Virtual Immunohistochemistry Staining for Histological Images Assisted by Weakly-supervised Learning

Jiahan Li[1], Jiuyang Dong[1], Shenjin Huang[1], Xi Li[3], Junjun Jiang[1], Xiaopeng Fan[1]*,
Yongbing Zhang[2]*
[1]Harbin Institute of Technology, [2]Harbin Institute of Technology, Shenzhen
[3]Department of Gastroenterology, Peking University Shenzhen Hospital
{jiahan.li, jiuyang.dong, shenjinhuang}@stu.hit.edu.cn,
{junjunjiang, fxp, ybzhang08}@hit.edu.cn, lixi122188@sina.com

Abstract

Recently, virtual staining technology has greatly promoted the advancement of histopathology. Despite the practical successes achieved, the outstanding performance of most virtual staining methods relies on hard-to-obtain paired images in training. In this paper, we propose a method for virtual immunohistochemistry (IHC) staining, named confusion-GAN, which does not require paired images and can achieve comparable performance to supervised algorithms. Specifically, we propose a multi-branch discriminator, which judges if the features of generated images can be embedded into the feature pool of target domain images, to improve the visual quality of generated images. Meanwhile, we also propose a novel patch-level pathology information extractor, which is assisted by multiple instance learning, to ensure pathological consistency during virtual staining. Extensive experiments were conducted on three types of IHC images, including a high-resolution hepatocellular carcinoma immunohistochemical dataset proposed by us. The results demonstrated that our proposed confusion-GAN can generate highly realistic images that are capable of deceiving even experienced pathologists. Furthermore, compared to using H&E images directly, the downstream diagnosis achieved higher accuracy when using images generated by confusion-GAN. Our dataset and codes will be available at https://github.com/jiahanli2022/confusion-GAN.

1. Introduction

Histopathological staining is a key step in clinical pathological analysis [2, 33]. As the most common histopathological staining method, H&E staining can clearly delineate the cellular structures in tissues. At the same time, immunohistochemistry (IHC) staining is a functional staining method used for evaluation of protein-specific expression, which greatly assists in tumor diagnosis and cancer prognosis [1]. For example, breast cancer with overexpression of human epidermal growth factor receptor 2 (HER2) tends to have aggressive clinical behaviour [24], while breast cancer with low expression of estrogen receptor (ER) often accompanies poor response to endocrine therapy [18]. Besides, glypican-3 (GPC3) is a recently discovered immunohistochemical protein that shows promising potential for indicating the presence of hepatocellular carcinoma.

Unfortunately, the process of IHC staining is laborious, time-consuming, and expensive [42]. Therefore, many researches have focused on utilizing virtual staining techniques to convert H&E images into IHC images [38, 41, 43]. The existing virtual IHC staining methods can be roughly divided into two categories: supervised learning-based methods and unsupervised learning-based methods. Supervised learning-based methods require a large amount of paired images for training. These H&E-IHC pairs typically come from adjacent layers of same tissues, accompanied by complex data processing and alignment. What's more, these methods cannot be used in scenarios where adjacent layers are not available. On the other hand, unsupervised methods often focus on the style of the generated images, ignoring the transmission of pathological information.

To address these aforementioned problems, we propose a virtual IHC staining framework called confusion-GAN, which utilizes weakly supervised learning to facilitate the unsupervised staining process of unpaired images. To explore the pathological information of H&E patches, a novel patch-level pathology information extractor which is assisted by multiple instance learning (MIL) is proposed. This method utilizes a dual-spherical loss proposed by us to push positive patches away from negative patches in the feature

*Corresponding author: Xiaopeng Fan, Yongbing Zhang space, aiming to achieve accurate acquisition of pathological information. When training the confusion-GAN, we constrain the pathological consistency between the staining results and the input images. In addition, we propose a multi-branch discriminator, which contains a confusion discriminator as well as a common single-image discriminator, aiming to enhance the fidelity of the staining results. Specifically, we first combine a generated image with multiple real images to create an image pool in each iteration. The confusion discriminator aims to distinguish the unique generated image from this pool. Simultaneously, the generator tries to hinder this judgment of the confusion discriminator.

To demonstrate the effectiveness of our proposed approach, we conducted extensive experiments on three datasets: 1) converting H&E staining images to HER2 staining images over the BCI dataset [24], 2) converting H&E staining images to ER staining images over the MIST dataset [18], and 3) convert H&E staining images to GPC3 staining images over our collected hepatocellular carcinoma immunohistochemical (HCI) dataset. In terms of image quality and pathological information transmission, our confusion-GAN can achieve superior performance compared to existing state-of-the-art methods and deceive human pathologists to some extent. Additionally, we compared the classification accuracy of ABMIL [13], a widely used algorithm for pathological image diagnosis, by using GPC3 images generated from our confusion-GAN and the original H&E images as the input data. The experimental results clearly indicate that the utilization of generated GPC3 images can significantly improve the diagnosis performance.

The contributions are summarized as follows:

- We propose an algorithm that utilizes weakly supervised learning to assist unsupervised virtual staining for the first time, which generates satisfactory images on three virtual IHC staining tasks.
- We propose a novel patch-level pathology information extractor (PPIE) for H&E image classification. As a plug-and-play method, PPIE is demonstrated to significantly improve the instance-level classification performance when combined with existing MIL methods.
- We propose a multi-branch discriminator (MBD) that can significantly enhance the fidelity of the staining results compared to only using a single-image discriminator.
- We collected a high-resolution hepatocellular carcinoma immunohistochemical dataset, which is contains both H&E and GPC3 staining images for the first time.

