Image-to-image translation for automatic ink removal in whole slide images

Jun Jiang,[a] Naresh Prodduturi,[a] David Chen,[a] Qiangqiang Gu,[a] Thomas Flotte,[a] Qianjin Feng,[b] and Steven Hart[a,*]

[a]Mayo Clinic, Health Science Research Department, Rochester, United States
[b]Southern Medical University, Guangzhou, China

Abstract

Purpose: Deep learning models are showing promise in digital pathology to aid diagnoses. Training complex models requires a significant amount and diversity of well-annotated data, typically housed in institutional archives. These slides often contain clinically meaningful markings to indicate regions of interest. If slides are scanned with the ink present, then the downstream model may end up looking for regions with ink before making a classification. If scanned without the markings, the information regarding where the relevant regions are located is lost. A compromise solution is to scan the slide with the annotations present but digitally remove them.

Approach: We proposed a straightforward framework to digitally remove ink markings from whole slide images using a conditional generative adversarial network based on Pix2Pix.

Results: The peak signal-to-noise ratio increased 30%, structural similarity index increased 20%, and visual information fidelity increased 200% relative to previous methods.

Conclusions: When comparing our digital removal of marked images with rescans of clean slides, our method qualitatively and quantitatively exceeds current benchmarks, opening the possibility of using archived clinical samples as resources to fuel the next generation of deep learning models for digital pathology.

 [DOI: 10.1117/1.JMI.7.XX.XXXXXX]

Keywords: ink removal; whole slide image; image to image translation.

Paper 20033R received Feb. 13, 2020; accepted for publication Sep. 21, 2020.

1 Introduction

The development of computer-aided diagnostics systems for pathology is a popular area of research due to the advent of deep learning. Deep learning can be used to automate many aspects of pathological medical diagnosis including bacterial recognition, cell type classification, and cell segmentation and counting.[1–3] Although there have been several breakthroughs in the field, there are still significant problems that hinder widespread adoption of AI in digital pathology.

Perhaps the most critical limitation is the need for clean, annotated data, which is the prerequisite to enable the training of deep neural networks. Although many healthcare institutions have large archives of physical slides, most of these slides have not been digitized as whole slide images (WSIs). Further hindering their use for training deep learning-based diagnostic system is that they are marked with ink from felt tip pens, which may introduce artifactual bias or block informative tissue features from the model training. Some WSIs also contain handwritten protected health information, which dramatically increases the difficulty in sharing that example within and among institutions.

*Address all correspondence to Hart Steven, E-mail: Hart.Steven@mayo.edu

2329-4302/2020/$28.00 © 2020 SPIE

Ideally, slides would be scanned before any physical markup, but this cannot be applied to a large number of historical slides and may cause prohibitive delays and disruptions to the existing pathology workflow. Alternatively, pen marks can be chemically removed, but this process risks damaging the tissue underneath.[4] Instead, we propose using a generative adversarial network (GAN) for image-to-image translation to digitally remove the on-slide annotations.

Image-to-image translation techniques have been used in other fields such as image restoration, image enhancement, and style transfer.[5–8] Digitally removing ink annotations would significantly increase the size of useable training data. Furthermore, such pipelines could be easily integrated into existing pathology workflows without major disruptions.

Several methods have been proposed to attempt ink removal from WSIs. These methods can be classified into three groups: image inpainting, color deconvolution, and image-to-image translation.

1.1 *Image Inpainting*

Image inpainting is designed to replace corrupt or damaged pixels with those of their neighbors and is used in photorestoration and computational photography.[6] In WSIs, inked regions can be segmented and converted to missing data, which is then filled in using the image inpainting technique. However, image inpainting assumes that missing areas can be retrieved from somewhere in background regions or image database with similar textures.[9] Even though data underneath the marker could provide information as to the histologic organization, it is ignored and new data are generated. The generated tissue could be dramatically different from the actual tissue covered by ink, which may lead to diagnostic errors.

1.2 *Color Deconvolution*

Color deconvolution is one of the most commonly used color normalization methods in diagnostic bright-field microscopy images.[10] By assuming a linear relation between stain concentration and absorbance,[11] these methods share the same idea that color images of multiple stained biological samples can be transformed into images representing the stain concentrations. However, this assumption is only valid under monochromatic conditions because the nonlinear characteristics of the absorbance formation may lead to significant deconvolution errors.[12] We have also found limitations based on the color of ink used.[13]

1.3 *Image-to-Image Translation*

The image-to-image translation is a class of vision and graphics problems with the goal of learning the mapping between an input image and an output image. This technique is commonly used to transfer art styles between images,[14] create hyper-realistic expressions on photographs of people,[15] or restore damaged images.[6] In the biomedical domain, image-to-image translation has been used to generate different contrast MRI images to enrich the dataset for model training[16] and denoise under-sampled images to create high-resolution CT images.[17] These techniques have also been applied to color normalization for pathological images. For example, a GAN-based stain-style transfer model was proposed by Cho et al.,[7] to normalize Hematoxylin and Eosin (H&E) stains across patients and institutes.

There are some stipulations that limit its application to ink removal. First, image transformation models may fail to preserve biological structure if trained with a set of representative reference images rather than images with the exact same content.[5] Second, compared with relatively large WSIs, pen markings only cover small areas. Arbitrarily applying normalization models to the entire slide may lead to pervasive color changes.

