Test-Time Stain Adaptation with Diffusion Models for Histopathology Image Classification

Cheng-Chang Tsai[1], Yuan-Chih Chen[2], and Chun-Shien Lu[1,2]

[1] Research Center for Info. Technology Innovation, Academia Sinica, Taiwan, ROC
[2] Institute of Information Science, Academia Sinica, Taiwan, ROC
{cctsai, willpower057, lcs}@iis.sinica.edu.tw

Abstract. Stain shifts are prevalent in histopathology images, and typically dealt with by normalization or augmentation. Considering training-time methods are limited in dealing with unseen stains, we propose a test-time stain adaptation method (TT-SaD) with diffusion models that achieves stain adaptation by solving a nonlinear inverse problem during testing. TT-SaD is promising in that it only needs a single domain for training but can adapt well from other domains during testing, preventing models from retraining whenever there are new data available. For tumor classification, stain adaptation by TT-SaD outperforms state-of-the-art diffusion model-based test-time methods. Moreover, TT-SaD beats training-time methods when testing on data that are inaccessible during training. To our knowledge, the study of stain adaptation in diffusion model during testing time is relatively unexplored.

Keywords: Stain Adaptation · Inverse Problem · Diffusion Model

1 Introduction

1.1 Background

Deep learning has significantly advanced computer-aided diagnosis based on histopathology images [1, 9, 13, 22], but can still encounter performance degradation when there is a distribution shift between the training data and testing data [14, 36]. There are many factors causing distribution shifts in histopathology images, such as digitization, blur, color, and stain [19, 50]. In this paper, we mainly focus on the distribution shifts resulting from stains. In the process of generating histopathology images, different stains must be applied to tissue sections to render the tissue visible through a microscope. However, many reasons, including tissue manipulation, staining, and scanning, lead to significant alterations in the chromatic appearance of histopathology images. We refer to these alterations in chromatic appearance as "stain shifts" that are commonly encountered cases of out-of-distribution in histopathology.

Moreover, stain shift is inherent in histopathology image classification. Most current classification works are slide-level [7, 27, 28, 30, 48]; however, we study

patch-level classification in this paper. The development of patch-level classification makes this paper able to concentrate on addressing stain shifts. Notably, the performance of a classifier will deteriorate if the input is out of the distribution with respect to the training data that the classifier has learned. For instance, two practical and common situations, where stain shifts become evident, are: (i) A new hospital joins, and (ii) A different staining protocol is applied. In both situations, incoming histopathology images are unseen data to the classifier.

The dominant methods for coping with stain shifts are either stain augmentation [6,40], expanding the model's capability of generalization by enriching the stains of the training data during training time, or stain normalization [32,39,43], mitigating the stain variations of inputs with reference to the normalized stains of training data during testing time. In this paper, we propose a test-time stain adaptation method with diffusion models [12, 21, 34, 41], dubbed TT-SaD, that uses the diffusion model trained on the training data, called the source data, to shift the stains of incoming inputs to the stains of source data.

1.2 Motivation

Based on the above concerns, we study how to adapt the incoming histopathology images during testing time with a diffusion model. The main reason that we choose a diffusion model as the generative model of TT-SaD is that it can overcome the problem of encountering unseen data, which is the primary limitation of stain augmentation methods. TT-SaD is also remarkably different from stain normalization in that it is not a mapping from target domain to source domain as stain normalization does; instead it is a model pushing data toward the source domain, as demonstrated in Figs. 6 and 7 of Sec. A.5 in Appendix.

The rationale behind TT-SaD is the design of sampling process in the context of diffusion model to shift the stains of inputs to those of source data as closely as possible. Specifically, our goal is to solve a nonlinear inverse problem formulated for stain adaptation by a diffusion model. With the classification task as the objective, the structure of inputs is important information since classifiers learn to classify inputs regarding their structure [14, 23, 35, 43]. As a result, TT-SaD should minimize the potential loss of structural similarity between the inputs and their stain-shifted ones as much as possible.

DiffPure [35] and DDA [14] are two methods similar to TT-SaD in that they all apply diffusion models to address the distribution shift problem in natural images. However, TT-SaD is developed for classification of histopathology images instead of natural images. Since histopathology images are stained with few dyes, we can exploit this characteristic, which was not considered in DiffPure and DDA, for the development of TT-SaD. In addition, there are several differences among them, as summarized in Table 1, including state-of-the-art methods exclusively developed for stains as well.

1.3 Our Contributions

Our contributions are summarized as follows:

Table 1: Comparisons among domain adaption methods in natural and histopathology images. "-" denotes "void" and "Unseen Target Domain" denotes the target domain is inaccessible during training.

Method	Image Type	Adaption	Single Training Domain	Stain Matrix	Structure Awareness	Unseen Target Domain	Test-Time	Stain Diversity
CycleGAN [51]	Natural	GAN	-	-	-	-	v	-
DiffPure [12]	Natural	Diffusion Model	v	-	-	v	v	-
DDA [14]	Natural	Diffusion Model	v	-	v	v	v	-
StainDiff [39]	Histopatholoy	Diffusion Model	-	-	-	-	v	-
Macenko *et al.* [32]	Histopatholoy	Normalization	v	v	v	v	v	-
Vahadane *et al.* [43]	Histopatholoy	Normalization	v	v	v	v	v	-
Mahapatra [33]	Histopatholoy	GAN	-	-	v	-	v	-
CAGAN [11]	Histopatholoy	pix2pix GAN	-	-	-	-	v	-
HistAuGAN [44]	Histopatholoy	DRIT++	-	-	v	-	-	v
Ke *et al.* [25]	Histopatholoy	CycleGAN	-	-	-	-	v	-
Stain Mixup [6]	Histopatholoy	Mixup [49]	-	v	v	-	-	v
RandStainNA [40]	Histopatholoy	Normalization	v	-	v	v	-	v
TTSA [47]	Histopatholoy	Mixup	v	v	-	v	v	v
TT-SaD (Ours)	Histopatholoy	Diffusion Model	v	v	v	v	v	v

– To our knowledge, we are the first to study the issue of stain adaptation from the perspective of an inverse problem.
– Our approach, test-time stain adaptation with diffusion models (TT-SaD), achieves stain adaptation with diffusion models during testing time.
– TT-SaD is promising in that it needs only a single domain for training but can adapt well from other domains without additional data, preventing hospitals from retraining models due to new data available.