2. Related Work

2.1. Virtual IHC Staining Techniques

Existing virtual IHC staining methods can be broadly divided into two categories: supervised and unsupervised learning based methods. Supervised learning based methods require a large number of H&E-IHC pairs as training images [8, 18, 24]. There are two ways of obtaining these paired images: (1) Carrying out axially consecutive scanning of tissues, actually staining H&E and IHC on adjacent slices, then performing alignment between H&E and IHC slides through complex data processing and registration to obtain the paired images for training. For example, Liu et al. [24] train their proposed PyramidP2P network over a registered H&E-HER2 dataset. However, the tissue contents from different layers are inherently difficult to be consistent, making pixel-level alignment impossible. Therefore, the inability to fully align will inevitably result in performance limitation. In addition, such supervised methods are unlikely to be applicable when adjacent layer data is unavailable. (2) Using images of a third modality other than H&E and IHC, such as microscopic fluorescence data, as a bridge, to generate pixel-wise aligned H&E and IHC training data [8]. Unfortunately, access to the third modality data will also add additional overhead.

The vast majority of existing unsupervised virtual IHC staining methods are based on image translation networks [5, 9, 21, 30, 33, 35, 40]. Although these networks can convert H&E-style images into IHC-style images, it does not mean that the Pos/Neg properties of the translated results will be consistent with that of their inputs. To keep the aforementioned pathological consistency, Liu et al. [23] and Boyd et al. [3] introduce expert annotations, which is an expensive solution.

2.2. Patch-level Classification Assisted by MIL

Currently, MIL are used for patch-level classification [4, 13, 16, 19, 32]. In the case of pathological images, each WSI is regarded as a bag, and each patch within the WSI is treated as an instance. A bag label is assigned to each bag, which is classified as positive if the bag contains at least one positive instance. These bag labels serve as the ground truths for training MIL networks in an end-to-end manner. Existing methods determine whether a patch is positive or not by comparing the attention scores (or instance-level predictions) of all patches [6, 20, 22, 29, 31, 34]. Unfortunately, these methods can only ensure accurate patch-level classification within a limited region of high attention scores. Therefore, we propose a novel patch-level pathology information extractor to compensate for the shortcomings of these methods.

3. Proposed Method

3.1. Framework

One significant difference between the virtual IHC staining and the image translation of natural images is that generated IHC images must maintain pathological consistency with

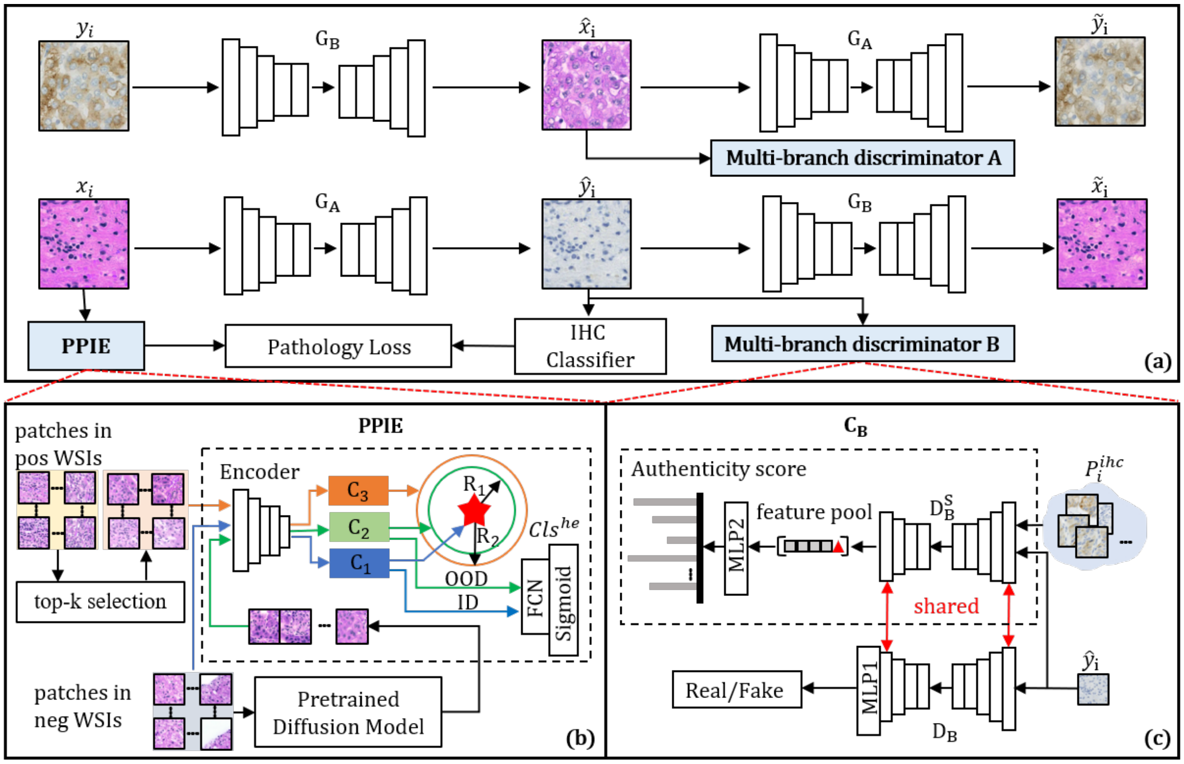

Figure 1. Structure of the proposed confusion-GAN. (a) The confusion-GAN contains two important modules: patch-level pathology information extractor (PPIE) and multi-branch discriminator (MBD). (b) The training process of PPIE. (c) The structure of the MBD.

their inputs while preserving texture consistency. From the perspectives of improving the quality of generated images and accurately transmitting pathological information, we propose confusion-GAN, which is illustrated in Fig. 1. The confusion-GAN contains two important modules, patch-level pathology information extraction module and multi-branch discriminator, which are employed to maintain the pathological consistency and generate more realistic images, respectively.

