OT-StainNet: Optimal Transport Driven Semantic Matching for Weakly Paired H&E-to-IHC Stain Transfer

Xianchao Guan[1, 2*], **Yifeng Wang**[1*], **Ye Zhang**[1], **Zheng Zhang**[1, 2†], **Yongbing Zhang**[1†]

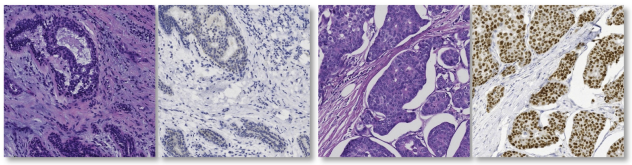
[1]Harbin Institute of Technology, Shenzhen, China
[2]Peng Cheng Laboratory
guanxianchao@stu.hit.edu.cn, darrenzz219@gmail.com, ybzhang08@hit.edu.cn

Abstract

Immunohistochemistry (IHC) examination is essential for characterizing tumor subtypes, providing prognostic information, and developing personalized treatment plans. However, IHC staining preparation is more complex and expensive compared to Hematoxylin and Eosin (H&E) staining, limiting its widespread clinical application. Transforming H&E images into IHC images presents a promising solution. In this paper, we propose OT-StainNet, a novel virtual IHC staining method. OT-StainNet employs a pre-trained diffusion model with richer prior knowledge as the generator and fine-tunes it with LoRA adapters through adversarial training. Given that adjacent images of the same tissue stained with H&E and IHC are not precisely aligned at the pixel level, existing methods struggle to fully utilize the supervisory information from weakly paired IHC images. To address this issue, we propose an optimal transport-driven semantic matching (OTSM) mechanism, establishing accurate semantic correspondences between H&E-IHC image pairs. By leveraging the real IHC features obtained through the OTSM mechanism, we design a semantic consistency constraint (SCC) to ensure that the correlations among virtual IHC features remain consistent with those among real IHC features, thereby preserving valuable correlation information during stain transfer. We validate OT-StainNet using four types of IHC staining across two datasets. Extensive experiments demonstrate the effectiveness of our method compared to state-of-the-art approaches.

Introduction

Histochemical staining is essential in pathological analysis. Hematoxylin and Eosin (H&E) is the most common and cost-effective staining agent, clearly delineating tissue cell structures to assist pathologists in making preliminary diagnoses (Mahbod et al. 2024). However, despite its widespread use, H&E staining often lacks the contrast necessary to accurately distinguish between various cancer subtypes, limiting its diagnostic utility in complex cases. In contrast, Immunohistochemistry (IHC) staining provides detailed insights into protein expression at the cellular level, both qualitatively and quantitatively, making it indispensable for the accurate diagnosis and staging of cancers, as well as for determining tumor prognosis (Ghahremani et al. 2022). For instance, in breast tissue pathology, pathologists assess biomarkers such as Ki67, ER, PR, and HER2 expression levels using IHC, facilitating precise diagnoses and personalized treatment.



(a) (b)

Figure 1: Examples of H&E and IHC-stained image. (a) shows the negative H&E-IHC image pairs. (b) shows the positive H&E-IHC image pairs which contains regions where specific biomarkers are expressed, providing crucial diagnostic and prognostic information.

However, compared to H&E staining, the histochemical staining process for IHC is more time-consuming and labor-intensive, typically requiring specialized histotechnologists and laboratory equipment (Gatenbee et al. 2023). These factors impede the widespread adoption of IHC staining in pathology. With advancements in digital pathology, virtual staining technologies based on generative models have emerged as a promising, cost-effective alternative. These technologies accelerate diagnosis, conserve tissue samples, and reduce patient waiting times, thereby holding significant clinical implications (Bai et al. 2023).

The goal of virtual staining is to simulate the style of the target staining while preserving the tissue structure of the source staining image (Guan et al. 2024). As a prominent branch of generative models, generative adversarial networks (GANs) have been widely used in virtual staining tasks due to their excellent image style transfer performance (Goodfellow et al. 2020). Existing GAN-based virtual staining methods can be roughly divided into two categories: unsupervised methods and fully supervised methods. Unsupervised methods primarily rely on CycleGAN (Zhu et al. 2017) or CUT (Park et al. 2020), using adversarial loss to learn the style of the target staining image and cycle consistency loss or contrastive learning loss to preserve the tissue structure details of the source staining image. Due to

*These authors contributed equally.
†Corresponding authors: Zheng Zhang and Yongbing Zhang.

the lack of effective supervisory information, these methods sometimes cannot guarantee the accuracy of the pathological characteristics of the virtual image (Zeng et al. 2022). Fully supervised methods predominantly rely on Pix2Pix (Isola et al. 2017), utilizing perfectly aligned input and ground truth to optimize the pixel-level loss between synthetic and real images. However, in histopathology, the irreversibility of the tissue staining process makes it physically infeasible to re-stain sections, posing a significant challenge in obtaining pixel-level aligned image pairs (Bai et al. 2023).

In clinical practice, pathologists cut consecutive sections from the same tissue, then stain and scan them separately (Rivenson et al. 2019). This process inevitably causes changes in tissue and cell structures between slide pairs, such as tissue loss, tearing, and artifacts, resulting in pixel-level inconsistencies (Li et al. 2023), as shown in Fig. 1. Despite these inconsistencies, significant pathological consistency information remains in adjacent layer tissue image pairs (Liu et al. 2021). However, existing fully supervised methods generate virtual images by calculating pixel-level loss, overlooking the adverse effects of structural inconsistencies on the model, which may introduce inaccurate pathological characteristics into virtual images.