To overcome this problem, a divide-and-conquer strategy was proposed by Ali et al.[18] in which classification, detection, and image generation models were concatenated into a patch processing pipeline. In this workflow, convolutional neural networks (CNNs) were trained to differentiate image patches based on the content of image patches, including ink-only (background with ink on it), tissue-only (tissues without ink on it), and inked-tissue (tissue covered by ink). Next, a Yolov3[19] ink detection model was trained to locate the ink area only on patches

with tissue and ink, followed by a cycle-GAN to color normalize inside the bounding box. Unfortunately, several limitations exist using the logic of this approach. First, image patches can be combinations of different image contents, which mean misclassification of one step will propagate into the following steps, leading to more error in the final result. Second, the output of ink detection was bounding boxes, but the shape of the ink could be nonrectangular, leading to over or under inclusiveness of tissues for normalization. Also, the cycle-GAN model was trained using two different data distribution (dense and sparse) pairs based on cell or cytoplasmic mass clusters, which does not account for uneven ink thickness in a single patch. Most important of all, this approach would be difficult to generalize because each step of this workflow requires a large number of annotations, which will be tedious if the model needs to be retrained to adopt different ink colors.

In this paper, we propose developing a deep learning model using conditional GANs (Pix2Pix) to automatically remove the on-slide annotations. This end-to-end model was trained using a unique dataset of clean (before pen marking) and marked (after pen marking) WSIs, thereby constraining the model only to normalize color rather than synthesizing tissue. Our contributions can be summarized as follows.

(1) A straightforward framework was proposed to automatically remove the on-slide annotations without an additional classification or segmentation step. This eliminates potential sources of error that may propagate to the final de-inked image.
(2) Introduction of patch blending to the restoration phase suppresses minor color discrepancies (i.e., edge-effects) between patches.
(3) Holistic evaluation of our model based on different image contents (inked-tissue, tissue-only, and ink-only) rather than indiscriminate patches provides a more real-work evaluation of the algorithm's performance.

2 Method

In brief, pairs of inked and clean WSI were co-registered to remove any difference offset and then grouped into pairs of image patches pairs. A Pix2Pix model was then trained for the image-to-image translation task using the registered pairs of inked and clean patches. In the testing phase, a similar patch extraction method was reapplied to marked WSIs, and the trained model was used to restore all of the image patches. Then the restored patches were blended to remove stitching effects and reconstructed back into WSIs. The process is described in Fig. 1.

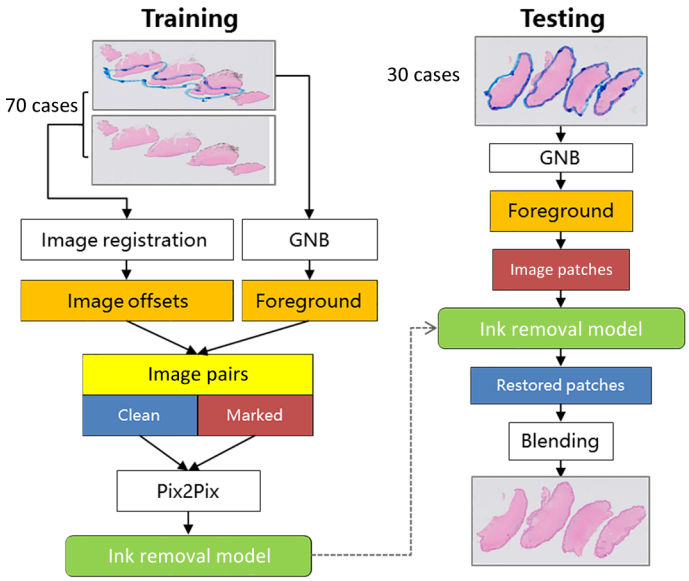

Fig. 1 Overview of the workflow.

2.1 *Pairwise Patch Extraction*

Only informative patches (patches with ink or tissue) were included in our research to maximize efficiency. The marked WSIs were down-sampled by a ratio of d_r to provide manageable image sizes. A binary foreground mask of the down-sampled WSI, in which foreground pixel locations correspond to coordinates from where informative patches with original resolutions can be extracted, was introduced. With manually annotated pixels, a Gaussian Naïve Bayes (GNB) model was trained to detect the foreground in down-sampled WSIs. This foreground detection strategy was considered to be robust to tissue brightness variation and artifacts, as it is shown in a quality control tool for digital pathology slides.[20–22]

To establish pixel correspondence, marked and inked slide pairs were registered to minimize any potential offset caused by the redigitization process using a modified rigid method that leverages the hierarchical nature of WSI.[23] With both image offset and foreground detection results, informative image pairs can be extracted.

Our ink removal model takes all of the foreground patches as input for both training and testing processes, which was different from the current benchmark in which only patches with ink were involved in color normalization.[18] To evaluate our model performance on patches with different image contents, a held-out testing dataset was annotated with QuPath.[24] The foreground of the testing dataset was divided into three categories based on image content: inked-tissue (tissue covered by ink in the patch), tissue-only (clean tissue in the patch), and ink-only (only ink in the patch). Annotations were parsed and saved to multilabel masks with the same down-sampling ratio d_r. The coordinate of where the patch was extracted was used to locate the corresponding pixel in the annotation masks, so the label (image content) of each patch could be determined.

2.2 *Image-to-Image Translation cGAN (Pix2Pix)*

Like all GANs, cGAN consists of a generator and a discriminator (Supplemental Fig. S1). The generator G is trained to generate realistic-looking data that mimic the input data. The job of the discriminator D is to predict if the data it is being shown come from a real or synthetic data set. There are many architectures available for cGANs,[14,24,25] but the Pix2Pix architecture[8] (in which a "U-Net" with skip connections between mirrored layers in the encoder and decoder was employed for the generator) was selected since it has been shown to be effective at capturing macroscopic and microscopic details in images.[26,27] However, unlike a GAN trained for common color normalization tasks, both the generator and discriminator observe both the marked (s: source) and clean (t: target) image. Denoting the input noise of G as z, the objective of our task is expressed as

$$\mathcal{L}_{c\mathrm{GAN}}(G,D)=\mathbb{E}_{s,t}[\log\ D(s,t)]+\mathbb{E}_{s,z}(\log\{1-D[s,G(s,z)]\}),$$