As shown in Fig. 1, the real H&E image x_i is first passed through the generator G_A to obtain the IHC-style image $\hat{y}_i$. Then, $\hat{y}_i$ is further remapped into the H&E-style image $\tilde{x}_i$ using the generator G_B. Similarly, we can obtain $\hat{x}_i$ and $\tilde{y}_i$ from real IHC image y_i.

During the process of generating IHC image from the H&E image, we devise a Pathology Loss to ensure the accurate transmission of pathological information. IHC classifier Cls_{ihc} and a PPIE module are designed to obtain the pathological information of the H&E image x_i and the generated IHC image $\hat{y}_i$, respectively. To pretrain a reliable IHC classifier, we simulate the annotation process of pathologists through color analysis to obtain patch-level pos/neg labels for actually captured IHC images: patches with positive exhibit protein-specific expression within more than 1% of the area (corresponding to yellow-brown coloration in the image). A light-weight image classfier Cls_{ihc} is then designed utilizing the pos/neg labels obtained by the previously described process, which consists of 6 convolutional layers activated by ReLU and a fully connected layer (FCN) for the final classification. The kernel size for the convolutional layers is 4, with a stride of 2 and padding of 1. Similarly, the pathological information of the H&E image is obtained utilizing the PPIE module, which will be elaborated in Sec. 3.2. Finally, we adopt the Pathology Loss to constrain the consistency of pathological information between x_i and $\hat{y}_i$:

$$L_{path} = -\frac{1}{n}\sum_{i=1}^{n}(PPIE(x_i)log(Cls^{ihc}(\hat{y}_i)) \\ +(1-PPIE(x_i))log(1-Cls^{ihc}(\hat{y}_i))), \tag{1}$$

where $PPIE(\cdot)$ represents the operation of PPIE module.

In addition, to enhance the quality of the generated images, we propose a multi-branch discriminator for both H&E-to-IHC and IHC-to-H&E process in the confusion-GAN. Detailed description of the multi-branch discriminator can be found in Sec. 3.3.

In general, the hybrid loss for the generator is as follows:

$$L_G = \alpha * L_{path} + \beta * L_G^C + \lambda * L_G^{Adv} \\ + \eta * L_G^{Cycle} + \iota * L_G^{Identity}, \tag{2}$$

where L_G^{Adv}, L_G^{Cycle}, and $L_G^{Identity}$ represent the adversarial loss, cycel loss, and identity loss similar to CycleGAN [44], while α, β, λ, η, and ι are hyper-parameters controlling the weighting coefficients. L_G^C is the loss function used to guide the generator optimization with the confusion discriminator, which will be described in detail in Sec. 3.3.

3.2. Patch-level Pathology Information Extraction

Due to memory limitations, virtual staining can only be performed at the patch level. Unfortunately, only WSI-level labels are available for majority cases, since one WSI usually contains gigapixel, impeding the patch-wise annotation in clinical. Therefore, it is necessary to extract patch-level pathological information (i.e., the pos/neg category of each patch) from the H&E WSI to ensure accurate virtual staining.

For H&E images that only have WSI level labels (lacking patch-level annotations), the classification of each patch is typically completed through attention scores or instance-level classification predictions within the MIL framework. In each positive WSI, there is a high likelihood that patches with the highest attention scores or instance-level classification scores will be positive. We refer to the action of selecting the top k patches of each positive WSI based on the highest attention scores (or instance-level prediction scores) as "top-k selection". When k is relatively small, it is ensured that the selected patches are all positive; when k is large, the selected collections may include many negative patches [26]. Therefore, the current methods have not been able to accurately classify all patches.

To address this problem, building upon the top-k selection, we propose a plug-and-play PPIE module that can improve the classification performance of H&E patches. Specifically, we first perform top-k selection to obtain X_{pos}^{he} from positive WSIs in the training set. Next, we crop a set of negative patches X_{neg}^{he} from negative WSIs. It should be noted all patches in a negative WSI are negative, unlike the cases of positive WSIs where both positive and negative patches may coexist. Besides, we utilized a state-of-the-art diffusion model BBDM [17] to learn how to generate a distribution similar to that of X_{pos}^{he} using X_{neg}^{he}. The set of images generated by BBDM is referred to as $X_{\widetilde{pos}}^{he}$.

On this basis, we borrow the idea of anomaly detection to tackle the classification problem of patch-level H&E images [10]. In each iteration of training the PPIE, we sample b images from X_{neg}^{he}, $X_{\widetilde{pos}}^{he}$, and X_{pos}^{he} respectively, and feed them into the encoder to obtain the feature sets $C_1 = \{c_{1i}\}_{i=1}^{b}$, $C_2 = \{c_{2i}\}_{i=1}^{b}$, and $C_3 = \{c_{3i}\}_{i=1}^{b}$, as shown in Fig. 1(b). Here, b represents the batch size during training. The encoder consists of 6 convolutional layers activated by ReLU and a fully connected layer with an output dimension of 512. The kernel size for the convolutional layers is 4, with a stride of 2 and padding of 1.

As is well known, the success of the anomaly detection depends on two aspects: 1) the samples in the in-distribution (ID) class should be as close as possible in the feature space; 2) the samples in the out-of-distribution (OOD) class should be as far as possible from samples in the ID class [39]. We use the loss function L_{ID} to force all the samples in the ID class to be distributed as close to the center as possible:

$$L_{ID} = \frac{1}{b}\sum_{i=1}^{b}||c_{1i} - O||^2, \tag{3}$$

where O represents the average features of all the ID data X_{neg}^{he}. It should be noted that O will be cumulatively updated during training. Besides, we propose a double-spherical loss L_{OOD} to ensure that the overall distribution of C_3 has a greater distance from O compared to the overall distribution of C_2:

$$\begin{aligned} L_{OOD} = &\frac{1}{b}\sum_{i=1}^{b} max(R_1 - ||c_{2i} - O||^2, 0) \\ &+\frac{1}{b}\sum_{i=1}^{b} max(R_2 - ||c_{3i} - O||^2, 0). \end{aligned} \tag{4}$$

