Exploiting Supervision Information in Weakly Paired Images for IHC Virtual Staining

Yueheng Li, Xianchao Guan, Yifeng Wang, and Yongbing Zhang(✉)

Harbin Institute of Technology (Shenzhen), Shenzhen, China
{22s051028,guanxianchao,wangyifeng}@stu.hit.edu.cn, ybzhang08@hit.edu.cn

Abstract. Immunohistochemical (IHC) staining plays a pivotal role in the evaluation of numerous diseases. However, the standard IHC staining process involves a series of time-consuming and labor-intensive steps, which severely hinders its application in histopathology. With the rapid advancement of deep learning techniques, virtual staining has promising potential to address this issue. But it has long been challenging to determine how to effectively provide supervision information for networks by utilizing consecutive tissue slices. To this end, we propose a weakly supervised pathological consistency constraint acting on multiple layers of GAN. Due to variations of receptive fields in different layers of the network, weakly paired consecutive slices have different degrees of alignment. Thus we allocate adaptive weights to different layers in order to dynamically adjust the supervision strengths of the pathological consistency constraint. Additionally, as an effective deep generative model, GAN can generate high-fidelity images, but it suffers from the issue of discriminator failure. To tackle this issue, a discriminator contrastive regularization method is proposed. It compels the discriminator to contrast the differences between generated images and real images from consecutive layers, thereby enhancing its capability to distinguish virtual images. The experimental results demonstrate that our method generates IHC images from H&E images robustly and identifies cancer regions accurately. Compared to previous methods, our method achieves superior results.

Keywords: Weakly supervised learning · Generative adversarial network · IHC virtual staining.

1 Introduction

Histological staining, a crucial step for tissue examination, is commonly used to enhance the visibility of different biological structures by modifying or intensifying their colors [1]. Among the histochemical stains, hematoxylin and eosin (H&E) can yield contrasting colors between nuclei and cytoplasm [2] and is the most widely used in the field of histopathology across the globe [3]. But it is difficult to observe all the information required for disease diagnosis only through H&E staining. In contrast, immunohistochemistry (IHC) staining, which is a

more advanced staining approach, can demonstrate antigens via specific antibodies in tissue slices [4]. It plays a significant role in distinguishing cancer subtypes. However, the standard IHC staining procedures comprise several time-consuming steps and require involvement of well-trained pathologists, leading to delay in disease treatment [5]. Thus there is a compelling need in clinical practice for an economical, rapid, and precise alternative to traditional IHC staining methods.

With the progression of deep learning technology, researchers have proposed IHC virtual staining to address these issues. Hong et al. achieved virtual staining from H&E to cytokeratin (CK) based on Pix2Pix [6]. But this supervised approach necessitates pixel-level paired input and ground truth images, which are hard to obtain clinically. In fact, when pathologists cut serial slices from the same tissue, it usually leads to distortion of cell and tissue structures inevitably. Mercan et al. used CycleGAN to accomplish virtual staining between H&E and Phosphohistone-H_3 (PHH_3) without the need for slice registration [7]. Nevertheless, unsupervised methods disregard critical pathological information of consecutive tissue slices, which often causes inaccuracies in staining results. Zeng et al. achieved PR virtual staining by registering consecutive slides and acquiring positive/negative labels for H&E images [8]. However, this method only uses label-level supervision instead of fully exploiting information in consecutive slices. In general, due to deformation between consecutive slices, current methods have not made full use of the pathological information.

In addition, as a classic generative model, GAN has been widely utilized in image translation tasks[9]. The efficacy of GAN heavily relies on the performance of the discriminator[10]. During the training process, the ability of the discriminator to distinguish between real and fake images frequently undergoes instability, resulting in ineffective feedback to the generator and causing unsatisfactory generated outcomes[11]. Therefore, it is a significant challenge to enhance the effectiveness of the discriminator.