where G tries to minimize this objective against an adversarial D that tries to maximize it, i.e.,

$$G^*=\arg\ \min_G \max_D \mathcal{L}_{\mathrm{cGAN}}(G,D).$$

Since previous works have found that using $L1$ distance encourages less blurring than $L2$,[28] our final objective is

$$G^*=\arg\ \min_G \max_D \mathcal{L}_{\mathrm{cGAN}}(G,D)+\lambda\mathcal{L}_{L1}(G),$$

where λ is a weighting parameter and

$$\mathcal{L}_{L1}(G)=\mathbb{E}_{s,t,z}[\|t-G(s,z)\|_1].$$

There are different perspectives on the input noise z of the generator. The simplest way is to add Gaussian noise to the input image.[29,30] However, some investigators have pointed out that the generator simply learned to ignore the noise if GAN is trained in this way.[8,31] Instead, noise was provided only in the form of dropout; it was applied on several layers of our generator at both training and testing time, as has been done before Ref. 8.

2.3 *Seamless Patch Blending*

Since the ink removal model was trained with marked and clean image pairs and no patch level neighborhood information was learned by our model, seam artifacts could be observed if the restored image patches were directly stitched back into a WSI (Fig. 6). To make the restored WSI visually realistic, seam artifacts should be eliminated.

Using larger patches for training may alleviate artifacts, but may lose some image details, while using small patches may promote sharpness, but suffer from tiling artifacts.[8] To balance this trade-off, Isola et al.[8] tried different patch sizes and retrained models to get the optimal performance. Instead of looping through a large number of different patch sizes, we introduced a simple but effective alpha blending strategy.[32]

Assume the restoration of a marked WSI was a $J \times K$ patch matrix, in which each element was an image patch (denoted by $P_{j,k}$). Image patches in this matrix were partially overlapping, and the overlapping area is noted as $P^O_{j,k}$.

In the horizontal direction, the blending result of two adjacent patches is formulated as

$$P^O_{j,k} = P^O_{j,k} * M_H + P^O_{j,k+1} * (1 - M_H),$$

where M_H denotes a horizontally gradual changing blending matrix, which is generated by repeating V_h for P_h times:

$$V_h = \left[0 : \frac{1}{O_w} : 1\right],$$

$$M_H = \left[\begin{Bmatrix} V_h \\ \vdots \\ V_h \end{Bmatrix} P_h\right],$$

where O_w denotes the width of overlapping area and P_h denotes the height of patch size.

Blended image patches were concatenated into image row R_j. Here we used $\|$ to denote the concatenation of image patches.

$$R_j = \Big\|_{k=0}^{K-1} P^O_{j,k}.$$

Similarly, image rows are blended with a vertically gradual changing blending matrix M_V, and the final result of seamless patch blending is noted as I

$$R^O_j = R^O_j * M_V + R^O_{j+1} * (1 - M_V),$$

$$I = \Big\|_{j=0}^{J-1} R^O_j,$$

where M_v denotes a vertically gradual changing blending matrix, which is the transpose of a matrix generated by repeating V_v for I_w times:

$$V_v = \left[0 : \frac{1}{O_h} : 1\right],$$

$$M_V = \left[\begin{Bmatrix} V_v \\ \vdots \\ V_v \end{Bmatrix} I_w\right]^{\mathrm{T}},$$

where O_h denotes the height of an overlapping area and I_w denotes the width of a restored image region (Fig. 2).

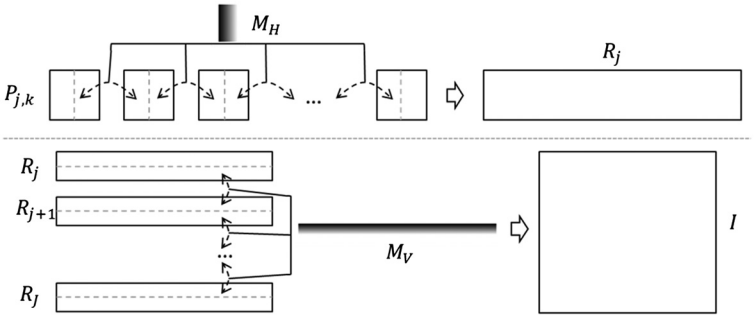

Fig. 2 Schematic of alpha patch blending. Adjacent patches were first blended with a horizontally gradual changing blending matrix, so image rows R_j could be generated. Then image rows were blended with a vertically gradual changing blending matrix, so final image I could be (optionally) reconstructed.

3 Experiment

3.1 *Data Preparation*

We compiled a set of 100 WSIs of skin tissue from routine biopsy cases. The WSIs were marked using blue and black markers. These were digitized using a Phillips Ultra Fast Scanner at a resolution of 0.25 μm/pixel with images exported as TIFF files (around 2 GB each). Following the initial digitization, we then chemically removed the ink and digitized the cleaned slides using the same scanner settings. The scanned clean and marked WSI pairs were randomly split into training ($n = 70$) and blind testing sets ($n = 30$). Each pair of WSIs was matched with an automatic rigid image registration method proposed in our previous work.[23] To extract informative image patches, tissues (foreground) were detected in down-sampled WSIs since the original WSIs were too large to be manipulated. The tissue detector (GNB) was trained with 30 marked images, which were annotated within QuPath.[33] Aimed at reusing the annotation in evaluating the ink removal model performance on different image contents in the testing phase, the foregrounds were labeled with "ink-only," "tissue-only," and "inked-tissue."

Foreground pixel detection was applied to a down-sampled WSI to identify regions where tissues were located. Down-sampling rates that are too high may have many blank patches. Too low of a down-sampling rate leads to low efficiency in downstream processing. We tried 32, 64, 128, and 256 down-sampling rates and found that 128 ($d_r = 1/128$) is the optimal for this dataset. The image patch size was set to 256×256. All of the image patches were extracted from the highest resolution of WSI pairs to maintain a high quality of restoration.