We set the hyperparameter R_2 to control the distance and ensure that R_2 is greater than R_1, as shown in Fig. 1(b). Employing L_{ID} and L_{OOD}, we can ensure that X_{neg}^{he} and $X_{\widetilde{pos}}^{he}$ have strong separability in the feature space. Finally, we treat X_{neg}^{he} and $X_{\widetilde{pos}}^{he}$ as ID and OOD classes respectively, and train the patch-level classifier Cls^{he} by cross entropy loss L_{cls}:

$$\begin{aligned} L_{cls} = &-\frac{1}{n}\sum_{i=1}^{s}(l_i log(Cls^{he}(x_i^{he})) \\ &+ (1 - l_i)log(1 - Cls^{he}(x_i^{he}))). \end{aligned} \tag{5}$$

When using PPIE for inference, negative patches are considered as ID class, while positive patches are considered as OOD class.

Overall, the hybrid loss function L_{PPIE} for training the PPIE module is as follows:

$$L_{PPIE} = \mu * L_{ID} + \tau * L_{OOD} + \upsilon * L_{cls}, \tag{6}$$

where μ, τ, and υ are hyper-parameters to control the weighting coefficients.

3.3. Multi-branch Discriminator

To achieve more realistic image generation, we propose a multi-branch discriminator, as shown in Fig. 1(c). Taking generating an IHC-style image from a H&E image as an example, we send the result $\hat{y}_i$ obtained by the generator G_A together with a reference pool P_i^{ihc} consisting of

$N-1$ real images from the same domain to the confusion discriminator C_B. C_B shares the feature extraction layer D_B^S with the single-image discriminator D_B, which helps eliminate the interference of the content of the reference pool images on the confusion discriminator. This one-to-many discrimination paradigm requires the generator's results to possess more common attributes of the current domain images, thereby forcing the generated distribution to approximate the real distribution more closely.

During the training of the C_B, we aim to accurately distinguish between the generated result and real images. Then we randomly insert $\hat{y}_i$ into a position in the list of P_i^{ihc}, and finally concatenate them in the channel dimension to obtain the input m_i^{ihc} for C_B. Correspondingly, we construct an N-dimensional vector l_i^{ihc} to represent the labels of each element in m_i^{ihc}, with 0 indicating real image and 1 indicating generated image. Next, we pass m_i^{ihc} into D_B^S to obtain a feature pool, which is fed into an MLP layer of C_B for liveness verification. C_B is then updated by minimizing:

$$L_{C_B} = \frac{1}{n}\sum_{i=1}^{n} ||C_B(m_i^{ihc}) - l_i^{ihc}||^2. \tag{7}$$

Similarly, during the process of IHC-to-H&E image translation, the loss function L_{C_A} of the confusion discriminator C_A is defined as follows:

$$L_{C_A} = \frac{1}{n}\sum_{i=1}^{n} ||C_A(m_i^{he}) - l_i^{he}||^2, \tag{8}$$

where m_i^{he} represents the concatenation result of the generated H&E image and $N-1$ real H&E images, and l_i^{he} represents the corresponding labels of m_i^{he}. The proposed confusion discriminators are trained using the following formula:

$$L_C = L_{C_B} + L_{C_A}. \tag{9}$$

The loss function used for training D_B and D_A is consistent with CycleGAN.

When training the generator, we obtain m_i^{ihc} and m_i^{he} in the same manner. However, in this stage, we want the confusion discriminator unable to identify the generator's results from these N images. Therefore, all labels in l_i^{ihc} and l_i^{he} are set to 0. The generators are updated based on the following loss functions:

$$\begin{aligned} L_G^C = & \frac{1}{n}\sum_{i=1}^{n} ||C_B(m_i^{ihc}) - l_i^{ihc}||^2 \\ & + ||C_A(m_i^{he}) - l_i^{he}||^2. \end{aligned} \tag{10}$$

In each iteration, we train the generators G_A and G_B, D_A and D_B, and our proposed C_A and C_B in an alternating manner, following a sequential order. This process is similar to what most generative adversarial networks do. Here G_A and G_B employ the unet256 network used in CycleGAN and the structures of D_A and D_B are also consistent with those in CycleGAN.

4. Experiments

4.1. Setup

Datasets. GPC3, discovered as a tumor marker in recent years, shows great potential for specific expression in Hepatocellular carcinoma (HCC). We have collected a HCC Immunohistochemical Image dataset named HCI, which includes both the H&E and GPC3 images. All data comes from Peking University Shenzhen Hospital. We selected 20 H&E-IHC image pairs from a pool of WSIs for testing and validation, and perform registration on them using the DeepHistReg [37]. Next, we conducted separate screenings for a total of 30 H&E images and 30 IHC images, and these images are used for training. Furthermore, we segment these 100 WSIs into approximately 1.4 million patches sized of 256×256 in a non-overlapping manner at 20x magnification. All H&E images have undergone staining normalization. We also verified our method on public datasets: BCI [24] and MIST [18].

Metrics. We evaluated the image quality of the staining results using objective metrics such as Frechet Inception Distance (FID) [27], Learnable Perceptual Image Patch Similarity (LPIPS), and Structural Similarity Index (SSIM) [11]. Additionally, we invited three pathologists to provide subjective evaluations of the staining results. In assessing the patch-level classification performance of the aforementioned PPIE, we utilized metrics such as area under curve (AUC) [25], False Positive Rate (FPR) [28], True Positive Rate (TPR), and Geometric Mean (G-mean) [7].

