Unpaired Stain Transfer Using Pathology-Consistent Constrained Generative Adversarial Networks

Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He

***Abstract*— Pathological examination is the gold standard for the diagnosis of cancer. Common pathological examinations include hematoxylin-eosin (H&E) staining and immunohistochemistry (IHC). In some cases, it is hard to make accurate diagnoses of cancer by referring only to H&E staining images. Whereas, the IHC examination can further provide enough evidence for the diagnosis process. Hence, the generation of virtual IHC images from H&E-stained images will be a good solution for current IHC examination hard accessibility issue, especially for some low-resource regions. However, existing approaches have limitations in microscopic structural preservation and the consistency of pathology properties. In addition, pixel-level paired data is hard available. In our work, we propose a novel adversarial learning method for effective Ki-67-stained image generation from corresponding H&E-stained image. Our method takes fully advantage of structural similarity constraint and skip connection to improve structural details preservation; and pathology consistency constraint and pathological representation network are first proposed to enforce the generated and source images hold the same pathological properties in different staining domains. We empirically demonstrate the effectiveness of our approach on two different unpaired histopathological datasets. Extensive experiments indicate the superior performance of our method that surpasses the state-of-the-art approaches by a significant margin. In addition, our approach also achieves a stable and good performance on unbalanced datasets, which shows our method has strong robustness. We believe that our method has significant potential in clinical virtual staining and advance the progress of computer-aided multi-staining histology image analysis.**

***Index Terms*— Histopathology, stain transfer, pathology consistency constraint, Ki-67, hematoxylin-eosin (H&E).**

I. Introduction

THE mortality rate of cancer has ranked second in the world, which is a great threat to human life. According to [1], cancer incidence is on the rise. Histopathology has been regarded as the gold standard for cancer diagnosis. Cancer examination is usually conducted by experienced pathologists through observing the examinee's tissue structure and cytopathic characteristics under a high-power microscopy, which is more objective and accurate than radiographic examinations.

In clinical practice, hematoxylin and eosin staining (H&E) is one of the most commonly used staining techniques in histopathology examination. Hematoxylin principally stains cell nuclei blue or dark-purple, and eosin stains the extracellular matrix and cytoplasm pink, with other structures taking on different shades, hues, and combinations of these colors, such as red blood cells which are stained intensely red [2], as shown in Fig.1 (a) and (c). Although H&E staining is available and cost-effective, it does not always provide enough contrast to differentiate normal cells and cancer cells. In these cases, more specific stains and methods are required. Immunohistochemistry (IHC) is a kind of molecular-level staining based on the principle of antigen-antibody binding [3]. Chemical reaction can bind the chromogen with labeled antibody to intracellular antigen. For example, Ki-67 protein is a cellular marker and is used in IHC examination, which is strictly associated with the growth fraction of a given cell population. In this process, Ki-67 positive tumor cells will be stained to be brown, and Ki-67 negative ones will be colored blue, as shown in Fig 1 (b) and (d). The fraction of Ki-67 positive tumor cells is often correlated with the clinical course of cancer. Hence, Ki-67 IHC examination provides a selective, high-contrast imaging of cells and tissue components.

As we all know, IHC examination is essential to confirm the malignancy type and provide key prognostic factors that direct the treatment offered. For example, ER, PR, and HER2 biomarker expression levels are evaluated by IHC for breast cancer. However, high accuracy IHC examination usually needs more time and labor than H&E test. Consider of time-cost of IHC test and the raising of cancer incident cases,

Manuscript received February 15, 2021; revised March 22, 2021; accepted March 24, 2021. Date of publication March 30, 2021; date of current version July 30, 2021. This work was supported in part by the National Science Foundation of China (NSFC) under Grant 61875102, Grant 81871395, and Grant 61675113; in part by the Science and Technology Research Program of Shenzhen City under Grant JCYJ20170816161836562, Grant JCYJ20170817111912585, Grant JCYJ20160427183803458, Grant JCYJ20170412171856582, and Grant JCY20180508152528735; and in part by the Oversea Cooperation Foundation, Graduate School at Shenzhen, Tsinghua University, under Grant HW2018007. *(Shuting Liu and Baochang Zhang contributed equally to this work.) (Corresponding authors: Tian Guan; Yonghong He.)*

Shuting Liu, Yiqing Liu, Tian Guan, and Yonghong He are with the Department of Life and Health, Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China (e-mail: guantian@sz.tsinghua.edu.cn; heyh@sz.tsinghua.edu.cn).

Baochang Zhang is with the Computer Aided Medical Procedures (CAMP), Technische Universität München, 85748 Munich, Germany.

Anjia Han and Huijuan Shi are with the Department of Pathology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, China.

Digital Object Identifier 10.1109/TMI.2021.3069874

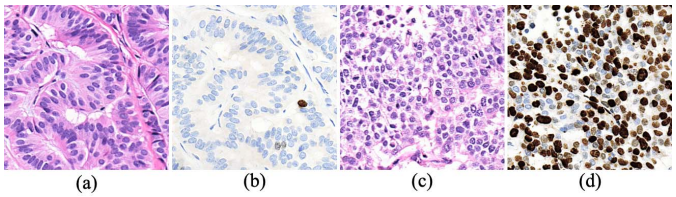

Fig. 1. Examples of H&E and Ki-67-stained image. (a) and (c) are the H&E-stained images; (b) and (d) are the corresponding Ki-67-stained images. (b) shows more Ki-67 negative tumor cells and (d) shows more Ki-67 positive tumor cells.

the workload of pathologists will increase gradually, which indeed needs more workforce, more efficient workflow and more advanced technologies. Meanwhile, both the high cost and the advanced technical skills required to perform the assay limit its utility in low- and middle- income countries. According to [4], only 1% of the cancer patients have access to IHC examination in some low-resource countries, which greatly hinds the process of global pathology diagnostic services for cancer. Furthermore, in many low- and middle-income countries, even when pathology services do exist, they are often under-resourced with a lack of trained health workforce, functional equipment and quality supplies, leading to unreliable quality and delay in diagnosis [5]. All of these issues show that developing an effective, efficient and low resource demanded method for cancer examination is significant.

Recently, researchers try to explore the correlation between H&E-stained images and Ki-67 IHC stained images, so as to reduce its cost. Liu *et al.* [6] developed a data-driven model to predict the positive or negative representation corresponding to Ki-67 staining of each cell from H&E-stained images, indicating a correlation between the two staining images. Therefore, there must be a way to model the pairwise relationship between H&E and Ki-67-stained images, achieving the generation of the one virtual stained image from another stained image. In histopathological image analysis field, the process of predicting one from another is called stain transfer, which aims to provide pathologists with different staining results of the same section, so as to improve the accuracy of cancer diagnosis. Meanwhile, compared with the traditional clinical staining process, the time cost will be reduced by tens or even hundreds of times, which will greatly improve the efficiency of cancer diagnosis. Hence, the generation of virtual Ki-67 staining sections from H&E staining images by computer assisted technology will be a feasible and novel solution for the existing issues.

In current histopathological practice, the staining process is almost irreversible, *i.e.*, once the tissue section is treated with a certain stain, it is hard to be restored to the pre-staining state by additional process, which makes it impossible to massive obtain pixel-level paired data. In such cases, the serial tissue sections are cut by pathologists from the same tissue block, each between $3\mu m$ and $5\mu m$ thick, and are stained with different methods and antibodies. This process introduces inevitable inter-slide variability in the cell and tissue structures. While performing serial cuts along the same axis, neighboring tissue sections are similar but not pixel-wise matching, which prevents co-location analyses across slides. Therefore, the pixel-wise unpaired data poses the first challenge for the stain transfer task.

During the last decade, many researchers are inspired by the profound ability of deep learning, which makes better use of contextual information and extracts powerful high-level features. Based on deep learning, a lot of creative and significant researches are conducted in many aspects of medical image analysis, such as X-ray [7], CT [8], PET [9], MRI [10] and histopathological image [11]. Usually, the performance of deep learning method is far superior to that of the traditional modeling method when the number of available samples is large [12]. However, the acquisition of the annotation is much more expensive for the training in the most of medical applications. As for histopathological image analysis, a whole slide image usually contains more than billions of pixels on the highest resolution level (e.g., 40x microscope), which poses great challenges to the pixel-level annotation. Furthermore, the H&E-stained image and the Ki67 stained image are not pixel-wise paired and they are hard to be registered completely by current registration methods.

Recently, Generative Adversarial Networks (GANs), as an important branch of deep learning, are usually used for data augmentation and style transfer [13]. However, GAN is mostly expected to generate various samples, which means that the network can be more *'creative'* and *'freedom'*. As for stain transfer, it is required to be *'regular'* and *'rigorous'*, rather than '*creative*' and '*freedom*'. Then, the second challenge posed by the stain transfer task is to ensure the preservation of microscopic structural details and the consistency of pathology, which are essential for the correct disease assessment.

In this study, we explore the potential of deep learning in unpaired image-to-image transformation in the field of histopathological analysis. To overcome the aforementioned two challenges in stain transfer task, we propose a novel adversarial learning method named *'PC-StainGAN'* for effective stain transfer between H&E domain and Ki-67 domain with a minimum annotation effort from pathologists. The major contributions of this article are as follows:

(1) We make a clear argument about the weakness of Cycle-GAN for pathological image analysis, especially stain transfer; and we proved the validity of our argument by experiment.

(2) We develop a novel and robust model to address 'H&E to Ki-67' unpaired stain transfer task, which merely requires a small number of simple annotations as expert knowledge to guide the model to learn pathological features, which significantly saves a lot of labor cost on annotations and facilitate the high performance.

