 88.9%) using a classifier based on ResNet-50 and VGG-16, respectively. The value of λ_{edg} was 4 during the stain translation, similar to Section 3.2. Note that since the class distribution was balanced, there was no need to calculate the F1-score and, therefore, the calculation of the accuracy suffices in this case.

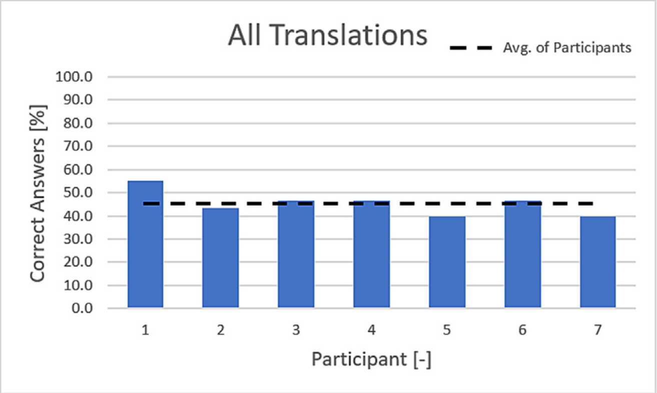


Fig. 7. Histogram of the percentage of correctly identified real images for the subjective test in total. The average score is indicated with a dashed line and approaches 50%.

Table 1
Levels of accuracy calculated on the test set, as depicted in Fig. 2, for the 2 networks ResNet-50 and VGG-16. The results were calculated for 2 different scenarios. The first one is for a classifier, the training set of which contained only 2000 real images. The training set in the second one consisted of 4000 images, half of which were the artificial images obtained from a stain translation.

Network	Acc. [%] – Real training set	Acc. [%] – Augmented training set
ResNet-50	82.1	90.1
VGG-16	79.6	88.9

Discussion

In this work, it has been shown that the proposed pipeline can effectively translate any stain to any other with the same network. The generated images are of high quality and their general morphological appearance remains plausible after translation. There are still minor visible changes during translation that could be improved in the future. However, the slight morphological changes of translated tissues can be disregarded and are not critical in this study, since the output images are exclusively used for training of other deep neural networks, and clearly improve the classifiers' accuracy.

Subjective tests on our medical and technical experts in the field of digital pathology suggest that the generated patches of histopathological images are technically plausible and cannot be distinguished from the real ones, as long as the spatial tissue context and the biological role of the stained structures are not being evaluated. In fact, the percentage of correctly distinguished patches based on real images from the synthetic ones approached 50% and this implies that the images are technically plausible because the participants answered the questions randomly. For all 6 different translations, the average scores also approach 50% with random deviations. While beyond the scope of this study, further research is warranted to test, e.g., with more questions per translation, if the artificial images also allow for biologically meaningful conclusions, and whether there is additional information provided by differences between the translations.

To prove the applicability of the generated images in the presented work, different networks for the task of breast cancer classification were trained once with manually annotated real images in a specific stain and once with the same images augmented with synthetic ones. The artificial images are the outputs of the translation of other annotated images to the same stain. Note that the focus of this research is not exclusively on breast cancer classification. The only reason for performing a binary classification here is to show that augmenting the training dataset with the synthesized images boosts the classification accuracy. Therefore, attempts were not made to find a classifier that performs better. A comparison of classifiers suggests that the proposed pipeline for image augmentation can effectively increase the classification accuracy. It confirms that the proposed workflow for a many-to-many stain translation with a single network enables an easy, realistic, and effective way for data generation to improve the performance of deep learning networks. It is important to note that there is still some risk in generating artificial images if they should be mistakenly used, e.g. for diagnostic purposes. In addition, our proposed method is necessary given the shortage of high-quality expert annotations in the field of digital pathology, acknowledging that the clear improvement of deep learning tasks by artificially generated images would be smaller if larger sets of annotated images were available. Thus, the intended application of our method is specifically in domains where available annotations are sparse.

In this work, the proposed method for image augmentation was evaluated by a classification task for a translation between H&E and CK5/14 stains. It could be repeated in future works for different translations to find the pair of stains that improves the classifier accuracy the most. In some cases, a defining an intermediate stain could also help, meaning that instead of converting stain x directly to stain y, it is converted to a third stain z first, and then y is obtained from z. In this case, z is the base stain that has the best translations and, therefore, the quality of x-to-y translation improves.

Another important application of this study regarding binary classification is that the classified patches could be connected to reconstruct the WSI. Then, it would be visible in the WSI which patches are malignant. This is of great importance for clinical practices and medical experts and proves that the proposed image augmentation and classification pipeline has added values.

In future works, the pipeline will be enhanced so that there is a one-to-one relationship between the output and source images and the maintenance of tissue morphology will be further improved. For this purpose, it may be helpful to consider other metrics such as Normalized Mutual Information (NMI), and to use them in the loss function or for evaluation. The addition of an identity loss function could also be helpful, as it motivates the generator to keep the images unchanged.[2] Other edge detectors, for example, methods based on total variation, could be employed for training the network instead of the Canny filter to further investigate their effect. Enhancing the network and motivating it to keep exactly the same structure as the input provides new opportunities. For example, annotation masks for segmentation would not change during translation and, therefore, the improvement of the segmentation task could also be evaluated in addition to the binary classification in this work. The segmentation task could be useful for a more exact detection of tumor tissues. In addition, image registration of consecutive and similar WSIs with different staining provides ground-truth images for comparison with the translation outputs and a quantitative metric for evaluation.

Another evaluation method in future works could be to insert unannotated patches of WSIs to our pipeline as the inputs and thereby obtain labeled patches as the output. After reconstructing the labeled version of the initial WSI by putting the patches back together, a medical expert could determine if the labels are correct. Reconstruction of the WSIs would be necessary, as medical experts need contextual information for evaluating the results.