2 Related Work

In this section, we review those methods that are proposed to handle stain shifts in medical imaging and that employ diffusion models to solve domain shifts.

2.1 Stain-related Methods for Medical Images

Hospitals are unable to obtain a large amount of data due to privacy, rare diseases, and cost issues. However, learning-based methods are susceptible to the batch effect [8] caused by the data provided by a single hospital. Therefore, they easily overfit to specific features, resulting in the issue of stain shift among different hospitals or even bias in different rounds of scanning processing.

Macenko *et al.* [32] proposed to find stain vectors for each histopathology image using color from a reference one, based on the linear combination of two stain vectors (*e.g.*, eosin and hematoxylin). Vahadane *et al.* [43] proposed to decompose a histopathology image into sparse and non-negative stain density map, followed by structure-preserving color normalization, which only combines the stain density map with stain color basis of a source histopathology image without altering the stain density map.

RandStainNA [40] is proposed to augment histopathology images in random color spaces (*i.e.*, HED, HSV, and LAB) by sampling the mean and standard deviation from a Gaussian distribution. Chang *et al.* [6] proposed Stain Mixup to mix the stain matrices, decomposed from histopathology images, of both target domain data and source domain data to train a stain robust model. To reduce the cost of training a new model, TTSA [47] only mixes the stain matrices during testing time. In [42], Tiard *et al.* trained a stain-invariant feature extractor by incorporating stain normalization and contrastive learning.

For GAN-based methods, Cong *et al.* [11] proposed CAGAN to map grayscale histopathology images to their RGB counterparts by a pix2pix GAN [24], trained by supervised learning on target data and self-supervised learning on source data. Mahapatra *et al.* [33] divided the latent code of a histopathology image into the texture and structural components, followed by a combination of the structural component of source data and the texture of target data, to deal with stain shift. Similar to Mahapatra *et al.* [33], Wagner *et al.* [44] used DRIT++ [26] to disentangle a histopathology image into shared content space and attribute space, and additionally maps histopathology images to different domains by their one-hot encoded domain vectors. To alleviate the mode collapse problem of GAN, Ke *et al.* [25] mixed the source data and target data to train an auxiliary model, producing stain-invariant latent feature that aids CycleGAN to stay away from the mode collapse problem by contrastive learning.

Note that the characteristic of our method, TT-SaD, relies on the use of source domain data only without accessing other domains data. Since it is impractical for hospitals to retrain the model whenever there are new data or a new scanner available, we mainly focus on the comparison of methods that adapt data from multiple target domains, including unseen domains, to a single source domain during testing time. On the other hand, we illustrate in Fig. 4 of Sec. A.1 in Appendix the different stain generalization methods for visual realization.

2.2 Applications of Diffusion Models

As mentioned in Sec. 2.1, most proposed methods are based on GANs [11, 25, 33, 44] that often lead to information loss or undesirable results [12, 25]. On the contrary, denoising diffusion probabilistic models (DDPMs) [21] represent a class of generative models known for their superior performance in image generation. In addition to image generation, diffusion models make a massive impact in the topics of inverse problems [10, 31, 45], adversarial purification [35], test-time adaption [14], and stain style transfer [39].

3 Preliminary

In this section, we first introduce the main problem and some notations in Sec. 3.1, followed by reviewing stain separation and diffusion models in Sec. 3.2 and Sec. 3.3, respectively, to make this paper self-contained.

3.1 Problem Formulation

The flowchart of our TT-SaD method is illustrated in Fig. 1. Suppose there are K hospitals in the stain adaptation problem. Given the histopathology images, $X_j = \{x_1^j, x_2^j, \ldots, x_{n_j}^j\}$, from the j-th hospital ($1 \leq j \leq K$), as depicted in Fig. 1(a), and their corresponding labels $Y_j = \{y_1^j, y_2^j, \ldots, y_{n_j}^j\}$, they constitute a dataset $\mathcal{D}_j = \{(x_i^j, y_i^j),\ i = 1, 2, \cdots, n_j\}$, where n_j is the number of histopathology images in $\mathcal{D}_j$. Without loss of generality, we will mostly omit the superscript or subscript for notation simplicity. Our goal is to shift the stain of an input image x^t in target domain to the source stain w_s by a diffusion model ϵ_θ trained on the dataset in source domain, denoted as $\mathcal{D}^s$.

3.2 Stain Separation

Stain separation [43], depicted in Figs. 1(b) and 1(c), isolates the stain matrix of a histopathology image with its structural information captured by the stain density map. More specifically, stain separation is to estimate the density of each stain at every pixel, which is beneficial for understanding the relation between the RGB value and stain density.

Given a histopathology image $x \in \mathbb{R}^{3\times n}$ in the RGB space, where n is the number of pixels, we convert x to its relative optical density v by the Beer-Lambert law [15]:

$$v = \mathrm{BL}(x) := -\log \frac{x}{I_0}, \tag{1}$$

where I_0 is the illuminating light intensity of an image, which is equal to 255 for 8-bit images in our case. After the process of conversion to the optical density space, v can be decomposed into a stain matrix $w \in \mathbb{R}^{3\times r}$ and a stain density map $h \in \mathbb{R}^{r\times n}$ as follows [43]:

$$v = wh, \tag{2}$$

where r represents the number of stains. To convert back to the RGB color space from the optical density space, we directly reverse the operations in Eq. (1) as:

$$x = \mathrm{BL}^{-1}(v) := I_0 \exp(-v). \tag{3}$$

We denote all stain matrices of X_j as $W_j = \{w_1^j, w_2^j, \cdots, w_{n_j}^j\}$, where w_i^j is the stain matrix of x_i^j. As illustrated in Fig. 1(b), when a so-called "domain center" is inferred from all data in X_j, we refer to it as the "domain overall center." While a domain center is inferred from the data in X_j with labels belonging to the tumor class only, we refer to it as the "domain tumor center."