(3) Pathology consistency constraint is first proposed to offset the weakness of *Cycle-GAN* and helps to translate the source image to target domain correctly. We also take fully advantage of structural similarity constraint and skip connection, and we find that structural similarity constraint cannot gain a great improvement on structural details preservation without the addition of skip connection.

(4) We demonstrate how we can successfully learn structural and pathology consistency stain transfer on different unpaired histopathological datasets, *i.e.,* neuroendocrine dataset and

breast dataset. Meanwhile, we provide massive qualitative and quantitative results on virtual Ki-67-stained image generation, which significantly presents the high performance and robustness of our method. Our code is available in: https://github.com/fightingkitty/PC-StainGAN.

II. Related Work

In fact, stain transfer can be regarded as a high demanding domain transfer or data synthesis task, which also requires that the generated image should be structurally and pathologically consistent with the source image. The existing works related to domain transfer can be divided into three categories: traditional Generative Adversarial Networks (GANs), conditional Generative Adversarial Networks (cGANs) and Image-to-image translation.

A. Generative Adversarial Networks (GANs)

GAN is first introduced by Goodfellow *et al.* who are inspired by the "two-gamer zero-sum game" in game theory [13]. A typical framework of GAN is conducted under the optimization of adversarial loss. At present, extraordinary success has been made using GANs in natural scene images processing tasks [14]–[17]. The GAN based model is also a very promising approach for medical image processing [18], [19]. Modelling the patterns in the histopathological images is a particularly complicated task for GANs, because morphologies of tissue have various texture patterns. In this circumstance, GANs are often used for data augmentation to generate more tissue regions. Due to the input of GANs is random noise, the generated samples are random and uncontrollable. Despite all that, the idea of adversarial loss is worth applying to stain transfer task.

B. Conditional GAN

Conditional GAN was proposed by Mirza *et al.* [20], where some additional conditions are incorporated into the generator and discriminator. These conditions can be manual annotation, some statistical information or other prior knowledge, which can help control the direction of generation results and improve the truthfulness of generated images. Currently, the idea of conditional image generation has also been successfully applied to pathological image analysis. For instance, Mahmood *et al.* proposed a method to overcome the diversity required in training data using synthetically generated data and utilized cGAN trained with synthetic and real data to achieve nuclei segmentation [21]. Bayramoglu *et al.* utilized cGAN to generate virtually H&E staining based on hyperspectral lung histology images [22]. Cho *et al.* [23] pointed out that performance of data-driven network for tumor classification varies with stain-style of histopathological images, and proposed a feature-preserving cGAN for stain transfer. As we all know, the size of histopathological image is pretty big, which makes it hard to obtain amounts of manual category annotations for training. Meanwhile, in the inference phase, the requirement of additional category information makes cGAN inefficient, which is the bottleneck of cGAN faced in the application of histopathological image analysis.

C. Image-to-Image Translation

Recent work has achieved impressive results in image-to-image translation, which can be divided into two parts according to whether the data is paired. For paired image-to-image translation, Pix2Pix is a successful variant of cGAN for high-resolution image-to-image translation, which is proposed by Isola *et al.* [16]. To alleviate the demand of paired data, Cycle-GAN, as an unpaired image-to-image translation framework, has been proposed by Zhu et al [24]. Cycle-GAN combines two GANs to learn the mapping between domain X and domain Y. A cycle consistency loss function is proposed to chain the two GANs together, which prompts them to reduce the distance between their possible mapping functions. Cycle-GAN has gained more and more attentions in stain normalization, cross modality transfer, *i.e.,* CT-to-MR [25], MR-to-CT [26], low dose CT denoising [27]. Gadermayr *et al.* [28] develop a fully-unsupervised segmentation approach exploiting Cycle-GAN to convert from the image to the label domain. Shaban *et al.* [29] proposed a StainGAN model based on Cycle-GAN for histopathology color normalization. Lahiani *et al.* [30] introduced an improved Cycle-GAN with the use of perceptual embedding consistency loss to generate virtual FAP-CK images from real stained H&E images. However, we find that the constraint ability of Cycle-GAN is not strong enough to ensure the consistency of pathology between the generated image and source image.

III. Method

In this part, we first restate the problem for stain transfer task. Then, we analyze the weakness of Cycle-GAN in this stain transfer task. Finally, we describe our proposed method for virtual Ki-67 generation from H&E-stained image, which can not only reserve important structural information, but also ensure the consistency of pathology for different staining methods.

A. Problem Setting

To illustrate our problem, we assume that we are given a H&E histopathology image dataset $\mathcal{D}_X$ and a Ki-67 histopathology image dataset $\mathcal{D}_Y$, which are unpaired. Firstly, we aim to find a style mapping between domain X and domain Y. In fact, a tissue section may contain various kinds of cells and cellular structures, like blood cell, positive tumor cell, negative tumor cell, stromal cell, nucleus, extracellular matrix and cytoplasm, which means that the domains X and Y can be divided into different subdomains. Therefore, the overall objective of the proposed method is to learn a mapping function F, which can not only achieve the low-level mapping (*i.e.,* looking like 'real') between the two domains X and Y, but also achieve a high-level mapping (*i.e.,* maintaining structural and pathological consistency) among these subdomains. It means that the desired mapping function F should meet the following equations,

$$\begin{cases} F(S_X) = S_Y \\ F(S_{Xi}) = S_{Yi};\ \ F(S_{Xi}) \neq S_{Yj}; i \neq ji, j = 1, 2, \cdots N \end{cases} \tag{1}$$

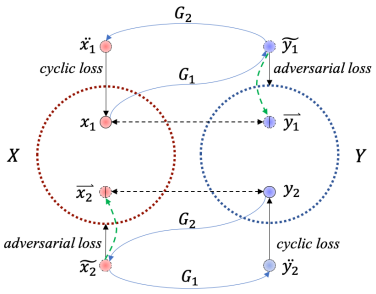

Fig. 2. Visual demonstration of Cycle-GAN's weakness in domain transfer. The green dotted-line arrows represent the constraint that Cycle-GAN lacks. X and Y are two different domains; x_1 and y_2 are the source images in corresponding domains; $\vec{y}_1$ and $\vec{x}_2$ are the target images of x_1 and y_2 respectively. $G_1 : X \rightarrow Y$ and $G_2 : Y \rightarrow X$ are the generators in Cycle-GAN; $\tilde{y_1}$ and $\tilde{x_2}$ are the generated images by corresponding generators; and $\ddot{x_1}$ and $\ddot{y_2}$ are the reconstructed images.

where $S_X \in \mathcal{D}_X$ is a H&E staining image sample; $S_Y \in \mathcal{D}_Y$ is a Ki-67 staining image sample; $X_i \in X$ is a subdomain in domain X; $Y_i, Y_j \in Y$ are two different subdomains in domain Y.

B. Weakness of Cycle-GAN

Cycle-GAN achieves style transfer between two different domains under the constrains of cycle consistency loss and adversarial loss. However, for complex texture domain transfer, Cycle-GAN shows incompetent and weak constraint. As shown in Fig.2, assuming that $\vec{y}_1 \in Y$ and $\vec{x}_2 \in X$ are the target images of x_1 and y_2 respectively, which are not existed in the given datasets $\mathcal{D}_Y$ and $\mathcal{D}_X$. Our goal is to build a generative model which can obtain the target staining image $\vec{y}_1$ from the image x_1. In fact, the cycle consistency loss is mainly employed to shorten the distance between the reconstructed image and the source one, *i.e.,* $min\{dist(x_1, \ddot{x_1})\}$ and $min\{dist(y_2, \ddot{y_2})\}$, which helps to reserve the structural content of source image and enforces that different source images will produce different images in target domain; the adversarial loss is mainly responsible for reducing the distance between the generated image and the target domain, *i.e.* $min\{dist(\tilde{y_1}, Y)\}$ and $min\{dist(\tilde{x_2}, X)\}$, which just achieves the low-level style mapping between the two domains X and Y and is powerless for the abovementioned high-level mapping. Therefore, Cycle-GAN lacks some constrains between the generated image and the target one, *i.e.,* $min\left\{dist\left(\tilde{y_1}, \vec{y}_1\right)\right\}$ and $min\left\{dist\left(\tilde{x_2}, \vec{x}_2\right)\right\}$, which are highlighted in Fig.2 by green dotted-line arrows.

C. Model Definition

1) Network Architecture: Although, it is infeasible to obtain the other staining results by re-staining sections in clinical practice. With the help of computer technology, we can model the stain transfer process into two phases: de-staining phase and re-staining phase. As shown in Fig.3, each generator is composed of an encoder-decoder architecture and a pathological representation network, where the feature extraction part of pathological representation network and the encoder part correspond to the de-staining phase; and the decoder part corresponds to the re-staining phase. Hence, the transformation of $G_1 : X \rightarrow Y$ is modeled as follows,

$$y = G_1(x) = G_{1_{de}}(G_{1_{en}}(x), P'_X(x)) \tag{2}$$

and the transformation of $G_2 : Y \rightarrow X$ is modeled as follows,

$$x = G_2(y) = G_{2_{de}}(G_{2_{en}}(y), P'_Y(y)) \tag{3}$$

where $G_{1_{en}}, G_{2_{en}}$ are the encoders of corresponding generators; $G_{1_{de}}, G_{2_{de}}$ are the decoders of corresponding generators; P'_X, P'_Y are the feature extractors of pathological representation networks P_X, P_Y, which are used to extract pathology-semantic features.