In addition to the limitation of not fully utilizing supervised information, the aforementioned GAN-based virtual staining methods face another major challenge: the need to train from scratch, which typically requires a large amount of data to achieve satisfactory generalization performance (Zhang et al. 2024). Recently, generative models like stable diffusion have demonstrated excellent performance in natural image generation by leveraging rich semantic knowledge and feature representations obtained through pre-training on large-scale datasets (Rombach et al. 2022; Li et al. 2024a). When applied to the pathology domain, the prior knowledge embedded in these pre-trained models aids in capturing the cellular and structural characteristics of histopathological images (Ferber et al. 2024; Zhang et al. 2024). Therefore, employing pre-trained models for the virtual staining of pathological images can enhance the accuracy and generalization of the model.

In this paper, we propose OT-StainNet, a novel H&E-to-IHC stain transfer network. To fully utilize valuable supervisory information in the adjacent layer IHC images, we design an optimal transport-driven semantic matching (OTSM) mechanism and a semantic consistency constraint (SCC). The contributions of our work can be summarized as follows:

- We propose OT-StainNet for IHC virtual staining, which leverages the rich prior knowledge embedded in diffusion models pre-trained on large-scale natural images while ensuring efficient inference.
- We introduce an optimal transport-driven semantic matching (OTSM) mechanism that establishes accurate semantic correspondences between H&E-IHC image pairs of adjacent layers that are not precisely paired at the pixel level.
- We design a semantic consistency constraint (SCC) to ensure consistency between the correlations among virtual IHC features and those among real IHC features, thereby preserving valuable correlation information during stain transfer.
- We evaluate OT-StainNet using four types of IHC staining on three datasets. Extensive experimental results demonstrate that OT-StainNet achieves promising performance compared to state-of-the-art approaches.

Related Work

Image-to-Image Translation

Image-to-image translation aims to transform a source domain image into the style of a target domain image while preserving its original content. Pix2Pix optimizes the pixel-level loss between the output and ground truth using a conditional GAN (Isola et al. 2017). To reduce the reliance on paired data, CycleGAN employs cycle consistency loss to maintain key properties between the input and output images (Zhu et al. 2017). CUT employs contrastive learning to maximize the mutual information between corresponding locations in the input and output images, implementing a one-sided image translation method (Park et al. 2020). These methods typically require training from scratch and large datasets to achieve satisfactory generalization. Recently, pre-trained generative models, such as stable diffusion, have outperformed GANs in various image generation tasks due to the prior knowledge acquired from training on large-scale natural images (Rombach et al. 2022; Li et al. 2024b). Among these, img2img-turbo (Parmar et al. 2024), a novel diffusion model-based method, achieves high-accuracy image translation by leveraging the internal knowledge of SD-Turbo (Sauer et al. 2023) while ensuring efficient inference. However, applying these methods directly to IHC virtual staining without paired data does not guarantee the generation of accurate pathological characteristics.

IHC Virtual Staining in Histopathological Analysis

With the rapid development of image-to-image translation in the field of computer vision, some applications have been successfully applied to IHC virtual staining. Liu *et al.* proposed a pathology representation network that utilizes expert-annotated positive region labels to achieve virtual staining of neuroendocrine and breast tissues from H&E to Ki67 (Liu et al. 2021). However, the performance of this method heavily depends on the quality and quantity of expert annotations. Zeng *et al.* designed a Pos/Neg classifier to ensure the accuracy of virtual IHC Pos/Neg (Zeng et al. 2022). This approach, however, only relies on Pos/Neg information from adjacent layer IHC images, which may lead to inaccurate identification of specific positive regions. To enhance IHC virtual staining accuracy, some studies have employed real IHC images from adjacent layers to supervise virtual IHC image generation. Liu *et al.* proposed a pyramid Pix2Pix image generation method to transform H&E images into HER2 images (Liu et al. 2022). Li *et al.* introduced an adaptive supervised PatchNCE loss for converting H&E staining into multiple IHC stainings (Li et al. 2023). Dai *et al.* developed a weakly supervised method to convert autofluorescence images into H&E images (Dai, Wong, and