3.3 Diffusion Models

DDPMs [21] is known to fit the distribution of a dataset so that the generated images are within the distribution of that dataset and consist of two processes,

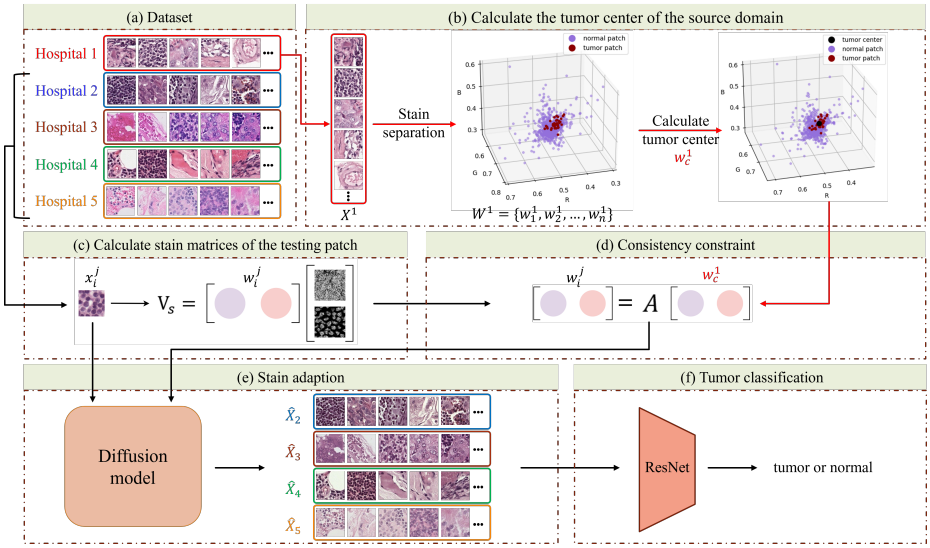


Fig. 1: Workflow of our TT-SaD Method

which are forward and reverse processes defined with T timesteps. Given a sample x_0 drawn from a data distribution $q(x_0)$, the forward process gradually adds Gaussian noises to x_0 to produce latent samples $x_1,\ x_2, \ldots,\ x_T$. Specifically, each step of the forward process adds Gaussian noise according to a fixed variance schedule given by β_t:

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}). \tag{4}$$

Ho *et al*. [21] pointed out that the forward process does not need to repeatedly apply Eq. (4) to sample x_t at an arbitrary timestep t but instead uses the following closed form:

$$\begin{aligned} q(x_t|x_0) &:= \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathbf{I}) \\ &= \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \end{aligned} \tag{5}$$

where $\alpha_t := 1-\beta_t$, $\bar{\alpha}_t := \prod_{s=1}^{t}\alpha_s$, and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$.

Since the reverse of $q(x_t|x_{t-1})$ is intractable, DDPMs learn parameterized $p_\theta(x_{t-1}|x_t)$ with learned mean and fixed variance:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2\mathbf{I}), \tag{6}$$

where $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$. One step of the reverse process is specified as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t\epsilon. \tag{7}$$

Denoising diffusion implicit models (DDIMs) [41] are a variant of DDPMs, generating higher-quality images using a much fewer number of steps. DDIMs

can use the same diffusion models trained for DDPMs without retraining. The most important part is that each step of the reverse process is specified as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}} \right) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2}\epsilon_\theta(x_t, t) + \sigma_t \epsilon. \quad (8)$$

It is noteworthy that the first term is referred to as a predicted x_0 and the second term is referred to as a direction pointing to x_t. This concept is crucial for our TT-SaD method.

4 Stain Adaptation with Diffusion Models

In this section, we will first provide a comprehensive explanation of how to cast stain shift as an inverse problem in Sec. 4.1 and then describe the details of our TT-SaD method, illustrated in Fig. 1, in Sec. 4.2.

4.1 Stain Shift as an Inverse Problem

We begin by studying how to formulate stain shift as an inverse problem. To this end, we exploit the stain separation method introduced in Sec. 3.2 to isolate stain matrices for stain shifts.

Suppose we have a desired stain matrix w_s, which represents the desired stain we want an input image x^t to possess. The relative optical density of x^t, denoted as v^t, can be decomposed for obtaining its stain matrix w_t through Eq. (2):

$$v^t = w_t h_t, \quad (9)$$

where h_t is the stain density map of x^t. To shift stains, we use a stain shift matrix $A \in \mathbb{R}^{3\times3}$ to map each stain of w_s to that of w_t as follows:

$$w_t = A w_s, \quad (10)$$

as depicted in Fig. 1(d). Obtaining a stain shift matrix A for mapping stains requires solving a system of linear equations. In this work, we obtain A through the pseudo-inverse $w_s^\dagger$ of w_s, *i.e.*, $A = w_t w_s^\dagger$. After obtaining a stain shift matrix, we can further formulate the inverse problem for stain shifts as follows:

$$x^t = \mathrm{BL}^{-1}(A(\mathrm{BL}(\hat{x}))), \quad (11)$$

where $\hat{x}$ is a histopathology image with the desired stain. Therefore, solving a stain adaptation problem is equivalent to solving an inverse problem explicitly formulated as in Eq. (11). We will address this inverse problem using diffusion models in this paper. Eq. (11) can be regarded as the condition imposed on all latent samples of a diffusion model to satisfy that the stains of the latent samples transformed by a stain shift matrix would be the stain of the input image [45], which is the rationale behind our TT-SaD method, as described in Sec. 4.2.