In the de-staining part, the patches with the size of $288 \times 288 \times 3$ is cropped from staining image as the input of encoder. The encoder then starts by a convolution with a kernel size of 7×7 and stride of 1. In order to maintain the spatial continuity and reserve more structural details, no pooling operation is adopted, instead, the convolution with a stride of 2 is employed. Thus, the input image is down-sampled from 288×288 to 36×36 after three convolutions with the kernel of 3 and stride of 2. The following module is a high-performance feature extractor which consists of five residual convolution blocks [31], as shown by orange bold arrow in Fig.3.

The primary goal of encoder is to extract principal structural and morphological content from source staining image. As for the pathological representation network, the purpose is to provide necessary pathological information so as to drive the generated image to be consistent with the original image in the pathological representation. On the basis of encoder, we select the features in different resolution layers (*i.e.,* 144×144, 72×72, 36×36) and adjust the size of these features to 72×72 by B-spline interpolation and down-sampling. Considering smaller computer memory consumption, a convolution with the stride of 2 and kernel of 3 is then used to fuse these features together. Similarly, the following operations are five residual convolution blocks to extract pathological features efficiently. Finally, a pathological representation map can be obtained by a 3×3 convolution and a 'Sigmoid' activation function.

In the re-staining phase, the extracted features first go through a five-residual-convolution-blocks module to fully integrate the features. Next, aiming to recover the resolution of the feature map from 36×36 to 288×288, deconvolution with the kernel of 3×3 and stride of 2 is utilized to up-sample the feature maps. Finally, the virtual staining image is obtained by a 7×7 convolution and a 'Tanh' activation function. Meanwhile, the skip connections are added between the encoder and decoder under the same resolution. The addition of skip connections allows us to share low level information between input and output, which also provides shortcuts for reserving and reconstructing structural details. It is worth to note that each convolution layer, as shown in Fig.3, is a series of operations, *i.e.,* convolution with a kernel of 3×3, instance normalization, and 'Leaky ReLU' activation layer.

2) Objective: The overall objective function of our model includes four loss types. Except for the adversarial loss and the cycle consistency loss described in [24], we add cycle structural consistency loss, pathological consistency loss $\mathcal{L}_{pathology}$

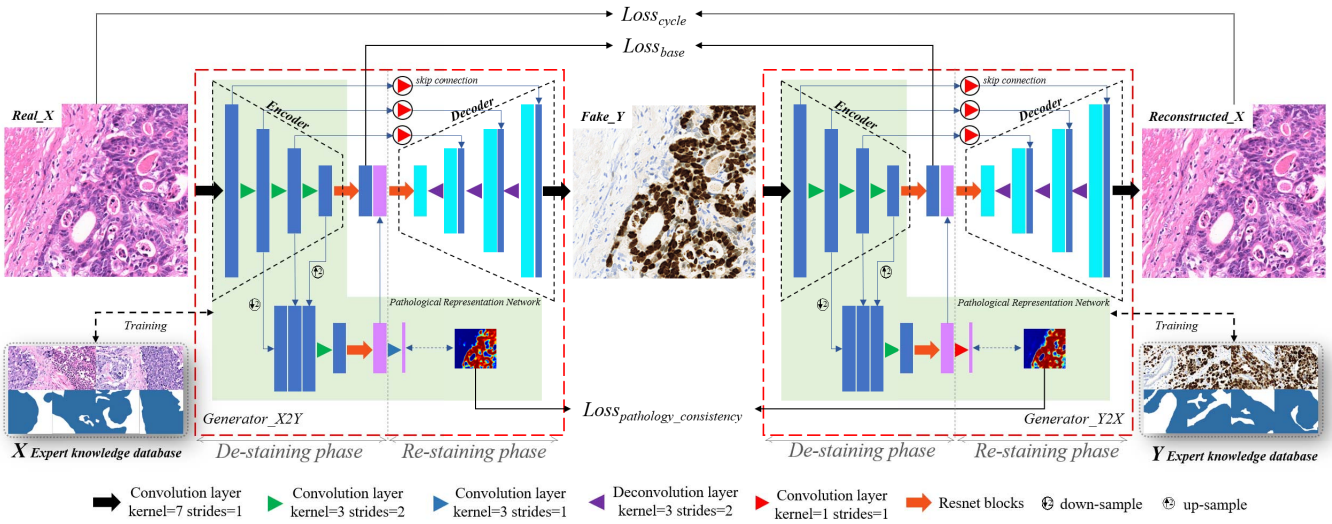

Fig. 3. The overview of the proposed method, which include two generators. Each generator is composed of an encoder-decoder architecture and a pathological representation network. The pathological representation network is co-trained by expert knowledge database and training dataset. The expert knowledge dataset is annotated by experienced pathologists, where blue areas are cancer lesion areas and white areas are the normal tissue areas. And the objective function includes: adversarial loss, cycle consistency loss, pathological consistency loss and base space aligned loss, where cycle consistency loss includes *L1* loss and structural similarity constraint *SSIM* loss.

and base space aligned loss $\mathcal{L}_{base}$.

$$\begin{aligned}&\mathcal{L}\left(G_1, G_2, D_X, D_Y, P_X, P_Y\right)\\&=\mathcal{L}_{adv}\left(G_1, D_Y\right)+\mathcal{L}_{adv}\left(G_2, D_X\right)\\&\quad+\lambda\mathcal{L}_{cycle}\left(G_1, G_2\right)+\beta\mathcal{L}_{pathology}\left(P_X, P_Y\right)\\&\quad+\gamma\mathcal{L}_{base}\left(G_1, G_2\right)\end{aligned}\tag{4}$$

where λ, β, γ are hyper-parameters that set the importance of each term in the optimization problem; $G_1 : X \rightarrow Y, G_2 : Y \rightarrow X$ are the two generators, D_X is the discriminator for X data, and D_Y is the discriminator for Y data; P_X is the pathological representation network with the input of X data, and P_Y is the pathological representation network with the input of Y data. Formally, we aim to solve the following optimization problem:

$$G_1^*, G_2^* = \arg\min_{G_1,G_2,P_X,P_Y}\max_{D_X,D_Y}\mathcal{L}\left(G_1, G_2, D_X, D_Y, P_X, P_Y\right)\tag{5}$$

The whole process involves playing a minimax game among these networks. Below, each term of the objective function is elaborated.

3) Transferring Stain Style: Instead of matching RGB values, the network aims to generate images with staining properties of the target domain. It means matching the distribution of the generated image to that of the target domain. Therefore, we adopt adversarial losses for the two mapping functions between the two domains X and Y. For the mapping function $G_1 : X \rightarrow Y$ and its corresponding discriminator D_Y, the adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_{adv}\left(G_1, D_Y\right)=&\mathbb{E}_y\left[\log D_Y\left(y\right)\right]\\&+\mathbb{E}_x\left[\log\left(1-D_Y\left(G_1\left(x\right)\right)\right)\right]\end{aligned}\tag{6}$$

where G_1 tries to generate an image $G_1(x)$ with the input of source image x that looks similar to one from domain Y, while D_Y aims to distinguish between the generated sample $G_1(x)$ and real sample y. G_1 aims to minimize this objective, while D_Y tries to maximize it. Similarly, the objective for the mapping function $G_2 : Y \rightarrow X$ and its corresponding discriminator D_X is defined as:

$$\begin{aligned}\mathcal{L}_{adv}\left(G_2, D_X\right)=&\mathbb{E}_x\left[\log D_X\left(x\right)\right]\\&+\mathbb{E}_y\left[\log\left(1-D_X\left(G_2\left(y\right)\right)\right)\right]\end{aligned}\tag{7}$$

4) Enhancing Structural Information Reservation: As mentioned in [13], adversarial loss alone is hard to ensure that the mapping function can accurately translate a specific input x to a desired output y. To further reduce the space of possible mapping functions, they introduced the cycle consistency loss, which implies that image transformation cycle should bring input image x back to the original image space. In fact, the controllability and specificity of the image transformation network are greatly improved by the addition of pathology-semantic features in our method, as mentioned in formula (2). Meanwhile, in addition to the abovementioned advantage of cycle consistency loss, we pay more attention to its structural preservation ability. To further enforce the generated image preserving more tissue structural details, a cycle structural consistency loss function is defined based on structural similarity (SSIM) [32] and introduced to the original cycle consistency loss. In many related studies, SSIM has been used to evaluate image quality, and it mainly evaluates three measures, *i.e.,* luminance, contrast and structure. For each pixel in the image, the SSIM is defined as:

$$ssim\left(a, b\right)=\frac{\left(2\mu_a\mu_b+c_1\right)\left(2\sigma_{ab}+c_2\right)}{\left(\mu_a^2+\mu_b^2+c_1\right)\left(\sigma_a^2+\sigma_b^2+c_2\right)}\tag{8}$$

where μ_a, μ_b are the mean of a sliding window $(N \times N)$ centered as the pixel, σ_a, σ_b are the standard derivations, σ_{ab} is the covariance. c_1 and c_2 are stabilizing factors variables to stabilize the division with weak denominator.