3.2 *Model Training*

Marked and clean image pairs were fed into a cGAN (Pix2Pix) to train the ink removal model. Training parameters were similar to Isola,[8] with a minibatch stochastic gradient descent[34] and the Adam solver,[35] with a learning rate of 0.0002 and momentum parameters $\beta_1 = 0.5$ and $\beta_2 = 0.99$. The weight of $L1$ loss of generator λ was set to 100. Implementation of our cGAN was borrowed from https://github.com/affinelayer/pix2pix-tensorflow. All of the code of our workflow can be found in a Github repository (https://github.com/smujiang/WSIPenMarkingRemoval).

Unfortunately, evaluating the convergence of a GAN is still an open problem. To prevent model collapse,[36] several preventive measures were adopted: first, an extremely large dataset (6,702,621 pair of image patches) was used for training, which greatly increased the data diversity. Second, checkpoints of generator and training curves were saved every 5000 steps. The maximum training epoch was set to 300, and the training process was stopped when the generator loss curve turned to flat and images reached acceptable image restoration according to qualitative assessment. The trained model can be downloaded from Google Drive.[37]

3.3 *Evaluation*

The quality of image normalization was evaluated quantitatively on a per-patch basis on the test set. Due to the complexity and subjectivity of the color normalization task, a combination of metrics was employed: peak-signal-to-noise ratio (PSNR),[38] structural similarity index (SSIM),[39] and visual information fidelity (VIF).[40] These metrics measure different aspects of model performance. Specifically, PSNR computes the mean squared error after restoration, which is slightly biased toward over smoothed results.[31] SSIM has been developed to take the similarity of the edges (high-frequency content) into account, making it more sensitive to structural changes. VIF was designed with information-theoretic settings and correlated well with human judgments of visual quality.[41] We calculated each metric before ("inked") and after ("restored") image restoration. We compared the performance of our proposed model against the method proposed by Ali et al.[18] The performance metrics were measured individually for each patch content type.

4 Results

4.1 *Model Training*

Since the ink removal model takes all foregrounds as input, unfocused areas, tissue damage, and other registration exceptions can encumber the performance of the model. The model reaches the highest performance when the benefits of data variance and the errors introduced by defective dataset get balanced. To estimate the optimal dataset size, models were trained with 10, 30, 50, and 70 pairs of WSIs with identical parameters. According to the evaluation curves of PSNR, SSIM, and VIF, our ink removal model got the highest performance when trained with 50 cases. As shown in Fig. 3, model performance increased from 10 cases to 30 cases but began to decline after 50 cases.

4.2 *Qualitative Assessment*

To qualitatively assess the performance of the model, restoration results of image patches containing different contents (i.e., dark marker, background, and no marker) were visually compared. As shown in Fig. 4, regardless of the tissue being covered by dark (patch 1) or light (patch 2) ink, or the color of ink being blue (patches 1 and 2) or black (patch 3), both Pix2Pix and the benchmark method of Ali et al.[18] can effectively remove pen marking in patches. However, dramatic differences can be observed in the images. The benchmark tends to enhance the edges of tissue and markedly changes the stain intensity of nuclei. These artifacts are also introduced into unmarked patches (patch 5), whereas our proposed method does not produce significant image alterations. The most striking result was in images that contain only inked-background. In this example (patch 4), a "hallucination" of tissue appears, even though there is no tissue located in the cleaned or marked image. In general, the Pix2Pix approach described here produces fewer qualitative artifacts than that of Ali et al.[18] regardless of whether the ink exists or not.

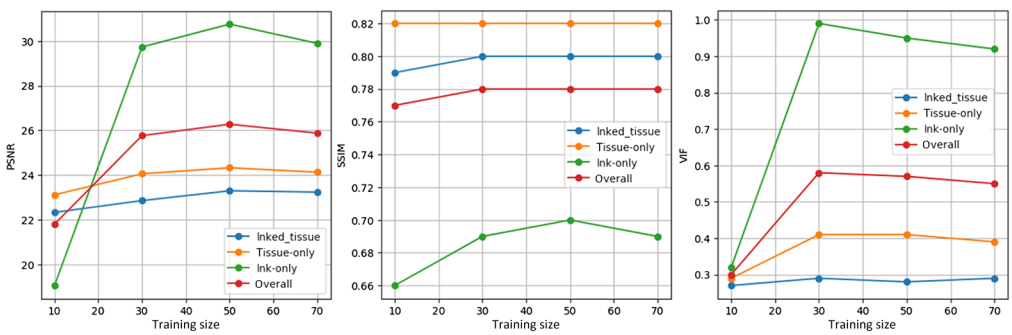

Fig. 3 Evaluation scores (PSNR, SSIM, and VIF) of models trained with increasing training sizes. Four curves (tissue_only, inked_tissue, ink_only, and overall) were included in each plot.