To address the aforementioned challenges, we propose the weakly supervised IHC virtual staining method. The main contributions are listed as follows: (1) We design the multi-layer weakly pathological consistency constraint to thoroughly exploit the information embedded in consecutive tissue slices, which preserves the pathological relationship between real and generated IHC images. (2) To address the challenge of varying degrees of alignment between consecutive slices in different regions, we propose adaptive weights for the pathological consistency constraint at different layers of the network. (3) In order to provide the discriminator with stronger supervision information, we design the discriminator contrastive regularization loss. It can guide the discriminator to compare real and fake images by using consecutive slices and enhance the discriminative capability of the discriminator.

2 Method

Fig. 1 illustrates the overall framework of our weakly supervised framework for IHC virtual staining. The weakly pathological consistency constraint with

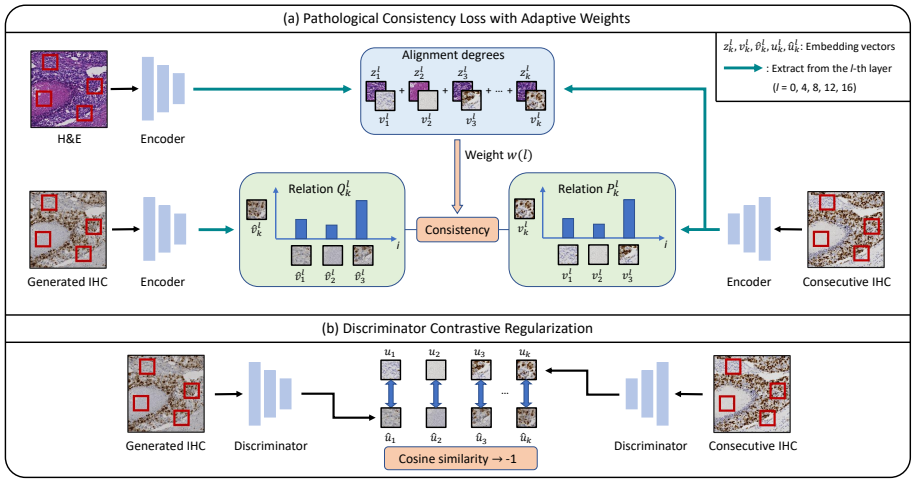

Fig. 1. The overall framework of our weakly supervised framework for IHC virtual staining.

adaptive weights is shown in Fig. 1(a). We obtain generated IHC images from H&E images through the generator. Then we acquire embedding vectors of H&E, generated IHC and consecutive IHC images through the same generator encoder and calculate alignment degrees of consecutive slices and pathological relations of IHC images. Fig. 1(b) presents our discriminator regularization loss. We obtain vectors of generated and consecutive IHC images through the discriminator in a similar way and constrain the cosine similarity of features at corresponding positions.

2.1 Weakly Pathological Consistency Constraint

It is paramount to ensure the accuracy of pathological information in IHC virtual staining, such as the positive/negative expressions. Though consecutive IHC slices cannot provide pixel-level supervision, it can provide pathological constraints for the generative process. To this end, we propose the multi-layer weakly pathological consistency constraint (WPCC) loss between consecutive slices. We adopt the method in CUT [12] to obtain embedding vectors. For a given patch y_k of a real IHC image Y, the pathological relation with another patch y_i is defined as:

$$P_k^l(i) = \frac{\exp\left(v_k^l \cdot v_i^l\right)}{\sum_{j=1}^{N} \exp\left(v_k^l \cdot v_j^l\right)}, \tag{1}$$

where v_k^l and v_i^l are the corresponding embedding vectors from the l-th layer of the generator encoder, the subscripts are the locations of patches and N is the

number of patch samples. Similarly, the pathological relation between the patch $\hat{y}_k$ and $\hat{y}_i$ of the generated image $\hat{Y}$ is given by:

$$Q_k^l(i) = \frac{\exp\left(\hat{v}_k^l \cdot \hat{v}_i^l\right)}{\sum_{j=1}^{N} \exp\left(\hat{v}_k^l \cdot \hat{v}_j^l\right)}. \tag{2}$$