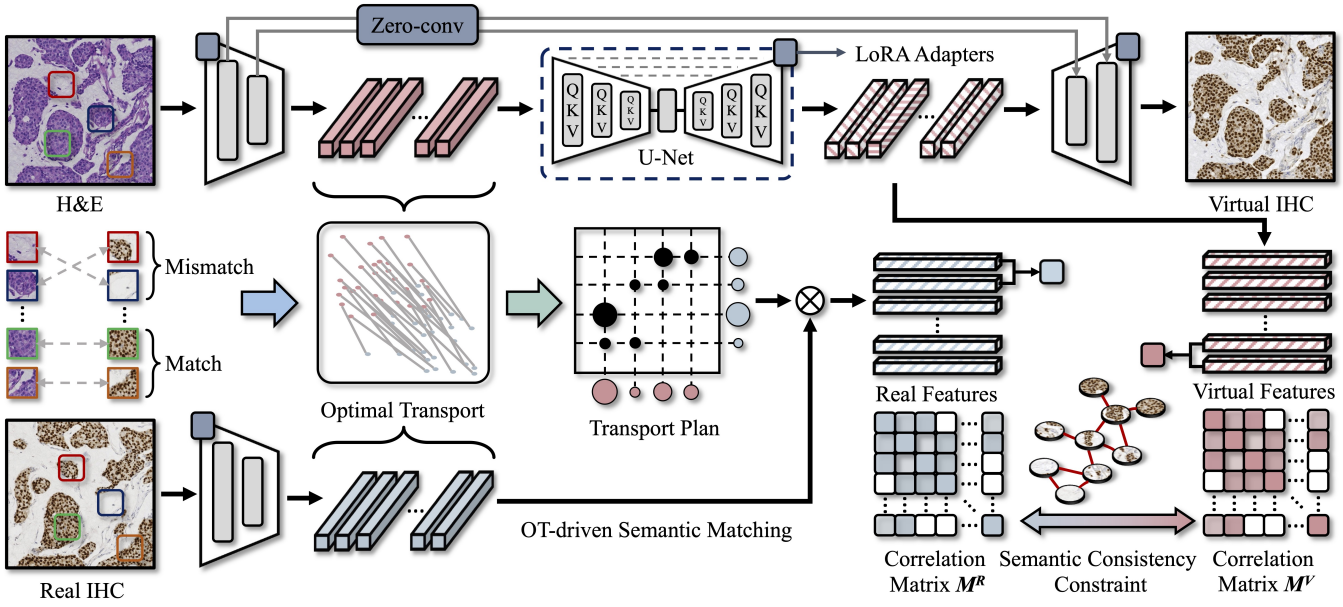


Figure 2: Overview of the OT-StainNet framework. Firstly, H&E and adjacent layer IHC images are encoded into feature vectors. Then, optimal transport is used to obtain real IHC features matching the H&E features. Finally, the semantic consistency constraint is applied to ensure the accuracy of pathological characteristics in the virtual IHC images generated by the one-step diffusion model. During training, the weights of the encoder, decoder, and U-Net are fixed, with only the LoRA adapters being trainable. Furthermore, skip connections and zero-convs between the encoder and decoder are incorporated to preserve the complex cell and tissue details of H&E images during stain transfer.

Wong 2023). Nonetheless, these methods generate virtual images by calculating pixel-level losses, overlooking local differences between adjacent layer H&E-IHC image pairs. Consequently, the virtual images generated by these methods may contain inaccurate pathological characteristics.

Methodology

Framework Overview

The OT-StainNet framework is illustrated in Fig. 2. Inspired by img2img-turbo (Parmar et al. 2024), we utilize the diffusion model SD-Turbo (Sauer et al. 2023) as the generator and fine-tune it with LoRA adapters using an adversarial learning objective. Furthermore, we streamline the latent diffusion model by integrating its three key components, the VAE encoder, VAE decoder, and UNet(Ronneberger, Fischer, and Brox 2015; Li et al. 2024c), into a unified, end-to-end network. This integration simplifies the architecture, reducing the need for separate module tuning and facilitating more efficient loss propagation. During training, we preserve the original weights of the VAE and UNet to retain the broad features and generalization capabilities acquired from pre-training on large-scale natural images. LoRA adapters allow us to fine-tune the pre-trained diffusion model by adding low-rank trainable parameters to specific layers while keeping the core model weights frozen. It helps OT-StainNet adapt to the unique characteristics of histopathological images without extensive retraining of the diffusion model. Additionally, to address the challenge of pixel-level misalignment between H&E and IHC images in adjacent layers, we propose an optimal transport-driven semantic matching (OTSM) mechanism to mine semantic correlations between H&E-IHC pairs, providing a foundation for supervision. Building on OTSM, we design a semantic consistency constraint (SCC) that extends beyond pixel-level supervision by enforcing consistency between the correspondences among virtual IHC features and those among real IHC features.

OT-driven Semantic Matching

In clinical practice, pathologists often stain consecutive tissue sections to obtain H&E and IHC slides. However, this process produces adjacent layer images that are inherently misaligned at the pixel level due to tissue distortions and variations. To avoid erroneous pathological characteristics in virtual IHC due to the supervisory information provided by mismatched real IHC images, we introduce an optimal transport-driven semantic matching (OTSM) mechanism to align semantic features between H&E-IHC image pairs.

As a solution to model the correspondences between two distributions, the goal of optimal transport is to retrieve a transport plan matrix $\mathbf{P}$ that minimizes the total cost matrix $\mathbf{C}$. Formally, the optimal transport from H&E features $\mathbf{F}^H = \left[\mathbf{f}_1^H, \mathbf{f}_2^H, ..., \mathbf{f}_n^H\right] \in \mathbb{R}^{n\times d}$ to IHC features $\mathbf{F}^I = \left[\mathbf{f}_1^I, \mathbf{f}_2^I, ..., \mathbf{f}_n^I\right] \in \mathbb{R}^{n\times d}$ can be defined by the discrete Kantorovich formulation(Kantorovich 2006) as follows:

$$\mathcal{W}(\mathbf{F}^H, \mathbf{F}^I) = \min_{\mathbf{P}\in\Pi(\mu_H,\mu_I)} \langle \mathbf{P}, \mathbf{C} \rangle. \quad (1)$$

Here, $\langle \mathbf{P}, \mathbf{C} \rangle$ denotes the Frobenius dot product of $\mathbf{P}$ and

$\mathbf{C}$. The matrix $\mathbf{C} \geq 0 \in \mathbb{R}^{n \times n}$ is the cost matrix defined by $\mathbf{C}_{u,v} = \mathcal{C}(\mathbf{f}_u^H, \mathbf{f}_v^I)$, where $\mathcal{C}(\cdot)$ is a ground distance metric, such as l_2-distance, that measures the distance of pairwise features in F^H and F^I. The set $\Pi(\mu_H, \mu_I) = \left\{\mathbf{P} \in \mathbb{R}_+^{n \times n} | \mathbf{P}\mathbf{1} = \mu_H, \mathbf{P}^\top \mathbf{1} = \mu_I\right\}$ represents the marginal constraints, ensuring total mass equality between the marginal distributions μ_H and μ_I for H&E and IHC features, respectively, where $\mathbf{1}$ is a vector of ones.

The marginal constraints of optimal transport ensure that the H&E features are consistent with the IHC features in terms of their global structure. Once the transport plan $\mathbf{P}$ is determined, capturing the matching correspondences between H&E and IHC features in the latent space, IHC features are aligned with H&E features through $\hat{\mathbf{F}}^I = \mathbf{P}^\top \mathbf{F}^I$. After alignment, the matched real IHC features are used to supervise the generation of virtual IHC images.

Semantic Consistency Constraint

There is rich pathological correlation information between IHC image features. To preserve this valuable correlation information during stain transfer and improve the accuracy of virtual IHC images, we design a semantic consistency constraint (SCC), thereby ensuring the consistency of pathological information between virtual and real IHC images.