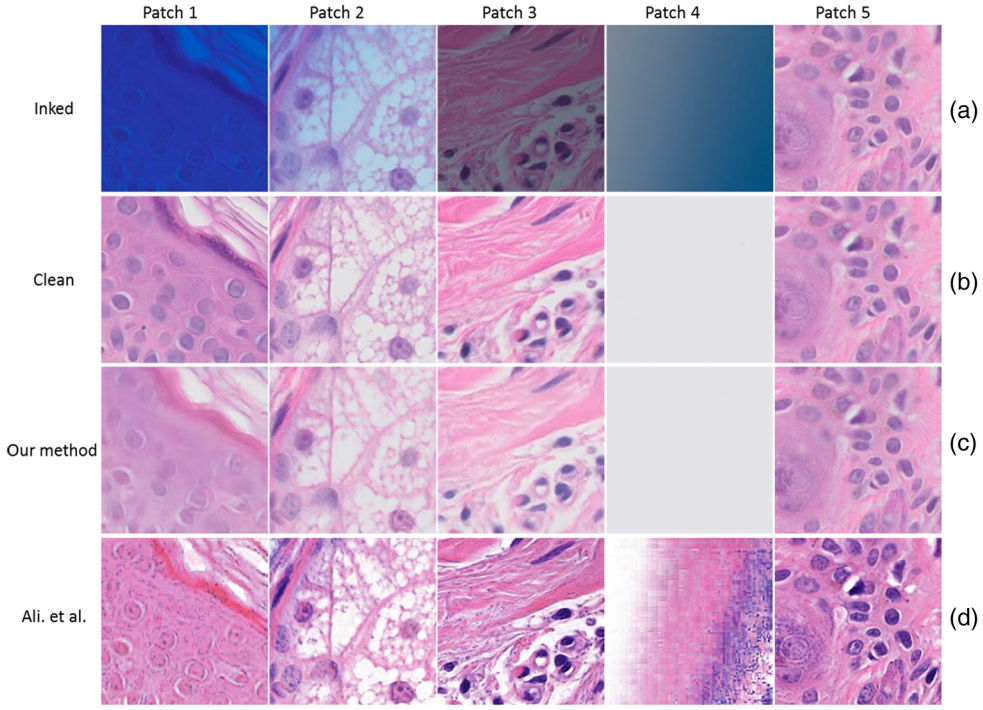

Fig. 4 Examples of image patch restoration. Each column represents an image patch containing different image contents. (a) Inked image patches from marked slides. Specifically, patches 1 to 3 represent image samples covered by dark blue, light blue, and light black, respectively; patch 4 is an ink-only image patch, and patch 5 is an example without ink. (b) The corresponding image patches from a clean slide. (c) The restored image patches from our model. (d) For comparison, we apply the restoration protocol from Ref. 18. Additional examples are provided in Fig. 7.

4.3 *Quantitative Validation*

To have a comparable quantitative evaluation, average and standard deviation of PSNR, SSIM, and VIF were compared with the benchmark method. For each model to be evaluated, metrics for measuring differences between each pair of clean images and the original marked image were calculated and denoted as "inked," whereas differences between clean images and restored images were calculated and denoted as "restored."

4.3.1 *Image content evaluation*

By comparing the evaluation metrics (PSNR, SSIM, and VIF) before and after ink removal, we can assess the qualitative improvement of image restoration. To evaluate the performance of the Pix2Pix model on different image contents, patches were divided into three groups: including inked-tissue, tissue-only, and ink-only. Metrics were stratified with respect to the image content. As shown in Table 1, for inked-tissue and ink-only patches, PSNR, SSIM and VIF significantly increased after image restoration, which indicates that the outputs of our model are more similar to the clean slides. Using the benchmark of Ali et al.,[18] we actually observed a decrease of image fidelity after cleaning, compared with our increase. For the tissue-only patches, both PSNR and SSIM slightly increased, but VIF changed only slightly in our model, supporting our claim that content remains relatively unchanged when ink is not present in an image.

To discover more characteristics of the Pix2Pix model, violin plots were introduced to enrich our assessment. As shown in Fig. 5, significant details can be observed. First, the distributions of all evaluation metrics for tissue_only patches are symmetrical, which indicates that our model preserves the image content if no ink presents. Second, for patches with ink (inked_tissue and ink_only), all three evaluation metrics move to higher values, which indicates that ink was effectively removed. In particular, for image patches with ink, SSIM distributions were dispersive before ink removal (which may be caused by uneven thickness of ink), but aggregate to higher values, which indicates that inked image patches were effectively restored.

Table 1 Evaluation metrics on three types of image content, including inked-tissue (denotes there is tissue covered by ink in image patches), tissue-only (denotes there is only clean tissue in image patches), and ink-only (denotes there is only ink in image patches). "Inked" denotes the metrics measuring the differences between the clean image and the marked image, whereas "restored" denotes the metrics measuring the differences between the clean image and the restored image.

		PSNR		SSIM		VIF	
	Patch content	Inked	Restored	Inked	Restored	Inked	Restored
Ali et al., 2019[a]	Inked_tissue	14.57 ± 4.35	17.06 ± 2.19	0.46 ± 0.26	0.66 ± 0.12	0.17 ± 0.18	0.13 ± 0.05
Our method	Inked_tissue	14.56 ± 4.37	22.51 ± 3.98	0.46 ± 0.26	0.79 ± 0.12	0.17 ± 0.18	0.26 ± 0.11
	Tissue_only	22.91 ± 6.06	23.62 ± 5.61	0.81 ± 0.18	0.84 ± 0.12	0.27 ± 0.17	0.27 ± 0.14
	Ink_only	14.31 ± 4.82	21.80 ± 5.38	0.45 ± 0.27	0.78 ± 0.16	0.18 ± 0.18	0.26 ± 0.26

[a]According to the workflow of benchmark, only include inked-tissues were fed into the patch restoration model.