Hence, our defined cycle consistency loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{cycle} = &\left\{\mathbb{E}_x\left[\|G_2\left(G_1\left(x\right)\right)-x\|_1\right]\right.\\&\left.+\mathbb{E}_y\left[\|G_1\left(G_2\left(y\right)\right)-y\|_1\right]\right\}\\&+\left\{\mathbb{E}_x\left[1-ssim\left(G_2\left(G_1\left(x\right)\right),x\right)\right]\right.\\&\left.+\mathbb{E}_y\left[1-ssim\left(G_1\left(G_2\left(y\right)\right),y\right)\right]\right\}\end{aligned} \tag{9}$$

5) Enforcing Pathological-Consistency Transformation: In our work, we model the stain transfer process into two phases: de-staining phase and re-staining phase, and we force the image x and corresponding image y mapping to same base space after de-staining phase, *i.e.,* $G_{1_{en}}(x) = o$; $G_{2_{en}}(y) = o$, which facilitate the image x and image y holding the same condensed anatomical information and pathological feature expression. Thus, we introduce a base space aligned loss function between the encoders of generators G_1,G_2. The addition of the base space aligned loss $\mathcal{L}_{base}$ based on L1-norm allows to minimize the distance between the base-space features in the two generators, which is defined as:

$$\begin{aligned}\mathcal{L}_{base} = &\mathbb{E}_x\left[\left\|G_{2_{en}}\left(G_1\left(x\right)\right)-G_{1_{en}}\left(x\right)\right\|_1\right]\\&+\mathbb{E}_y\left[\left\|G_{1_{en}}\left(G_2\left(y\right)\right)-G_{2_{en}}\left(y\right)\right\|_1\right]\end{aligned} \tag{10}$$

where G_1 and G_2 correspond to the generators in the model, $G_{1_{en}}$ and $G_{2_{en}}$ correspond to the encoders of the first and second generator respectively and $\|\|_1$ is the L1 distance.

For multi staining histopathological image analysis, the most important thing is ensuring the consistency of pathology. In each generator, a pathological representation network is added, which is used to dig the pathological representation heatmap of the input image. In order to further improve the pathological consistency between the input image and generated image, we introduce an additional loss function to reduce the difference between the pathological representation heatmaps from input image x and generated image y respectively. Meanwhile, since the learned pathological heatmap is expected to be accordance with the diagnosis of the clinicians, the pathological representation networks P_X, P_Y are also trained by the expert knowledge databases K_X, K_Y respectively. And pathological consistency loss is finally defined as:

$$\begin{aligned}\mathcal{L}_{pathology} = &\mathbb{E}_x\left[\|P_X(x)-P_Y(G_1\left(x\right))\|_1\right]\\&+\mathbb{E}_y\left[\|P_Y(y)-P_X(G_2\left(y\right))\|_1\right]\\&+w_1*\mathbb{E}_{k\_x}\left[\|P_X(k\_x)-label(k\_x)\|_1\right]\\&+w_2*\mathbb{E}_{k\_y}\left[\|P_Y(k\_y)-label(k\_y)\|_1\right]\end{aligned} \tag{11}$$

where P_X, P_Y are the pathological representation networks for domains X, Y respectively; k_x, k_y are the samples from expert knowledge databases K_X, K_Y respectively, and $label\,(*)$ is the expert annotation of the sample $*$; w_1, w_2 are the weights for supervised learning process, which are set to 2.0.

D. Model Implementation

Our model is implemented with Python based on the open-source deep learning library Pytorch on a computer with Intel Core i7-6850k CPU, 128 GB RAM, and three NVidia GTX 1080-Ti GPUs. In the training phase, the patch size is set to 288 × 288 × 3. While in the inference phase, a bigger input size of 576 × 576 is used to overcome the tiling artifact problem by using overlapping tiles. In addition, two data augmentation strategies are employed during training phase, including random cropping and random flipping.

In our model, all the parameters in the convolutional layers are initialized according to [33]. The hyper-parameters in the overall loss are chosen similarly to [30] and fixed as: $\lambda = 10, \beta = 5, \gamma = 5$. Adam optimizer [34] is utilized to minimize the overall loss. The whole model is trained end-to-end using backpropagation. The batch size of training dataset is set to 2, the batch size of expert knowledge database is set to be 8, and the learning rate is set as 0.0002 initially and decreases using exponential decay with the decay rate of 0.9 and the decay epoch of 2.

IV. Experiments and Results

In this paper, two datasets are utilized to evaluate the performance of the proposed method. The relevant information and the corresponding results are as follows.

A. Experimental Datasets

For neuroendocrine tumor dataset, 150 pieces of H&E staining images with the size of 3000 × 3000 were collected and the same quantity of Ki-67 staining images were sampled, which were unpaired and used as the training sets. For the expert knowledge database, 15 images were selected from H&E training set and Ki-67 training set respectively. And the expert knowledge dataset is annotated by experienced pathologists, where cancer lesion areas are labelled as '1' and the normal tissue areas are labelled as '0', as shown in Fig.3. In addition, we collected 42 pairs H&E and Ki-67 staining images with the size of 3000 × 3000, which were basically similar in tissue structure but not pixel-level matched. In order to evaluate our method properly, the 42 pairs H&E and Ki-67 staining images are registered to each other using an alignment technique, and then the sizes of paired images were cropped to 1500 × 1500. It is worth to note that the intersection of training set and testing set are empty.

Similarly, in breast dataset, there are 160 H&E staining images and 160 Ki-67 staining images are collected as training sets with the size of 3000 × 3000. Then, 20 images were selected as expert knowledge database from H&E training set and Ki-67 training set respectively, which are annotated by pathologists. As for the testing set, 54 pairs H&E and Ki-67 staining images with the size of 1500 × 1500 were collected, and these paired images were coarsely aligned using registration technology.

Due to memory limitations, all the training images in the both datasets were split into 288 × 288 tiles with 144 overlap. After tiling, our neuroendocrine training dataset contains about 54000 H&E 288 × 288 RGB tiles and 54000 Ki-67 288 × 288 RGB tiles; our breast training dataset contains about 57000 H&E 288 × 288 RGB tiles and 57000 Ki-67 288 × 288 RGB tiles.

B. Experimental Design and Baseline Method

In this paper, due to the unavailability of pixel-level paired data, a commonly used unpaired image-to-image translation method 'Cycle-GAN' is regarded as the baseline method. In order to better understand our work, we show a series of experiments on neuroendocrine dataset, where you can learn the model evolution process and the superiority of our proposed method. For example, (*i*) through the study of skip connection and cycle structural consistency loss on image details reservation, you can know why they are employed; (*ii*) through the study of pathological consistency constraint on the performance of virtual Ki-67 generation, you can learn why our proposed pathological consistency constraint is important for stain transfer task; (*iii*) through the experiment on unbalanced training sets and the experiment of expert knowledge degradation, you can further confirm the superiority of our proposed method. Furthermore, we explore the generality of the proposed method by evaluating its performance on another histopathology dataset, *i.e.,* breast dataset.

C. Evaluation Metrics

Actually, for the stain transfer task of $X \rightarrow Y$, there are four types of images, *i.e.,* source image X, generated image $\tilde{Y}$, reconstructed image $\ddot{X}$ and referenced image Y.

1) For $X \leftrightarrow \ddot{X}$: According to [35], they have summarized various evaluation metrics for different medical image tasks using GAN, including domain transfer. In this work, four metrics are selected from their review work as our quantitative evaluation metrics, *i.e., SSIM*, multi-scale SSIM (*MS-SSIM*), mean absolute error (*MAE*) and peak signal to noise ratio (*PSNR*).

2) For $X \leftrightarrow \tilde{Y}$: Due to inherent distance between two different stain domains X and Y, it is unjustified to use these metrics mentioned in section IV.C.1. Therefore, based on the definition on SSIM, Contrast-Structure Similarity (CSS) is proposed to evaluate how much structural information is preserved from the source image. CSS can be regarded as a variant of SSIM, which mainly evaluates the similarity of samples on contrast and structure, rather than intensity. The definition of CSS is as follow:

$$CSS\left(a,b\right)=\frac{\left(2\sigma_{ab}+c\right)}{\left(\sigma_a^2+\sigma_b^2+c\right)} \tag{12}$$

where σ_a, σ_b are the standard derivations, σ_{ab} is the covariance. c is a stabilizing factor variables to stabilize the division with weak denominator. The meanings of all the variables are same as these in SSIM.

3) For $\tilde{Y} \leftrightarrow Y$: In order to evaluate the relationship of the generated Ki-67-stained images and referenced Ki-67-stained image, the positive and negative Ki-67-stained areas of each image are calculated after color deconvolution and threshold segmentation, which is shown in Fig.4. Then, Pearson correlation coefficient (Pearson-R) is used to evaluate the pathological correlation between the generated and the referenced.

In addition, a channel-level evaluation metric named 'Perceptual Hash Value' (PHV) is designed to thoroughly evaluate the similarity of the generated image and the referenced one.

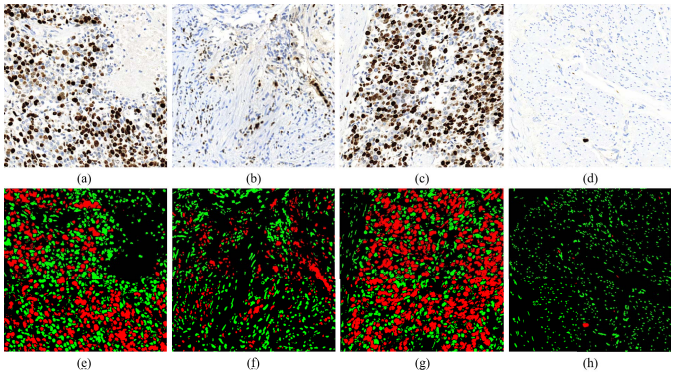

Fig. 4. Ki-67 positive and negative stained area segmentation results. (a), (b), (c) and (d) are the Ki67 stained images; (e), (f), (g) and (h) are the corresponding segmentation masks. Red: Ki-67 positive stained area; Green: Ki-67 negative stained area.

