Technical Note

On the Acceptance of "Fake" Histopathology: A Study on Frozen Sections Optimized with Deep Learning

Mario Siller[1], Lea Maria Stangassinger[2], Christina Kreutzer[3], Peter Boor[4,5], Roman D. Bulow[4,5], Theo J. F. Kraus[6], Saskia von Stillfried[4,5], Soraya Wolfl[7], Sebastien Couillard-Despres[3], Gertie Janneke Oostingh[2], Anton Hittmair[8], Michael Gadermayr[1]

[1]Department of Information Technology and System Management, Salzburg University of Applied Sciences, Salzburg, Austria, [2]Department of Biomedical Sciences, Salzburg University of Applied Sciences, Salzburg, Austria, [3]Institute of Experimental Neuroregeneration, Spinal Cord Injury and Tissue Regeneration Center Salzburg, Paracelsus Medical University, Salzburg, Austria, [4]Institute of Pathology, RWTH Aachen University Hospital, Aachen, Germany, [5]Center for Integrated Oncology Aachen Bonn Cologne Duesseldorf (CIO ABCD), Aachen, Germany, [6]Institute of Pathology, University Hospital Salzburg, Paracelsus Medical University, Salzburg, Austria, [7]Patholab Salzburg, Salzburg, Austria, [8]Department of Pathology and Microbiology, Kardinal Schwarzenberg Klinikum, Schwarzach, Austria

Submitted: 30-April-2021 **Revised:** 01-August-2021 **Accepted:** 20-September-2021 **Published:** 05-January-2022

Abstract

Background: The fast acquisition process of frozen sections allows surgeons to wait for histological findings during the interventions to base intrasurgical decisions on the outcome of the histology. Compared with paraffin sections, however, the quality of frozen sections is often strongly reduced, leading to a lower diagnostic accuracy. Deep neural networks are capable of modifying specific characteristics of digital histological images. Particularly, generative adversarial networks proved to be effective tools to learn about translation between two modalities, based on two unconnected data sets only. The positive effects of such deep learning-based image optimization on computer-aided diagnosis have already been shown. However, since fully automated diagnosis is controversial, the application of enhanced images for visual clinical assessment is currently probably of even higher relevance. **Methods:** Three different deep learning-based generative adversarial networks were investigated. The methods were used to translate frozen sections into virtual paraffin sections. Overall, 40 frozen sections were processed. For training, 40 further paraffin sections were available. We investigated how pathologists assess the quality of the different image translation approaches and whether experts are able to distinguish between virtual and real digital pathology. **Results:** Pathologists' detection accuracy of virtual paraffin sections (from pairs consisting of a frozen and a paraffin section) was between 0.62 and 0.97. Overall, in 59% of images, the virtual section was assessed as more appropriate for a diagnosis. In 53% of images, the deep learning approach was preferred to conventional stain normalization (SN). **Conclusion:** Overall, expert assessment indicated slightly improved visual properties of converted images and a high similarity to real paraffin sections. The observed high variability showed clear differences in personal preferences.

Keywords: Frozen sections, generative adversarial networks, histology, paraffin sections, thyroid cancer, whole slide imaging

Background

Digital whole slide scanners are capable of effectively digitizing specimen slides, showing both microscopic detail and the large context, without significant manual effort. In best case, the whole processing pipeline is automated, including archiving and linking to the clinical software.[1] Apart from more efficient storage, this digitization enables the application of digital image processing approaches with the goal of facilitating clinical workflows. Without additional functionality, costs and effort for digitization are hard to argue, as pathologists are used to conventional microscopy and diagnostic accuracy and convenience are not automatically improved.[2-4] Recently, numerous

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/jpi.jpi_53_21

Address for correspondence: Dr. Michael Gadermayr, Department of Information Technology and System Management, Salzburg University of Applied Sciences, Urstein Sud 1, 5412 Puch, Austria.
E-mail: michael.gadermayr@fh-salzburg.ac.at

How to cite this article: Siller M, Stangassinger LM, Kreutzer C, Boor P, Bulow RD, Kraus TJ, *et al*. On the acceptance of "fake" histopathology: A study on frozen sections optimized with deep learning. J Pathol Inform 2022;13:6.
Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2022/13/1/6/335003

image analysis approaches were developed to support pathologists during clinical routine, consisting of automated segmentation,[5,6] stain normalization (SN),[7] and classification approaches.[8] For this purpose, state-of-the-art deep-learning approaches showed a particularly high performance[9] and often outperformed classical techniques.