Implementation Details. We trained the PPIE using the Adam optimizer with a learning rate of 0.001. The dual-sphere distances R_1 and R_2 in Eq. (4) are set to 150 and 262.5 respectively. The hyper-parameters μ, τ, and υ in Eq. (6) are all set to 1. When training the virtual staining network, α, β, λ, η and ι in Eq. (2) are set to 1, 1, 1, 10, and 5 respectively. We set the reference pool size N as 32 and kept the other configuration of the staining network consistent with CycleGAN. In the downstream experiments, we trained the ABMIL model on a re-partitioned dataset for a maximum of 15 epochs. For all experiments, we trained and tested our confusion-GAN on patches of size 256×256.

4.2. Staining Performance

HCI. As shown in Fig. 2, we compared our method with three unsupervised image translation methods, namely CycleGAN [44], UGATIT [15], and DCD [12], on the HCI dataset. The first column in Fig. 2 displays the input H&E images, while the last column shows actually stained IHC (GPC3) images for reference. These images are from corre-

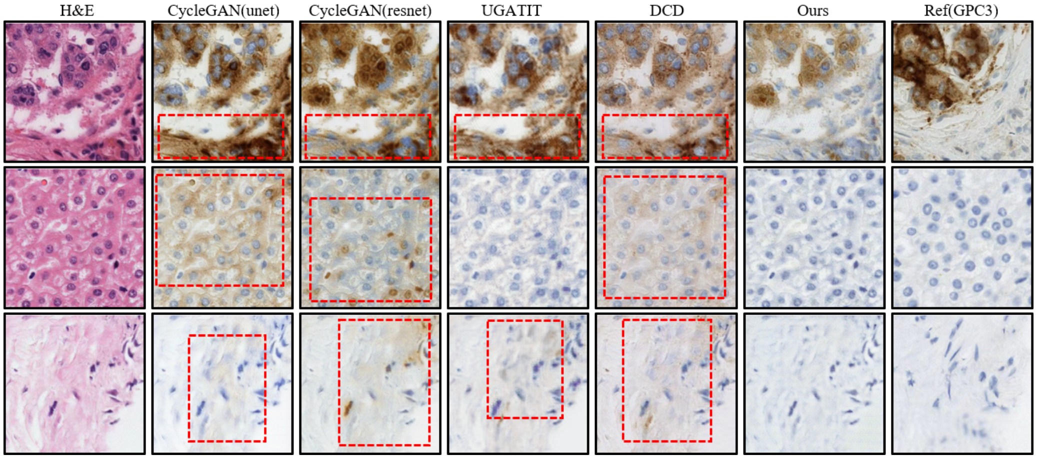

Figure 2. Comparison of the abilities of different unsupervised algorithms to preserve pathological consistency on HCI. The red dashed box indicates the area where the generated image cannot maintain consistent pathological information with the input H&E image.

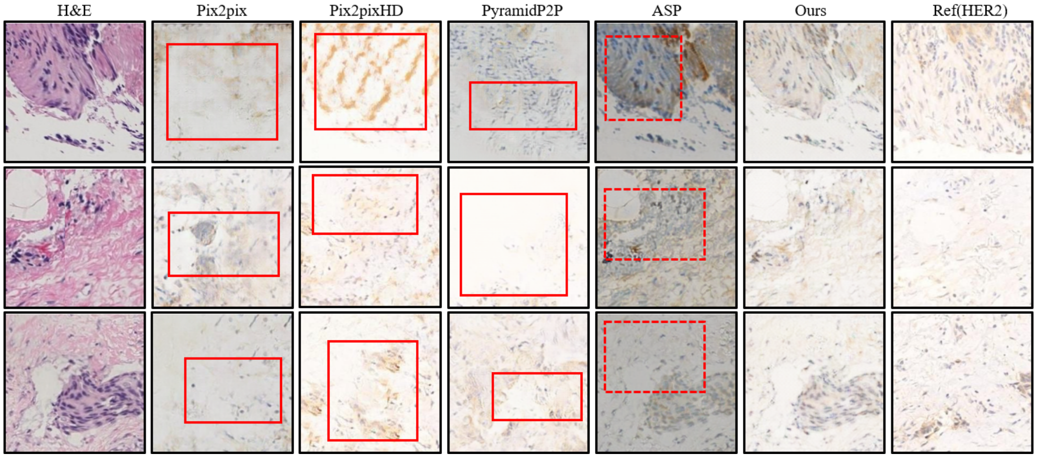

Figure 3. Comparing the ability of the state-of-the-art supervised methods and ours to simultaneously preserve the texture and pathological information of the input image on BCI. The red solid rectangles indicate regions with texture errors, while the red dashed rectangles represent regions with pathological information errors.

sponding regions of adjacent slices in the same tissue. We marked in red dashed lines the regions where the pathological information is inconsistent with the input. Obviously, although the competing methods CycleGAN, UGATIT, and DCD can maintain the texture details of the input images well, there are significant misstaining areas since the consistency of pathological information can not be ensured before and after staining. This phenomenon demonstrates the effectiveness of the proposed PPIE module. Besides, we use three objective metrics, FID, LPIPS, and SSIM to evaluate these virtual staining methods. As shown in Tab. 1, our method achieves the best performance in all the three metrics, indicating that the output distribution of the confusion-GAN is closest to the distribution of the real images.

Method	FID↓	LPIPS↓	SSIM↑
CycleGAN(unet)	105.4	0.636	0.1870
CycleGAN(resnet)	90.9	0.622	0.2086
UGATIT	95.3	0.602	0.2223
DCD	91.3	0.612	0.2266
Ours	**77.3**	**0.570**	**0.2300**

Table 1. The evaluation of different algorithms on HCI.