Generally, prevalent fully supervised methods utilize l_1-distance to ensure the generator's output closely matches the ground truth, which only supervises the generation of virtual IHC images at the local feature level, neglecting the semantic correlation among features at the group level.

To model the semantic correlation among IHC features effectively, we calculate the relational function between any two features within features of each IHC image to obtain the correlation matrix $\hat{\mathbf{M}}^I$, representing the semantic correspondences among different features. The correlation matrix $\hat{\mathbf{M}}^I$ is defined as follows:

$$\hat{\mathbf{M}}^I = \begin{bmatrix} \mathcal{D}(\hat{\mathbf{f}}_1^I, \hat{\mathbf{f}}_1^I) & \mathcal{D}(\hat{\mathbf{f}}_1^I, \hat{\mathbf{f}}_2^I) & \cdots & \mathcal{D}(\hat{\mathbf{f}}_1^I, \hat{\mathbf{f}}_n^I) \\ \mathcal{D}(\hat{\mathbf{f}}_2^I, \hat{\mathbf{f}}_1^I) & \mathcal{D}(\hat{\mathbf{f}}_2^I, \hat{\mathbf{f}}_2^I) & \cdots & \mathcal{D}(\hat{\mathbf{f}}_2^I, \hat{\mathbf{f}}_n^I) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{D}(\hat{\mathbf{f}}_n^I, \hat{\mathbf{f}}_1^I) & \mathcal{D}(\hat{\mathbf{f}}_n^I, \hat{\mathbf{f}}_2^I) & \cdots & \mathcal{D}(\hat{\mathbf{f}}_n^I, \hat{\mathbf{f}}_n^I) \end{bmatrix}, \quad (2)$$

where $\mathcal{D}(\cdot, \cdot)$ denotes the relational function. Here, we utilize cosine similarity to calculate $\mathcal{D}(\cdot, \cdot)$.

Then, we employ the same strategy to obtain the correlation matrix $\tilde{\mathbf{M}}^I$ for the virtual IHC feature group generated by the one-step diffusion model. To provide more effective semantic-level supervision for the generation of virtual IHC images, we design a semantic consistency loss $\mathcal{L}_{sc}$, defined as follows:

$$\mathcal{L}_{sc} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| \hat{\mathbf{M}}_{ij}^I - \tilde{\mathbf{M}}_{ij}^I \right|. \quad (3)$$

Under the constraint of semantic consistency, virtual IHC features are compelled to retain the semantic correlation information present in real IHC features during generation.

Loss Functions

OT-StainNet is trained based on the CycleGAN framework, with loss terms including adversarial loss $\mathcal{L}_{\text{adv}}$, cycle consistency loss $\mathcal{L}_{\text{cycle}}$, identity loss $\mathcal{L}_{\text{id}}$, and semantic consistency loss $\mathcal{L}_{\text{sc}}$. The adversarial loss $\mathcal{L}_{\text{adv}}$ is defined as:

$$\begin{aligned} \mathcal{L}_{\text{adv}} = \mathbb{E}_y \log D_Y(y) + \mathbb{E}_x \log(1 - D_Y(G(x))) \\ + \mathbb{E}_x \log D_X(x) + \mathbb{E}_y \log(1 - D_X(G(y))), \end{aligned} \quad (4)$$

where $\mathbb{E}$ denotes the expectation, X and Y represent the source and target domains, respectively, G is the generator, and D is the discriminator, which uses the CLIP (Radford et al. 2021) model as its backbone.

The cycle consistency loss $\mathcal{L}_{\text{cycle}}$ ensures that the tissue structure details of the source staining image are preserved during stain transfer. It is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{cycle}} = \mathbb{E}_x \left[\mathcal{L}_{\text{rec}} \left(G(G(x)), x \right) \right] \\ + \mathbb{E}_y \left[\mathcal{L}_{\text{rec}} \left(G(G(y)), y \right) \right], \end{aligned} \quad (5)$$

where $\mathcal{L}_{\text{rec}}$ a combination of l_1-distance and LPIPS(Zhang et al. 2018).

The identity loss $\mathcal{L}_{\text{id}}$ ensures that the generator preserves the images of the target domain. It is defined as follows:

$$\begin{aligned} \mathcal{L}_{\text{id}} = \mathbb{E}_x \left[\mathcal{L}_{\text{rec}} \left(G(x), x \right) \right] \\ + \mathbb{E}_y \left[\mathcal{L}_{\text{rec}} \left(G(y), y \right) \right]. \end{aligned} \quad (6)$$

The total loss for OT-StainNet is expressed as follows:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{cycle}} \mathcal{L}_{\text{cycle}} + \lambda_{\text{id}} \mathcal{L}_{\text{id}} + \lambda_{\text{sc}} \mathcal{L}_{\text{sc}}, \quad (7)$$

where λ_{adv}, λ_{cycle}, λ_{id}, and λ_{sc} are hyperparameters that weight the relative importance of the various loss terms.

Algorithm 1: OT-StainNet Training Algorithm

Input: H&E image dataset X, adjacent layer IHC image dataset Y.