So-called generative adversarial networks have been developed to translate one specific imaging modality into another. Based on a solely data-driven approach, a model can be trained to translate, for example, MR images into corresponding virtual CT images.[10] This methodology can also easily be applied to digital pathology to translate between different stains[6] or even between two preparation techniques.[11] Image translation between different stains is clearly a controversial challenge, as special features are often only visible when specific staining methods are used. As the acquired information strongly depends on the staining used, a translation could easily result in so-called hallucinations. A consequence of hallucination artifacts is that the images look realistic, but they do not correspond to the real underlying tissue as parts are randomly generated. Cohen *et al*.[12] showed that hallucination can remove tumor tissue in magnetic resonance images, which is highly unwanted in diagnostic settings. The choice of training data has a strong impact on the behavior of generative adversarial networks with regard to hallucination. However, as long as the fake image data are used as an intermediate representation during automated processing, a thorough evaluation would immediately identify unintentionally introduced biases. In previous work, this was achieved by evaluating classification scores, such as F1-score, precision, and recall.[6,13] Introduced bias could be identified as an increase in one and a decrease in the other measure, in turn decreasing the overall F1-score. Although research on automated image analysis and decision support systems showed high potential of image translation techniques,[6,13] these techniques are not yet applied in clinical workflows. For an effective and safe application, it is essential to exclude that the side effects of deep-learning techniques (such as hallucination artifacts) can affect the diagnosis.[4] This requires an extended clinical study with a large number of experts performing the diagnosis of a diverse spectrum of digital image data to perform a similar evaluation as for automated approaches.[6,13] Intra-rater variability exhibits an additional difficulty of expert examinations.

In this work, we focus on the translation of frozen sections to virtual formalin-fixed and paraffin-embedded (FFPE) sections. The FFPE material is most commonly used in diagnostic histopathology. It is compatible with a large variety of staining methods and allows thin sectioning (down to a few micrometers) with a high visual quality. Frozen sections [Figure 1A] are typically generated during interventions (e.g., cancer resections) to achieve information on malignancies as fast as possible. The preparation time for paraffin sections is normally too long to be used for this specific purpose. Frozen sections allow surgeons to wait for the histological results during the intervention in order to base the further procedures on the outcome of this histology. A drawback of frozen sections is their quality. Compared with paraffin sections, the image quality of frozen sections is typically lower, leading to a higher rate of misdiagnoses during clinical routine.[14-16] The cellular structure is, by far, more pronounced in paraffin sections, as these are fixed before embedding in paraffin. This is not the case in frozen sections, with the result of partly indiscernible or damaged tissue features.[17,18] Due to fast acquisition times, state-of-the-art whole slide scanners are effectively applied under time constraints in case of frozen-section pipelines.[19] For the classification of thyroid cancer diagnosis, considered in this work, it has been shown that the classification performance (F1-score) of computer-aided diagnosis clearly increases, independently of the setting of the classification model.[11] However, as pathologists are specifically trained to compensate artifacts due to improper acquisition, it is not clear as to whether a similar improvement can be obtained in case of manual diagnosis. However, this work indicates that image translation does not lead to strong hallucination artifacts misleading the diagnostic pipeline (as a model trained on real paraffin sections can be effectively applied to virtual paraffin sections).

In this article, we study how expert pathologists assess fake pathology [Figure 1, particularly (b)]. We make use of so-called generative adversarial networks to perform a virtual translation from the domain of frozen sections to the superior domain of paraffin sections. Based on real frozen sections, virtually improved frozen sections, and real paraffin sections, we conducted two investigations: First, based on the opinion of expert pathologists from different institutions, we assessed whether generative adversarial networks improved the visual quality of scanned specimen slides. Second, we evaluated whether pathologists were able to identify whether sections were virtually generated or real. As image material, we used a data set showing two different thyroid cancer categories, namely papillary carcinoma and follicular nodules.

Methods

Sample collection, preparation, and imaging

The data set consisted of totally 80 whole slide images, that is, 40 slides were available for each section type. All images were acquired during clinical routine at the Kardinal Schwarzenberg Hospital. They were diagnosed by an expert pathologist with more than 20 years of experience. A total of 42 (21 per modality) slides were labeled as papillary carcinoma whereas 38 (19 per modality) were labeled as follicular carcinoma. The mean and median age of patients at the date of dissection was 47 and 50 years, respectively. The data set comprised 13 male and 27 female patients. As we focus on visual assessment of

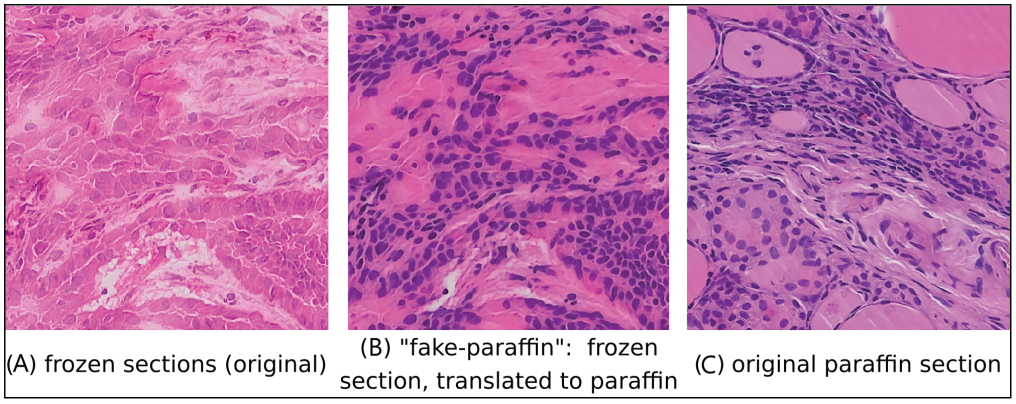

Figure 1: Original frozen section (A) and a corresponding optimized "fake-paraffin" patch showing characteristics of paraffin sections (translated with the CUT setting, see the section "Virtual frozen-to-paraffin translation"). An example of a paraffin section is shown on the right side (C)

image characteristics and not on diagnosis, the exact type of pathology is secondary. For the frozen sections, fresh tissues were frozen at −15 °C; slides were cut (thickness 5μm) and stained immediately with hematoxylin and eosin. For the paraffin sections, tissues were fixed in 4% phosphate-buffered formalin for 24 h. Subsequently, FFPE tissue was cut (thickness 2μm) and stained with hematoxylin and eosin. Images were digitized with an Olympus VS120-LD100 slide loader system. Overviews at a 2x magnification were generated to manually define scan areas. Focus points were automatically defined and adapted when needed (we used the Olympus extended focus imaging [EFI] setting). Scans were performed with a 20x objective, leading to a resolution of 344.57 nm/pixel. The image files were stored in the Olympus vsi format based on lossless image compression to avoid compression artifacts.

