SGCD: Stain-Guided CycleDiffusion for Unsupervised Domain Adaptation of Histopathology Image Classification

Hsi-Ling Chen
Miin-Wu School of Computing
National Cheng Kung University, Taiwan
nn6114027@gs.ncku.edu.tw

Chun-Shien Lu
Institute of Information Science
Academia Sinica, Taiwan
lcs@iis.sinica.edu.tw

Pau-Choo Chung
Department of Electrical Engineering
National Cheng Kung University & National Chung Cheng University, Taiwan
pcchung@ee.ncku.edu.tw

Abstract

The effectiveness of domain translation in addressing image-based problems of Unsupervised Domain Adaptation (UDA) depends on the quality of the translated images and the preservation of crucial discriminative features. However, achieving high-quality and stable translations typically requires paired data, which poses a challenge in scenarios with limited annotations in the target domain. To address this issue, this paper proposes a novel method termed Stain-Guided Cycle Diffusion (SGCD), employing a dual diffusion model with bidirectional generative constraints to synthesize highly realistic data for downstream task fine-tuning. The bidirectional generative constraints ensure that the translated images retain the features critical to the downstream model in properly controlling the generation process. Additionally, a stain-guided consistency loss is introduced to enhance the denoising capability of the dual diffusion model, thereby improving the quality of images translated between different domains using latents from one domain and a diffusion model trained on another. Experiments conducted on four public datasets demonstrate that SGCD can effectively enhance the performance of downstream task models on the target domain.

1 Introduction

Machine learning is powerful for aiding pathologists in analyzing histopathology slides and diagnosing cancer. However, in medical imaging, models trained on one dataset often struggle to generalize across different hospitals or laboratories due to variations in sample preparation, staining protocols, and digitization processes Howard et al. (2021). These inconsistencies create domain shifts between the training domain (source domain) and real-world application settings (target domain), leading to a drop in model performance. In scenarios where the source domain is fully labeled but the target domain lacks annotations, Unsupervised Domain Adaptation (UDA) Wilson and Cook (2020) seeks to bridge this gap by aligning the distributions of two domains, allowing models trained on the source domain to perform effectively in the target domain.

Traditional stain normalization-based UDA methods Chang et al. (2021); Vahadane et al. (2016); Zhou et al. (2019) align image distributions by decomposing an input image into a stain color matrix and a stain density map, using a reference image's stain color matrix for normalization. However, their

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

performance is highly dependent on selecting an appropriate reference image, which requires domain expertise to ensure it accurately represents the target domain. Moreover, annotating Whole Slide Images (WSIs) is time-consuming and demands expert interpretation, adding complexity to domain adaptation. In histopathology, positive and negative samples often share similar morphological features, making it challenging to distinguish critical diagnostic details. Preserving subtle structural information is crucial for reliable cancer diagnosis, yet it is easily lost during domain adaptation. While generative model-based UDA methods Chang et al. (2021); Figueira et al. (2020); Xing et al. (2019) transform images across domains, they primarily emphasize statistical feature alignment, often at the expense of fine-grained structural details. For instance, STRAP Yamashita et al. (2021) employs AdaIN Huang and Belongie (2017) to normalize feature distributions, and SST Cho et al. (2017) utilizes Kullback-Leibler divergence for feature alignment. However, according to Khamankar et al. (2023), these techniques tend to overlook structural integrity, which is crucial for accurate diagnosis in histopathology.

While contrastive (*e.g.*, CluSiam Wu et al. (2023)) and continual learning (*e.g.*, ConSlide Huang et al. (2023)) enhance feature representations using unlabeled data, they do not directly tackle domain discrepancies. GAN-based methods address this by generating realistic samples to align source and target domains, thereby reducing domain shifts Chiou et al. (2024). Dual consistency models like HistAuGAN Wagner et al. (2021) and ContriMix Nguyen et al. (2024) further enhance alignment by extracting domain-invariant content through encoder-decoder designs. However, this architectural dependence limits their ability to disentangle domain-specific and pathology-relevant features Li et al. (2023b). For instance, MultiPathGAN Nazki et al. (2023) shows that while high-level structures can be modeled, semantic alignment remains a challenge. ContriMix's reliance on accurate content and attribute encoders also constrains its adaptation performance Nguyen et al. (2024). Other approaches, such as Region-Guided CycleGAN Boyd et al. (2022) and CAGAN Cong et al. (2022), utilize ROI localization or histogram loss but are sensitive to ROI accuracy or reference quality. Additionally, GANs commonly suffer from mode collapse, limiting sample diversity and their domain adaptation efficacy.

Thus, diffusion models Ho et al. (2020) have emerged as a promising alternative to GANs for image translation in UDA problems, offering more stable and controlled training while improving diversity. While proper diffusion modeling requires paired data to ensure reliable domain transformation through direct supervision—enabling the model to learn exact correspondences between the source and target domains—such data are often extremely difficult to obtain in real-world scenarios, particularly in the medical domain. To address this limitation, our study proposes a Stain-Guided CycleDiffusion (SGCD) architecture with bi-directional generation constraints to synthesize highly realistic data for downstream task fine-tuning. The dual-diffusion model is based on the stain-based conditional constraints and semantic constraints, which allows the semantic information of the predicted images to be refined backward and forward from the initial generation step, ensuring that important discriminative features are preserved in the generated images, and thus achieving higher UDA performance. Meanwhile, the stain-guided consistency loss is also proposed, which can enhance the denoising ability of the dual-diffusion model in the domain translation.

The contributions of this study include:

- The proposed SGCD is a dual diffusion framework with bidirectional generative constraints that preserves semantic information during domain translation to enhance downstream task performance in the target domain.
- The stain-guided consistency loss mitigates the reliance on paired data, thereby improving the model's applicability in real-world scenarios.
- The results obtained on four public pathology test sets show that SGCD can generate higher-quality images, which further ensures the performance of downstream task models on the target domain.

2 Related Work

This section reviews the related work on three key approaches underlying the proposed SGCD method. In Table 1, we compare SGCD with existing stain UDA methods based on various aspects, including whether they require paired training data or specific reference images for adaptation, rely on auxiliary

Method	Input Image	Generative Model	Paired Data or Reference Image	Handling Less Heterogeneity	Auxiliary Models or Data
Vahadane Vahadane et al. (2016)	WSI	X	V	V	X
Stain mix-up Chang et al. (2021)	WSI	X	X	V	X
StainNet Kang et al. (2021)	WSI	GAN	X	X	X
StainDiffShen and Ke (2023)	WSI	Diffusion Model	X	V	V
BBDM Li et al. (2023a)	Natural Image	Diffusion Model	V	X	X
A-Bridge WANG et al. (2024)	Natural Image	Diffusion Model	V	X	X
HistAuGAN Wagner et al. (2021)	WSI	GAN	X	V	V
G-SAN Li et al. (2023b)	WSI	GAN	X	V	X
STRAP Yamashita et al. (2021)	WSI	X	X	V	X
Ours (SGCD)	WSI	Diffusion Model	X	V	X

Table 1: Comparision of different stain UDA methods. SGCD does not require specific reference images and can be directly applied to the target domain without the need for image normalization.

models or incorporate additional input information, and are capable of handling the less heterogeneity of medical images.