Output: Optimal parameters θ_G and θ_D of the generator and discriminator.

for each training iteration **do**
  Select an H&E image $x \in X$ and a corresponding IHC image $y \in Y$.
  Stage 1: Update Discriminator Parameters
  Generate the target staining images $G(x)$.
  Calculate $\mathcal{L}_{\text{adv}}$ using Eq. (4) and update discriminator parameters θ_D based on backpropagation.
  Stage 2: Update Generator Parameters
  Apply the encoder to obtain H&E features $\mathbf{F}^H$, adjacent IHC features $\mathbf{F}^I$, and virtual IHC features $\tilde{\mathbf{F}}^I$.
  Compute the semantic consistency loss:
    a. For each H&E-IHC image pair, use optimal transport to obtain the transport plan matrix $\mathbf{P}$, then derive the aligned real IHC features $\hat{\mathbf{F}}^I$.
    b. Compute the correlation matrices $\hat{\mathbf{M}}^I$ and $\tilde{\mathbf{M}}^I$ using Eq. (2). Subsequently, calculate the semantic consistency loss $\mathcal{L}_{\text{sc}}$ using Eq. (3).
  Compute the total loss $\mathcal{L}_{\text{total}}$ using Eq. (7) and update generator parameters θ_G based on backpropagation.
end for

Datasets	Methods	H&E→Ki67			H&E→ER			H&E→PR			H&E→HER2		
		SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓
MIST	Pix2Pix-Turbo	0.127	15.358	102.622	0.129	15.153	82.705	0.125	15.218	85.742	0.104	14.965	94.940
	CycleGAN-Turbo	0.133	14.856	82.351	<u>0.144</u>	14.331	47.170	0.129	13.895	94.517	0.100	13.831	107.738
	CUT	0.113	14.061	41.493	0.112	13.646	46.199	0.109	13.755	40.973	0.095	13.688	<u>40.664</u>
	PyramidP2P	0.135	**15.540**	86.585	0.132	<u>15.479</u>	89.660	0.133	<u>15.626</u>	74.396	0.116	<u>15.713</u>	116.002
	PRGAN	0.127	14.228	<u>39.151</u>	0.126	13.928	44.222	0.099	13.198	50.017	0.102	13.976	47.018
	ASP	0.134	15.091	50.992	0.116	14.297	<u>39.783</u>	0.117	14.612	43.330	**0.125**	14.504	48.957
	U-Frame	<u>0.145</u>	15.499	42.114	0.135	15.197	40.658	<u>0.140</u>	15.444	<u>38.955</u>	0.111	15.049	49.308
	Ours	**0.158**	<u>15.516</u>	**36.286**	**0.148**	**16.056**	**35.072**	**0.142**	**15.647**	**34.222**	<u>0.118</u>	**15.842**	**36.304**
IHC4BC	Pix2Pix-Turbo	0.210	18.154	203.989	0.468	20.384	152.751	0.548	<u>25.535</u>	190.576	0.183	12.793	105.109
	CycleGAN-Turbo	<u>0.223</u>	<u>18.671</u>	54.151	<u>0.505</u>	20.053	57.105	0.584	25.338	76.615	<u>0.204</u>	13.328	57.802
	CUT	0.117	18.371	31.908	0.339	15.556	65.941	0.315	13.690	162.275	0.195	13.133	53.826
	PyramidP2P	0.212	18.465	102.747	0.454	20.444	126.608	0.540	24.785	98.775	0.199	13.615	75.651
	PRGAN	0.170	17.218	28.705	0.453	19.228	33.119	0.515	24.612	64.407	0.201	13.434	<u>43.959</u>
	ASP	0.141	16.324	27.846	0.404	**21.392**	170.404	<u>0.609</u>	24.858	141.260	0.203	<u>13.625</u>	61.339
	U-Frame	0.180	17.584	**25.101**	0.464	19.458	<u>32.651</u>	0.537	25.438	<u>35.700</u>	0.202	13.575	47.205
	Ours	**0.241**	**18.855**	<u>26.377</u>	**0.526**	<u>21.379</u>	**30.480**	**0.635**	**25.614**	**31.076**	**0.206**	**13.664**	**41.700**

Table 1: Comparison of stain transfer performance using Pix2Pix-Turbo (Parmar et al. 2024), CycleGAN-Turbo (Parmar et al. 2024), CUT (Park et al. 2020), PyramidP2P (Liu et al. 2022), PRGAN (Zeng et al. 2022), ASP (Li et al. 2023), U-Frame (Dai, Wong, and Wong 2023), and our method across two benchmark datasets. The best and second-best scores are indicated in **bold** and <u>underlined</u>, respectively.

Training Details

To comprehensively describe the training process of OT-StainNet, we present the update procedures for the discriminator and generator in Algorithm 1.

Experiments and Results

Experimental Settings

Datasets We present the results of H&E to IHC stain transfer across two benchmark datasets: MIST (Li et al. 2023) and IHC4BC (Akbarnejad et al. 2023). Both datasets contain adjacent layer pairs of H&E and IHC-stained images obtained from consecutive tissue sections for four different IHC stains: Ki67, ER, PR, and HER2. For each IHC type, the MIST dataset has approximately $4{,}000$ patches for training and 1,000 patches for testing; the IHC4BC dataset has approximately $20{,}000$ patches for training and $1{,}000$ patches for testing. These images are resized to 256×256 pixels before being input into OT-StainNet. For each stain type, we use $1{,}000$ test images.

Implementation Details Our method is implemented in Python using PyTorch on a computer equipped with four NVIDIA RTX 3090 GPUs. All experiments are conducted at an image resolution of 256×256 pixels. OT-StainNet is trained for $30{,}000$ iterations. During training, we use the Adam optimizer with a learning rate of 1×10^{-5}. The hyperparameters in Eq. 7 are set as follows: $\lambda_{\text{adv}} = 0.5$, $\lambda_{\text{cycle}} = 10$, $\lambda_{\text{id}} = 1$, and $\lambda_{\text{sc}} = 1{,}000$.

Evaluation Metrics We use several evaluation metrics to assess the performance of H&E to IHC stain transfer, including the Structural Similarity Index Measure (SSIM) (Wang et al. 2004), Peak Signal-to-Noise Ratio (PSNR) (Hore and Ziou 2010), and Fréchet Inception Distance (FID) (Heusel et al. 2017).

Comparative Results and Discussion

We evaluate the performance of our method against other state-of-the-art (SOTA) methods both quantitatively and qualitatively using the MIST and IHC4BC datasets.