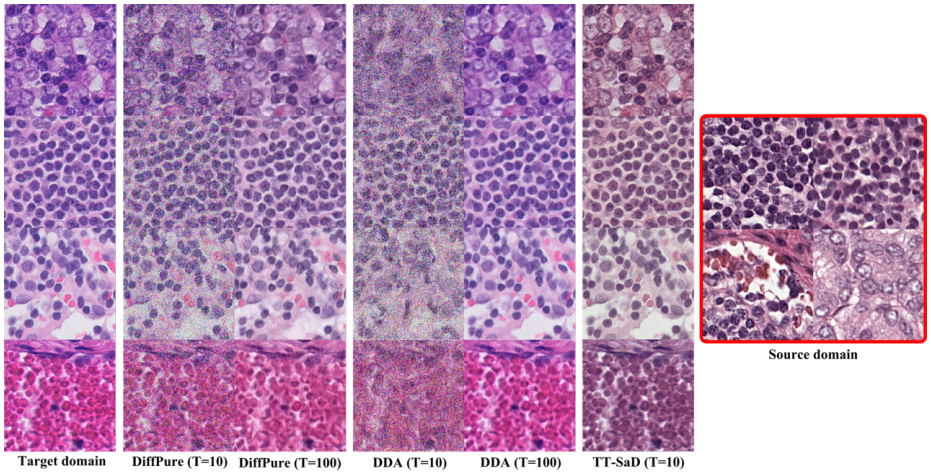
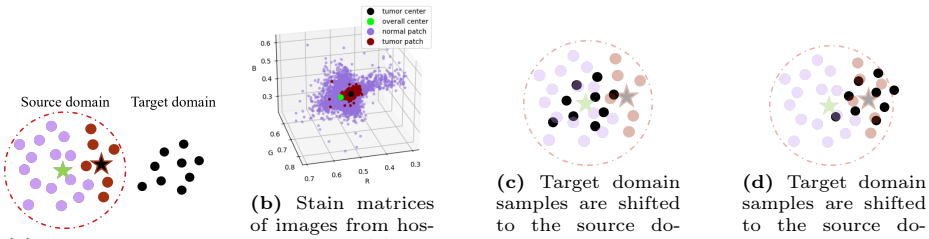


Fig. 2: Visualization of stain adaptation via different test-time methods based on diffusion models. It is evident that our TT-SaD method generates images with a color closer to that of images in the source domain qualitatively.

4.2 Stain Adaptation with Diffusion Models

As pointed out by Guo *et al.* [14], there is a trade-off between class information preservation and domain translation when choosing different timesteps of the forward process of diffusion models. To address this trade-off, DDA [14] takes $N < T$ forward steps, where T is the total timesteps of the exploited diffusion model, and employs a low-pass filter to not only preserve class information but also translate domains. Inspired by this, we take $N < T$ forward steps to generate the latent sample x_N through Eq. (5). Nevertheless, according to our observations, employing a low-pass filter is insufficient to guarantee that the outputs of a diffusion model closely replicate the structure of their input images to a significant degree, as shown in Tables 2 and 3 quantitatively and Fig. 2 qualitatively. To minimize the potential loss of structural similarity between inputs and their stain-shifted ones during the reverse process of diffusion models, we exploit the inverse problem introduced in Sec. 4.1 for bridging the outputs and corresponding input images to replicate the structure.

Inspired by DDNM [45], we add the consistency constraint to each step of the reverse process of diffusion models. Such consistency constraint is formulated as a linear inverse problem with an explicit form of the operator. The consistency constraint is imposed to make the output $\hat{x}$ of our TT-SaD method and the input image x strictly satisfy Eq. (11). Two inevitable issues then arise: (i) Unlike the degradation operators introduced in [45], the stain shift matrix A depends on both the input images and the desired stain. Thus, Eq. (11) does not have any explicit form. (ii) Eq. (11) is not a linear inverse problem, which violates the assumption made in [45].



(a) Source and target domain samples

(b) Stain matrices of images from hospital 1 in CAMELYON17

(c) Target domain samples are shifted to the source domain by the domain overall center.

(d) Target domain samples are shifted to the source domain by the domain tumor center.

Fig. 3: Illustration of selecting a domain center. Previous works [11, 47] used all data in the source domain to calculate the domain overall center. We demonstrate the differences that may affect the classification results between the domain overall center and the domain tumor center in (c) and (d), respectively. Purple and brown dots indicate the normal and tumor images in the source domain, respectively, while black dots indicate images in the target domain. Green and dark stars denote the domain overall center and domain tumor center, respectively. The red dash-circle indicates the classifier trained on the source data.

For the first issue, we need both the stain matrices of input images and the desired stain to construct the stain shift matrix A. The stain matrices of input images can be obtained by the stain separation introduced in Sec. 3.2. As for the stain matrix of the desired stain, an intuitive approach to obtaining it is to infer a representative stain matrix from the training dataset of tumor classifier. Previous works [11, 47] tend to use all training data to calculate the domain overall center. Although it works, we have observed some issues, as illustrated in Fig. 3. As a result of the class-imbalanced nature of tumor classification, selecting a domain center from all data may result in a situation, as depicted in Fig. 3(c), where stains of histopathological images belonging to the tumor class could be lost. Therefore, we propose to select a domain center from all histopathological images belonging to the tumor class, as depicted in Fig. 1(b) and Fig. 3(d), where the stains of shifted images could not be biased by the selected domain center. After determining the dataset used for selecting a domain center, we can infer it by the process subsequently described.

Let $X = \{x^1, x^2, \ldots, x^S\}$ denote the dataset used to infer a domain center, where S is the number of images in X, and let $W_1, W_2, \ldots, W_S$ denote the corresponding stain matrices of data in X. We first obtain the mean stain matrix of X as:

$$\overline{W} = \frac{1}{S} \sum_{i=1}^{S} W_i. \tag{12}$$

Then, x^j is said to be the domain center of X if its stain matrix W_j is the solution to the optimization problem:

$$\min_{W_j \in \{W_1, W_2, \ldots, W_S\}} \left\| W_j - \overline{W} \right\|. \tag{13}$$

After obtaining W_j, a stain shift matrix A can be obtained by solving Eq. (10) with W_j being w_s and the stain matrix of an input image being w_t, respectively.