Virtual frozen-to-paraffin translation

We denote the domain of frozen sections as F ($f \in F$) and the domain of paraffin sections as P ($p \in P$). For frozen-to-paraffin translation, a fully convolutional neural network was trained to perform translation T from domain F to domain P. Although underlying tissue is intended to remain stable, image characteristics such as stain intensity, sharpness, and contrast are intended to be adapted to domain P, showing a higher perceptual quality.

As perfectly corresponding pairs consisting of a frozen and a paraffin cut are not achievable, conventional fully convolutional neural networks alone[20] are not applicable here. Instead, generative adversarial networks, allowing unpaired training, were employed. These networks are based on at least one generator and at least one discriminator. The generator (which is a fully convolutional network) performs the translation, whereas a discriminator ensures that the generated image is similar to real samples from domain P.

Although the similarity to real samples can be easily obtained in unpaired settings by means of a generative adversarial model, a challenge is to ensure that translation does not change the underlying tissue structure. In case of Cycle-GAN[21] (CG), this is enforced by learning full cycles, for example, from F to P and back to F_0. The cycle-loss compares the original images from domain F with the reconstructed image (F_0). The CG makes use of two generators and two discriminators, whereas the contrastive unpaired translation (CUT) approach[22] learns only one direction and ensures similarity using a feature-based loss with advantages in case that the mapping is ambiguous.[22] We investigated both CG[21] and CUT[22] individually. The mathematical formulations of CG and CUT are explained in more detail in the following subsection:

CG formulation

In case of cycle-GAN, two generative models, $T: F \rightarrow P$ and T': $P \rightarrow F$ and two discriminators, D_F and D_P are trained while optimizing the cycle consistency loss L_c as well as the adversarial loss L_d. T and T' are forced by D_F and D_P to generate fake images that look similar to real images, whereas D_F and D_P aim at distinguishing between translated and real samples. The generators aim at minimizing this adversarial objective against the discriminators that try to maximize it. The cycle-consistency loss forces that the output after translation from F to P and back to F (and vice versa) is similar to the input. For comparison, a pixel-wise $L1$ loss is utilized. This pixel-wise loss exhibits a limitation in case of ambiguous mappings.[23]

CUT Formulation

The goal of CUT[22] is to optimize a loss criterion consisting of a weighted sum of a GAN loss L_{GAN}, a patch similarity loss $L_{PatchNCE}(T,H,F)$ forcing corresponding patches to share content, and an additional regularization term $L_{PatchNCE}(T,H,P)$. In summary, the loss can be formulated as follows:

$$L = L_{GAN}(T,H,F,P) + \lambda_F L_{PatchNCE}(T,H,F) + \lambda_G L_{PatchNCE}(T,H,P), \quad (1)$$

with T being the generator, D being the discriminator, H being a two-layer perceptron, and λ_F and λ_G being scalar weights. For further details, we refer to the original publication. We used the Pytorch default reference implementation[21,22] with a patch size of 256 × 256 pixels.

Stitching artifacts

A challenge in case of whole slide images is that patch-wise processing (as the images are too large to fit into GPUs memory) can lead to clearly visible stitching (tiling) artifacts. For that purpose, we additionally investigated a method to reduce these artifacts. For that purpose, CG was combined with an additional perceptual embedding consistency (PEC) loss,[24] which introduces a penalty in the generators' latent space. The PEC loss forces the generators to learn a semantic content and contrast free features in the latent space, allowing a homogenization of the output contrast when the new style is added to the semantic features in the decoder block. This enables improved contrast normalization within the patch automatically, relaxing the issue of discrepancies in border regions.

Conventional stain normalization

As reference, we also investigated a conventional SN technique. For that purpose, we employed the approach introduced by Reinhard *et al.*[25] Compared with the deep learning-based techniques CG and CUT, this method only performs translation based on single pixels and does not incorporate any neighboring information, in turn limiting the potential of change. Although these methods are not developed to translate between modalities (frozen-to-paraffin), but rather for SN, they serve as a baseline.

Expert study

In order to find out whether experts are able to distinguish between real paraffin sections and translated frozen sections and to find out whether image translation is capable of improving the perceived visual quality of frozen sections, we asked six pathologists (four male, two female) to perform the following two experiments. The pathologists' clinical experiences vary from 3 years to 30 years (3, 6, 10, 12, 20, 30). Five of them self-assessed their practical experience with digital pathology in research or clinical diagnostics as significant (four) or high (one). Only one pathologist self-assessed his/her experience as low. We did not explicitly indicate that stitching artifacts occur due to image translation approaches as this would guide them to focus on these artifacts only and not on any other visual characteristics. An overview of the setting is shown in Figure 2.