Subsequently, we enforce the pathological consistency of the l-th layer between all the corresponding patches of the real and fake IHC images by using Jensen-Shannon Divergence (JSD):

$$\mathcal{L}_{\text{WPCC}}(l) = \sum_{k=1}^{N} JSD(P_k^l || Q_k^l). \tag{3}$$

2.2 Adaptive Weights for WPCC

Our WPCC loss is computed at multiple layers, which allows the model to capture pathological information at various levels of the image. However, due to differences in alignment, consecutive IHC slices cannot always provide precise pathological constraints. Consecutive slices usually appear similar overall, yet they have notable differences at finer scales. Embedding vectors from deeper layers correspond to larger receptive fields in input images, so the weights for the WPCC loss at different layers should not be simply equal. For the regions with better alignment, larger weights are supposed to be assigned, and vice versa. Therefore, we introduce adaptive weights for the WPCC loss. We use the Pearson correlation coefficient to measure the degrees of alignment, and the weight for the l-th layer is:

$$w_l = \sum_{k=1}^{N} \left(1 + \frac{cov(z_k^l, v_k^l)}{\sigma_{z_k^l} \sigma_{v_k^l}}\right), \tag{4}$$

where z_k^l denotes the vector of the image patch x_k in the consecutive H&E slice, cov is the covariance and $\sigma_{z_k^l}$ and $\sigma_{v_k^l}$ are the standard deviations. Thus the adaptive pathological consistency constrain between consecutive slices is defined as:

$$\mathcal{L}_{\text{APCC}} = \sum_{l=1}^{L} \frac{w_l}{W} \cdot \mathcal{L}_{\text{WPCC}}(l), \tag{5}$$

where W is a normalization coefficient. The adaptive weights can mitigate the influence of misaligned patches in consecutive layers. At the early stage of training, the generator cannot extract meaningful representations. Therefore, we employ adaptive weights linearly starting from the 41st epoch during training.

2.3 Discriminator Contrastive Regularization

A discriminator capable of extracting more powerful representations contributes to improving the performance of the generator. For a real IHC image and the

fake IHC image which is generated from the consecutive H&E slice, though they appear visually similar, the discriminator is expected to distinguish between them. Thus the discriminator needs to exploit their contrasting feature information. To enhance the discriminative ability of the discriminator and improve generation quality, we propose the discriminator regularization loss. Following the approach of obtaining embedding vectors in the generator, we denote all but the last layer of the discriminator as D_0 and a small two-layer MLP as H. Thus we obtain vectors $u_k = H(D_0(y_k))$ and $\hat{u}_k = H(D_0(\hat{y}_k))$, where y_k and $\hat{y}_k$ represent patches from the real and fake IHC images at the corresponding location respectively. Our discriminator regularization loss is determined as:

$$\mathcal{L}_{\text{Reg}} = \frac{1}{N} \cdot \sum_{k=1}^{N} \left\| \frac{u_k \cdot \hat{u}_k}{\|u_k\|_2 \, \|\hat{u}_k\|_2} - (-1) \right\|_2 . \tag{6}$$

In addition, we adopt the PatchNCE loss in CUT, which encourages the generator to maximize the mutual information between corresponding patches of the input and output images. Finally, our total loss formulations are as follows:

$$\mathcal{L}_{\text{G}} = \mathcal{L}_{\text{Adv}} + \lambda_{\text{APCC}} \mathcal{L}_{\text{APCC}} + \lambda_{\text{PatchNCE}} \mathcal{L}_{\text{PatchNCE}}, \tag{7}$$

$$\mathcal{L}_{\text{D}} = \mathcal{L}_{\text{Adv}} + \lambda_{\text{Reg}} \mathcal{L}_{\text{Reg}}, \tag{8}$$

where $\mathcal{L}_{\text{Adv}}$ represents the conventional adversarial loss of GAN.

3 Experiments

3.1 Experimental Setup

Datasets. In our experiments, we used the following datasets: the Breast Cancer Immunohistochemical (BCI) challenge dataset [13] and MIST dataset [14]. The BCI dataset covers various levels of HER2 expression. The MIST dataset contains IHC staining data for HER2, PR, ER, and Ki67. The datasets are divided into training and testing sets as illustrated in Table 1.