As for the second issue, we relax the linearity assumption to incorporate Eq. (11) into the reverse process of diffusion. Although the whole inverse problem is nonlinear, we find there is still a linear function inside Eq. (11). To reach this linearity, we first propose to transform from the RGB color space to the optical density space. As one can see, after this transformation, the inverse problem in Eq. (11) becomes

$$V = A(\mathrm{BL}(\hat{x})), \tag{14}$$

where V is the relative optical density of an input image x and A is the stain shift matrix with respect to x. Eq. (14) suggests that we can impose the consistency constraint in the optical density space instead of the RGB color space.

Another important issue needs to be noted is that although a typical strategy is to impose a condition, Eq. (11), on all latent samples during the reverse process of diffusion models, it is noteworthy that all latent samples are actually noisy versions of the output of diffusion models with different noisy levels. Therefore, to impose the consistency constraint more reasonably, we will turn to impose Eq. (11) on the predicted output at each step of the reverse process. At timestep t, $x_{0;t}$ denotes the predicted output of a diffusion model, which depends on the inference noise at the current timestep.

Specifically, at timestep t of the reverse process, we first reference a predicted $x_{0;t}$ as:

$$x_{0;t} = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t)}{\sqrt{\bar{\alpha}_t}}. \tag{15}$$

To ensure that the output x_0 conforms to Eq. (11), the following step is performed:

$$\hat{x}_{0;t} = \mathrm{BL}^{-1}(A^\dagger(\mathrm{BL}(x)) + (\mathbf{I} - A^\dagger A)\mathrm{BL}(x_{0;t})), \tag{16}$$

where x is the input image and $A^\dagger$ is the pseudo-inverse of A. Eq. (16) can be regarded as a step that only shifts the stains because both x and $x_{0;t}$ are transformed by $A^\dagger$ and $\mathbf{I} - A^\dagger A$, respectively, after conversion to the optical density space. After imposing the consistency constraint on the predicted output, we can generate the latent sample at timestep $t-1$ using Eq. (8).

In summary, as illustrated in Fig. 1, we first take N forward steps of diffusion model from the target domain by adding the noise at each step, and then we denoise this latent sample in an iterative manner with the consistency constraint as the condition to shift the input image to the one with the desired stain. To further understand TT-SaD, we provide another interpretation and the pseudo code in Sec. A.2 of Appendix.

5 Experimental Results

We describe the datasets adopted for experiments in Sec. 5.1, experimental settings and performance evaluation metrics in Sec. 5.2 and Sec. 5.3, respectively, and main results in Sec. 5.4.

5.1 Datasets

Both MitosAtypia14 [38] and CAMELYON17 [5] were adopted to conduct stain adaptation experiments because they are open-access datasets containing whole-slide images (WSIs) stained with hematoxylin and eosin (H&E) dyes. Since all WSIs are H&E-stained, the number r of stains, as specified in Eq. (2), in the experiments was equal to 2. In our stain adaptation problem, WSIs in MitosAytpia14 were scanned by two scanners, denoted as "Aperio Hospital" and "Hamamatsu Hospital," respectively, and WSIs in CAMELYON17 were collected from five distinct medical centers, denoted as Hospital 1 $\sim$ Hospital 5, respectively. For each hospital, we extracted histopathology images/patches of size 256×256 from WSIs at $40\times$ magnification following the procedure mentioned in [16] to generate a dataset for that hospital. In Sec. A.3 of Appendix, we show a few sample images of each hospital in Fig. 5 and the numbers of all types of images in each hospital in Table 7.

5.2 Experimental Setup

Our experiment consists of two evaluations: (i) Stain adaptation quality and (ii) Tumor classification. For the evaluation of stain adaptation quality, one hospital was set as the source domain, while the others in the same dataset were the target domains. However, for the evaluation of tumor classification, only CAMELYON17 was used in that Hospital 1 was set as the source domain while the others serve as the target domains. We trained a diffusion model on the source domain for the evaluation of stain adaptation quality on the target domains and also trained a ResNet-50 [18] classifier for the evaluation of tumor classification with the AdamW [29] optimizer and a learning rate of 0.001, as suggested in [6]. We mainly compared TT-SaD with DDA [14] and DiffPure [35], using the same diffusion model, for both evaluations since we focus on comparing test-time methods with access to data in the source domain only during training, which is the real-world scenarios mentioned in Sec. 1 and Sec. 2.1. Additionally, we provided a comparison between TT-SaD and two selected training-time methods, Stain Mixup [6] and RandStainNA [40], demonstrating that TT-SaD outperforms training-time methods, whether they used additional data or not. For the timestep T in diffusion models, we provided two cases: 10 and 100. The forward step N was set to $\frac{T}{2}$ for TT-SaD, while for DiffPure and DDA, the default settings in the papers were adopted.

5.3 Evaluation Metrics

We followed [52] to adopt Structural Similarity Index (SSIM) [46], Wasserstein Distance (WD) [37], and Fréchet Inception Distance (FID) [20] for evaluating the quality of histopathology images generated by stain shift methods with consideration of two factors: (i) How well generated images reproduce the visual appearance of images from the target domain? and (ii) How well a generated image preserves the structure of a target image? In addition, we also adopted

Table 2: Evalutation of stain adaptation quality on MitosAtypia14 dataset.