2.1 Stain Normalization

When scanning histologically stained tissue samples, a histopathology image $x \in \mathbb{R}^{3\times n}$ with n pixels in RGB space is converted to its relative optical density via the Beer-Lambert (BL) law Gavrilovic et al. (2013): $BL(x) = -\log \frac{x}{I_0} = WH$, where I_0 is the illumination intensity (255 for 8-bit images), and $W \in \mathbb{R}^{3\times s}$ and $H \in \mathbb{R}^{s\times n}$ represent the stain color matrix and stain density map, respectively, for s stains. BL law supports stain normalization by reconstructing a target image using the source's stain density and the target's stain color. However, relying on a single reference image Chang et al. (2021); Rabinovich et al. (2003); Vahadane et al. (2016) may introduce color artifacts due to staining and digitization variability. ContriMix Nguyen et al. (2024) builds on this with optical-style transfer to synthesize images for domain adaptation, but its performance is limited by encoder accuracy and the difficulty of designing content and attribute encoders for diverse datasets.

2.2 Generative Adversarial Network (GAN)

Numerous GAN-based approaches have been developed for UDA in histopathology Vasiljević et al. (2023); Guan et al. (2024); Nazki et al. (2023); Wagner et al. (2021). Similar to stain normalization methods Hetz et al. (2024); Salehi and Chalechale (2020); Nishar et al. (2020), these approaches often convert target domain images into the source domain to enable direct application of source-trained models. StainGAN Shaban et al. (2019) first adopted a CycleGAN-based architecture Zhu et al. (2017) for stain normalization, while StainNet Kang et al. (2021) enhanced performance and efficiency via knowledge distillation using StainGAN outputs. Alternatively, model generalization techniques Figueira et al. (2020); Xing et al. (2019) transform annotated source images into the target domain for training. HistAuGAN Wagner et al. (2021) disentangles content and style to manipulate color properties, but despite producing realistic structures, such GANs often struggle with semantic consistency Nazki et al. (2023).

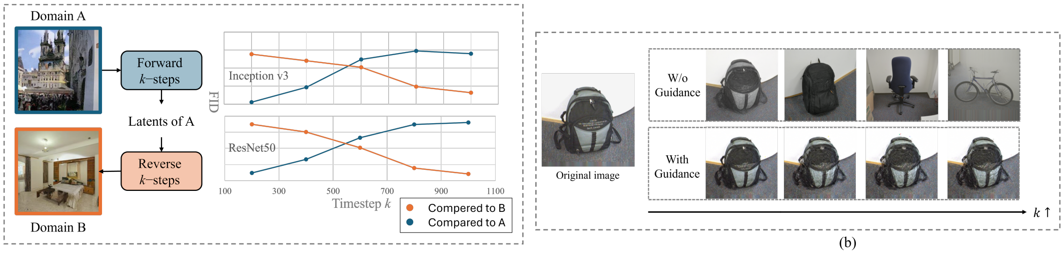
2.3 Denoising Diffusion Probabilistic Models (DDPM)

Denoising Diffusion Probabilistic Models (DDPM) Ho et al. (2020) consist of a forward and a reverse phase of small, reversible transformations. In the forward phase, noise is progressively added to the input image until it approximates a normal distribution $\mathcal{N}(0, \mathbf{I})$. Let x_t be the latent at step t and let D be the diffusion model. The forward process is defined as:

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

where β_t^D denotes the noise schedule. The reverse process gradually removes the noise to recover the original data, modeled as:

$$p(x_{t-1}|x_t, D) = \\ \mathcal{N}\left(x_{t-1}; \frac{1}{\sqrt{1-\beta_t^D}}\left(x_t - \frac{\beta_t^D}{\sqrt{1-\bar{\beta}_t^D}} D(x_t, t)\right), \beta_t^D \mathbf{I}\right), \tag{2}$$



(a)

(b)

Figure 1: Illustrative example. (a) Similarity of the generated image with domain A and domain B, respectively. After adding a specific degree of noise to the image from domain A, the reverse process is performed via a diffusion model trained in domain B. The generated images are more similar to domain B when more noise is added. For less noise, it is more similar to domain A. The results show the same trend when using Inception V3 and ResNet50 as feature extractors for FID metrics. (b) The reverse process guided by the additional condition preserves the desired categorical features regardless of the increase in added noise.

with $\bar{\beta}_t^D = \prod_{s=1}^{t}(1 - \beta_s^D)$. Recent works such as BBDM Li et al. (2023a) and A-Bridge WANG et al. (2024) enable image-to-image translation by modifying the noise schedule. SynDiff Özbey et al. (2023) incorporates paired generators and discriminators into the reverse process for source-target domain translation with large-step sampling. StainDiff Shen and Ke (2023) further adapts diffusion for stain normalization in histopathology images, but its reliance on auxiliary networks to preserve fine structural details may hinder robustness in handling rare or subtle patterns.

3 Preliminary and Motivation

Let D_A be the diffusion model trained on domain A and let k be the given timestep. Referring to Eq. (1), the forward process for an initial image x_0 is defined as: $f_{D_A}(x_0, k) = \prod_{t=1}^{k} q(x_t|x_{t-1}, D_A)$. From Eq. (2), the reverse process for diffusion model D_A and a noisy image x_k can be defined as: $r_{D_A}(x_k, k) = p(x_k)\prod_{t=1}^{k} p(x_{t-1}|x_t, D_A)$. Two experiments were performed to justify the motivation and intuition underlying the proposed SGCD method: (1) An investigation into the relationship between the images generated by the diffusion model and the actual target domain images; and (2) A demonstration of the use of additional constraints to ensure that the generated images retain specific, important features.

3.1 Similarity of Generation Distributions to Target Distributions

As previously described, the diffusion model's forward process adds noise to input images, while the reverse process removes it to reconstruct the data. This enables domain-specific image generation by applying the reverse process to noise using a model trained on the target domain Su et al. (2022). However, it remains unclear whether a latent from domain A can yield similar results when denoised by a model trained on domain B. To examine this, we used two public diffusion models from Google Google (2022a,b), trained on the LSUN bedroom and church datasets Yu et al. (2015), representing domains A and B, respectively (denoted D_A and D_B). Images $x_A \sim A$ were corrupted at various noise levels k and then denoised using D_B, i.e., $r_{D_B}(f_{D_A}(x_A, k), k)$. As shown in Figure 1(a), FID scores Heusel et al. (2017) computed with Inception v3 Szegedy et al. (2016) and ResNet50 He et al. (2016) reveal that higher noise levels led to outputs resembling domain B, but at the cost of losing key characteristics of the original domain A images.