Experiment 1

Quality assessment: First of all, we asked the pathologists to rank sets of five randomly selected image patches with a size of 2500 × 2500 pixels containing an original frozen section (F), a stain normalized version (SN), and translated images obtained with three different deep learning-based configurations (see the section "Virtual frozen-to-paraffin translation"). To exclude patches showing background in large image areas, only patches showing majorly tissue were included (90%). Overall, each pathologist assessed 20 of these corresponding quintets. Each of the images within a quintet was assigned with a rank between one and five. Due to the level of measurement of a ranking, it was not allowed (from a statistic point of view) to compute and analyze differences between ranks. For that purpose, only positive (lower rank) or negative (higher rank) effect was extracted as information and not the difference of the rank (which would introduce bias). Each quintet allows one to perform 10 different comparisons. Overall, 10 × 20 (quintets) × 6 (pathologists) =1,200 comparisons could be performed based on the available data.

Experiment 2: Real vs. fake: Second, we asked the experts to select the real image out of randomly selected pairs

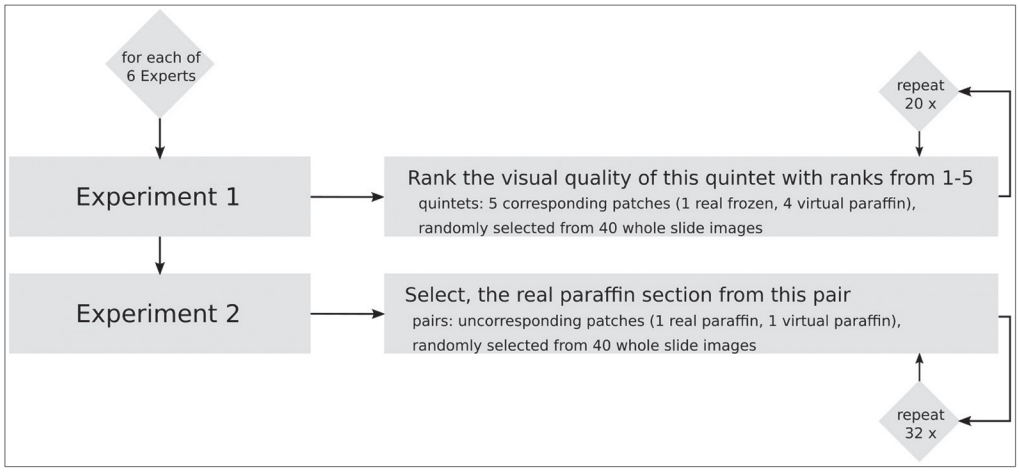

Figure 2: Overview of the two conducted experiments, performed by each of the six experts. In experiment 1, five corresponding images, containing one real and four virtual counterparts, were ranked. In experiment 2, the expert has to decide which image out of a non-corresponding pair is real

consisting of a real paraffin and a fake frozen-to-paraffin image (i.e., a frozen section translated to paraffin). These pairs were non-corresponding (as perfectly corresponding pairs cannot be generated). Overall, we assessed 32 of these pairs (eight for each image translation setting). Overall, with six pathologists, we obtained 32 × 6 = 192 comparisons. As for experiment 1, randomly selected 2500 × 2500 pixel patches showing majorly tissue (90%) were included.

Configurations

In this study, we made use of three different image translation models, namely CG, CUT, and Cycle-GAN, including the perceptual embedding consistency loss[24] (CG-PEC) to prevent tiling artifacts. These settings were chosen, as CG is a commonly used powerful technique and CUT is an enhancement, showing improved performance in case of ambiguous mappings between the domains. As both methods are prone to tiling artifacts, CG-PEC is a natural extension of CG. The perceptual embedding consistency loss requires two generators and can therefore only be applied in case of cycle-GAN. For all approaches, we used a learning rate of 0.0002 and trained for 10 epochs. For training, an unpaired data set consisting of 512 patches per WSI was created. Overall, 20 frozen and 20 paraffin WSIs were utilized, resulting in 23,552 patches overall. Further details are shown in Table 1.

Results

Figure 4 shows the outcome of experiment 1, individually for each image translation method and for each expert. The score indicates the portion of samples that were assessed as visually more appropriate than a real reference sample (i.e., the portion of translated images showing a lower rank than the corresponding reference images). Subfigure (a) shows the assessment of image translation compared with original samples (original = reference sample), whereas (b) shows the assessment of deep learning-based image translation compared with basic SN = reference sample. For example, the left top column contains the information whereby expert 1 assessed that the quality in case of SN in 40% of samples was higher than the quality of the original image. The mean (arithmetic mean) values can be similarly interpreted (e.g., in 57% of samples, over all experts, the quality of SN was assessed as higher compared with the original images).