Method	Aperio to Hamamatsu				Hamamatsu to Aperio			
	SSIM↑	FID↓	KID↓	WD↓	SSIM↑	FID↓	KID↓	WD↓
$T = 10$								
DiffPure	0.178961	223.507067	0.277424	0.004422	0.149235	208.153927	0.273480	0.005107
DDA	0.056028	279.529698	0.360294	0.004685	0.057248	265.458139	0.356771	0.005265
TT-SaD	**0.922287**	**43.931259**	**0.047004**	**0.001059**	**0.897057**	**21.447456**	**0.019419**	**0.001094**
$T = 100$								
DiffPure	0.595026	52.814048	0.055277	0.001198	0.571443	48.222087	0.060194	0.001748
DDA	0.594029	51.691766	0.053648	**0.001100**	0.586576	43.125872	0.048243	0.002053
TT-SaD	**0.962218**	**29.523483**	**0.030251**	0.001487	**0.956795**	**24.633128**	**0.023857**	**0.000608**

Kernel Inception Distance (KID) [2] here.[3]On the other hand, for tumor classification, accuracy (ACC) [4], AUROC [17], and AUPRC [3] were used to evaluate the classification performance.

5.4 Main Results

Evaluation of Stain Adaption Quality. We first verify that TT-SaD achieves the goal of stain adaptation in Tables 2 and 3 with respect to MitosAtypia14 and CAMELYON17. First, as shown in Tables 2 and 3, TT-SaD achieves much better image structure-preserving after stain shift than both DDA and DiffPure in terms of SSIM. Second, we compare the distance between the source domain and the distribution of images generated from DiffPure, DDA, and TT-SaD, respectively, in terms of FID and KID. The results show that TT-SaD can fit the distribution of source domain much better than DiffPure and DDA. Finally, we compare the discrepancy in color appearance between the source domain and the distribution of images generated from PiffPure, DDA, and TT-SaD, respectively, in terms of WD. The results show that TT-SaD can fit the distribution of source domain in terms of color appearance much better than DiffPure and DDA. In summary, the above evaluations in terms of SSIM, FID, KID, and WD reveal that TT-SaD exhibits much better stain adaptation than other test-time methods based on diffusion models.

Evaluation of Tumor Classification. We verify stain adaption in tumor classification, as depicted in Fig. 1(f), in three aspects on the CAMELYON17 dataset. First, we compare our method, TT-SaD, with two "training-time" stain augmentation methods, including Stain Mixup [6] and RandStainNA [40]. It can

[3] SSIM evaluates the degradation of structural information in processed images. WD, computed between two one-dimensional discrete distributions, is used to measure the discrepancy between the color appearances of generated and source images. In contrast to WD, FID allows the assessment of not only color but also texture or structure similarity between datasets. KID is an improved version of FID that relaxes the Gaussian assumption.

Table 3: Evaluation of stain adaptation quality on CAMELYON17 dataset.

Target Domain	Metrics	$T = 10$		$T = 100$		$T = 10$
		DiffPure	DDA	DiffPure	DDA	TT-SaD
Hospital 1	SSIM↑	0.174811	0.081332	0.626188	0.662186	**0.912751**
	FID↓	261.017452	349.336953	74.720898	85.566728	**55.795800**
	KID↓	0.348538	0.490933	0.088709	0.104124	**0.062019**
	WD↓	0.005199	0.005396	0.001530	0.001530	**0.000633**
Hospital 2	SSIM↑	0.171325	0.080708	0.621270	0.652846	**0.922293**
	FID↓	267.715239	339.900880	71.464745	72.219387	**60.823601**
	KID↓	0.370639	0.496060	0.084869	0.084191	**0.070901**
	WD↓	0.005139	0.005331	0.001293	0.001706	**0.000890**
Hospital 3	SSIM↑	0.182460	0.086871	0.629850	0.682776	**0.905450**
	FID↓	250.139553	289.847952	120.775310	139.110711	**107.935508**
	KID↓	0.308900	0.366130	0.137194	0.164843	**0.116940**
	WD↓	0.004029	0.004214	0.001058	0.001285	**0.000830**
Hospital 4	SSIM↑	0.176054	0.082357	0.630521	0.672238	**0.920977**
	FID↓	257.771510	339.650079	60.743493	67.347708	**53.115137**
	KID↓	0.347907	0.486585	0.068074	0.077611	**0.058618**
	WD↓	0.005010	0.005220	0.001598	0.001884	**0.001320**
Hospital 5	SSIM↑	0.196455	0.095217	0.644095	0.685063	**0.925894**
	FID↓	222.289948	286.107657	100.581779	108.990788	**78.117246**
	KID↓	0.283990	0.377037	0.112548	0.128344	**0.085603**
	WD↓	0.004679	0.004900	0.000972	0.000940	**0.000840**

seen from Table 4 that TT-SaD can adapt well to the testing data that is out-of-distribution of training data and remarkably outperform Stain Mixup and RandStainNA even our method used fewer training datasets. Please also refer to Table 8 of Sec. A.4 in Appendix for more scenarios.

Second, we verify tumor classification for TT-SaD and other "test-time" adaptation methods, including DDA [14] and DiffPure [35], based on diffusion models. In this case, the source domain is fixed to be "Hospital 1" and other hospitals are target domains. We can observe from Table 5 that our method, TT-SaD, mostly obtain the best results. In addition, our method with the use of domain tumor center is slightly better than the one with domain overall center. Please note that due to the extreme imbalanced nature of the CAMELYON17 dataset, AUPRC is usually quite low for all methods. The visualizations of stain matrix distributions of all hospitals and those after stain shift are, respectively, illustrated in Figs. 6 and 7 of Sec. A.5 in Appendix.

Table 4: Tumor classification comparisons between TT-SaD (with domain tumor center) and other training-time stain augmentation methods. "v" denotes the use of a specific dataset.

	Training Data		Testing Data					
Method	Hospital 1	Hospital 2	Hospital 3	Hospital 4	Hospital 5	ACC	AUROC	AUPRC
w/o any augmentation	v		v	v	v	58.71	77.93	38.70
Stain Mix-Up	v	v	v	v	v	73.47	77.87	43.89
RandStainNA	v	v	v	v	v	75.23	15.86	2.94
TT-SaD (Ours)	v		v	v	v	**81.54**	**89.89**	**44.20**

Table 5: Results of tumor classification for diffusion model-based test-time adaptation methods. The source domain is Hospital 1. "DOC" denotes domain overall center. "DTC" denotes domain tumor center. $T = 10$ for TT-SaD.