3.2 Reverse Process Guided by Conditions

In diffusion models, conditional constraints can retain critical information during forward and reverse processes. For instance, Gao *et al.* Gao et al. (2023) applied low-pass filtering to preserve image outlines throughout denoising, enabling corrupted categories to be inferred from restored images. To explore this mechanism further, we conducted a second experiment using the Office31 dataset Saenko et al. (2010), which includes three domains and 31 categories. Specifically, the Amazon (domain

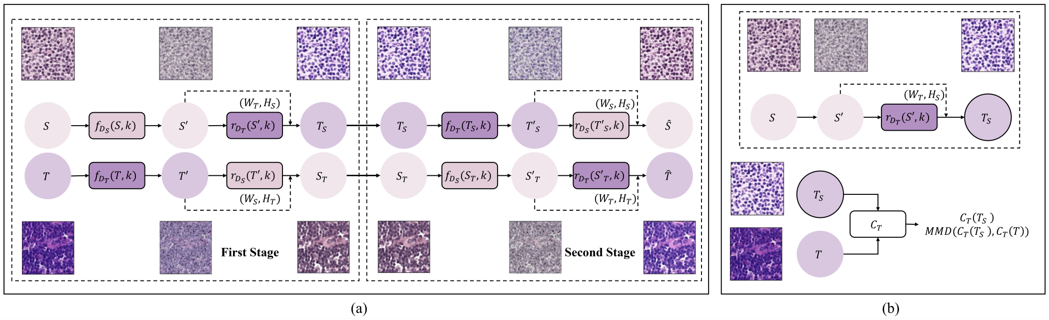


Figure 2: Overview of the proposed method. (a) shows the two-stage conversion of proposed SGCD architecture. (b) shows the training flow of target classifier C_T, where both the annotated image converted by S and the target image are used to fine-tune the classifier to be applied to the target domain.

A) and Webcam (domain B) domains were selected to observe changes in categorical features. Algorithm 1 in Appendix A.1 describes the condition-guided reverse process. Unlike Gao et al. (2023), we use a Canny edge detector Canny (1986) as $\phi(\cdot)$, guiding $\hat{x}_{t-1}$ along the gradient minimizing the difference between $\phi(\hat{x}_0)$ and $\phi(x_0^A)$, thus preserving texture features. As shown in Figure 1(b), the unconditional case yields textures unrelated to the original image, while conditional guidance ensures generated features resemble the source.

4 Proposed Method

4.1 Stain-Guided CycleDiffusion (SGCD)

Figure 2 illustrates the basic structure of the proposed SGCG method. Based on the similarity results shown in Figure 1(a), a dual diffusion model pre-trained in the source and target domains is utilized to perform the reverse process, effectively converting the input images into their corresponding domains. However, such a transformation does not guarantee the preservation of key features. In SGCD, this issue is addressed by a dual diffusion model combined with bidirectional constraints and a stain-guided consistency loss.

To better simulate the target domain and improve the performance of the downstream task model, a two-stage conversion cyclic framework is utilized to train the dual diffusion model. Without this cyclic framework, the diffusion model would merely convert data from one domain to another with no additional control, potentially leading to inconsistencies and loss of important features. However, the proposed cyclic framework allows the final reconstruction results to be incorporated as additional constraints, thereby enhancing the consistency between the translations from domain A to domain B and domain B to domain A, respectively. As a result, the model's ability to generate realistic target images is improved. For a given diffusion model D and a latent x_t at time t, an estimate image at step 0 is obtained by: $\hat{x}_0(x_t) = a_t \cdot x_t - b_t \cdot D(x_t, t)$, where $a_t = \sqrt{1/\bar{\beta}_t^D}$ and $b_t = \sqrt{1/\bar{\beta}_t^D - 1}$.

And given a reference image I_{ref}, the stain-guided process is:

$$\begin{aligned} G(x_t, I_{\text{ref}}) &= \hat{x}_{t-1} - \nabla_{x_t} \|\hat{x}_0(x_t) - I_{\text{ref}}\| \\ &\text{for } \hat{x}_{t-1} \sim p_(x_{t-1}|x_t, D), \end{aligned} \tag{3}$$

which can be iterated by: $x_t^{(n+1)} = G(x_t^{(n)}, I_{\text{ref}}), \quad x_t^{(0)} = x_t.$

In Eq. (3), image $\hat{x}_{t-1}$ is moved along the gradient that minimizes the distance between $\hat{x}_0(x_t)$ and I_{ref}. In the first stage, the diffusion model D_T trained using the target domain T converts the source images in S to a set of target-style images, denoted as T_S, for a given timestep k (*i.e.*, $S \rightarrow T$). Meanwhile, the diffusion model D_S in source domain S converts the target images in T to a set of source-style images, denoted as S_T, for a given timestep k (*i.e.*, $T \rightarrow S$). The first stage can be summarized as:

$$S' = \{f_{D_S}(x,k) \mid x \in S\} \quad T' = \{f_{D_T}(x,k) \mid x \in T\}, \tag{4}$$

$$T_S = \{r_{D_T}(x^{(k_G)}, k') \mid x \in S'\}, \quad S_T = \{r_{D_S}(x^{(k_G)}, k') \mid x \in T'\}, \quad k' = k - k_G. \tag{5}$$

Eq. (4) denotes the forward noise addition processes on S and T, respectively. In Eq. (5), T_S denotes the target-style image generated from the source image and S_T denotes the source-style image produced from the target image. T_S uses a reference image constructed using the target stain color matrix W_T and the source stain density map H_S, while S_T uses a reference image constructed using the source stain color matrix W_S and the target stain density map H_T, respectively.

The second stage transforms the outputs of the first stage back to the original source and target domains (*i.e.*, $S \to T \to S$ and $T \to S \to T$). It is formulated as:

$$T'_S = \{f_{D_T}(x, k) \mid x \in T_S\}, \quad S'_T = \{f_{D_S}(x, k) \mid x \in S_T\}, \tag{6}$$

$$\hat{S} = \{r_{D_S}(x^{(k_G)}, k') \mid x \in T'_S\}, \quad \hat{T} = \{r_{D_T}(x^{(k_G)}, k') \mid x \in S'_T\}, \quad k' = k - k_G. \tag{7}$$

Eq. (6) denotes the forward noise addition processes on T_S and S_T, respectively. In Eq. (7), $\hat{S}$ denotes the source image reconstructed from the target-style image and $\hat{T}$ denotes the target image reconstructed from the source-style image. $\hat{S}$ uses a reference image constructed using the source stain color matrix W_S and the source stain density map H_S, while $\hat{T}$ uses a reference image constructed using the target stain color matrix W_T and the target stain density map H_T, respectively.

4.2 Training of Dual Diffusion Model

Recall that the results of the second experiment (Figure 1(b)) show that the use of only a diffusion model to convert images from one domain to another may result in the loss of important features. To effectively realize the conversion between different domains, while simultaneously ensuring that the detailed information in the pathological images is preserved during the conversion process, SGCD utilizes bidirectional constraints and stain-guided consistency (SGC) loss to enforce the diffusion model's generative process in both forward and backward directions.

A stain-guided constraint is applied at each reverse step from k to k_G, where k_G is the hyperparameter to control the range of stain-guide constraint. Specifically, for the route $S \to T_S$ in Figure 2(a)), the reference image $I^{T_S}_{ref}$ is used to guide the generation of T_S. Each step in the reverse process is moved along the gradient that is close to the reference image, ensuring that the final converted image at step 0 is as similar as possible to the reference image. An analogous procedure is employed for the route $T \to S_T \to \hat{T}$ and $T_S \to \hat{S}$. The detailed steps of the $S \to T \to S$ conversion process are shown in Algorithm 2 in the Appendix.