Compared Methods The methods compared in this study can be broadly classified into two categories: 1. Diffusion-based methods, including Pix2Pix-Turbo (Parmar et al. 2024) and CycleGAN-Turbo (Parmar et al. 2024); 2. GAN-based methods, including CUT (Park et al. 2020), PyramidP2P (Liu et al. 2022), PRGAN (Zeng et al. 2022), ASP (Li et al. 2023), and U-Frame (Dai, Wong, and Wong 2023).

Quantitative Assessment We quantitatively evaluate the performance of our method against state-of-the-art (SOTA) methods using the MIST and IHC4BC datasets. All comparison methods were implemented strictly according to the conditions specified in their respective papers or by utilizing their open-source code. The results are presented in Table 1. Our method significantly outperforms the comparison methods across SSIM, PSNR, and FID metrics, demonstrating its effectiveness for IHC virtual staining.

Unsupervised methods such as CycleGAN-Turbo and CUT, as well as the semi-supervised method PRGAN, exhibit inferior performance in SSIM and PSNR. This is primarily due to the limited supervisory information available from adjacent layer IHC images. In contrast, fully-supervised methods like Pix2Pix-Turbo and PyramidP2P

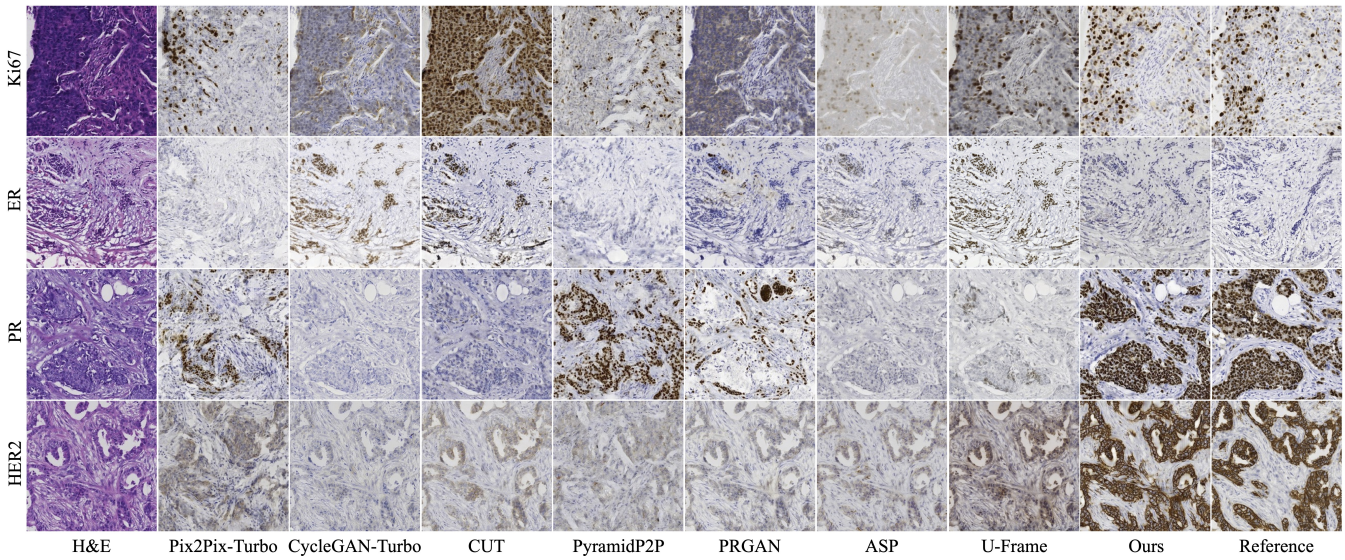


Figure 3: Virtual IHC image results of different methods on the MIST dataset.The first column displays the H&E images. Columns 2 to 9 show images virtually stained by different methods. The last column presents the adjacent layer real IHC images.

perform better on these metrics. However, Pix2Pix-Turbo and PyramidP2P require optimization of the L1 loss between virtual and adjacent layer weakly paired IHC images, which results in poor FID performance. Although ASP and U-Frame account for the weakly paired characteristics of adjacent layer image pairs, they still generate virtual images by calculating pixel-level loss, which can lead to inaccuracies in pathological characteristics. In comparison, our method achieves superior results. First, OT-StainNet establishes accurate semantic correspondences between H&E and IHC images using OTSM. Second, OT-StainNet effectively supervises virtual IHC images by leveraging correlations among real IHC features through SCC.

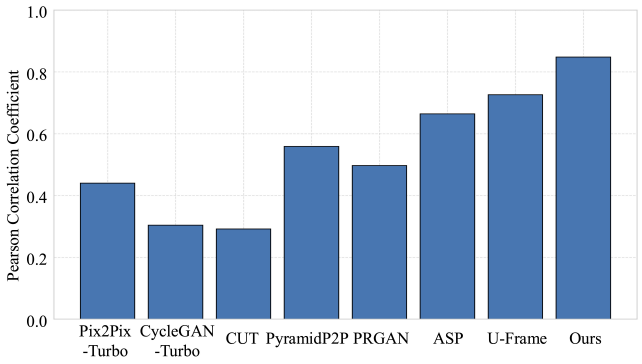


Figure 4: Pearson correlation coefficient between the size of the positive area in virtual IHC images generated by different methods and the corresponding real IHC images from adjacent layers.

Qualitative Assessment To further explore the visual differences in virtual staining images generated by different methods, we present the results of our method and other competing methods on the MIST dataset in Fig. 3. Compared to other methods, the virtual IHC images generated by our method most closely resemble the real IHC images of adjacent layers in terms from staining intensity, pathological characteristics, and tissue structure.

Fig. 3 demonstrates that existing methods, such as CycleGAN-Turbo, CUT, and PRGAN, generate IHC images with inaccurate styles due to insufficient supervisory information. Methods like Pix2Pix-Turbo and PyramidP2P fail to preserve the tissue structure in H&E-stained images because of forced optimization of pixel-level losses between virtual images and weakly paired IHC images from adjacent layers. Additionally, ASP and U-Frame neglect tissue misalignment and deformation between H&E-IHC image pairs, leading to virtual IHC images with erroneous pathological characteristics. In contrast, with OTSM and SCC, virtual IHC images generated by OT-StainNet are more similar to the referenced adjacent layer IHC images.