In both subexperiments ((a) and (b)), CG, CUT, and CG-CUT showed similar mean scores of 0.59, 0.60, and 0.60 in case of (a) and 0.55, 0.50, and 0.55 in case of (b). The individual scores of the experts showed high variability ranging from 0.30 (CUT, E2) to 0.95 (CG-PEC, E3) in subexperiment (a) and from 0.40 (CUT, E5) to 0.80 (CGPEC, E3) in subexperiment (b). Scores averaged per expert were between 0.51 (E1) and 0.87 (E3) for subexperiment (a) and between 0.45 (E5) and 0.68 (E3) for subexperiment (b). To test for significance, we collected the four image translation approaches and the experts (a) and obtained a p-value of 0.008 (Wilcoxon signed-rank test). For subexperiment (b), significance could not be shown ($p > 0.05$).

From data obtained from experiment 2, the detection accuracy (real vs. fake sample) was computed. A score of 0.5 refers to random guessing, whereas 1.0 corresponds to perfect detection. Averaged over the experts, we obtained scores of 0.97, 0.78, and 0.62 for CUT, CG-PEC, and CG. Averaged over the methods, we obtained scores of 0.79.

Translating a patch with a size of 256 × 256 took 0.06 s on average on an NVIDIA RTX-2080. On an Intel Xeon Silver 4114 CPU 2.20GHz, we measured an average computing time of 0.14 s.

We did not notice stability issues during neural network training, such as mode collapse. Repeated training of CG and CUT with different initialization resulted in similar output images.

Discussion

The goal of this study was to assess how expert pathologists rate "fake" paraffin sections generated from real frozen sections. We noticed clear differences between the three investigated deep learning-based methods [Figure 1]. Although CUT led to high contrast and sharp details, it also showed the strongest tiling artifacts. In case of CG-PEC, tiling artifacts almost disappeared. However, contrast was slightly lower and the efficiency of translation was assessed as lower, that is, the image characteristics were easily distinguishable from those of real paraffin samples.

Table 1: Configurations of the three investigated deep learning-based image translation models

	CG[21]	CUT[22]	CG-PEC[21,24]
Patch size	256 × 256	256 × 256	256 × 256
Batch size	1	1	1
Regularization	Batch normalization	Batch normalization	Batch normalization
Generator(s)	U-Net	U-Net	U-Net
Discriminator	Patch-GAN	Patch-GAN	Patch-GAN
Weights	$\lambda cyc_X = \lambda cyc_Y = 10$ $\lambda_{id} = 0.5$	$\lambda_X = \lambda_Y = 1$ $\lambda_{GAN} = 1$	$\lambda cyc_X = \lambda cyc_Y = 10$ $\lambda embd_X = \lambda embd_Y = 10$ $\lambda_{id} = 0.5$