Let C_S and C_T be the source and target classifier, both pre-trained on S. Task constraints $-\sum y \cdot C_S(\hat{S})$ and $-\sum y \cdot C_T(T_S)$ are applied to preserve the crucial feature information required for downstream tasks from step 0 to k, thereby enabling the downstream model to produce consistent results. Additionally, to ensure the latents from the source (target) domain can be converted into the target (source) domain, a consistency constraint is imposed on the guided reconstructed images $\hat{S}$ and $\hat{T}$ for further improving the quality of the converted images. Therefore, the Stain-Guided Consistency (SGC) loss is defined as:

$$\begin{aligned} \text{loss}_{SGC} = \|S - \hat{S}\|_2 - \sum y \cdot C_S(\hat{S}) + \|T - \hat{T}\|_2 \\ - \sum y \cdot C_T(T_S). \end{aligned} \tag{8}$$

Since the pre-trained diffusion model is capable of generating images corresponding to the training domain directly, the guiding processes in $S \to T \to S$ and $T \to S \to T$ do not require paired images or specified reference images. Thus, our method does not require specific reference images and paired data, making it more adaptable to a wider variety of applications, as described in Table 1. The two-stage conversion process yields complete $S \to S$ and $T \to T$ cycles. Thus, the round-trip cyclic process can be used to fine-tune pre-trained diffusion models, enabling them to generate images with distributions similar to the training domain, even when the input is perturbed (*i.e.*, different from the training domain). This then allows the converted source domain images to be used to train downstream task models. A more detailed discussion of the task losses is provided in the next section.

4.3 Training Strategy

An alternating training approach is used to update the diffusion models and classifiers iteratively. C_S is fixed during all the training phases to force the diffusion model to produce the correct image during the training phase. C_T is the desired target model, whose classifier head will be fine-tuned through the alternating training. In particular, classifiers C_S and C_T are first fixed, and the two-stage conversion process introduced in the previous section is performed using Eq. (8) to fine-tune the diffusion models with the reconstructed images in $\hat{S}$ and $\hat{T}$, and the target-style images in T_S. In the next step, the two diffusion models are fixed and the classifier C_T is trained using the task loss in Eq. (9). The generalization ability of the classifier C_T is gradually enhanced using the images converted by the boosted diffusion model and source images with annotations such that it can progressively adapt to operating in the target domain. Furthermore, given the availability of unlabeled target domain images, the maximum mean discrepancy (MMD) Gretton et al. (2012) loss is additionally employed to reduce the distribution distance between the converted images and the real target images. Thus, the task loss is defined as:

$$\text{loss}_{task} = -\sum y \cdot C_T(T_s) + \text{loss}_{\text{MMD}}(C_T(T_s), C_T(T)), \tag{9}$$

where loss_{MMD} represents the MMD loss, which is used to measure the distance between the two embedding feature distributions. Given the annotated target-style images T_s, the cross entropy loss is used to fine-tune the target classifier C_T. The training process of C_T is shown in Figure 2(b).

5 Experiment Results

5.1 Datasets

SGCD was evaluated on four open datasets: Camelyon17 Bejnordi et al. (2017), Camelyon16 Bejnordi et al. (2017), Camelyon17-WILDS Koh et al. (2021), and MITOS & ATYPIA14 Racoceanu et al. (2014). The details of the four datasets are provided in Sec. A.2 of the Appendix.

5.2 Setting

The experiments were implemented on NVIDIA V100 GPU with Python 3.10.12 and Pytorch 2.4.0. The Adam optimizer was employed with a learning rate of $2e-4$ and batch size of 4. The total timestep T of the diffusion models was set to 1000. Stain guidance was applied from timestep 600 to 100. The remaining timesteps used the standard reverse diffusion process in Eq. (2).

For the balanced dataset Camelyon17-WILDS, following the WILDS benchmark, DenseNet121 Huang et al. (2017) was used as the backbone of classifiers C_S and C_T, and the models were evaluated using the average accuracy. For the imbalanced datasets Camelyon16 and Camelyon17, ResNet50 He et al. (2016) was used as the backbone of the classifiers, and the Area Under the Curve (AUC) was adopted as the evaluation metric. For the MITOS & ATYPIA14 dataset, visualizations of the generated images were provided, and their quality was evaluated using the SSIM Wang et al. (2004) and PSNR metrics.

5.3 Comparison of General UDA Methods

To investigate the distinction between the medical image-specific UDA methods and general UDA methods, experiments were conducted on Camelyon17-WILDS. Table 2 presents the comparison results. Among the comparison methods, Connect Later Qu and Xie (2024) and SwAV Caron et al. (2020) were initially trained on the target domain to enhance their clustering capability inside it, followed by fine-tuning on the labeled source domain. Regarding the difference between the various comparison models, Connect Later simulates the target data by augmenting the source data, while AFN Xu et al. (2019) aims to achieve domain invariance between the source and target domains. Simprov Tahir et al. (2022) adopts knowledge distillation to enable the student model to adapt to the target domain. RLSbench Garg et al. (2023) refines the estimation of the target domain distribution, making it more closely aligned with that of the actual target. Designed specifically for medical imaging, ContriMix Nguyen et al. (2024), STRAP Yamashita et al. (2021), and our SGCD demonstrate better performance than general UDA approaches. Nevertheless, Connect Later, through its tailored augmentation approach and subsequent model fine-tuning, demonstrates a marginally

Method	Test ACC	Test AUC
Connect Later Qu and Xie (2024)	95.0	98.7
SwAV Caron et al. (2020)	91.4	95.2
AFN Xu et al. (2019)	83.2	91.3
Simprov Tahir et al. (2022)	92.8	-
RLSbench Garg et al. (2023)	86.8	-
STRAP Yamashita et al. (2021)	93.7	98.1
ContriMix Nguyen et al. (2024)	94.6	-
Ours (SGCD)	94.7	98.6

Table 2: Histopathology classification results for Camelyon17-WILDS.

superior performance as a result of an enhanced feature-level alignment between the source and target domains.

5.4 Histopathology Classification

For histopathology classification, Vahadane et al. (2016), Macenko et al. (2009), and Reinhard et al. (2001) are the classical stain normalization methods, while Stain Mix-Up Chang et al. (2021) uses stain-normalized images as augmented data to train the classifiers for improved generalization. StainNet Kang et al. (2021) and MultiPathGAN Nazki et al. (2023) utilize GANs for stain normalization, further enhancing the image quality. BCD-net Yang et al. (2023) estimates more accurate color matrices and stain density maps using two models, leading to improved stain normalization results. SPA Xiao et al. (2024), an advanced UDA method for general images, enhances in-domain classification and cross-domain alignment using latent feature matching.