Training Details. Our model is implemented based on PyTorch and trained on an NVIDIA RTX 3090 GPU. We use the Adam optimizer ($\beta_1 = 0.5$, $\beta_2 = 0.999$) and set the batch size to 1 for training. We randomly crop images to the resolution of 512×512 during training. The number of patch samples N is set to 256. The hyperparameter settings are as follows: $\lambda_{\text{APCC}} = 0.4$, $\lambda_{\text{PatchNCE}} = 1$, $\lambda_{\text{Reg}} = 0.2$.

Table 1. The number of image pairs in each dataset.

Dataset	HER2_{BCI}	$\text{HER2}_{\text{MIST}}$	ER_{MIST}	PR_{MIST}	$\text{Ki67}_{\text{MIST}}$
Training Set	3896	4642	4153	4139	4361
Testing Set	977	1000	1000	1000	1000

Table 2. Comparison of CycleGAN, CUT, Pyramid Pix2Pix, ASP and our method. We bold the highest value and underline the second highest value. KID scores have been multiplied by 1000.

Dataset	Method	PSNR ↑	SSIM ↑	FID ↓	KID ↓
$HER2_{BCI}$	CycleGAN	14.201	0.424	63.7	45.3
	CUT	17.322	0.438	65.0	15.1
	Pyramid Pix2Pix	**21.160**	0.477	80.1	46.4
	ASP	17.869	<u>0.492</u>	<u>54.3</u>	<u>10.7</u>
	Ours	<u>19.132</u>	**0.499**	**50.1**	**9.8**
$HER2_{MIST}$	CycleGAN	12.396	0.181	120.1	96.2
	CUT	14.091	0.182	89.2	48.0
	Pyramid Pix2Pix	12.931	<u>0.201</u>	59.9	56.2
	ASP	<u>14.192</u>	0.190	<u>51.1</u>	<u>12.4</u>
	Ours	**15.973**	**0.231**	**40.1**	**9.1**
ER_{MIST}	CycleGAN	11.903	0.181	88.7	61.3
	CUT	12.031	0.183	47.1	13.8
	Pyramid Pix2Pix	12.108	0.191	80.8	62.7
	ASP	**13.991**	<u>0.206</u>	<u>41.2</u>	<u>5.8</u>
	Ours	<u>13.901</u>	**0.209**	**34.9**	**3.9**
PR_{MIST}	CycleGAN	12.988	0.187	78.6	45.2
	CUT	13.564	0.192	53.2	15.8
	Pyramid Pix2Pix	<u>14.428</u>	<u>0.224</u>	79.2	59.8
	ASP	14.325	0.216	<u>44.5</u>	<u>10.2</u>
	Ours	**15.936**	**0.248**	**34.2**	**7.8**
$Ki67_{MIST}$	CycleGAN	12.917	0.201	100.8	87.6
	CUT	13.697	0.212	53.1	30.1
	Pyramid Pix2Pix	13.987	<u>0.248</u>	89.8	69.7
	ASP	<u>14.824</u>	0.241	<u>50.9</u>	<u>19.1</u>
	Ours	**16.093**	**0.262**	**31.1**	**15.2**

Evaluation Metrics. We use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) as evaluation metrics to measure the quality of generated images. PSNR is computed by the error between the corresponding pixels of two images and SSIM can measure the differences in brightness, contrast and structure. FID and KID estimate the distributions of real and fake images in deep network spaces.

3.2 Comparative Experiment Results and Analysis

We compare our method with CycleGAN [15], CUT [12], Pyramid Pix2Pix [13] and Adaptive Supervised PatchNCE (ASP) [14]. Table 2 presents the quantitative experimental results for all methods. Compared with other methods, our method achieves the best performance in the majority of metrics. For qualitative evaluations, as illustrated in Fig. 2, our method can identify cancer regions accurately and maintain the pathological consistency between the input and generated images. While other methods are hard to distinguish between negative and positive images. We speculate that the suboptimal performance of ASP may be attributed to the fact that similarities between the generated and real IHC