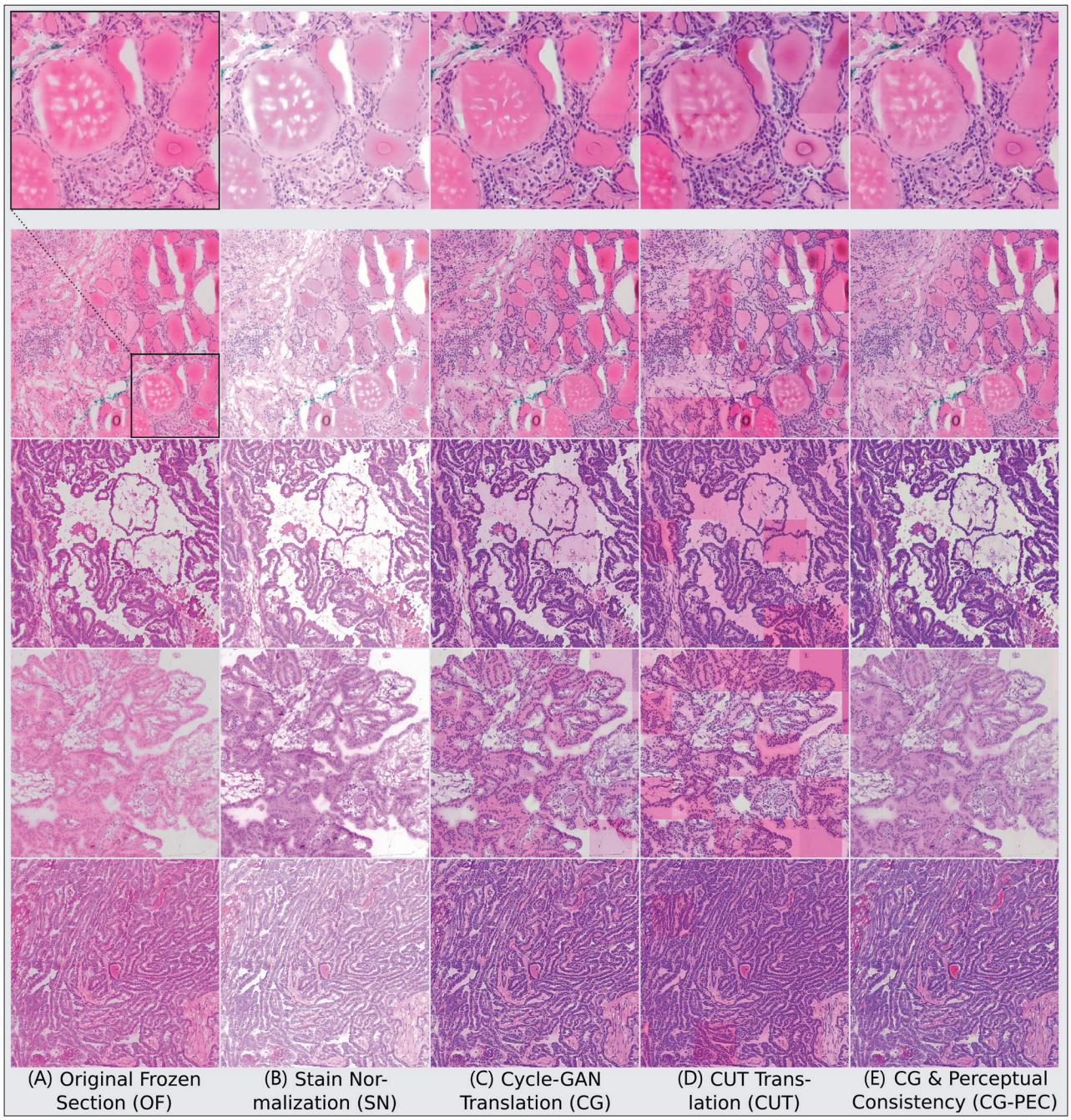

Figure 3: Corresponding tissue showing original frozen sections (A: OF) and virtually enhanced sections achieved with stain normalization (B: SN) and deep learning-based approaches (C: CG, D: CUT, E: CG-PEC). The top row shows a magnified version of the second row. Please zoom in to view the images in a higher resolution

Visually analyzed, CG delivers a good trade-off between CUT and CG-PEC, including slight tiling artifacts in combination with effective translation of underlying image characteristics. A comparison of three state-of-the-art deep learning-based techniques and a conventional stain-normalization method showed that perceived image quality can be improved on average. Overall (on average), we obtained improvements for each setting. However, we also noticed a clear difference between the pathologists. High inter-rater variability was observed and, in addition, also personal preferences with respect to the specific approaches could be observed. Interestingly, on average, all methods showed similar scores. Obviously, the experts do not automatically (although some experts do) downgrade images showing artifacts (as in case of CUT). Others even upgrade CUT, which was showing the strongest artifacts, probably due to the clearly visible image characteristics within the tiles.

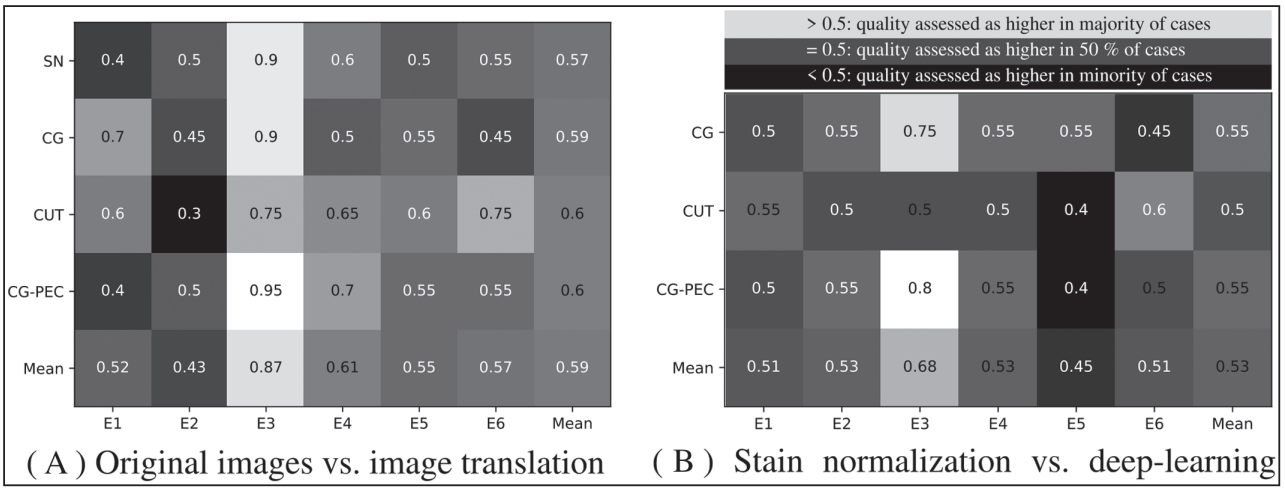

Figure 4: Experiment 1: Portion of samples, assessed as visually more appropriate than the original samples (A) and then the stain normalized samples (B), respectively. For example, from (A), we can extract that expert 3 (E3) assessed stain normalization (SN) as superior to the original image in 90% (0.9) of the cases

Strong tendencies were obtained in experiment 2. Interestingly, the fake images processed with CG were wrongly assessed as real samples in 38% of pairs, which comes close to random guessing. Even though this setting contains artifacts, it was hard to distinguish CG-altered images from real paraffin pathology. The CUT is supposed to be easily identified as fake due to the strong artifacts. The CG-PEC is probably not effective enough in changing the characteristics from the frozen to the paraffin modality. The generated sections still show more high similarity to original frozen sections (and not paraffin sections). These characteristics automatically reveal a fake image. The CG is, thus, proposed to deliver the best trade-off between improved image characteristics (providing that the image should look similar to paraffin sections) and a reasonable degree of tiling artifacts.

Finally, we need to refer to the aspect of inter-site variability. In our study, we focused on data from a single site only, to limit complexity by fixing this variable (*ceteris paribus*). As both CG and CUT automatically perform normalization, data from different institutions would probably be aligned. To optimize image quality in a multisite study, an important variable is given by the utilized training data. Making use of a data subset showing a particularly high visual quality as a target stain (here paraffin) has the potential to increase quality while decreasing variability.