Table 3 presents the classification results. It is observed that the traditional stain normalization methods exhibit a poorer performance and are susceptible to the influence of the reference images, resulting in a less stable performance. StainNet and MultiPathGAN, benefiting from the excellent image generation capabilities of GAN architectures, achieve promising results on many domains. G-SAN Li et al. (2023b) improves the feature alignment in GAN to further enhance the classification accuracy. HistAuGAN Wagner et al. (2021) and ContriMix Nguyen et al. (2024) are both augmentation methods but are inherently constrained by the diversity of input data or the availability of source-domain samples, leading to performance discrepancies when encountering unseen data. Stain Mix-Up enhances model generalization by using augmented data, but perturbed data in highly similar domains may lead the model to deviate from the target domain. Connect Later performs better in the balanced dataset, Camelyon17-WILDS, than the imbalanced dataset, Camelyon17, primarily due to its sensitivity to augmentation hyperparameters. BCD-Net, which focuses on solving blind color deconvolution problems for histological images, and SPA, which employs latent feature matching, can both preserve more critical class information in the images, and thus yield better performance. A more thorough evaluation is presented in Sec. A.3 of the Appendix.

Table 4 presents the classification results for Camelyon16. In comparison to traditional stain normalization and GAN-based methods (*e.g.*, StainGAN and StainNet), SGCD exhibits a higher AUC score.

5.5 Ablation Studies

The visual results of various stain transfer methods are presented and discussed in Sec. A.4 of the Appendix. In addition, the validation of proposed two-step conversion process is examined in Sec. A.5 of the Appendix.

5.6 Remarks on Stability, Reproducibility, and Generalizability

SGCD aims to enable diffusion models to accept specific images as input instead of noise. This adaptation is accomplished through fine-tuning with consistency constraints, avoiding the need for a complex training framework.

Method	C_2	C_3	C_4	C_5	Average
No adaptation	83.8	64.5	85.0	73.6	76.7
Vahadane *et al.* Vahadane et al. (2016)	79.5	88.1	86.4	67.3	80.3
Mackenko *et al.* Macenko et al. (2009)	63.1	86.9	71.8	78.8	75.2
Reinhard *et al.* Reinhard et al. (2001)	82.9	85.9	81.6	88.6	84.8
Stain Mix-Up Chang et al. (2021)	87.2	82.6	86.9	68.3	81.3
StainNet Kang et al. (2021)	83.6	89.5	86.2	87.7	86.8
MultiPathGAN Nazki et al. (2023)	85.0	69.8	90.7	80.3	81.5
BCD-net Yang et al. (2023)	89.0	92.4	91.8	87.9	90.3
Connect Later Qu and Xie (2024)	88.9	82.3	93.0	84.1	87.1
SPA Xiao et al. (2024)	88.7	92.3	94.7	92.7	92.1
HistAuGAN Wagner et al. (2021)	90.5	90.3	91.9	85.0	89.4
G-SAN Li et al. (2023b)	87.9	84.7	92.7	82.5	87.0
ContriMix Nguyen et al. (2024)	89.0	90.3	92.0	88.5	90.0
Ours (SGCD)	89.1	94.9	98.1	93.9	**94.0**

Table 3: Histopathology classification results for Camelyon17 under the condition that C_1 is the source domain and others are regarded as the target domain individually. Here, AUC (%) was adopted as the evaluation metric.

Method	Test AUC (%)
No Adaptation	75.9
Reinhard *et al.* Reinhard et al. (2001)	89.3
Mackenko *et al.* Macenko et al. (2009)	90.3
Vahadane *et al.* Vahadane et al. (2016)	88.2
StainGAN Shaban et al. (2019)	90.5
StainNet Kang et al. (2021)	93.5
Ours (SGCD)	**95.8**

Table 4: Histopathology classification for Camelyon16.

- Stability: SGCD fine-tunes a pre-trained diffusion model using consistency constraints to guide the adaptation process. Since the process involves only fine-tuning, it is inherently stable.
- Reproducibility: The fine-tuning process involves only one hyperparameter, *i.e.*, the timestep k introduced in Sec. 4.2. In addition, the proposed two-step approach preserves both structural integrity and distribution consistency in the S→T and T→S transformations, as validated in Sec. A.5 of Appendix.
- Generalizability: The consistency constraints allow SGCD to generalize effectively across diverse pathological domains, as shown in Tables 2∼ 5 for images from different staining protocols and Table 6 and Figure 5 of the appendix for images from diverse scanners.

6 Conclusion

An innovative stain-guided cyclic diffusion (SGCD) model has been proposed to effectively solve the problem of model performance degradation caused by domain distribution differences in histopathology images. SGCD consists of: (1) bidirectional generative constraints to maintain feature consistency, (2) a SGC loss to improve the quality the synthesized images, and (3) high-quality target domain synthesized images that preserve crucial discriminative features and enhance the generalization ability of downstream task models. The experimental results have confirmed the superiority of SGCD for adaptive tasks in the pathology image domain.

Limitations. We acknowledge that the cyclic bi-directional training, while crucial for maintaining semantic integrity without paired data, introduces additional computational demands compared to traditional UDA techniques. For more efficient deployment in practice, future efforts will focus on optimizing sampling schedules to reduce inference steps, and exploring smaller, more efficient diffusion model architectures.

Future Work. Our initial focus on binary classification within H&E staining establishes foundational efficacy. Future studies will explore applying SGCD to more complex scenarios, including multi-class classification and adaptation between entirely different staining protocols (e.g., H&E to immunohistochemistry). We plan to tackle highly challenging cross-organ domain adaptation tasks and generalize the methodology to other medical imaging domains, such as immunology problems, where domain heterogeneity is a significant challenge.

7 Acknowledgement

This work was supported by the National Science and Technology Council (NSTC) with Grant NSTC 114-2634-F-006-002. We thank to National Center for High- erformance Computing (NCHC) for providing computational and storage resources.

References

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.

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

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

Jia-Ren Chang, Min-Sheng Wu, Wei-Hsiang Yu, Chi-Chung Chen, Cheng-Kung Yang, Yen-Yu Lin, and Chao-Yuan Yeh. Stain mix-up: Unsupervised domain generalization for histopathology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 117–126, 2021.

Chien-Yu Chiou, HUNG-WEN TSAF, WEI-JONG YAN, MING TING St, MARIA GABRANIN KUO-SHENG CHENCP, MENG-LING WU, PAU-CHOO CHUNG, et al. Acc-gan: Cross scanner robustness with annotation consistency guided cycle-gan. *Journal of Information Science & Engineering*, 40(3), 2024.

Hyungjoo Cho, Sungbin Lim, Gunho Choi, and Hyunseok Min. Neural stain-style transfer learning using gan for histopathological images. *arXiv preprint arXiv:1710.08543*, 2017.

Cong Cong, Sidong Liu, Antonio Di Ieva, Maurice Pagnucco, Shlomo Berkovsky, and Yang Song. Colour adaptive generative networks for stain normalisation of histopathology images. *Medical Image Analysis*, 82: 102580, 2022.

Gonçalo Figueira, Yaqi Wang, Lingling Sun, Huiyu Zhou, and Qianni Zhang. Adversarial-based domain adaptation networks for unsupervised tumour detection in histopathology. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 1284–1288. IEEE, 2020.

Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven test-time adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Saurabh Garg, Nick Erickson, James Sharpnack, Alex Smola, Sivaraman Balakrishnan, and Zachary Chase Lipton. Rlsbench: Domain adaptation under relaxed label shift. In *International Conference on Machine Learning*, pages 10879–10928. PMLR, 2023.

Milan Gavrilovic, Jimmy C Azar, Joakim Lindblad, Carolina Wählby, Ewert Bengtsson, Christer Busch, and Ingrid B Carlbom. Blind color decomposition of histological images. *IEEE transactions on medical imaging*, 32(6):983–994, 2013.

Google. DDPM for bedroom image generation. `https://huggingface.co/google/ddpm-bedroom-256`, 2022a. [Online; accessed 01-July-2024].

Google. DDPM for church image generation. `https://huggingface.co/google/ddpm-ema-church-256`, 2022b. [Online; accessed 01-July-2024].

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

Xianchao Guan, Yifeng Wang, Yiyang Lin, Xi Li, and Yongbing Zhang. Unsupervised multi-domain progressive stain transfer guided by style encoding dictionary. *IEEE Transactions on Image Processing*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Martin J Hetz, Tabea-Clara Bucher, and Titus J Brinker. Multi-domain stain normalization for digital pathology: A cycle-consistent adversarial network for whole slide images. *Medical Image Analysis*, 94:103149, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017.

Yanyan Huang, Weiqin Zhao, Shujun Wang, Yu Fu, Yuming Jiang, and Lequan Yu. Conslide: Asynchronous hierarchical interaction transformer with breakup-reorganize rehearsal for continual whole slide image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21349–21360, 2023.

Hongtao Kang, Die Luo, Weihua Feng, Shaoqun Zeng, Tingwei Quan, Junbo Hu, and Xiuli Liu. Stainnet: a fast and robust stain normalization network. *Frontiers in Medicine*, 8:746307, 2021.

Vaibhav Khamankar, Sutanu Bera, Saumik Bhattacharya, Debashis Sen, and Prabir Kumar Biswas. Histopathological image analysis with style-augmented feature domain mixing for improved generalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 285–294. Springer, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023a.

Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. A laplacian pyramid based generative h&e stain augmentation network. *IEEE Transactions on Medical Imaging*, 2023b.

Marc Macenko, Marc Niethammer, James S Marron, David Borland, John T Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE international symposium on biomedical imaging: from nano to macro*, pages 1107–1110. IEEE, 2009.

Haseeb Nazki, Ognjen Arandjelovic, In Hwa Um, and David Harrison. Multipathgan: Structure preserving stain normalization using unsupervised multi-domain adversarial network with perception loss. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, pages 1197–1204, 2023.

Tan H Nguyen, Dinkar Juyal, Jin Li, Aaditya Prakash, Shima Nofallah, Chintan Shah, Sai Chowdary Gullapally, Limin Yu, Michael Griffin, Anand Sampat, John Abel, Justin Lee, and Amaro Taylor-Weiner. Contrimix: Scalable stain color augmentation for domain generalization without domain labels in digital pathology. In *MICCAI Workshop on Computational Pathology with Multimodal Data (COMPAYL)*, 2024.

Harshal Nishar, Nikhil Chavanke, and Nitin Singhal. Histopathological stain transfer using style transfer network with adversarial loss. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 330–340. Springer, 2020.

Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023.

Helen Qu and Sang Michael Xie. Connect later: Improving fine-tuning for robustness with targeted augmentations. In *Forty-first International Conference on Machine Learning*, 2024.

Andrew Rabinovich, Sameer Agarwal, Casey Laris, Jeffrey Price, and Serge Belongie. Unsupervised color decomposition of histologically stained tissue samples. *Advances in neural information processing systems*, 16, 2003.

Daniel Racoceanu, Jessica Calvo, Elham Attieh, Gilles Le Naour, and A. Gloaguen. Detection of mitosis and evaluation of nuclear atypia score in breast cancer histological images. 2014.

Erik Reinhard, Michael Adhikhmin, Bruce Gooch, and Peter Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.

Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 213–226. Springer, 2010.

Pegah Salehi and Abdolah Chalechale. Pix2pix-based stain-to-stain translation: A solution for robust stain normalization in histopathology images analysis. In *2020 International conference on machine vision and image processing (MVIP)*, pages 1–7. IEEE, 2020.

M Tarek Shaban, Christoph Baur, Nassir Navab, and Shadi Albarqouni. Staingan: Stain style transfer for digital histological images. In *2019 Ieee 16th international symposium on biomedical imaging (Isbi 2019)*, pages 953–956. IEEE, 2019.

Yiqing Shen and Jing Ke. 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*, pages 549–559, 2023.

Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *The Eleventh International Conference on Learning Representations*, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Anique Tahir, Lu Cheng, Ruocheng Guo, and Huan Liu. Distributional shift adaptation using domain-specific features. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 5593–5597. IEEE, 2022.

Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE transactions on medical imaging*, 35(8):1962–1971, 2016.

Jelica Vasiljević, Friedrich Feuerhake, Cédric Wemmert, and Thomas Lampert. Histostargan: A unified approach to stain normalisation, stain transfer and stain invariant segmentation in renal histopathology. *Knowledge-Based Systems*, 277:110780, 2023.

Sophia J Wagner, Nadieh Khalili, Raghav Sharma, Melanie Boxberg, Carsten Marr, Walter De Back, and Tingying Peng. 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*, pages 257–266. Springer, 2021.

PEIYONG WANG, Bohan Xiao, Qisheng He, Carri Glide-Hurst, and Ming Dong. Score-based image-to-image brownian bridge. In *ACM Multimedia 2024*, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46, 2020.

Weiyi Wu, Chongyang Gao, Joseph DiPalma, Soroush Vosoughi, and Saeed Hassanpour. Improving representation learning for histopathologic images with cluster constraints. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21404–21414, 2023.

Zhiqing Xiao, Haobo Wang, Ying Jin, Lei Feng, Gang Chen, Fei Huang, and Junbo Zhao. Spa: a graph spectral alignment perspective for domain adaptation. *Advances in Neural Information Processing Systems*, 36, 2024.

Fuyong Xing, Tell Bennett, and Debashis Ghosh. Adversarial domain adaptation and pseudo-labeling for cross-modality microscopy image quantification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 740–749. Springer, 2019.

Ruijia Xu, Guanbin Li, Jihan Yang, and Liang Lin. Larger norm more transferable: An adaptive feature norm approach for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1426–1435, 2019.

Rikiya Yamashita, Jin Long, Snikitha Banda, Jeanne Shen, and Daniel L. Rubin. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Transactions on Medical Imaging*, 40(12):3945–3954, 2021.

Shuowen Yang, Fernando Pérez-Bueno, Francisco M. Castro-Macías, Rafael Molina, and Aggelos K. Katsaggelos. Deep bayesian blind color deconvolution of histological images. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 710–714, 2023.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Niyun Zhou, De Cai, Xiao Han, and Jianhua Yao. Enhanced cycle-consistent generative adversarial network for color normalization of h&e stained images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 694–702, 2019.

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

A Appendix

A.1 Pseudo Code of Proposed Method-SGCD

Algorithms 1 and 2 show the pseudo-codes of the proposed SGCD method.