This metric is inspired by perceptual loss [36]. The feature maps of the generated images are extracted using a pre-trained Resnet-101 network. Similarly, the feature maps of referenced images are also extracted. Hence, the PHV is defined as follows,

$$PHV=\frac{1}{N}\sum H\left[\left|avg\left(F_i\left(\tilde{y}\right)\right)-avg\left(F_i\left(y\right)\right)\right|-T\right] \tag{13}$$

where N is total channel number of the extracted features. $F_i\left(\cdot\right)$ represents the feature maps extracted from the i_{th} layer of Resnet-101. $\tilde{y}$ and y are the generated stained image and refereced image respectively. $avg\left(\cdot\right)$ is the average pooling operation which is used to turn the 3-D features to a 1-D vector, and $H\left[\cdot\right]$ is the unit step function and T is a preset threshold.

D. Experimental Results on Neuroendocrine Dataset

1) Importance of SSIM Constraint and Skip Connection: Based on the analysis of cycle-consistency loss in Cycle-GAN [24], we speculate that the addition of skip connection and SSIM constraint (*i.e.,* cycle structural consistency loss) could further improve structural details preservation. Therefore, in order to study the importance of SSIM constraint and skip connection for stain transfer, '*Cycle-GAN*', '*Cycle-GAN+skip-connection*', '*Cycle-GAN+SSIM Constraint*' and '*Cycle-GAN+skip-connection+SSIM Constraint*' are trained with same training set. Table I shows their averaged structure preservation performances by the comparison of source image and reconstructed image $X \leftrightarrow \ddot{X}$. Comparing with *Cycle-GAN*, the performance of *Cycle-GAN+SSIM Constraint* on *SSIM* and *MS-SSIM* metrics has a slight rise, but the performances on *MAE* and *PSNR* metrics are worse than that of *Cycle-GAN*, which means the addition of SSIM constraint indeed guide the network to pay more attention on structural similarity but the improvement on structural details preservation is negligible. Under the same constraint condition as *Cycle-GAN*, we can find that the performance of *Cycle-GAN+skip-connection* gains an improvement on each metric, which indicates that the short paths offered by the skip connection facilitate the broadcast of structural information. From Table I, it is obvious that

TABLE I
QUANTITATIVE EVALUATION RESULTS FOR STRUCTURAL INFORMATION PRESERVATION, $X \leftrightarrow \ddot{X}$: BETWEEN SOURCE IMAGES AND RECONSTRUCTED IMAGE AND $X \leftrightarrow \tilde{Y}$: BETWEEN SOURCE IMAGE AND GENERATED IMAGE

	$X \leftrightarrow \ddot{X}$				$X \leftrightarrow \tilde{Y}$
	MAE	*SSIM*	*MS-SSIM*	*PSNR*	*CSS*
Cycle-GAN	5.59±1.00	94.40±0.88	97.43± 0.30	29.66±1.42	78.99±4.91
Cycle-GAN+SC	4.73±0.52	96.83±0.96	98.48±0.41	31.23±0.93	82.07±3.29
Cycle-GAN+SSIM	6.23±1.11	96.25±0.50	98.39±0.16	29.52±1.33	80.92±3.37
Cycle-GAN+SC+SSIM	3.60±0.78	98.46±0.26	99.33±0.09	33.62±1.30	87.43±2.63

NOTE: 'SC' MEANS SKIP-CONNECTION; 'SSIM' MEANS SSIM CONSTRAINT.

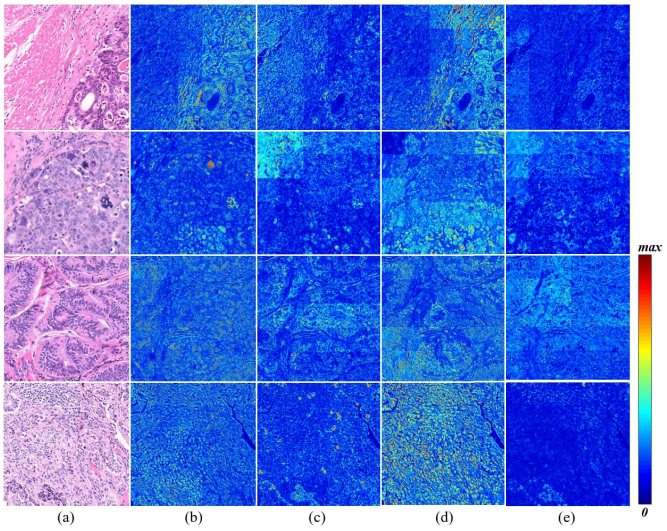

Fig. 5. The distance maps of different methods between the reconstructed image and the source image. The columns from left to right correspond to (a) original image and distance maps of (b) *Cycle-GAN*, (c) *Cycle-GAN+skip-connection*, (d) *Cycle-GAN+SSIM constraint*, and (e) *Cycle-GAN+skip-connection+SSIM constraint*.

Cycle-GAN+skip-connection+SSIM Constraint makes an arresting improvement, which confirms the structural perseveration ability of skip connection and SSIM constraint. Meanwhile, comparing *Cycle-GAN+SSIM Constraint* and *Cycle-GAN+skip-connection+SSIM Constraint*, we can conclude that SSIM constraint will be powerless to structural details preservation if without the addition of skip connection. More intuitive results are shown in Fig.5, which shows the distance maps of different methods between the reconstructed images and the source images. The darker the color blue, the smaller distance value between the reconstructed image and the source image it has. Observing Fig.5, It can be further confirmed that the reconstructed images are closer to the source images with the addition of the skip connection and SSIM constraint.

Meanwhile, we use the metric 'CSS' to evaluate how much structural information has been preserved by the generated Ki-67-stained image from the source H&E-stained image. And the quantitative evaluation results are shown in the column X $\leftrightarrow \tilde{Y}$ of Tab.I. It is obvious that the addition of skip connection and SSIM constraint gains a significant improvement on structural information preservation (*e.g.,* the rise of *CSS* from 78.99 to 87.43). In a word, all the experimental results indicate that the addition of SSIM constraint and skip connection is important for the structural preservation. Hence, skip connection and SSIM constraint are employed in our work.

2) Importance of Pathology Consistency Constraint: Currently, many unpaired stain transfer works are based on Cycle-GAN, but they ignore the serious weakness of Cycle-GAN in pathological image analysis. In our work, pathology consistence constraint is proposed to address this issue. Before studying the importance of pathology consistence constraint, we would like to mention the work of Lahiani *et al.* According to their work, they did massive experiments and found that base space aligned loss is conducive to reducing the tiling artifact [30]. Hence, we adapt their work as base space aligned loss in our work, and the difference is that we use L1-norm rather than L2-norm.

In order to study the importance of pathology consistence constraint, Cycle-GAN with the addition of skip connection, SSIM constraint and base space aligned loss $\mathcal{L}_{base}$ is regarded as the main comparison method, i.e., PC-StainGAN without pathology consistence constraint. Meanwhile, some other unpaired stain transfer frameworks are trained using the same training sets and hardware device. The virtual Ki-67-stained images generated by different methods are presented in Fig.6 for a visual comparison. We can intuitively observe that, with the constraint of pathology consistency, virtual Ki-67-stained images generated by our proposed PC-StainGAN are more similar to the referenced Ki-67-stained images, while the results of other methods present obvious mistakes in pathological representation.

In addition, the Ki-67 positive and negative stained areas are calculated under a sliding window with the size of 750 × 750 and the step of 375. Generally, for Ki-67-stained image analysis, the pathologists pay more attention to the Ki-67-stained positive cells. Therefore, the correlation between the generated images and referenced ones is studied via scatter diagram, which is presented in Fig.7. The red dotted line represents the distribution trend of scattered blue points, and the regression equation and the square of the correlation coefficient are also shown in Fig.7. It is clearly observed that our methods show stronger correlation between the generated images and referenced images. Meanwhile, the quantitative evaluation results are shown in Table II. For Ki-67-stained positive area, the *Pearson-R* of our proposed *PC-StainGAN* gains 0.9755, which are much higher than comparison methods (*e.g.,* 0.6974 of *Cycle-GAN* and 0.6381 of *PC-StainGAN without pathology consistence constraint*). In addition, observing the scores of *PHV* calculated from different layers, our method also significantly outperforms the comparison methods (*e.g., PC-StainGAN* with the *PHV-II* score of 91.32 *vs.* 86.36 of *PC-StainGAN without pathology consistence constraint*). All the experimental results validate that the pathological consistency constraint plays an important role in virtual Ki-67-stained image generation, which helps the generated virtual Ki-67-stained image have highly similar pathological representation with the real Ki-67-stained image.