Target Domain	Metrics	w/o Adaptation	$T = 10$		$T = 100$		TT-SaD (Ours)	
			DiffPure	DDA	DiffPure	DDA	w/ DOC	w/ DTC
Hospital 2	ACC	74.13	73.81	75.86	81.04	75.35	98.04	**98.84**
	AUROC	93.90	81.70	80.38	**94.86**	94.41	93.30	93.22
	AUPRC	4.41	1.86	1.52	5.90	4.70	7.30	**10.26**
Hospital 3	ACC	71.26	**72.56**	67.24	54.89	69.05	57.84	58.12
	AUROC	82.50	83.63	83.20	80.68	83.40	85.95	**86.58**
	AUPRC	22.19	23.52	22.76	20.27	23.01	29.54	**30.58**
Hospital 4	ACC	97.89	96.52	87.57	98.02	97.87	99.12	**99.35**
	AUROC	**98.81**	98.72	98.27	98.38	98.67	98.69	98.70
	AUPRC	**89.11**	87.37	56.48	81.31	86.93	87.07	88.43
Hospital 5	ACC	6.99	6.93	6.68	10.97	7.51	85.77	**87.16**
	AUROC	52.47	53.37	52.78	56.50	52.82	83.70	**84.41**
	AUPRC	4.80	4.89	4.83	5.23	4.84	13.02	**13.59**
Average	ACC	62.57	62.46	59.34	61.23	62.45	85.19	**85.87**
	AUROC	81.92	79.36	78.66	82.60	82.33	90.41	**90.73**
	AUPRC	30.13	29.41	21.40	28.18	29.87	34.23	**35.71**

Table 6: Results of tumor classification between TT-SaD and other test-time adaptation methods based on diffusion models. This result is the average of the cases that each hospital is in turn set as the source domain and the others play the target domains.

Methods	ACC	AUROC	AUPRC
DiffPure	65.54	71.13	20.17
DDA	64.75	69.35	16.52
TT-SaD (Ours)			
w/ domain overall center	71.55	**75.05**	**26.91**
w/ domain tumor center	**72.06**	74.01	26.82

Finally, we evaluate the scenario that each hospital was set in turn as the source domain and the others play the target domains. The results are averaged and shown in Table 6. It can be observed that our method, TT-SaD, performs better than DiffPure and DDA. In addition, the use of domain overall center and domain tumor center obtain comparable results.

6 Conclusion & Limitation

In this paper, to our knowledge, we are the first to study the issue of test-time stain adaptation with diffusion models (TT-SaD), formulated as an image inverse problem. A unique advantage is that TT-SaD only needs a single domain but can adapt well from other domains without needing their data for training, preventing hospitals from retraining models whenever there are new data available. Limitations of TT-SaD include slow inference and adaptation quality constrained by the performance of diffusion models.

Acknowledgements

This work was supported by the National Science and Technology Council (NSTC), Taiwan, ROC, under Grants NSTC 112-2634-F-006-002 and MOST 110-2221-E-001–020-MY2. We also thank Taiwan Cloud Computing (TWCC) for providing computational and storage resources.

References

1. Bera, K., Schalper, K.A., Rimm, D.L., Velcheti, V., Madabhushi, A.: Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. Nature reviews Clinical oncology **16**(11), 703–715 (2019)
2. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. In: International Conference on Learning Representations (2018)
3. Boyd, K., Eng, K.H., Page, C.D.: Area under the precision-recall curve: point estimates and confidence intervals. In: In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD. pp. 451–466 (2013)
4. Brodersen, K.H., Ong, C.S., Stephan, K.E., , Buhmann, J.M.: The balanced accuracy and its posterior distribution. In: Proceedings of the ICPR. pp. 3121–3124 (2010)
5. Bándi, P., et al., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Ehteshami Bejnordi, B., Lee, B., Paeng, K., Zhong, A., Li, Q., Zanjani, F.G., Zinger, S., Fukuta, K., Komura, D., Ovtcharov, V., Cheng, S., Zeng, S., Thagaard, J., Dahl, A.B., Lin, H., Chen, H., Jacobsson, L., Hedlund, M., Çetin, M., Halıcı, E., Jackson, H., Chen, R., Both, F., Franke, J., Küsters-Vandevelde, H., Vreuls, W., Bult, P., van Ginneken, B., van der Laak, J., Litjens, G.: From detection of individual metastases to classification of lymph node status at the patient level: The camelyon17 challenge. IEEE Transactions on Medical Imaging **38**(2), 550–560 (2019). https://doi.org/10.1109/TMI.2018.2867350
6. Chang, J.R., Wu, M.S., Yu, W.H., Chen, C.C., Yang, C.K., Lin, Y.Y., Yeh, C.Y.: Stain mix-up: Unsupervised domain generalization for histopathology images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 117–126. Springer (2021)
7. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F.: Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155 (2022)
8. Chen, Y., Zee, J., Smith, A., Jayapandian, C., Hodgin, J., Howell, D., Palmer, M., Thomas, D., Cassol, C., Farris III, A.B., et al.: Assessment of a computerized quantitative quality control tool for whole slide images of kidney biopsies. The Journal of pathology **253**(3), 268–278 (2021)
9. Chen, Y.C., Lu, C.S.: Rankmix: Data augmentation for weakly supervised learning of classifying whole slide images with diverse sizes and imbalanced categories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23936–23945 (2023)
10. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14367–14376 (2021)