BCI and MIST. BCI and MIST datasets provide paired H&E-IHC patch images, where the acquisition of these patches relies on alignment with the sliced images from adjacent layers within the same tissue. However, the tissue contents from different adjacent layers are inherently impossible to be completely consistent, making pixel-level alignment impossible. We compared our method with four fully supervised methods (Pix2pix [14], Pix2pixHD [36],

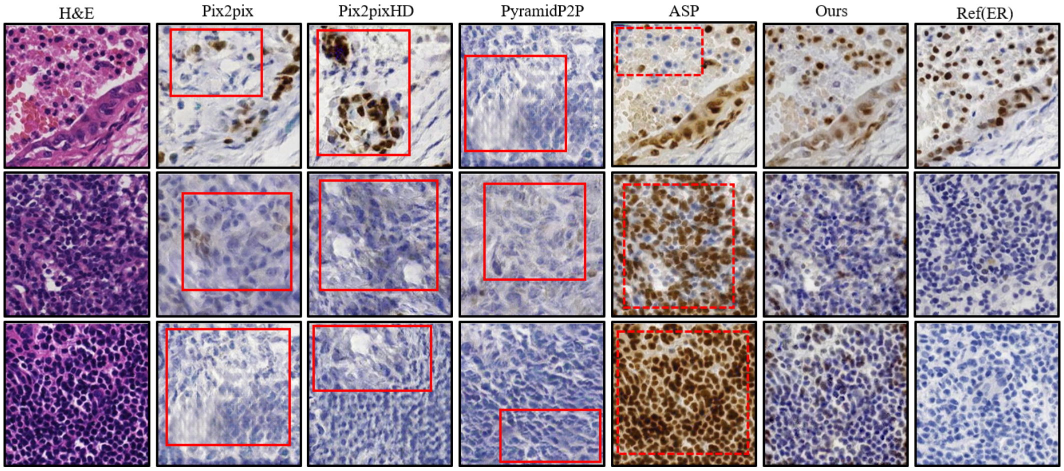

Figure 4. Comparing the ability of the state-of-the-art supervised methods and ours to simultaneously preserve the texture and pathological information of the input image on MIST. The red solid rectangles indicate regions with texture errors, while the red dashed rectangles represent regions with pathological information errors.

PyramidP2P [24], ASP [18]) on these two datasets, as shown in Fig. 3 and Fig. 4. We also marked the areas with pathological information errors in the staining results with dashed lines. Additionally, we marked the areas with content errors in the staining results with solid lines. From Fig. 3 and Fig. 4, it can be observed that Pix2pix, Pix2pixHD, and PyramidP2P methods are unable to maintain the textures of the input images, often resulting in significant presence of artifacts in the generated results. This phenomenon is attributed to the significant discrepancy in content between the input H&E images and the reference IHC images used as the ground truth during the training process. The ASP based on contrastive learning demonstrates good performance in preserving content. However, the misalignment of supervised information can still introduce some errors in the pathological information of the staining results. In contrast, our proposed confusion-GAN has achieved better results than these supervised learning methods.

We used FID, LPIPS, and SSIM for objective evaluation. As can be seen from Tab. 2, our confusion-GAN achieved the best FID and LPIPS scores on both datasets, which indicating the effectiveness of our algorithm. Regarding the SSIM evaluation metric, which is directly calculated in the pixel space, our method still achieved the highest score on MIST dataset. Meanwhile, our method can achieve the best visual effects on both BCI and MIST, especially considering the fidelity of textures in the generated results. It is worth mentioning that, due to the lack of complete WSIs in MIST and BCI, we assigned labels to IHC images using the color analysis mentioned in Sec. 3.1, and then transfer the labels of IHC images to the corresponding H&E images. Regarding to BCI, we directly utilized the provided patch-level labels, which guides the training of confusion-GAN.

Methods	BCI			MIST		
	FID↓	LPIPS↓	SSIM↑	FID↓	LPIPS↓	SSIM↑
Pix2pix	110.7	0.480	**0.4957**	80.9	0.529	0.2127
Pix2pixHD	106.8	0.505	0.4766	131.0	0.564	0.2129
PyramidP2P	99.6	0.505	0.4663	97.5	0.538	0.2061
ASP	54.3	0.503	0.4923	41.2	0.532	0.2062
Ours	**49.2**	**0.478**	0.4579	**39.8**	**0.516**	**0.2174**

Table 2. Comparing the performance of our method and supervised algorithms on BCI and MIST.

4.3. Ablation Study

We explored the working mechanism of PPIE and analyzed how PPIE and the MBD affect the staining results.

Method	k	AUC↑	FPR↓	TPR↑	G-mean↑
ABMIL	10	0.713	0.277	0.641	0.680
	30	**0.782**	**0.226**	**0.726**	**0.750**
	100	0.637	0.408	0.589	0.591
	wo	0.732	0.323	0.686	0.681
Maxpool	10	0.682	0.324	0.666	0.671
	30	**0.731**	**0.257**	**0.689**	**0.715**
	100	0.706	0.267	0.667	0.699
	wo	0.680	0.328	0.613	0.680
ClamMb	10	0.766	0.238	0.691	0.726
	30	**0.770**	**0.219**	**0.708**	**0.744**
	100	0.732	0.286	0.627	0.669
	wo	0.726	0.338	0.687	0.675
Meanpool	10	0.745	0.253	0.671	0.708
	30	**0.768**	**0.239**	**0.728**	**0.744**
	100	0.709	0.254	0.619	0.679
	wo	0.758	0.324	0.710	0.693

Table 3. Patch-level classification results obtained from different top-k patches on HCI.

Effect of k in the top-k selection. We selected four classic MIL methods (ABMIL [13], MaxPool, CLAM-MB[26], and MeanPool) for top-k selection in PPIE training. As shown in Tab. 3, we evaluate the classification performance of PPIE using AUC, FPR, TPR, and G-mean. We found that the classification performance is best when k is set to 30. This is because when k is too small, the PPIE cannot obtain sufficient training data. On the other hand, when k is too large, the top-k selection results will contain a large number of negative patches, which hinders the effectiveness of the double-spherical loss. Among these four MIL methods, we obtained similar conclusions and improved the patch-level classification performance, achieving gains of 5.0%, 5.1%, 4.4%, and 1.0% respectively.