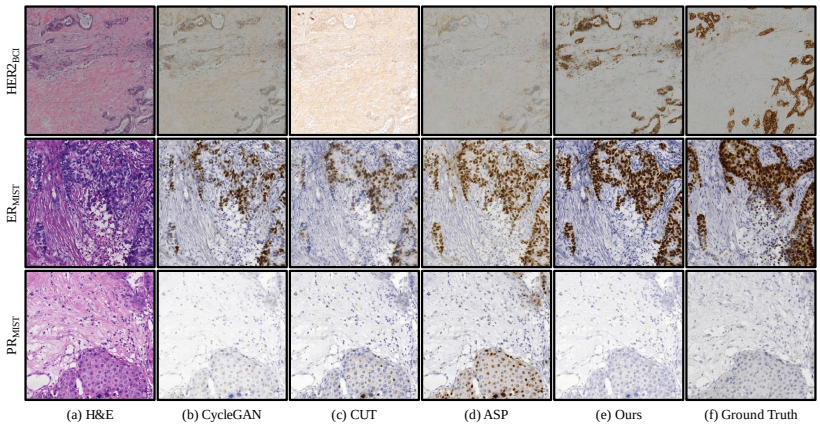

Fig. 2. Virtual staining image results of comparative experiments.

images cannot represent the degrees of alignment between consecutive slices accurately. This is because positive/negative expressions of the generated images impact their similarities largely. Therefore, our method utilizes H&E and real IHC images to calculate the degrees of alignment.

3.3 Ablation Experiment Results and Analysis

We compare the quantitative performance of models trained on different settings in Table 3, where $WPCC$ denotes weakly pathological consistency constraint, *Ada* denotes adaptive weights for WPCC, and *Reg* denotes the discriminator contrastive regularization loss. Owing to the page limit, we present the results of HER2 in BCI and ER in MIST here. When none of these three components are used, our method is simplified to CUT. By integrating all the components,

Table 3. Ablation experiment for several different settings. We bold the highest value and underline the second highest value. KID scores have been multiplied by 1000.

Dataset	Settings			PSNR ↑	SSIM ↑	FID ↓	KID ↓
	WPCC	Ada	Reg				
$HER2_{BCI}$	×	×	×	17.322	0.438	65.0	15.1
	✓	×	×	18.132	0.452	58.3	12.5
	✓	×	✓	**19.298**	0.463	<u>55.2</u>	<u>10.7</u>
	✓	✓	×	18.812	<u>0.481</u>	56.5	12.1
	✓	✓	✓	<u>19.132</u>	**0.499**	**50.1**	**9.8**
ER_{MIST}	×	×	×	12.031	0.183	47.1	13.8
	✓	×	×	12.912	0.188	42.7	9.2
	✓	×	✓	13.762	0.189	<u>39.7</u>	<u>5.2</u>
	✓	✓	×	<u>13.838</u>	<u>0.192</u>	40.8	7.8
	✓	✓	✓	**13.901**	**0.209**	**34.9**	**3.9**

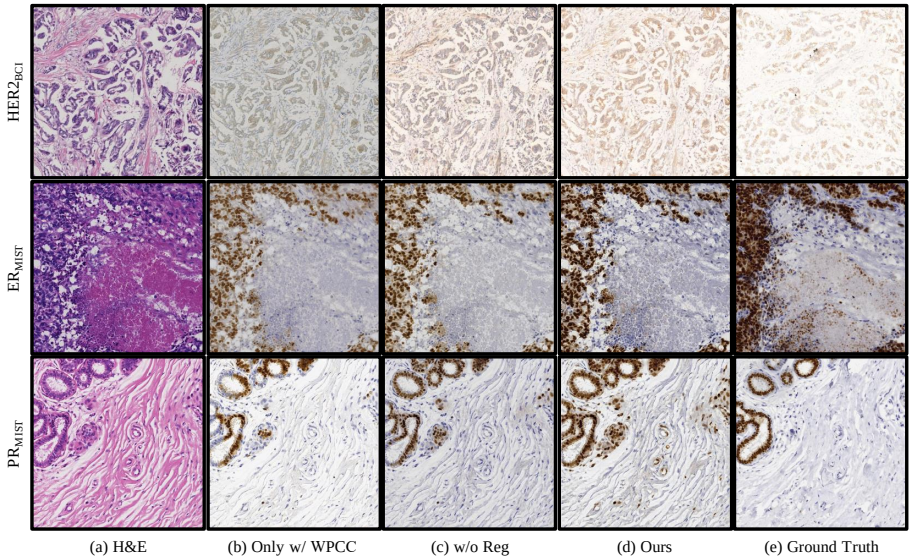

Fig. 3. Virtual staining image results of ablation experiments.