11. Cong, C., Liu, S., Di Ieva, A., Pagnucco, M., Berkovsky, S., Song, Y.: Colour adaptive generative networks for stain normalisation of histopathology images. Medical Image Analysis **82**, 102580 (2022)
12. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
13. Dimitriou, N., Arandjelović, O., Caie, P.D.: Deep learning for whole slide image analysis: an overview. Frontiers in medicine **6**, 264 (2019)
14. Gao, J., Zhang, J., Liu, X., Darrell, T., Shelhamer, E., Wang, D.: Back to the source: Diffusion-driven adaptation to test-time corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11786–11796 (2023)
15. Gavrilovic, M., Azar, J.C., Lindblad, J., Wählby, C., Bengtsson, E., Busch, C., Carlbom, I.B.: Blind color decomposition of histological images. IEEE Transactions on Medical Imaging **32**(6), 983–994 (2013). https://doi.org/10.1109/TMI.2013.2239655
16. Guo, Z., Liu, H., Ni, H., Wang, X., Su, M., Guo, W., Wang, K., Jiang, T., Qian, Y.: A fast and refined cancer regions segmentation framework in whole-slide breast pathological images. Scientific reports **9**(1), 882 (2019)
17. Hanley, J.A., McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (roc) curve. In: Radiology. pp. 29–36 (1982)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. He, Z., He, J., Ye, J., Shen, Y.: Artifact restoration in histology images with diffusion probabilistic models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 518–527. Springer (2023)
20. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
21. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
22. Humphries, M., Maxwell, P., Salto-Tellez, M.: Qupath: The global impact of an open source digital pathology system. Computational and Structural Biotechnology Journal **19**, 852–859 (2021)
23. Irshad, H., Veillard, A., Roux, L., Racoceanu, D.: Methods for nuclei detection, segmentation, and classification in digital histopathology: A review—current status and future potential. IEEE Reviews in Biomedical Engineering **7**, 97–114 (2014). https://doi.org/10.1109/RBME.2013.2295804
24. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
25. Ke, J., Shen, Y., Liang, X., Shen, D.: Contrastive learning based stain normalization across multiple tumor in histopathology. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 571–580. Springer (2021)
26. Lee, H.Y., Tseng, H.Y., Huang, J.B., Singh, M., Yang, M.H.: Diverse image-to-image translation via disentangled representations. In: Proceedings of the European conference on computer vision (ECCV). pp. 35–51 (2018)

27. Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., Yang, L.: Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7454–7463 (2023)
28. Lin, T., Yu, Z., Hu, H., Xu, Y., Chen, C.W.: Interventional bag multi-instance learning on whole-slide pathological images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19830–19839 (2023)
29. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
30. Lu, M.Y., Chen, B., Zhang, A., Williamson, D.F., Chen, R.J., Ding, T., Le, L.P., Chuang, Y.S., Mahmood, F.: Visual language pretrained multiple instance zero-shot transfer for histopathology images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19764–19775 (2023)
31. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., Van Gool, L.: Repaint: Inpainting using denoising diffusion probabilistic models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11461–11471 (2022)
32. Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Guan, X., Schmitt, C., Thomas, N.E.: A method for normalizing histology slides for quantitative analysis. In: 2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro. pp. 1107–1110 (2009). https://doi.org/10.1109/ISBI.2009.5193250
33. Mahapatra, D., Korevaar, S., Bozorgtabar, B., Tennakoon, R.: Unsupervised domain adaptation using feature disentanglement and gcns for medical image classification. In: European Conference on Computer Vision. pp. 735–748 (2022)
34. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021)
35. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. In: International Conference on Machine Learning. pp. 16805–16827. PMLR (2022)
36. Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., Lawrence, N.D.: Dataset shift in machine learning. MIT Press, Cambridge, MA, USA (2009)
37. Ramdas, A., García Trillos, N., Cuturi, M.: On wasserstein two-sample testing and related families of nonparametric tests. Entropy **19**(2), 47 (2017)
38. Roux, L.: Detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. In: 22nd International Conference on Pattern Recognition, Stockholm, Sweden (2014)
39. Shen, Y., Ke, J.: Staindiff: Transfer stain styles of histology images with denoising diffusion probabilistic models and self-ensemble. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 549–559. Springer (2023)
40. Shen, Y., Luo, Y., Shen, D., Ke, J.: Randstainna: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 212–221. Springer (2022)
41. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
42. Tiard, A., Wong, A., Ho, D.J., Wu, Y., Nof, E., Soatto, S., Nadeem, S.: Stain-invariant self supervised learning for histopathology image analysis. arXiv preprint arXiv:2211.07590 (2022)

43. Vahadane, A., Peng, T., Sethi, A., Albarqouni, S., Wang, L., Baust, M., Steiger, K., Schlitter, A.M., Esposito, I., Navab, N.: Structure-preserving color normalization and sparse stain separation for histological images. IEEE Transactions on Medical Imaging **35**(8), 1962–1971 (2016). https://doi.org/10.1109/TMI.2016.2529665
44. Wagner, S.J., Khalili, N., Sharma, R., Boxberg, M., Marr, C., de Back, W., Peng, T.: Structure-preserving multi-domain stain color augmentation using style-transfer with disentangled representations. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24. pp. 257–266. Springer (2021)
45. Wang, Y., Yu, J., Zhang, J.: Zero-shot image restoration using denoising diffusion null-space model. In: The Eleventh International Conference on Learning Representations (2022)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
47. Xu, C., Wen, Z., Liu, Z., Ye, C.: Improved domain generalization for cell detection in histopathology images via test-time stain augmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 150–159. Springer (2022)
48. Zhang, H., Meng, Y., Zhao, Y., Qiao, Y., Yang, X., Coupland, S.E., Zheng, Y.: Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18802–18812 (2022)
49. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
50. Zhang, Y., Sun, Y., Li, H., Zheng, S., Zhu, C., Yang, L.: Benchmarking the robustness of deep neural networks to common corruptions in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 242–252. Springer (2022)
51. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
52. Zingman, I., Frayle, S., Tankoyeu, I., Sukhanov, S., Heinemann, F.: A comparative evaluation of image-to-image translation methods for stain transfer in histopathology. In: Medical Imaging with Deep Learning. pp. 1509–1525. PMLR (2024)