3) Against Unbalanced Training Dataset: In order to study the robustness of our proposed *PC-StainGAN* on unbalanced datasets, we divided our training set into three subsets with

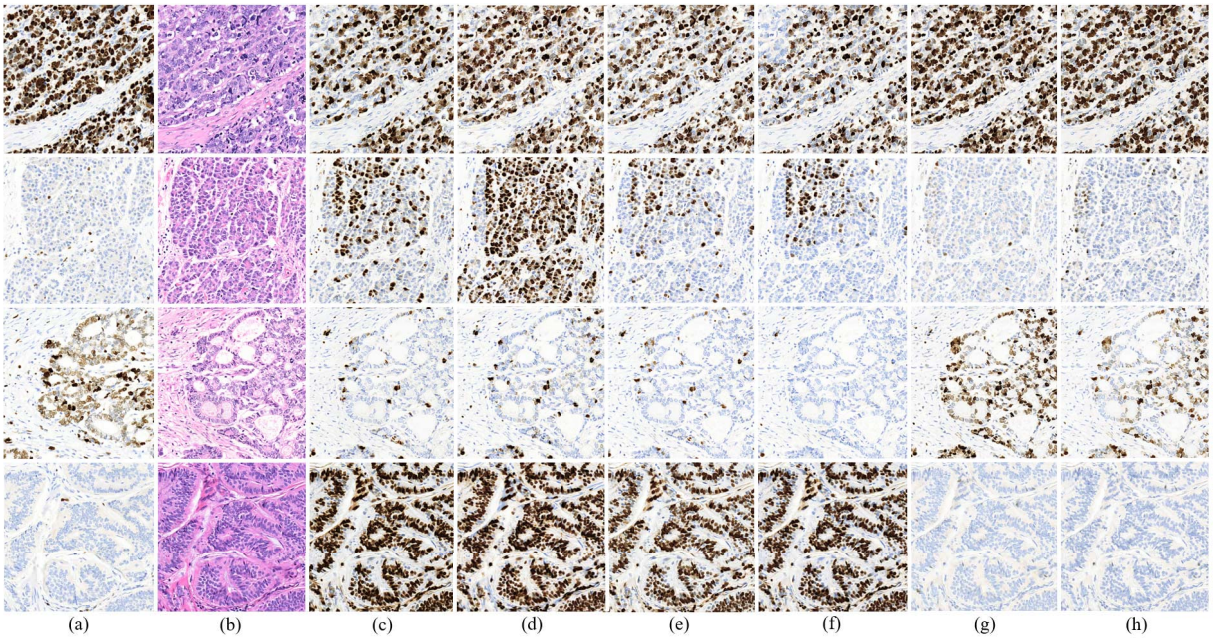

Fig. 6. Virtual Ki-67-stained image results of different methods on neuroendocrine dataset. These columns from left to right are: (a) referenced Ki-67-stained images, (b) source H&E-stained image, and generated Ki-67-stained images by (c) *Cycle-GAN [25], [30]*, (d) *Lahiani et al's method [31]*, (e) *Cycle-GAN+skip-connection+SSIM constraint*, (f) *PC-StainGAN without pathology consistence constraint*, (g) *PC-StainGAN* and (h) *PC-StainGAN-loose*.

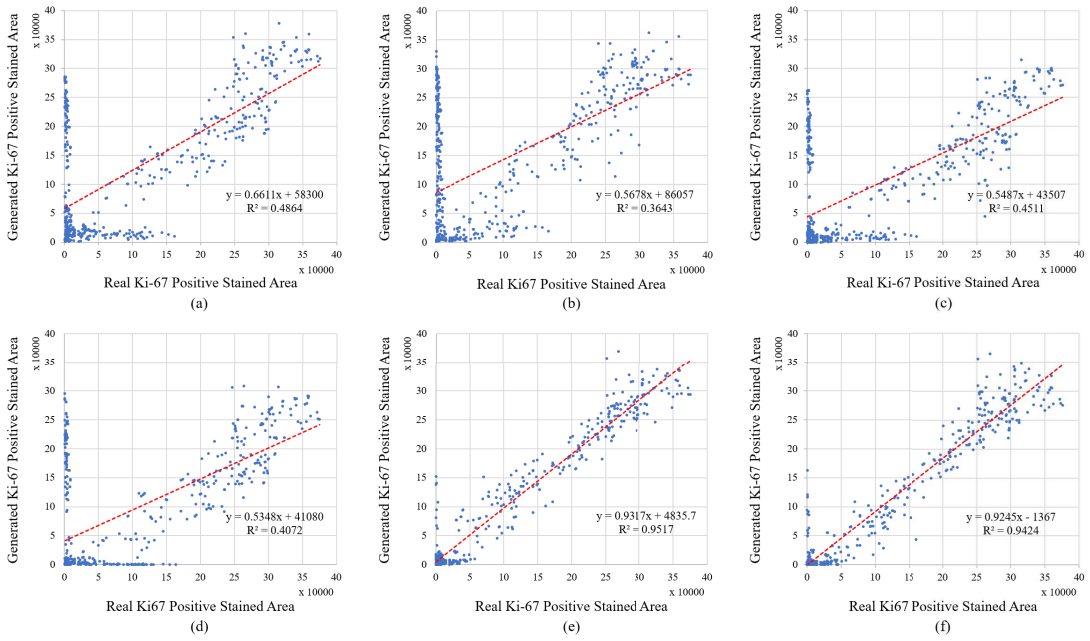

Fig. 7. Scatter diagrams of different methods of the Ki-67 positive stained area between the generated and the reference images. (a) *Cycle-GAN [25], [30]*, (b) *Lahiani et al's method [31]*, (c) *Cycle-GAN+skip-connection+SSIM constraint*, (d) *PC-StainGAN without pathology consistence constraint*, (e) *PC-StainGAN* and (f) *PC-StainGAN-loose*.

different Ki-67-stained positive/negative area ratios, *i.e.,* subset I with ratio of 1:1, subset II with ratio of 1:3 and subset III with ratio of 3:1. Each training subset contains about 72 H&E-stained images and 72 Ki-67-stained images both with the size of 3000 × 3000. Then, our proposed *PC-StainGAN*, *Cycle-GAN* [24], [29] and *Lahiani et al*'s method [30]. are trained with the three predesigned training subsets respectively. A visual comparison is presented in Fig.8. It is obvious that *PC-StainGAN* achieves a stable and correct performance on both unbalanced and balanced training datasets, which fully verifies the robustness of our method. From Fig.8, we also find that the referenced Ki-67-stained image shows more negative cells and the results produced by both *Cycle-GAN* and *Lahiani et al*'s method are similar to the referenced image only when the training set contains more negative samples. Whereas, when there are more positive samples in the training set, their

TABLE II
QUANTITATIVE EVALUATION RESULTS OF VIRTUAL KI-67-STAINED IMAGE GENERATION ON NEUROENDOCRINE DATASET

	Pearson-R (*positive*)	*Pearson-R* (*negative*)	*PHV I* (*T*=0.005) %	*PHV II* (*T*=0.005) %	*PHV III* (*T*=0.005) %
Cycle-GAN [25,30]	0.6974	0.2908	70.74±14.11	85.14±9.39	86.06±7.60
Lahiani et al [31]	0.6035	0.1966	69.33±13.57	84.61±7.99	84.79±6.87
Cycle-GAN+SC+SSIM	0.6716	0.2283	71.33±11.84	85.81±8.31	86.81±7.36
PC-StainGAN without PC	0.6381	0.2144	70.58±12.84	86.36±9.12	86.45±7.50
PC-StainGAN	0.9755	0.8665	79.52±10.57	91.32±7.42	90.46±4.85
PC-StainGAN-loose	0.9707	0.8609	78.46±10.94	89.72±8.32	89.50±6.48

Note:'*PHV n*' means it is calculated based on the layer '*n*' in the Resnet-101.

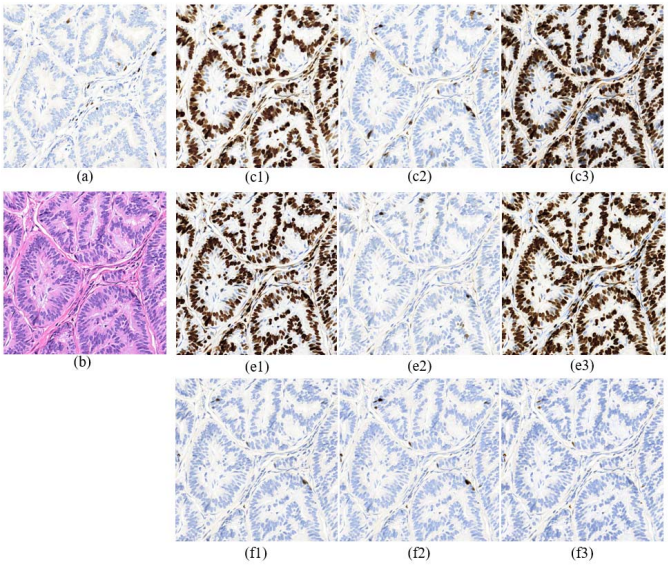

Fig. 8. Experiment results of our method and Cycle-GAN on unbalanced datasets. Referenced Ki-67-stained images (a), original input H&E-stained image (b), generated Ki-67-stained images by Cycle-GAN trained on *Subset I* (c1), *Subset II* (c2) and *Subset III* (c3), generated Ki-67-stained images by Lahiani et al's method [31] trained on *Subset I* (e1), *Subset II* (e2) and *Subset III* (e3), generated Ki-67-stained images by our method trained on *Subset I* (f1), *Subset II* (f2) and *Subset III* (f3).

results turn to be positive property, which means *Cycle-GAN* and *Lahiani et al*'s method are volatile and pretty sensitive to the imbalance of training dataset. Meanwhile, the quantitative evaluation results are listed in Table III. Under the attack of unbalanced training set, the performance of our *PC-StainGAN* just shows a slight fluctuation, (*e.g.,* the standard deviations are 0.0042 for positive part and 0.0088 for negative part), while the standard deviations of *Cycle-GAN* reach to 0.0845 for positive part and 0.0402 for negative part, and the standard deviations of *Lahiani et al*'s method reach to 0.0957 for positive part and 0.0386 for negative part. All the experimental results indicate the superiority of *PC-StainGAN*.