In addition to classification, the presented methodology could have other applications, such as segmentation. In some cases, tumors may be more visible in one specific stain than other stains. That specific stain could act as the base stain of the pipeline, similar to the concept of "easy-to-segment" stain referred to in Gadermayr et al.[3] Therefore, a segmentation network could be trained on the base stain and all other stains could then be translated to the base stain to become applicable to the trained segmentation tool. In contrast to Gadermayr et al.,[3] a many-to-many translation, instead of pairwise translations, is feasible in this work, which facilitates this concept tremendously.

This work could be more generalized by taking other human body organs into account as well. In the current settings, all images are related to the breast and according to the StarGAN-v2 nomenclature,[7] different stains represent domains and there is no diversity or different styles in this work. Another setting for a more generalized pipeline could be to choose different human organs as domains and then having different stains as styles, without omitting the style diversification term in the loss function.

Another suggestion is to consider different tumor types or different cancer grades as other parameters of translation and to train stain translators separately for each group since different tumors or grades of cancer result in different appearances of the tissues. In this work, fat tissues have been disregarded but they could be important in the case of cancer in other organs. They could also be considered separately for training the stain translators. For this purpose, fat-preserving staining agents must be used for the WSIs so that the fat tissues do not dissolve and appear as empty space.

Conclusions

This paper proposes a pipeline for effectively translating any available histopathological stain to any other stain of interest with a single network for image augmentation. The core findings of this proof-of-concept study are as follows: First, the synthesized stain images, generated by a structure preserving Star-GAN method, are indistinguishable from the real ones. Second, compared to a vanilla breast cancer classification without stain augmentation, the accuracy of our proposed classification with stain augmentation increases considerably from 82.1% to 90.1% using ResNet-50, and from 79.6% to 88.9% using VGG-16. In future work, once larger multi-stain pathological datasets are available, we intend to continue this work to investigate whether the proposed stain-augmented classification could also be applied to support further medical histology classification tasks. Also, considering all possible stain types commonly applied in a certain pathological field, it could be quantified to which extent both the selected real stains and the augmented virtual stains contribute to the final classification task. This would provide insights into whether an optimal combination of both real and virtual stains exists for a pathological classification task of interest.

Authors' Contributions

MB, BH, and DM planned the presented research. JL, FF, and DM wrote grant proposals and acquired fundings. MB conducted the experiments, statistical analyses, programming, and training of the networks and discussed the results with BH and DM. NS and FF performed medical image acquisition and annotation. MB wrote the first draft of the manuscript and generated the figures. NS, BH, JL, FF, and DM critically reviewed the manuscript and figures. All authors have read and approved the final version of the article.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported by the Federal Ministry of Education and Research (BMBF) of Germany under the grant no. FKZ 01IS21067A-C as well as FKZ 01ZX1608A.

References

1. Barisoni L, Lafata KJ, Hewitt SM, Madabhushi A, Balis UGJ. Digital pathology and computational image analysis in nephropathology. Nat Rev Nephrol 2020;16:669–685.
2. Bouteldja N, Klinkhammer BM, Schlaich T, Boor P, Merhof D. Improving unsupervised stain-to-stain translation using self-supervision and meta-learning. Journal of Pathology Informatics 2022;13, 100107. https://doi.org/10.1016/j.jpi.2022.100107.
3. 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 2019:2293–2302.
4. Gupta L, Klinkhammer BM, Boor P, Merhof D, Gadermayr M. GAN-based image enrichment in digital pathology boosts segmentation accuracy. Medical Image Computing and Computer Assisted Intervention – MICCAI; 2019.
5. 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.
6. Choi Y, Choi M, Kim M, Ha J-W, Kim S, Choo J. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2018.
7. Choi Y, Uh Y, Yoo J, Ha J-W. StarGAN v2: diverse image synthesis for multiple domains. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2020.
8. World Health Organization (WHO). "Cancer," [Online]. Available. https://www.who.int/news-room/fact-sheets/detail/cancer. [Accessed 6 7 2022].
9. Yari Y, Nguyen H, Nguyen TV. Accuracy improvement in binary and multi-class classification of breast histopathology images. IEEE Eighth International Conference on Communications and Electronics (ICCE); 2021.
10. Canny J. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intel 1986;PAMI-8(6):679–698.
11. OpenCV (Open Source Computer Vision). "Canny Edge Detection," [Online]. Available. https://docs.opencv.org/4.x/da/d22/tutorial_py_canny.html. [Accessed 22 May 2022].
12. Schaadt NS, Alfonso JCL, Schönmeyer R, et al. Image analysis of immune cell patterns in the human mammary gland during the menstrual cycle refines lymphocytic lobulitis. Breast Cancer Res Treat 2017;164(2):305–315.
13. Schaadt NS, Schönmeyer R, Forestier G, et al. Graph-based description of tertiary lymphoid organs at single-cell level. PLOS Computat Biol 2020:1-18. https://doi.org/10.1371/journal.pcbi.1007385.
14. Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernet 1979;9(1):62–66.
15. Goode A, Gilbert B, Harkes J, Jukic D, Satyanarayanan M. OpenSlide: a vendor-neutral software foundation for digital pathology. J Pathol Inform 2013;4(27).
16. Paszke AP, Gross S, Massa F, et al. PyTorch: an imperative style, high-performance deep learning library. 33rd Conference on Neural Information Processing Systems (NeurIPS); 2019.
17. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Computer Vision and Pattern Recognition; 2014.
18. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016.
19. Chollet F. Keras. [Online]. Available: https://keras.io/api/applications/ 2015.