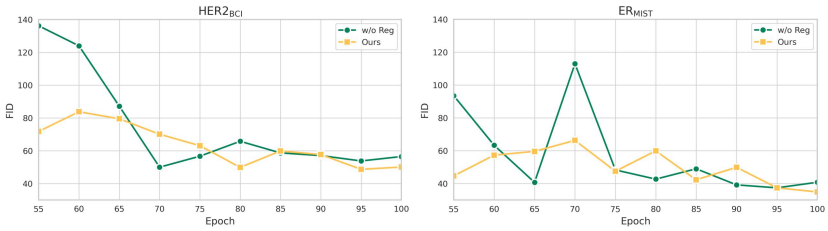

Fig. 4. The fluctuation of FID scores for generated images during training.

the results exhibit the optimal performance in most of the metrics. In Fig. 3, we present the visual results under different settings. By adding these three components, our results are the most similar to the ground truth images. Fig. 4 illustrates the fluctuation of FID scores for generated images during training. It shows that our discriminator contrastive regularization loss can stabilize the training process.

4 Conclusion

In this paper, we propose a weakly supervised IHC virtual staining method which utilizes pathological information from consecutive tissue slices. Considering the inconsistency between weakly paired consecutive images, we introduce the weakly pathological consistency constrain with adaptive weights at multiple layers of the network. And we propose the contrastive regularization loss in the

discriminator to alleviate the issues of training instability in GAN. Our method does not require finely aligned datasets and generates IHC images precisely. We provide qualitative and quantitative experimental results on two datasets and they demonstrate that our method outperforms existing methods.

Acknowledgments. 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).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ye Zhang, Ziyue Wang, Yifeng Wang, Hao Bian, Linghan Cai, Hengrui Li, Lingbo Zhang, and Yongbing Zhang. Boundary-aware contrastive learning for semi-supervised nuclei instance segmentation. *arXiv preprint arXiv:2402.04756*, 2024.
2. Ada T Feldman and Delia Wolfe. Tissue processing and hematoxylin and eosin staining. *Histopathology: methods and protocols*, pages 31–43, 2014.
3. Xianchao Guan, Yifeng Wang, Yiyang Lin, Xi Li, and Yongbing Zhang. Unsupervised multi-domain progressive stain transfer guided by style encoding dictionary. *IEEE Transactions on Image Processing*, 2024.
4. JA Ramos-Vara and MA Miller. When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry—the red, brown, and blue technique. *Veterinary pathology*, 51(1):42–87, 2014.
5. 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.
6. Yiyu Hong, You Jeong Heo, Binnari Kim, Donghwan Lee, Soomin Ahn, Sang Yun Ha, Insuk Sohn, and Kyoung-Mee Kim. Deep learning-based virtual cytokeratin staining of gastric carcinomas to measure tumor–stroma ratio. *Scientific Reports*, 11(1):19255, 2021.
7. Caner Mercan, GCAM Mooij, David Tellez, Johannes Lotz, Nick Weiss, Marcel van Gerven, and Francesco Ciompi. Virtual staining for mitosis detection in breast histopathology. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1770–1774. IEEE, 2020.
8. 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.
9. Chanyong Jung, Gihyun Kwon, and Jong Chul Ye. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18260–18269, 2022.
10. Minsu Ko, Eunju Cha, Sungjoo Suh, Huijin Lee, Jae-Joon Han, Jinwoo Shin, and Bohyung Han. Self-supervised dense consistency regularization for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18301–18310, 2022.

11. Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020.
12. Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.
13. 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.
14. 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.
15. 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.