4) Learn With Degraded Expert Knowledge: Under the pathology consistence constraint, the pathological representation networks are also trained by expert knowledge, which helps the learned pathological features are accordant with the diagnosis of the clinicians. Hence, we are also interested in the impact of expert knowledge on the performance of *PC-StainGAN*. We degrade the annotations in the expert knowledge by converting the pixel-level annotation to image-level annotation. If the cancer lesion area is bigger than the normal

TABLE III
QUANTITATIVE EVALUATION RESULTS ON UNBALANCED TRAINING DATASETS

	Subset I (*P*:*N*=1:1)		*Subset II* (*P*:*N*=1:3)		*Subset III* (*P*:*N*=3:1)		*Mean±SD*	
	Pearson-R (*positive*)	*Pearson-R* (*negative*)	*Pearson-R* (*positive*)	*Pearson-R* (*negative*)	*Pearson-R* (*positive*)	*Pearson-R* (*negative*)	*Pearson-R* (*positive*)	*Pearson-R* (*negative*)
Cycle-GAN [25,30]	0.6812	0.2438	0.4783	0.2295	0.5443	0.1522	0.5679±0.0845	0.2085±0.0402
Lahiani et al [31]	0.6209	0.1161	0.4341	0.2107	0.6504	0.1668	0.5684±0.0957	0.1645±0.0386
PC-StainGAN	0.9327	0.7747	0.9249	0.7861	0.9347	0.7964	0.9307±0.0042	0.7857±0.0088

Note: 'SD' means standard deviation.

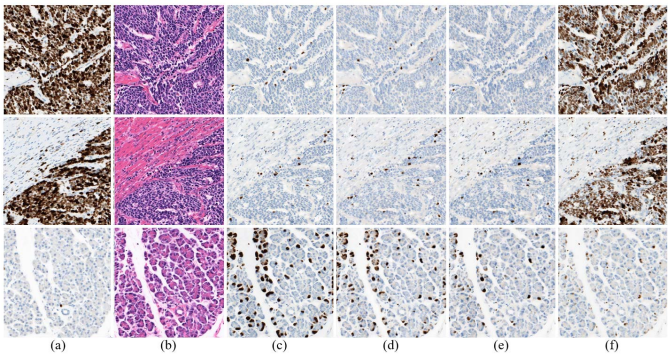

Fig. 9. Experiment results of different methods on breast dataset. These columns from left to right are: (a) referenced Ki-67-stained images, (b) source H&E-stained image, and generated Ki-67-stained images by (c) *Cycle-GAN* [25], [30], (d) *Lahiani et al*'s method [31], (e) *Cycle-GAN+skip-connection+SSIM constraint* and (f) *PC-StainGAN*.

tissue area, we will label the image as a positive sample, otherwise, we will label it as a negative sample. And *PC-StainGAN-loose* is trained, where the pathological representation network is adjusted as pathological classification network to tell the pathological property of given image, *i.e.,* 'positive' or 'negative'. Observing Fig. 6, it is obvious that the virtual Ki-67-stained images generated by *PC-StainGAN-loose* are also pretty close to the referenced Ki-67-stained images. From Table II, we can find that the performance of *PC-StainGAN-loose* is also better than the other comparison method except *PC-StainGAN*, which indicates that the performance of our method has not decreased significantly with the degradation of supervised information. Compared with *PC-StainGAN*, *PC-StainGAN-loose* also gains comparable performance, only has a slight decrease in performance, *e.g., PC-StainGAN* with positive the *Pearson-R* score of 0.9755 *vs.* 0.9707 of *PC-StainGAN-loose.*

E. Experimental Results on Breast Dataset

To verify the generalization of the proposed method, we also evaluate our method on breast dataset. The virtual Ki-67-stained images generated by different methods are shown in Fig.9. It is clearly observed that the Ki-67-stained images generated by *PC-StainGAN* are more similar to the referenced ones and present consistent pathological information. However, the comparison methods lead to absolutely opposite results in pathological representation, which can be confirmed by comparing the column (a) with the columns (c-e) in Fig.9. In addition, Table IV presents some quantitative evaluation results between the generated Ki-67-stained images and the referenced ones. Observing the results of comparison methods

TABLE IV
QUANTITATIVE EVALUATION RESULTS OF VIRTUAL KI-67-STAINED IMAGE GENERATION ON BREAST DATASET

	Pearson-R (*positive*)	*Pearson-R* (*negative*)	*PHV I* (*T=0.005*) %	*PHV II* (*T=0.005*) %	*PHV III* (*T=0.005*) %
Cycle-GAN [25,30]	-0.2752	-0.1929	67.84±11.89	82.68±7.02	82.56±6.94
Cycle-GAN+SC+SSIM	0.4731	-0.1514	76.93±15.68	88.44±8.57	86.76±8.13
Lahianiet al [31]	-0.3197	-0.2196	69.19±14.07	84.67±8.02	81.43±9.47
PC-StainGAN	0.9578	0.8146	81.37±8.61	92.21±4.28	90.99±3.70

Note:'*PHV n*' means it is calculated based on the layer '*n*' in the Resnet-101.

TABLE V
QUANTITATIVE EVALUATION RESULTS OF DIFFERENT SETTINGS OF HYPERPARAMETER β

PC-StainGAN	$X \leftrightarrow \hat{X}$				$X \leftrightarrow \hat{Y}$	$\hat{Y} \leftrightarrow Y$	
	MAE	*SSIM*	*MS-SSIM*	*PSNR*	*CSS*	*Pearson-R* (*positive*)	*Pearson-R* (*negative*)
β=0	3.40±0.57	97.50±0.34	99.14±0.16	33.52±1.55	85.58±2.69	0.6381	0.2144
β=5	4.20±0.44	97.21±0.53	98.65±0.26	31.82±0.95	84.87±2.94	0.9755	0.8665
β=10	4.58±0.50	96.81±0.47	97.91±0.18	32.02±1.31	82.87±2.72	0.9785	0.8672
β=15	4.72±0.57	96.64±0.45	97.81±0.22	31.66±1.38	81.97±3.47	0.9801	0.9162

in the Table V, we find that their results have an extremely weak correlation with the referenced images, even a negative correlation. For example, the *Pearson-R* for Ki-67-stained positive area of *Cycle-GAN* gains -0.2752 and -0.1929 for negative area. As for *PC-StainGAN,* the generated Ki-67-stained images are strongly associated with the referenced images in both the positive and negative area (*e.g.,* 0.9578 for positive part and 0.8146 for negative part). In addition, all the quantitative results of our method are vastly superior to the comparison methods. Through the experiments on breast dataset, we further validate the efficiency and generalization of our proposed method.

F. The Selection of Hyper-Parameter β Setting

In our work, the proposed pathology consistency constraint plays an important role in stain transfer. Hence, we conduct an additional experiment to study the weight setting of pathological consistency loss. We set the choice space of β to be [0,5,10,15]. The experimental results are presented in Table V. From Table V, it is obvious that our method always achieves satisfied results when $\beta > 0$, which verifies the high-efficacy of pathology consistency constraint again. Meanwhile, it seems that with the increasing of β, the ability of pathological representation consistency gets some slight improvement but the ability of structural information preservation has declined slightly. Hence, based on final clinical requirements, we can make a trade-off between the performances of structural information preservation and pathological representation consistency. Here, we set β to be 5, and according to current experimental results, current hyper-parameter settings have achieved the satisfactory results.

V. DISCUSSION AND FUTURE WORK

In this section, we want to discuss some issues: first, the choice of unpaired stain transfer; then, the unreliability of *Cycle-GAN* in clinical stain transfer; and finally, the clinical value of *PC-StainGAN*. Meanwhile, we also want to state the imperfection of our current work and provide some insightful thinking in future work.

A. The Choice of Unpaired Stain Transfer

Histopathology plays an essential role in cancer diagnosis and regards as the gold standard in many medical protocols. In clinical practice, H&E staining is most commonly used staining techniques, which can distinguish various tissue and lesions in different colors. However, H&E staining is always hard to provide enough contrast to differentiate normal cells and diseased cells. Ki-67 staining is an excellent marker to study the clinical course of cancer, which is used in IHC examination. However, the Ki-67-IHC examination is very time-consuming and expensive, which is current bottleneck for the process of global pathology diagnostic services for cancer. In many low- and middle-income countries, cancer patients do not have the chance to access to IHC examination service, and the supply of pathologists severely falls short of demand. Hence, stain transfer will be highly welcome to ease the workload and address the issue of medical manpower shortage on underserved areas.

Many studies have stated image-to-image translation is able to covert between different stains. Furthermore, we want to discuss that successful stain transfer techniques should have the following properties: (1) they should ensure the preservation of fine details and microscopic structural information; (2) they should not change pathological properties of the tissue or cell.

In addition, unlike paired stain transfer or segmentation tasks where a large number of pixel-wise annotations are provided to design fully supervised learning algorithms, so accurately unpaired stain transfer will be more challenging. Meanwhile, serial tissue sections are cut by pathologists from the same tissue block, but they usually have significant structural differences, and deformation is inevitable during making slices in clinical practice. Hence, massive unpaired medical data is undeveloped and easily available, and unpaired stain transfer will be more competitive in many application scenarios.

B. The Unreliability of Cycle-GAN

Based on the intuitive results, we notice that the virtual stained images generated by *Cycle-GAN* look highly like the real ones, and the underlying image content remains unchanged. But there is still a big difference between the generated image and reference one, which is caused by the inherent weakness of *Cycle-GAN*. In section III.B, a clear argument about the weakness of *Cycle-GAN* for stain transfer is presented. In fact, many current related works do not realize this weakness and mainly focus on the structural consistency. As for successful stain transfer, the constraints of most *Cycle-GAN*-based work are not strong enough to ensure the consistency of pathological representation between the generated image and source image, where the adversarial loss just enforces the generated images look like a real image in target domain. According to quantitative results, we find that *Cycle-GAN* is volatile and its results can be easily misguided by the distribution of training dataset. Hence, both

qualitative analysis and quantitative evaluation demonstrate that the low clinical reliability of *Cycle-GAN* in virtual stain generation.