A limitation of image translation approaches in general is that damaged or cracked tissue (which is more frequent in frozen sections) cannot be compensated by means of current approaches. Although local modifications such as contrast enhancement or color changes were performed well, geometric transformations (to compensate cracks) were not noticed. This is a result of the underlying neural network architectures in combination with the optimization criteria.

Shortcomings and follow-up clinical study

A follow-up clinical study is inevitable to measure the effect on diagnosis accuracy. This has not been performed here due to the following issues. First, the combination of a limited number of whole slide images and the relatively high accuracy of experts of about 95% complicate a statistically sound analysis of classification scores.[26] If on average 5% of images are misclassified by an expert, 2 out of 40 images are wrongly labeled. Accordingly, in case of an improvement of 20% (error rate), this would on average lead to a decrease of 0.4 misclassified images. For that reason, literature also suggests a minimum number of images for such a study of 60.[4] In addition, bias could be introduced, if both corresponding frozen and virtual-paraffin sections would be shown to the same experts (in any order). This reduces the number of frozen sections per expert from 40 to 20 and requires that (less powerful) unpaired tests are applied to test for significance. Alternatively, a 2-week washout could be applied to circumvent this procedure.[4] Anyway, based on these aspects, in combination with a performed sample size estimation, we concluded that such a study is not promising at that point of time with the available image material. In future, we strive toward achieving the additionally needed slides to conduct a diagnostic study.

Computational effort and practical considerations

Slide scanning time strongly depended on the size of the section. Scanning an individual section took on average roughly 30 min. For clinical application, however, there is a high potential to increase acquisition speed, which has high relevance in case of frozen sections. For example, decreased magnification would increase speed quadratically (due to two dimensional images). The lost resolution could be compensated with the investigated deep-learning techniques. In addition, the auto-focus approach can be adjusted to be clearly more efficient

(without the EFI setting). Based on feasible scanning times with 20× in the range of 1–2 min for an area of $10mm \times 10mm$,[27] we did not observe a strong limitation for future clinical application on frozen sections. The effect of more efficient imaging was not studied here. Focus was on obtaining optimum image quality without focusing on this optimization step to show whether the method is suitable in principle. For clinical application, experiments are needed to obtain the best trade-off between image quality and computational efficiency.

The training phase of image translation models requires powerful computation resources, whereas the clinically relevant inference phase is quite efficient. Fully convolutional networks, such as the used U-Net, ideally exploit the parallel computing potential of GPUs. However, modern CPUs also allow fast computation. A patch with a size of 256 × 256 pixels can be processed in 0.06 s on average on a modern powerful consumer GPU (Nvidia RTX-2080). The same task on a Workstation CPU (Intel Xeon Silver 4114 CPU 2.20GHz) takes 0.14 s on average. For an $8k \times 8k$ pixel whole slide images, the overall processing time increases to 66 s (GPU) and 140 s (CPU) in case of patch-wise processing. For a $16k \times 16k$ pixel whole slide images, the overall processing time increases to 264 s (GPU) and 559 s (CPU). For the latter realistic case, time consumption is in the range of 5 to 10 min, which is not negligible, but also not an insurmountable hurdle. A GPU is helpful, but not necessarily needed. To increase efficiency even further, dedicated architectures could be considered, optimizing the trade-off between complexity and performance.[28] Further potential for a speedup is given by decreasing the overall image resolution or by reducing the parameters of the architecture. Increasing the image resolution by a factor 2 in each direction results in an approximate speedup of factor 4. The source code for image translation is available and easily applicable to image patches (for trained models, please contact the corresponding author).[22]

Conclusions

In this work, we obtained an overview of the principal assessment of fake histological images, optimized based on deep-learning techniques. Quantitative evaluation identified an extraordinarily high variability between pathologists. Particularly unnatural tiling artifacts, which vary between different techniques, have the potential to reduce acceptance and reveal virtual samples. Overall, improved image characteristics outweighted the effect of tiling artifacts, at least up to a certain extent. Average scores showed that deep learning-based image translation was rated higher, compared with both original frozen sections and stain-normalized images. The standard CG is assessed here as the best-suited method that provides a good trade-off between effective translation and limited artifacts.

It is important to state that this study did not investigate the impact on diagnostic performance. To make sound statements on the improvement of these scores, a large clinical setting including a larger data set is needed. The obtained knowledge of principally high acceptance of pathologists provides motivation for such a clinical study and informs about promising approaches.

Financial support and sponsorship

This work was partially funded by the County of Salzburg under grant number FHS2019-10-KIAMed and by the Austrian Agency for International Cooperation in Education and Research (OeAD-GmbH) HR02/2018.

Conflicts of interest

There are no conflicts of interest.