Algorithm 1 Reverse Process Guided by Conditions

```
Input: $x_0^A$: Image from $A$, $D_B$: DDPM pretrained on $B$, $\phi(\cdot)$: Canny edge detector
Parameter: Timestep $k$
Output: Converted image $x_B$
$x_k \sim f_{D_A}(x_0^A, k)$ using Eq. (1).
for $t \leftarrow k...1$ do
    $\hat{x}_{t-1} = p(x_{t-1}|x_t, D_B)$ using Eq. (2) .
    $\hat{x}_0 = \sqrt{\frac{1}{\alpha_t}} \cdot x_t - \sqrt{\frac{1}{\alpha_t} - 1} \cdot D_B(x_t, t)$.
    $x_{t-1} = \hat{x}_{t-1} - \nabla_{x_t} \|\phi(\hat{x}_0) - \phi(x_0^A)\|$.
end for
return $x_B = x_0$
```

Algorithm 2 $S \rightarrow T \rightarrow S$ conversion of SGCD

```
Input: $x_0^S$: Image from $S$, $x_0^T$: Image from $T$, $D_S$: DDPM pretrained on $S$, $D_T$: DDPM pretrained on $T$
Parameter: Timestep $k$, Guide range $k_G$
Output: Converted image $x_S$
 1: Compute $W_T$ and, $H_T$ from $x_0^T$, and $W_S$ and $H_S$ from $x_0^S$, respectively, using BL law in 2.1.
 2: Get noisy image $x_k \sim f_{D_S}(x_0^S, k)$ using Eq. (1).
 3: Get reference image $I_{ref} = I_0 \exp(-W_T H_S)$.
 4: for $t \leftarrow k...k_G$ do
 5:     $x_{t-1} = G_{D_T}(x_t, I_{ref})$ using Eq. (5).
 6: end for
 7: Initialization for the next stage $x_0^{T_S} = r_{D_T}(x_{k_G}, k_G)$.
 8: Get noisy image $x_k \sim f_{D_T}(x_0^{T_S}, k)$ using Eq. (1).
 9: Get reference image $I_{ref} = I_0 \exp(-W_S H_S)$.
10: for $t \leftarrow k...k_G$ do
11:     $x_{t-1} = G_{D_S}(x_t, I_{ref})$ using Eq. (7).
12: end for
13: return $x_S = r_{D_S}(x_{k_G}, k_G)$
```

In Algorithm 2, Lines 2 and 8 represent the forward processes on S and T, respectively. Lines 4$\sim$6 represent the Stain-Guided reverse process on T, and Lines 10$\sim$12 represent the Stain-Guided reverse process on S. Lines 1 derives stain-guided reference images from BL law in Sec. 2.1, which are used to guide the reverse processes to S and T, respectively. To ensure that the image can be transformed to the corresponding domain by the diffusion model, a hyperparameter k_G is employed to specify the guidance. Furthermore, by adding stain guidance at each step within a specified range in the reverse process, it can be ensured that the converted image x_0 is as similar as possible to the reference image I_{ref} (especially in terms of the stain color and stain density map), thereby encouraging each step-generated image to retain similar features to the stain-guided reference image. Similar steps are applied to $T \rightarrow S \rightarrow T$.

A.2 Datasets

The effectiveness of SGCD was evaluated on four open datasets: Camelyon17 Bejnordi et al. (2017), Camelyon16 Bejnordi et al. (2017), Camelyon17-WILDS Koh et al. (2021), and MITOS & ATYPIA14 Racoceanu et al. (2014). The details of the four datasets are described below. Camelyon17 is obtained from five hospitals, denoted by C_1 to C_5, in the Netherlands. In the present study, C_1 was taken as the source domain and the others were taken as the target domain.

Camelyon16 is obtained from two hospitals, Radboud University Medical Center (RUMC) and UniversityMedical Center Utrecht (UMCU), in the Netherlands. RUMC contains 249 WSIs, 99 of which have tumor annotations, while UMCU contains 150 WSIs, 60 of which have tumor annotations. In the experiments, RUMC and UMCU were set as the source domain and target domain, respectively. Camelyon17-WILDS is a balanced version of Camelyon17. Given the extremely small number of lesion areas compared to normal ones in Camelyon17, the ratio of positive to negative samples derived from WSI patches is unbalanced. By comparison, Camelyon17-WILDS provides a more equitable ratio of positive and negative samples. Furthermore, it groups the images from hospitals with similar characteristics into a training set, with the data in the remaining hospitals serving as the validation and test sets. MITOS & ATYPIA14 is obtained from the same slide samples scanned by two scanners, namely Aperio Scanscope XT (A) and Hamamatsu Nanozoomer 2.0-HT (H). A training set was constructed consisting of $10,000$ patches randomly selected from the first 184 WSIs of the two scanners. Furthermore, 500 patches from the remaining 100 WSIs from the scanners were selected at random for testing. The A domain was taken as the source domain, and the H domain was taken as the target domain.

A.3 Thorough Evaluations in Histopathology Classification

A complete evaluation was conducted on the Camelyon17 dataset in addition to Table 3 to validate the efficacy of SGCD further. Table 5 presents the results of a cross-hospital domain adaptation experiment in that each of the five hospitals in Camelyon17 was in turn assigned as the source domain while the remaining hospitals served as the target domain. For example, when hospital C_2 was the source domain, hospitals C_1, C_3, C_4, and C_5 were treated as the target domains. It is observed that the proposed method, SGCD, generally demonstrates superior performance in almost all cases and the best result averagely, indicating that it enables diffusion models to generate more realistic and high-quality images, which can be effectively fine-tuned for downstream task models.

Method	C_1	C_3	C_4	C_5	Average
No adaptation	78.4	66.0	79.5	64.6	72.1
Vahadane *et al.* Vahadane et al. (2016)	79.8	77.7	83.1	78.8	79.9
Mackenko *et al.* Macenko et al. (2009)	75.9	71.0	85.1	73.3	76.3
Reinhard *et al.* Reinhard et al. (2001)	79.0	78.2	85.5	76.9	79.9
Stain Mix-Up Chang et al. (2021)	89.1	73.3	80.5	88.9	83.0
StainNet Kang et al. (2021)	84.8	85.8	81.5	88.0	85.0
MultiPathGAN Nazki et al. (2023)	88.4	80.4	87.3	89.1	86.3
BCD-net Yang et al. (2023)	86.8	82.8	85.8	87.6	85.8
Connect Later Qu and Xie (2024)	88.8	85.1	96.7	91.0	90.4
SPA Xiao et al. (2024)	90.1	83.6	96.8	92.0	90.6
HistAuGAN Wagner et al. (2021)	86.6	86.5	95.7	82.4	87.8
G-SAN Li et al. (2023b)	88.1	84.8	85.1	85.5	85.9
ContriMix Nguyen et al. (2024)	86.8	85.8	94.0	88.9	88.9
Ours (SGCD)	91.2	87.3	95.3	93.3	**91.8**

(a) Experiment results when C_2 as source domain.