Effect of PPIE and MBD. To investigate how PPIE and MBD impact the performance, we conducted an ablation experiment on HCI, as shown in Fig. 5. By comparing the second, third, and fourth columns in Fig. 5, it can be observed that the PPIE plays a crucial role in accurately transmitting pathological information. On the other hand, the MBD can enhance the clarity of cell nuclei and surrounding tissues, affecting the overall style of the generated results. What's more, based on the qualitative analysis of the indicators in Tab. 4, we found that the method combining PPIE and the MBD achieved the best results.

PPIE	✗	✗	✓	✓
MBD	✗	✓	✗	✓
FID↓	105.4	82.4	103.3	**77.3**
LPIPS↓	0.636	0.588	0.574	**0.570**
SSIM↑	0.1870	0.2223	0.2294	**0.2300**

Table 4. Exploring the impact of PPIE and MBD on HCI.

4.4. Down-stream Classification Performance

We conducted downstream experiments related to the HCC diagnosis, and compared the classification performance achieved by using the H&E and generated GPC3 slides as inputs of the ABMIL model, respectively. When using H&E images as input achieves an AUC of 55.6%, while using the generated images as input achieves an AUC of 88.9%. This indicates that compared to H&E images, our generated GPC3 images can improve the classification accuracy of HCC.

4.5. Subjective Evaluation by Pathologists

We invited three pathologists to provide subjective evaluations on two aspects: transmission of pathological information and the quality of the images. We randomly selected 50 pairs of generated GPC3 images along with their corresponding adjacent tissues and then asked pathologists to evaluate whether the protein expression levels were consistent between the two. As shown in the first row of Tab. 5, in most cases, pathologists believed that the protein expression levels were consistent between them. This indicates that our confusion-GAN can effectively maintain pathological consistency. Afterwards, we randomly selected 50 GPC3 images and 50 generated GPC3 images for pathologists to distinguish which are generated images. From the second row of the Tab. 5, it can be seen that doctors cannot distinguish the generated images well, indicating that our confusion-GAN can generate images comparable to real images.

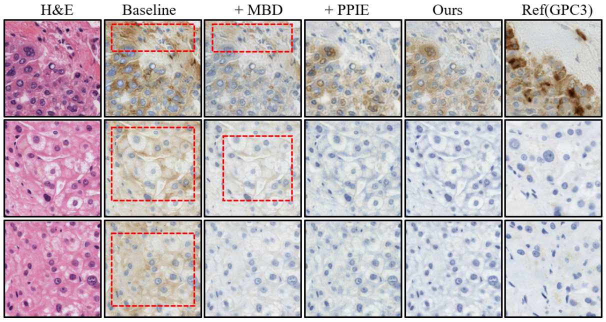

Figure 5. Ablation study: effect of PPIE and MBD on HCI. The red box indicates regions with pathological information errors.

Accuracy	pathologist1	pathologist2	pathologist3
Neg/Pos	0.86	0.80	0.84
Fake/True	0.62	0.48	0.52

Table 5. Subjective evaluations of pathologists on HCI.

5. Conclusion

In this paper, we propose a method called confusion-GAN, which is the first to utilize weakly-supervised learning to assist in virtual IHC staining of unpaired pathological images. Extensive experiments have demonstrated that utilizing patch-level pathological information to guide the training of staining network is crucial for achieving accurate virtual IHC staining. Furthermore, we propose a multi-branch discriminator for achieving high-fidelity image generation. Our approach achieved the state-of-the-art results on multiple datasets. To advance the diagnosis of HCC, we have also proposed the first GPC3 dataset and demonstrated that the images generated by confusion-GAN are more helpful in HCC detection compared to H&E images. We believe that unsupervised methods like confusion-GAN, which do not require fine annotation or complex preprocessing, will be crucial in virtual staining research in the future.

Acknowledgement. This work was supported in part by the National Natural Science Foundation of China (62031023&62331011), in part by the Shenzhen Science and Technology Project (GXWD20220818170353009), and in part by the Fundamental Research Funds for the Central Universities (Grant No. HIT.OCEF.2023050).

References

[1] Bijie Bai, Xilin Yang, Yuzhu Li, Yijie Zhang, Nir Pillar, and Aydogan Ozcan. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1):57, 2023. 1

[2] Marian Boktor, Benjamin R Ecclestone, Vlad Pekar, Deepak Dinakaran, John R Mackey, Paul Fieguth, and Parsin Haji Reza. Virtual histological staining of label-free total absorption photoacoustic remote sensing (ta-pars). *Scientific Reports*, 12(1):10296, 2022. 1

[3] Joseph Boyd, Irène Villa, Marie-Christine Mathieu, Eric Deutsch, Nikos Paragios, Maria Vakalopoulou, and Stergios Christodoulidis. Region-guided cyclegans for stain transfer in whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 356–365. Springer, 2022. 2

[4] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 2

[5] Tanujit Chakraborty, Shraddha M Naik, Madhurima Panja, Bayapureddy Manvitha, et al. Ten years of generative adversarial nets (gans): A survey of the state-of-the-art. *arXiv preprint arXiv:2308.16316*, 2023. 2

[6] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 2

[7] Yanyan Chen, Kuaini Wang, and Ping Zhong. One-class support tensor machine. *Knowledge-Based Systems*, 96:14–28, 2016. 5

[8] Kevin de Haan, Yijie Zhang, Jonathan E Zuckerman, Tairan Liu, Anthony E Sisk, Miguel FP Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, et al. Deep learning-based transformation of h&e stained tissues into special stains. *Nature communications*, 12(1):4884, 2021. 2