Pathology Consistency Assessment Pathologists generally focus on the positive areas of IHC staining for image analysis. Therefore, the accuracy of virtual IHC images can be effectively evaluated by examining the correlation between the positive area sizes of virtual IHC images and those of real IHC images in adjacent layers.

We obtain the corresponding positive area based on the stain characteristics of the IHC patch. For each IHC patch, the hematoxylin channel (blue) and DAB channel (brown) are separated by stain deconvolution. In the DAB channel, we calculate the size of the positive area in each IHC patch and determine the Pearson correlation coefficient between the positive area sizes of virtual images generated by different methods and real images in adjacent layers in the IHC4BC test set, as shown in Fig. 4. The results show that

Strategy	H&E→Ki67			H&E→ER			H&E→PR			H&E→HER2		
	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓	SSIM↑	PSNR↑	FID↓
w/o all	0.133	14.856	82.351	0.144	14.331	47.170	0.129	13.895	94.517	0.100	13.831	107.738
w/o OTSM	0.138	15.214	47.331	0.138	15.283	42.488	0.120	15.013	56.491	0.113	15.211	55.944
w/o SCC	0.150	14.877	39.980	**0.151**	15.616	39.856	0.131	15.352	40.176	0.107	15.472	39.882
Ours	**0.158**	**15.516**	**36.286**	0.148	**16.056**	**35.072**	**0.142**	**15.647**	**34.222**	**0.118**	**15.842**	**36.304**

Table 2: The quantitative results of the effects of FMM and SCC on OT-StainNet. The best score is indicated in **bold**.

our method demonstrates stronger pathology consistency between the virtual and reference images compared to other methods.

Ablation Study and Analysis

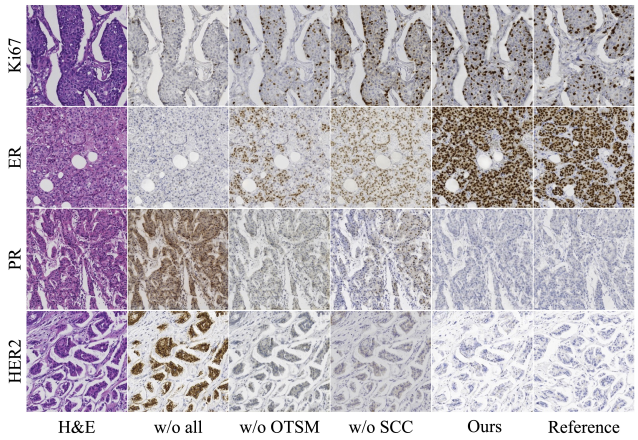


Figure 5: The visualization of the effects of FMM and SCC on OT-StainNet. "w/o SCC" denotes replacing SCC with the standard L1 loss.

The implementation of OT-StainNet primarily relies on the OT-driven semantic matching (OTSM) mechanism and the semantic consistency constraint (SCC). To investigate the impact of OTSM and SCC on virtual staining performance, we performed ablation experiments on the MIST dataset. The results are presented in Table 2, where "w/o SCC" denotes replacing SCC with the standard L1 loss. Both OTSM and SCC individually enhance virtual staining performance, with OTSM providing a more substantial improvement. The combination of OTSM and SCC yields the highest stain transfer performance. This is because OTSM establishes accurate dense semantic correspondences between H&E and IHC images, while SCC leverages the correlation among real IHC features to provide more effective semantic-level supervision for virtual IHC.

To investigate the effect of OTSM and SCC in OT-StainNet on the visual performance of virtual IHC, we present qualitative ablation results on the MIST dataset in Figure 5. As shown, when OT-StainNet is not constrained by OTSM and SCC, the model cannot utilize the supervisory information provided by the adjacent layer IHC images, resulting in the inability to generate accurate IHC image styles. Adding OTSM or SCC individually improves virtual staining performance, but the virtual IHC images still exhibits some erroneous pathological features. When OT-StainNet incorporates both OTSM and SCC, the virtual IHC images more closely resemble the reference adjacent layer IHC images, demonstrating the crucial role of OTSM and SCC in enhancing the transfer performance of IHC staining.

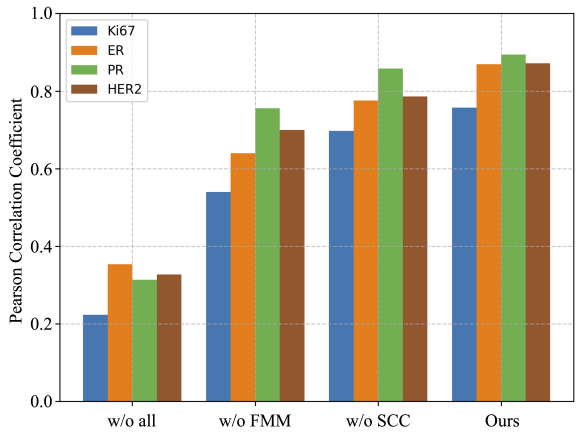


Figure 6: The effects of FMM and SCC in OT-StainNet on the Pearson correlation coefficient between the positive area sizes of virtual IHC images and adjacent real IHC images.

To further investigate the effect of OTSM and SCC on the accuracy of positive regions in virtual IHC images, we calculated the Pearson correlation coefficient between the size of positive regions in virtual images for each IHC staining type and the corresponding real images from adjacent layers under various experimental conditions on the IHC4BC dataset, as shown in Figure 6. The results demonstrate that OTSM and SCC significantly enhance the pathological consistency between virtual and real IHC images.