C. The Clinical Reliability of PC-StainGAN

In order to offset the weakness of *Cycle-GAN*, the pathology consistency constraint is introduced in *PC-StainGAN*. According to the unbalanced dataset experiment, compared with *Cycle-GAN*, our proposed *PC-StainGAN* always achieves a stable and correct performance on both unbalanced and balanced training datasets, which fully verifies the robustness of our method. Meanwhile, according to the study on degraded expert knowledge, when only the image-level annotation is available, our method still can achieve a high performance, which further verifies the superior of our model. Last but not least, an experiment is conducted to study the importance of pathology consistency constraint, where our evaluation scores are much higher than those of comparison methods. It can be confirmed that our proposed *PC-StainGAN* shows strong correlation between the generated images and referenced images, and our results have highly similar pathological representation with the referenced Ki-67-stained image. All of these experimental results demonstrate that our approach significantly outperforms state-of-the-art techniques and closely matches the performance of real Ki-67 examination results. As you can see, our massive experiments are enough to prove that our model has a high-level performance in clinical reliability.

D. Imperfection and Future Work

Although our method has achieved satisfactory results in the experiments, there are still some imperfections and a big space for further improvement. In our work, instead of adjusting the architecture of our model according to the dataset used in this study, we used high-level intuition based on the expected properties of the model; Due to the limitation of computation ability, the value of hyperparameters is determined through simple manual tuning rather than some hyper-parameter optimization algorithms. Hence, if the framework is tailored into different architectures according to specific tasks and datasets, or the hyperparameters are searched by hyper-parameter optimization algorithm, the performance of our method will be further improved. As for our future work, we will strive for the early application of the proposed method to clinical practice. First, we will further explore the potential of the proposed method with other efficient architectures, for instance, EfficientNet [37], Non-local-Block [38]; optimize the hyper-parameters of the objective and teste the applicability of our method in variant clinical datasets. Second, we will extend our method into multi-domain transfer, which is still a challenge and very important in practice. Third, data privacy remains a major barrier to access, and federated learning offers a way to counteract this data dilemma and its associated governance and privacy concerns by enabling collaborative learning without centralizing the data. Hence, we will change current local training manner and train our method with federated learning mode, which can distill and share the knowledge among AI agents in a robust and privacy-preserved fashion.

VI. Conclusion

In conclusion, we proposed a novel adversarial learning method based on unpaired data for effective stain transfer between H&E domain and Ki-67 domain. Most of the aforementioned issues and demands have been well tackled in this paper. First, two pathological representation networks are first proposed to learn the pathological features from Ki-67-stained images and H&E-stained images. Also, the pathological consistency loss function is designed to constrain stained images with the same pathological properties in both H&E and Ki-67 straining domain. Meanwhile, just about 10% training data is annotated by pathologists as expert knowledge to ensure the features learned by pathological representation networks are correct, which significantly saves a lot of labor cost on annotations. Finally, we employ skip connection and structural consistency loss to further improve the preservation of structural details. In order to validate the efficacy of the proposed method, we first evaluated our method on a neuroendocrine cancer dataset for comprehensive analysis, and then we tested our method on a breast cancer dataset. Extensive experiments validated the superiority of the proposed method which significantly outperformed the state-of-the-art methods on two datasets.

Acknowledgment

The authors would like to thank Xi Li (Gastroenterology Department, Peking University Shenzhen Hospital, China) and Aiping Zheng (Pathology Department, Peking University Shenzhen Hospital, China). Both datasets are collected from Peking University Shenzhen Hospital.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA, A Cancer J. Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018.

[2] C. T. Soares, U. Frederigue-Junior, and L. A. de Luca, "Anatomopathological analysis of sentinel and nonsentinel lymph nodes in breast cancer: hematoxylin-eosin versus immunohistochemistry," *Int. J. Surgical Pathol.*, vol. 15, no. 4, pp. 358–368, Oct. 2007.

[3] R. A. Sheikh, B. H. Min, and S. Yasmeen, "Correlation of Ki-67, p53, and Adnab-9 immunohistochemical staining and ploidy with clinical and histopathologic features of severely dysplastic colorectal adenomas," *Digestive Diseases Sci.*, vol. 48, no. 1, pp. 223–229, 2003.

[4] F. Anglade, D. A. Milner, and J. E. Brock, "Can pathology diagnostic services for cancer be stratified and serve global health?" *Cancer*, vol. 126, no. S10, pp. 2431–2438, May 2020.

[5] World Health Organization. (2019). *Guide for Establishing a Pathology Laboratory in the Context of Cancer Control*. [Online]. Available: https://apps.who.int/iris/rest/bitstreams/1266241/retrieve

[6] Y. Liu *et al.*, "Predict ki-67 positive cells in H&E-stained images using deep learning independently from IHC-stained images," *Frontiers Mol. Biosciences*, vol. 7, p. 183, Aug. 2020.

[7] Y. Wang, L. L. Sun, and Q. Jin, "Enhanced diagnosis of pneumothorax with an improved real-time augmentation for imbalanced chest X-rays data based on DCNN," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Apr. 19, 2019, doi: 10.1109/TCBB.2019.2911947.

[8] Z. Tang, K. Chen, M. Pan, M. Wang, and Z. Song, "An augmentation strategy for medical image processing based on statistical shape model and 3D thin plate spline for deep learning," *IEEE Access*, vol. 7, pp. 133111–133121, 2019.

[9] J. Yang, D. Park, G. T. Gullberg, and Y. Seo, "Joint correction of attenuation and scatter in image space using deep convolutional neural networks for dedicated brain 18F-FDG PET," *Phys. Med. Biol.*, vol. 64, no. 7, Apr. 2019, Art. no. 075019.

[10] B. Zhang *et al.*, "Cerebrovascular segmentation from TOF-MRA using model- and data-driven method via sparse labels," *Neurocomputing*, vol. 380, pp. 162–179, Mar. 2020.

[11] X. Wang *et al.*, "Weakly supervised deep learning for whole slide lung cancer image analysis," *IEEE Trans. Cybern.*, vol. 50, no. 9, pp. 3950–3962, Sep. 2020.

[12] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, Jun. 2017.

[13] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[14] S. C. M. Arjovsky and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, 2017, pp. 214–223.

[15] X. Huang, Y. Li, and O. Poursaeed, "Stacked generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5077–5086.

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[17] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[18] C. Baur, S. Albarqouni, and N. Navab, "MelanoGANs: High resolution skin lesion synthesis with GANs," 2018, *arXiv:1804.04338*. [Online]. Available: http://arxiv.org/abs/1804.04338

[19] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*. [Online]. Available: http://arxiv.org/abs/1710.10196

[20] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[21] F. Mahmood *et al.*, "Deep adversarial training for multi-organ nuclei segmentation in histopathology images," *IEEE Trans. Med. Imag.*, vol. 39, no. 11, pp. 3257–3267, Nov. 2020.

[22] N. Bayramoglu, M. Kaakinen, L. Eklund, and J. Heikkila, "Towards virtual H&E staining of hyperspectral lung histology images using conditional generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 64–71.

[23] H. Cho, S. Lim, G. Choi, and H. Min, "Neural stain-style transfer learning using GAN for histopathological images," 2017, *arXiv:1710.08543*. [Online]. Available: http://arxiv.org/abs/1710.08543

[24] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[25] J. Jiang, Y. C. Hu, and N. Tyagi, "Tumor-aware, adversarial domain adaptation from CT to MRI for lung cancer segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2018, pp. 777–785.

[26] J. M. Wolterink, A. M. Dinkla, and M. H. Savenije, "Deep MR to CT synthesis using unpaired data," in *Proc. Int. Workshop Simulation Synth. Med. Imag.*, 2017, pp. 14–23.

[27] Z. Li, S. Zhou, J. Huang, L. Yu, and M. Jin, "Investigation of low-dose CT image denoising using unpaired deep learning methods," *IEEE Trans. Radiat. Plasma Med. Sci.*, vol. 5, no. 2, pp. 224–234, Mar. 2021.

[28] M. Gadermayr, L. Gupta, V. Appel, P. Boor, B. M. Klinkhammer, and D. Merhof, "Generative adversarial networks for facilitating stain-independent supervised and unsupervised segmentation: A study on kidney histology," *IEEE Trans. Med. Imag.*, vol. 38, no. 10, pp. 2293–2302, Oct. 2019.

[29] M. T. Shaban, C. Baur, N. Navab, and S. Albarqouni, "Staingan: Stain style transfer for digital histological images," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 953–956.

[30] A. Lahiani, N. Navab, and S. Albarqouni, "Perceptual embedding consistency for seamless reconstruction of tilewise style transfer," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2019, pp. 568–576.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[32] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[33] L. Liu, X. Liu, J. Gao, W. Chen, and J. Han, "Understanding the difficulty of training transformers," 2020, *arXiv:2004.08249*. [Online]. Available: http://arxiv.org/abs/2004.08249

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[35] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Med. Image Anal.*, vol. 58, Dec. 2019, Art. no. 101552.

[36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.

[37] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*. [Online]. Available: http://arxiv.org/abs/1905.11946

[38] Y. Li *et al.*, "Neural architecture search for lightweight non-local networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10297–10306.