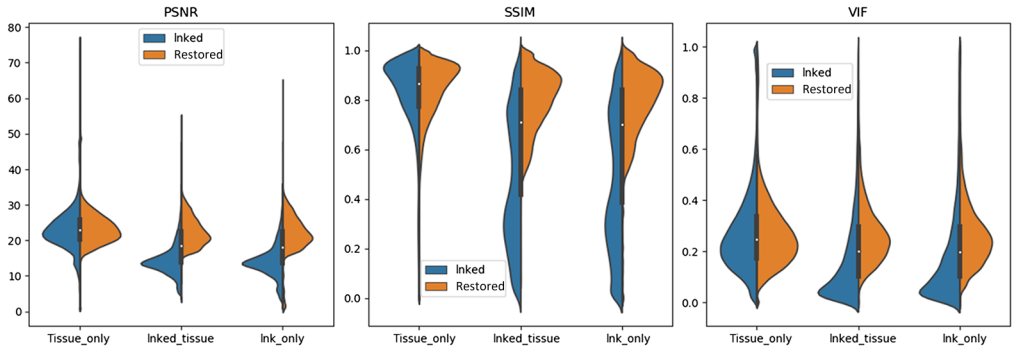

Fig. 5 Violin plots (PSNR, SSIM, and VIF) for model assessment on different image contents. Three different image contents (tissue_only, inked_tissue, and ink_only) were included in each plot. Blue parts indicated the distributions of evaluation metrics calculated before ink removal (inked). Orange parts indicated the distributions of evaluation metrics calculated after ink removal (restored).

4.4 *Postprocessing*

One important part that the previous benchmark did not formally recognize is the introduction of tiling artifacts after image restoration. As shown in the middle of Fig. 6, tiling artifacts can be obvious if the restored patches were directly stitched back into WSIs. The reason could be that our ink removal model did not take into account neighborhood information, from where the uniform hue, saturation, and brightness were expected to be learned. As shown in Fig. 6(a), with our patch blending strategy, texture information near patch edges can be used from adjacent tiles, and no tiling artifacts can be observed.

We also quantitatively evaluated the image quality with and without the patch blending strategy. Evaluations were conducted on the ROI level since we have difficulty writing processed image patches into large WSI due to the limitation of image IO packages. We arbitrarily selected 25 ROIs from 5 cases (2 with black ink and 3 with blue ink), and each ROI (size: 384×384 pix) was circulated from inked-tissue (tissue covered by ink). Images (including image before restoration, image generated by direct stitching, and image generated by alpha blending) were compared with their clean counterparts (extracted from clean WSIs) to calculate the evaluation metrics. The averaged evaluation metrics and the standard deviation are shown in Table 2, from which we observe notable improvement in all of the metrics after ink removal even if the patches were directly stitched together, and applying patch blending could further improve image quality, especially from PSNR aspect.

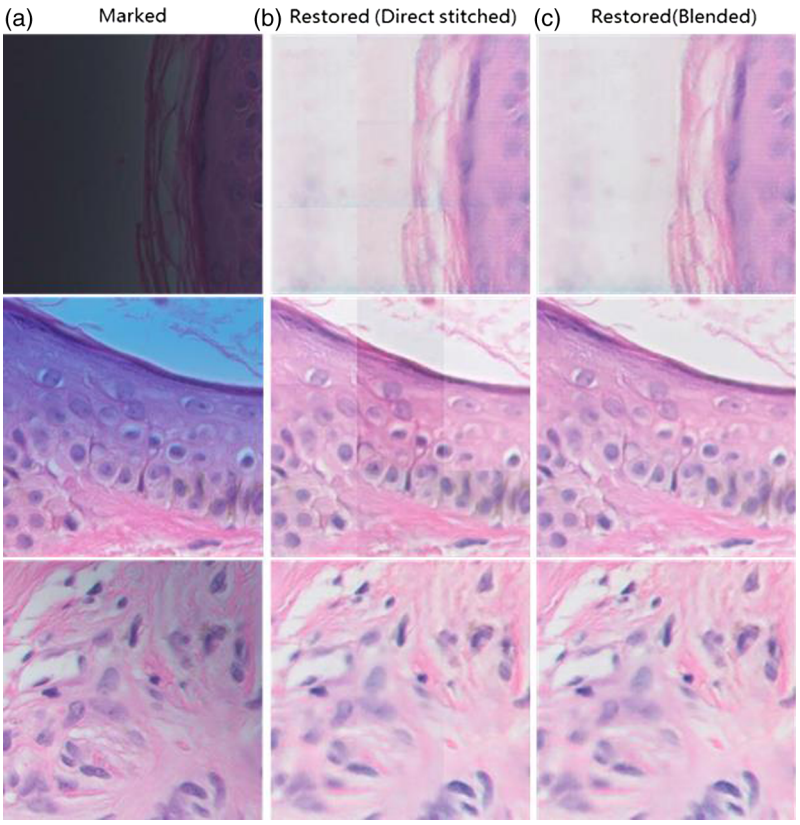

Fig. 6 Visualization of patch blending, each row is an example: (a) an inked region of marked WSI; (b) reconstructed region by direct stitching restored image patches; and (c) reconstructed region by blending restored image patches.

Table 2 Evaluation metrics variation with versus without patch blending. Values (mean and std) in each cell were the statistic of evaluation metrics (PSNR, SSIM, and VIF), which were calculated by comparing image samples with their counterparts. The first row denotes comparing between inked image (image before restoration) and clean image; the second row denotes comparing between directly stitched restored image (without patch blending) and clean image; and the third row denotes comparing between restored image (with patch blending) and clean image.

	PSNR	SSIM	VIF
Before restoration	11.955 ± 4.65	0.479 ± 0.24	0.143 ± 0.11
Direct stitched	22.600 ± 2.52	0.840 ± 0.08	0.226 ± 0.08
Blended	22.647 ± 2.51	0.840 ± 0.08	0.227 ± 0.08

5 Discussion

In this work, we described an end-to-end model for on-slide annotation removal using a Pix2Pix model. The proposed workflow quantitatively yielded better results compared with prior models. This is particularly noticeable in patches with ink-only and inked-tissue patches, where the Ali et al. method[18] often generated synthetic textures or tissues that are clearly not present in the underlying slide. This aspect of image integrity is particularly important to medical imaging, in which confidence in the diagnostic and prognostic power of a test is strongly tied to the perceived integrity.