Conclusion

In this paper, we propose a novel IHC virtual staining method, OT-StainNet, which employs an optimal transport-driven semantic matching (OTSM) mechanism to establish accurate semantic correspondences between H&E and IHC image pairs that are not precisely paired at the pixel level. This approach avoids errors in pathological characteristics caused by optimizing pixel-level loss between synthetic and

real images, providing new insights for the histopathology image virtual staining research community. Leveraging the real IHC features obtained by the OTSM mechanism, we introduce a semantic consistency constraint (SCC) to provide effective semantic-level supervision for the generation of virtual IHC images. The OT-StainNet is not inherently limited to specific tissue types and can, in theory, be applied to various staining techniques, including special stains (e.g., MAS, PAS and PASM) and fluorescence stains. However, the performance of OT-StainNet is related to the quality of the dataset, and it may not be as expected when the training images are contaminated (the cell structures in the images cannot be clearly identified). In the future, we plan to extend this method to various IHC stainings of other tissue types, thereby expanding the application of OT-StainNet.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62031023, 62331011), in part by the Shenzhen Science and Technology Program (Grant Nos. GXWD20220818170353009, RCYX20221008092852077), and in part by the Fundamental Research Funds for the Central Universities (Grant No. HIT.OCEF.2023050).

References

Akbarnejad, A.; Ray, N.; Barnes, P. J.; and Bigras, G. 2023. Predicting ki67, er, pr, and her2 statuses from h&e-stained breast cancer images. *arXiv preprint arXiv:2308.01982*.

Bai, B.; Yang, X.; Li, Y.; Zhang, Y.; Pillar, N.; and Ozcan, A. 2023. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1): 57.

Dai, W.; Wong, I. H.; and Wong, T. T. 2023. A weakly supervised deep generative model for complex image restoration and style transformation. *Authorea Preprints*.

Ferber, D.; Wölflein, G.; Wiest, I. C.; Ligero, M.; Sainath, S.; Laleh, N. G.; El Nahhas, O. S.; Müller-Franzes, G.; Jäger, D.; Truhn, D.; et al. 2024. In-context learning enables multimodal large language models to classify cancer pathology images. *arXiv preprint arXiv:2403.07407*.

Gatenbee, C. D.; Baker, A.-M.; Prabhakaran, S.; Swinyard, O.; Slebos, R. J.; Mandal, G.; Mulholland, E.; Andor, N.; Marusyk, A.; Leedham, S.; et al. 2023. Virtual alignment of pathology image series for multi-gigapixel whole slide images. *Nature communications*, 14(1): 4502.

Ghahremani, P.; Li, Y.; Kaufman, A.; Vanguri, R.; Greenwald, N.; Angelo, M.; Hollmann, T. J.; and Nadeem, S. 2022. Deep learning-inferred multiplex immunofluorescence for immunohistochemical image quantification. *Nature machine intelligence*, 4(4): 401–412.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Guan, X.; Wang, Y.; Lin, Y.; Li, X.; and Zhang, Y. 2024. Unsupervised Multi-Domain Progressive Stain Transfer Guided by Style Encoding Dictionary. *IEEE Transactions on Image Processing*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*, 2366–2369. IEEE.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kantorovich, L. V. 2006. On the Translocation of Masses. *Journal of mathematical sciences*, 133(4).

Li, C.; Feng, B. Y.; Liu, Y.; Liu, H.; Wang, C.; Yu, W.; and Yuan, Y. 2024a. Endosparse: Real-time sparse view synthesis of endoscopic scenes using gaussian splatting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 252–262. Springer Nature Switzerland Cham.

Li, C.; Liu, H.; Liu, Y.; Feng, B. Y.; Li, W.; Liu, X.; Chen, Z.; Shao, J.; and Yuan, Y. 2024b. Endora: Video generation models as endoscopy simulators. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 230–240. Springer Nature Switzerland Cham.

Li, C.; Liu, X.; Li, W.; Wang, C.; Liu, H.; Liu, Y.; Chen, Z.; and Yuan, Y. 2024c. U-kan makes strong backbone for medical image segmentation and generation. *arXiv preprint arXiv:2406.02918*.

Li, F.; Hu, Z.; Chen, W.; and Kak, A. 2023. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 632–641. Springer.

Liu, S.; Zhang, B.; Liu, Y.; Han, A.; Shi, H.; Guan, T.; and He, Y. 2021. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE transactions on medical imaging*, 40(8): 1977–1989.

Liu, S.; Zhu, C.; Xu, F.; Jia, X.; Shi, Z.; and Jin, M. 2022. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1815–1824.

Mahbod, A.; Polak, C.; Feldmann, K.; Khan, R.; Gelles, K.; Dorffner, G.; Woitek, R.; Hatamikia, S.; and Ellinger, I. 2024. NuInsSeg: A fully annotated dataset for nuclei instance segmentation in H&E-stained histological images. *Scientific Data*, 11(1): 295.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. 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*, 319–345. Springer.

Parmar, G.; Park, T.; Narasimhan, S.; and Zhu, J.-Y. 2024. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rivenson, Y.; Wang, H.; Wei, Z.; de Haan, K.; Zhang, Y.; Wu, Y.; Günaydın, H.; Zuckerman, J. E.; Chong, T.; Sisk, A. E.; et al. 2019. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature biomedical engineering*, 3(6): 466–477.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Sauer, A.; Lorenz, D.; Blattmann, A.; and Rombach, R. 2023. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Zeng, B.; Lin, Y.; Wang, Y.; Chen, Y.; Dong, J.; Li, X.; and Zhang, Y. 2022. Semi-supervised pr virtual staining for breast histopathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 232–241. Springer.

Zhang, Q.; Li, J.; Liao, P.; Hu, J.; Guan, T.; Han, A.; and He, Y. 2024. Leveraging Pre-trained Models for FF-to-FFPE Histopathological Image Translation. *arXiv preprint arXiv:2406.18054*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.