[9] Michael Gadermayr, Maximilian Tschuchnig, Lea Maria Stangassinger, Christina Kreutzer, Sebastien Couillard-Despres, Gertie Janneke Oostingh, and Anton Hittmair. Improving automated thyroid cancer classification of frozen sections by the aid of virtual image translation and stain normalization. *Computer Methods and Programs in Biomedicine Update*, 3:100092, 2023. 2

[10] Xiaoyuan Guo, Judy W Gichoya, Saptarshi Purkayastha, and Imon Banerjee. Margin-aware intraclass novelty identification for medical images. *Journal of Medical Imaging*, 9(1): 014004–014004, 2022. 4

[11] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 5

[12] Tie Hu, Mingbao Lin, Lizhou You, Fei Chao, and Rongrong Ji. Discriminator-cooperated feature map distillation for gan compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20351–20360, 2023. 5

[13] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 2, 8

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

[15] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019. 5

[16] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14318–14328, 2021. 2

[17] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1952–1961, 2023. 4

[18] Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. *arXiv preprint arXiv:2303.06193*, 2023. 1, 2, 5, 7

[19] Weijian Li, Viet-Duy Nguyen, Haofu Liao, Matt Wilder, Ke Cheng, and Jiebo Luo. Patch transformer for multi-tagging whole slide histopathology images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 532–540. Springer, 2019. 2

[20] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023. 2

[21] Kechun Liu, Beibin Li, Wenjun Wu, Caitlin May, Oliver Chang, Stevan Knezevich, Lisa Reisch, Joann Elmore, and Linda Shapiro. Vsgd-net: Virtual staining guided melanocyte detection on histopathological images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1918–1927, 2023. 2

[22] Kangning Liu, Weicheng Zhu, Yiqiu Shen, Sheng Liu, Narges Razavian, Krzysztof J Geras, and Carlos Fernandez-Granda. Multiple instance learning via iterative self-paced supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3355–3365, 2023. 2

[23] Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative

adversarial networks. *IEEE Transactions on Medical Imaging*, 40(8):1977–1989, 2021. 2

[24] Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1815–1824, 2022. 1, 2, 5, 7

[25] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008. 5

[26] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 4, 8

[27] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *Advances in Neural Information Processing Systems*, 33:11913–11924, 2020. 5

[28] Alvin I Mushlin, Ruth W Kouides, and David E Shapiro. Estimating the accuracy of screening mammography: a meta-analysis. *American journal of preventive medicine*, 14(2): 143–153, 1998. 5

[29] Linhao Qu, Manning Wang, Zhijian Song, et al. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35:15368–15381, 2022. 2

[30] Jesus Salido, Noelia Vallez, Lucía González-López, Oscar Deniz, and Gloria Bueno. Comparison of deep learning models for digital h&e staining from unpaired label-free multispectral microscopy images. *Computer Methods and Programs in Biomedicine*, 235:107528, 2023. 2

[31] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021. 2

[32] Zhuchen Shao, Yifeng Wang, Yang Chen, Hao Bian, Shaohui Liu, Haoqian Wang, and Yongbing Zhang. Lnpl-mil: Learning from noisy pseudo labels for promoting multiple instance learning in whole slide image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21495–21505, 2023. 2

[33] Lulin Shi, Ivy HM Wong, Claudia TK Lo, Lauren WK Tsui, and Terence TW Wong. Unsupervised multiple virtual histological staining from label-free autofluorescence images. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2023. 1, 2

[34] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4078–4087, 2023. 2

[35] Elena Y Trizna, Aleksandr M Sinitca, Asya I Lyanova, Diana R Baidamshina, Pavel V Zelenikhin, Dmitrii I Kaplun, Airat R Kayumov, and Mikhail I Bogachev. Brightfield vs fluorescent staining dataset–a test bed image set for machine learning based virtual staining. *Scientific Data*, 10(1):160, 2023. 2

[36] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 6

[37] Marek Wodzinski and Henning Müller. Deephistreg: Unsupervised deep learning registration framework for differently stained histology samples. *Computer methods and programs in biomedicine*, 198:105799, 2021. 5

[38] Georg Wölflein, In Hwa Um, David J Harrison, and Ognjen Arandjelović. Hoechstgan: Virtual lymphocyte staining using generative adversarial networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4997–5007, 2023. 1

[39] Xuan Xia, Xizhou Pan, Nan Li, Xing He, Lin Ma, Xiaoguang Zhang, and Ning Ding. Gan-based anomaly detection: A review. *Neurocomputing*, 493:497–535, 2022. 4

[40] Zhaoyang Xu, Xingru Huang, Carlos Fernández Moro, Béla Bozóky, and Qianni Zhang. Gan-based virtual re-staining: a promising solution for whole slide image analysis. *arXiv preprint arXiv:1901.04059*, 2019. 2

[41] Bowei Zeng, Yiyang Lin, Yifeng Wang, Yang Chen, Jiuyang Dong, Xi Li, and Yongbing Zhang. Semi-supervised pr virtual staining for breast histopathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–241. Springer, 2022. 1

[42] Ranran Zhang, Yankun Cao, Yujun Li, Zhi Liu, Jianye Wang, Jiahuan He, Chenyang Zhang, Xiaoyu Sui, Pengfei Zhang, Lizhen Cui, et al. Mvfstain: multiple virtual functional stain histopathology images generation based on specific domain mapping. *Medical Image Analysis*, 80:102520, 2022. 1

[43] Yijie Zhang, Luzhe Huang, Tairan Liu, Keyi Cheng, Kevin de Haan, Yuzhu Li, Bijie Bai, and Aydogan Ozcan. Virtual staining of defocused autofluorescence images of unlabeled tissue using deep neural networks. *Intelligent Computing*, 2022. 1

[44] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 4, 5