The proposed model has significant advantages over prior work in both architecture and training data. Zanjani et al.,[42] proposed a similar concept by defining the color normalization as a learning generative model that is able to generate various color copies of the input image

through a nonlinear parametric transformation. They claimed that their model preserved structures after color conversion and that only the chromatic information can be subject to change. However, their training data use nonmatched pairs of patches, thereby rendering the model unable to decouple anatomic structure from color. Ali et al.[18] separated the problem space into smaller predefined patches, subsequently developing a combination patch classification, detection, and normalization pipeline. However, such methods are very sensitive to the performance of the classification or detection steps, from which errors can propagate through the processing pipeline. Even though we have already mitigated the adverse impact of classification or detection errors (only inked-tissues were included in performance comparison), the performance of our model still surpassed this benchmark. There are several reasons that contribute to this. Unlike step-wise models,[18] our end-to-end model does not propagate errors caused by the failure of the early stages. Problems of classifying patches into the correct content type or errors in segmentation will not affect the normalization. Since there are no extra classification steps in our workflow, fewer total parameters need to be optimized, making it faster to train and possibly easier to generalize to new ink colors. The model does not require annotated patch types or tissue segmentations, alleviating the burden for manual data annotation. The data used for this work are also unique. Most models are trained using unmatched patches because of the difficulty in getting matched clean and marked images. Therefore, the models naturally want to learn how to "morph" the tissue from the marked to clean image. Decoupling this is difficult and requires either complex training tricks or novel architectures. We have used a dataset of matched clean and marked slides that are uniquely suited for this task. Therefore, we minimize the network's ability to learn textural transformations unrelated to markings.

One limitation of this work is the relative blurriness of some color normalized images. Despite adopting a weighted $L1$ loss to control for blurry images, this did not address image sharpness in all cases. One cause contributing to this may be a natural property of generative models having a difficult time reproducing high-frequency textures,[43] especially in image patches with tissue covered by thick ink (Fig. 7, sample 1). Another cause is likely the artifacts of our data, including (1) misalignment of clean and marked WSIs (Fig. 7, sample 2), which can

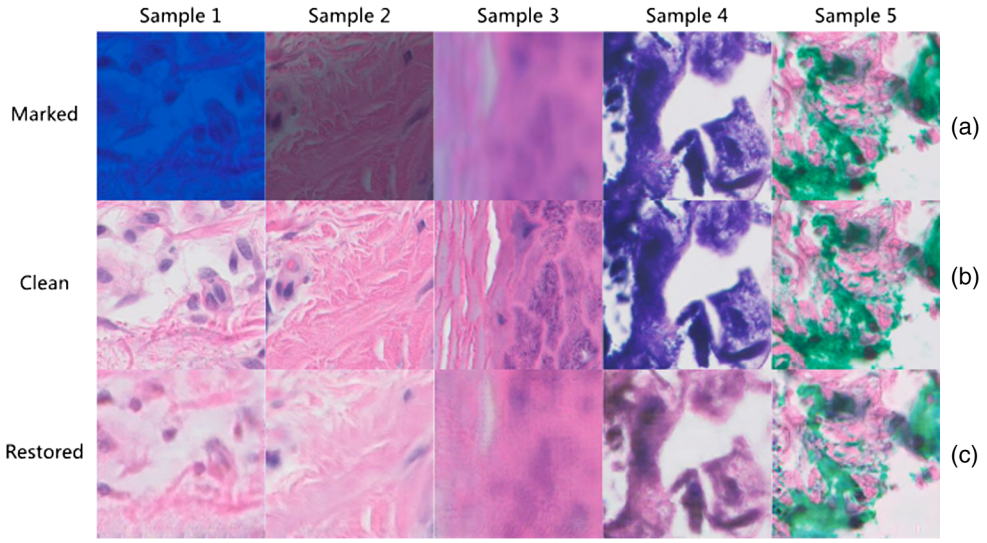

Fig. 7 Failed cases of image restoration, each column consists of a sample from (a) marked slides, (b) clean slides, and (c) the restoration results. The failure of sample 1 may be caused by the natural property of generative models having a difficult time reproducing high-frequency textures. The slight misalignment in sample 2 increased the difficulty of creating mapping from pixels in the marked image to pixels in the clean image, which may lead to blurry restoration. In sample 3, it is impossible for our model to restore a sharp image from an unfocused marked image. Dark blue artifacts in sample 4 were converted into H&E purple, controversial tissue was generated. However, the green artifacts in sample 5 were totally ignored by our model, but the model seemed to generate tissue on the left side of the restored image since there is misalignment in this case.

be caused by tissue fold, tissue damage, and other registration exceptions; (2) poorly focused clean images (Fig. 7, sample 3), which are due to the focus failure of the scanner. It may be possible to train a model to sharpen these poorly focused images;[17] however, that is outside the scope of this work; and 3) stain artifacts (Fig. 7, samples 4 and 5), which can be caused by inappropriately treating slides. It seems that the trained model may try to generate tissue/cells from the artifacts. However, the large number of patches tempers the effects of these noisy training data. Therefore, we believe the contribution of blurry training data to the model's performance is small.

6 Conclusion

Our end-to-end model for on-slide annotation removal exceeds current benchmarks, opening the possibility of leveraging archived slides for deep learning applications while preserving the clinical annotations introduced during diagnosis. Specifically, the model presented here is shown to provide two-fold higher quality (based on PSNR, SSIM, and VIF) than other methods. Although training the large Pix2Pix model requires significant computation time, the strategy we propose could readily be extended to remove ink from other pen colors not present in our initial dataset. Future experiments will be necessary to test the assumption that these digitally cleaned slides provide the same prediction/segmentation results as those that do not contain any clinical markup.

7 Appendix

6

Disclosures

We confirm that there are no known conflicts of interest associated with this publication.