References

1. Eccher A, Fontanini G, Fusco N, Girolami I, Graziano P, Rocco EG, *et al*. Digital slides as an effective tool for programmed death ligand 1 combined positive score assessment and training: Lessons learned from the "programmed death ligand 1 key learning program in head-and-neck squamous cell carcinoma". J Pathol Inform 2021;12:1.
2. Girolami I, Pantanowitz L, Marletta S, Brunelli M, Mescoli C, Parisi A, *et al*. Diagnostic concordance between whole slide imaging and conventional light microscopy in cytopathology: A systematic review. Cancer Cytopathol 2020;128:17-28.
3. Eccher A, Girolami I. Current state of whole slide imaging use in cytopathology: Pros and pitfalls. Cytopathology 2020;31:372-8.
4. Evans AJ, Brown RW, Bui MM, Chlipala EA, Lacchetti C, Milner DA, *et al*. Validating whole slide imaging systems for diagnostic purposes in pathology: Guideline update from the college of American pathologists in collaboration with the American Society for Clinical Pathology and the Association For Pathology Informatics. Archives of Pathology & Laboratory Medicine 2021. https://pubmed.ncbi.nlm.nih.gov/34003251/
5. Halicek M, Shahedi M, Little JV, Chen AY, Myers LL, Sumer BD, *et al*. Head and neck cancer detection in digitized whole-slide histology using convolutional neural networks. Sci Rep 2019;9:14043.
6. 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;38:2293-302.
7. Bentaieb A, Hamarneh G. Adversarial stain transfer for histopathology image analysis. IEEE Trans Med Imaging 2018;37:792-802.
8. Hou L, Samaras D, Kurc TM, Gao Y, Davis JE, Saltz JH. Patch-based convolutional neural network for whole slide tissue image classification. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016;2016:2424-33.
9. Deng S, Zhang X, Yan W, Chang EI, Fan Y, Lai M, *et al*. Deep learning in digital pathology image analysis: a survey. Front Med 2020;14:470-87.
10. Wolterink JM, Dinkla AM, Savenije MHF, Seevinck PR, van den Berg CAT, Isgum I. Deep MR to CT synthesis using unpaired data. In: Proceedings of the International MICCAI Workshop Simulation and Synthesis in Medical Imaging (SASHIMI'17). 2017:14-23.
11. Gadermayr M, Tschuchnig M, Merhof D, Kramer N, Truhn D, Gess B. An asymmetric cycle: Consistency loss for dealing with many-to-one mappings in image translation: A study on thigh MR scans. In: Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI). 2021:1182-6.

12. Cohen JP, Luck M, Honari S. How to cure cancer (in images) with unpaired image translation. In: Proceedings of the Conference on Medical Imaging with Deep Learning (MIDL'18). 2018.
13. Gadermayr M, Tschuchnig M, Stangassinger LM, Kreutzer C, Couillard-Despres S, Oostingh GJ, *et al*. Frozen-to-paraffin: Categorization of histological frozen sections by the aid of paraffin sections and generative adversarial networks. In: Proceedings of the 6th International MICCAI Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI). 2021:99-109.
14. Najah H, Tresallet C. Role of frozen section in the surgical management of indeterminate thyroid nodules. Gland Surg 2019;8:112-7.
15. Osamura RY, Hunt JL. Current practices in performing frozen sections for thyroid and parathyroid pathology. Virchows Arch 2008;453:433-40.
16. Huber GF, Dziegielewski P, Matthews TW, Warshawski SJ, Kmet LM, Faris P, *et al*. Intraoperative frozen-section analysis for thyroid nodules. Arch Otolaryngol Head Neck Surg 2007;133:874.
17. Leteurtre E, Leroy X, Pattou F, Wacrenier A, Carnaille B, Proye C, *et al*. Why do frozen sections have limited value in encapsulated or minimally invasive follicular carcinoma of the thyroid? Am J Clin Pathol 2001;115:370-4.
18. Udelsman R, Westra WH, Donovan PI, Sohn TA, Cameron JL. Randomized prospective evaluation of frozen-section analysis for follicular neoplasms of the thyroid. Ann Surg 2001;233:716-22.
19. Pantanowitz L, Farahani N, Parwani AV. Whole slide imaging in pathology: Advantage, limitations, and emerging perspectives. Pathol Lab Med Int 2015;1.
20. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer Aided Interventions (MICCAI'15). 2015:234-41.
21. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the International Conference on Computer Vision (ICCV'17). 2017:2223-32.
22. Park T, Efros AA, Zhang R, Zhu JY. Contrastive learning for conditional image synthesis. In: Proceedings of the European Conference on Computer Vision (ECCV'20). 2020.
23. Zhang Y. XOGAN: one-to-many unsupervised image-to-image translation. ArXiv 2018. abs/1805.07277 abs/1805.07277.
24. Lahiani A, Navab N, Albarqouni S, Klaiman E. Perceptual embedding consistency for seamless reconstruction of tilewise style transfer. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'19). 568-76.
25. Reinhard E, Ashikhmin M, Gooch B, Shirley P. Color transfer between images. IEEE Comput Graphics Appl 2001;21:34-41.
26. Mandell DL, Genden EM, Mechanick JI, Bergman DA, Biller HF, Urken ML. Diagnostic accuracy of fine-needle aspiration and frozen section in nodular thyroid disease. Otolaryngol Head Neck Surg 2001;124:531-6.
27. Rojo MG, García GB, Mateos CP, García JG, Vicente MC. Critical comparison of 31 commercially available digital slide systems in pathology. Int J Surg Pathol 2006;14:285-305.
28. Beheshti N, Johnsson L. Squeeze u-net: A memory and energy efficient image segmentation network. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE 2020:1495-504.