Acknowledgments

This work was funded by Leon Lowenstein Foundation and Mayo Clinic Office of Artificial Intelligence.

References

1. A. Ferrari, S. Lombardi, and A. Signoroni, "Bacterial colony counting with convolutional neural networks in digital microbiology imaging," *Pattern Recognit.* **61**, 629–640 (2017).
2. W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Comput. Methods Biomech. Biomed. Eng.* **6**(3), 283–292 (2018).
3. M. Längkvist and A. Loutfi, "Unsupervised feature learning for electronic nose data applied to bacteria identification in blood," in *NIPS Workshop Deep Learn. and Unsupervised Feature Learn.* (2011).
4. E. R. A. Van Hove et al., "An alternative paper based tissue washing method for mass spectrometry imaging: localized washing and fragile tissue analysis," *J. Am. Soc. Mass. Spectrom.* **22**(10), 1885 (2011).
5. M. T. Shaban et al., "StainGAN: stain style transfer for digital histological images," in *IEEE 16th Int. Symp. Biomed. Imaging*, pp. 953–956 (2019).
6. J. Yu et al., "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 5505–5514 (2018).
7. H. Cho et al., "Neural stain-style transfer learning using GAN for histopathological images," arXiv:1710.08543 (2017).

8. P. Isola et al., "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1125–1134 (2017).
9. Y. Zhao et al., "Guided image inpainting: replacing an image region by pulling content from another image," in *IEEE Winter Conf. Appl. Comput. Vision*, pp. 1514–1523 (2019).
10. A. Vahadane et al., "Structure-preserving color normalization and sparse stain separation for histological images," *IEEE Trans. Med. Imaging* **35**(8), 1962–1971 (2016).
11. P. A. Bautista and Y. Yagi, "Staining correction in digital pathology by utilizing a dye amount table," *J. Digital Imaging* **28**, 283–294 (2015).
12. P. Haub and T. Meckel, "A model based survey of colour deconvolution in diagnostic brightfield microscopy: error estimation and spectral consideration," *Sci. Rep.* **5**, 12096 (2015).
13. T. Flotte et al., "Spectral unmixing of microscopic slides with annotations using multispectral imaging and a linear unmixing algorithm for producing an image of the annotation and an image of the histologic stain in one scan," Laboratory Investigation, pp. 1452–1452 (2020).
14. J.-Y. Zhu et al., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, pp. 2223–2232 (2017).
15. Y. Choi et al., "StarGAN: unified adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 8789–8797 (2018).
16. C. Han et al., "GAN-based synthetic brain MR image generation," in *IEEE 15th Int. Symp. Biomed. Imaging*, pp. 734–738 (2018).
17. C. You et al., "CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE)," *IEEE Trans. Med. Imaging* **39**(1), 188–203 (2020).
18. S. Ali et al., "Ink removal from histopathology whole slide images by combining classification, detection and image generation models," arXiv:1905.04385 (2019).
19. J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," arXiv:1804.02767 (2018).
20. A. Janowczyk et al., "HistoQC: an open-source quality control tool for digital pathology slides," *JCO Clin. Cancer Inf.* **3**, 1–7 (2019).
21. P. Bándi et al., "Resolution-agnostic tissue segmentation in whole-slide histopathology images with convolutional neural networks," *PeerJ* **7**, e8242 (2019).
22. D. Bug, F. Feuerhake, and D. Merhof, "Foreground extraction for histopathological whole slide imaging," in *Bildverarbeitung für die Medizin*, H. Handels et al., Eds., pp. 419–424, Springer, Berlin, Heidelberg (2015).
23. J. Jiang et al., "Robust hierarchical density estimation and regression for re-stained histological whole slide image co-registration," *PLoS One* **14**(7), e0220074 (2019).
24. L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 1415–1424 (2017).
25. T. Schlegl et al., "f-AnoGAN: fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.* **54**, 30–44 (2019).
26. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
27. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: a deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017).
28. T. F. Chan and S. Esedoglu, "Aspects of total variation regularized L1 function approximation," *SIAM J. Appl. Math.* **65**(5), 1817–1837 (2005).
29. X. Wang and A. Gupta, "Generative image modeling using style and structure adversarial networks," *Lect. Notes Comput. Sci.* **9908**, 318–335 (2016).
30. M. Sabokrou et al., "Adversarially learned one-class classifier for novelty detection," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, pp. 3379–3388 (2018).
31. M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," arXiv:1511.05440 (2015).
32. T. Porter and T. Duff, "Compositing digital images," *ACM Siggraph Comput. Graphics* **18**(3), 253–259 (1984).

33. P. Bankhead et al., "QuPath: open source software for digital pathology image analysis," *Sci. Rep.* **7**(1), 1–7 (2017).
34. L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Springer, pp. 177–186 (2010).
35. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
36. D. Berthelot, T. Schumm, and L. Metz, "Began: boundary equilibrium generative adversarial networks," arXiv:1703.10717 (2017).
37. https://drive.google.com/file/d/1kqmhp1IBpJlrY3KObD8O2FOFE4ya7iaG/view?usp=sharing.
38. A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *20th Int. Conf. Pattern Recognit.*, pp. 2366–2369 (2010).
39. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
40. H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *First Int. Workshop Video Process. and Quality Metrics Consumer Electron.*, p. 2 (2005).
41. H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.* **15**(11), 3440–3451 (2006).
42. F. G. Zanjani et al., "Stain normalization of histopathology images using generative adversarial networks," in *IEEE 15th Int. Symp. Biomed. Imaging*, pp. 573–577 (2018).
43. Y. Sun et al., "Digital radiography image denoising using a generative adversarial network," *J. X-Ray Sci. Technol.* **26**(4), 523–534 (2018).

Biographies of the authors are not available.