Method	C_1	C_2	C_3	C_5	Average
No adaptation	75.5	76.0	67.4	60.1	69.8
Vahadane *et al.* Vahadane et al. (2016)	85.0	81.3	76.9	76.5	79.9
Mackenko *et al.* Macenko et al. (2009)	83.2	76.5	72.5	76.5	77.2
Reinhard *et al.* Reinhard et al. (2001)	83.8	78.2	78.8	77.7	79.6
Stain Mix-Up Chang et al. (2021)	85.7	80.6	87.8	85.2	84.8
StainNet Kang et al. (2021)	87.6	88.9	85.0	85.9	86.9
MultiPathGAN Nazki et al. (2023)	88.8	90.1	84.7	86.1	87.4
BCD-net Yang et al. (2023)	90.5	90.7	91.1	92.2	91.1
Connect Later Qu and Xie (2024)	94.5	93.0	91.7	91.3	92.6
SPA Xiao et al. (2024)	93.8	92.0	92.2	93.6	92.9
HistAuGAN Wagner et al. (2021)	87.5	86.5	95.7	82.4	88.0
G-SAN Li et al. (2023b)	88.9	87.2	82.6	83.7	87.0
ContriMix Nguyen et al. (2024)	90.6	86.8	91.3	85.4	88.5
Ours (SGCD)	95.7	94.6	93.6	95.4	**94.8**

(b) Experiment results when C_4 as source domain.

Method	C_1	C_2	C_4	C_5	Average
No adaptation	74.9	77.2	79.5	83.4	78.8
Vahadane *et al.* Vahadane et al. (2016)	82.8	76.3	81.9	91.9	83.2
Mackenko *et al.* Macenko et al. (2009)	76.1	74.0	82.6	87.3	80.0
Reinhard *et al.* Reinhard et al. (2001)	77.8	72.2	85.3	84.4	79.9
Stain Mix-Up Chang et al. (2021)	90.7	77.0	88.6	95.2	87.9
StainNet Kang et al. (2021)	82.1	78.1	88.8	94.8	86.0
MultiPathGAN Nazki et al. (2023)	84.8	84.4	91.7	94.6	88.9
BCD-net Yang et al. (2023)	88.3	85.9	88.7	96.8	89.9
Connect Later Qu and Xie (2024)	92.9	84.9	92.5	95.5	91.5
SPA Xiao et al. (2024)	94.2	87.6	97.7	96.6	94.0
HistAuGAN Wagner et al. (2021)	88.9	86.5	92.1	93.6	90.3
G-SAN Li et al. (2023b)	86.7	87.7	88.5	93.1	89.0
ContriMix Nguyen et al. (2024)	89.0	84.3	94.8	93.7	90.5
Ours (SGCD)	94.3	89.3	95.7	97.5	**94.2**

(c) Experiment results when C_3 as source domain.

Method	C_1	C_2	C_3	C_4	Average
No adaptation	65.2	65.9	73.8	69.4	68.6
Vahadane *et al.* Vahadane et al. (2016)	79.8	79.1	72.8	73.3	76.3
Mackenko *et al.* Macenko et al. (2009)	76.7	77.3	74.0	70.0	74.5
Reinhard *et al.* Reinhard et al. (2001)	74.0	73.9	72.2	70.1	72.6
Stain Mix-Up Chang et al. (2021)	85.7	81.9	81.8	77.9	81.8
StainNet Kang et al. (2021)	82.1	79.9	80.6	76.0	79.7
MultiPathGAN Nazki et al. (2023)	88.1	82.1	87.2	83.7	85.3
BCD-net Yang et al. (2023)	89.0	82.4	88.9	89.8	87.5
Connect Later Qu and Xie (2024)	88.7	80.6	91.3	90.1	87.7
SPA Xiao et al. (2024)	90.0	81.7	93.5	91.1	89.1
HistAuGAN Wagner et al. (2021)	88.7	81.7	85.7	89.0	86.3
G-SAN Li et al. (2023b)	90.2	80.3	81.0	88.7	85.1
ContriMix Nguyen et al. (2024)	90.1	81.4	84.7	93.1	87.8
Ours (SGCD)	92.5	82.6	93.5	95.2	**91.0**

(d) Experiment results when C_5 as source domain.

Table 5: Histopathology classification results for Camelyon17 were obtained under a series of experimental settings. Here, one of the four hospitals $C_2 \sim C_5$ was designated as the source domain, while the remaining four ones were the target domains. Here, AUC (%) was adopted as the evaluation metric.

A.4 Visual Results

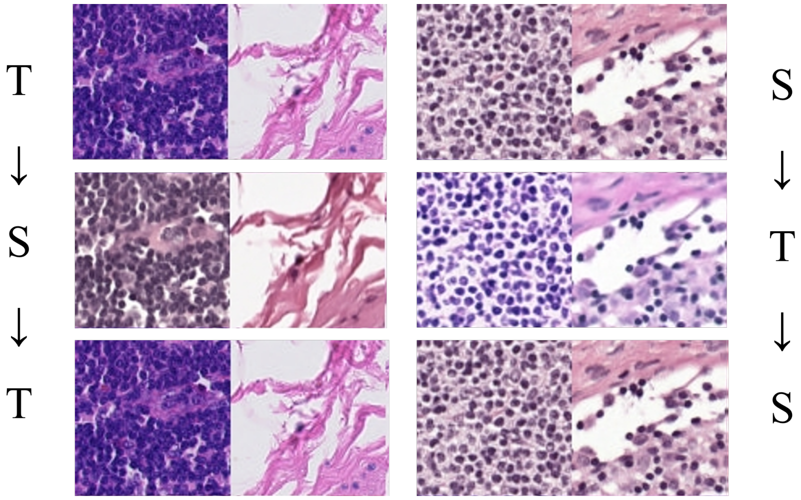


Figure 3: Samples generated by SGCD used to train diffusion models D_S and D_T.

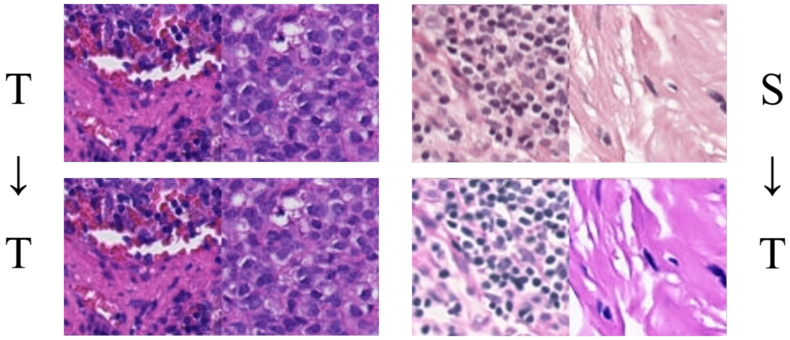


Figure 4: Samples generated by SGCD used to train target classifier C_T.

Method	SSIM	PSNR (dB)
Vahadane normalization Vahadane et al. (2016)	0.63	12.7
Mackenko normalization Macenko et al. (2009)	0.66	13.5
Reinhard normalization Reinhard et al. (2001)	0.61	13.6
StainGAN Shaban et al. (2019)	0.71	17.1
Ours (SGCD)	**0.88**	**27.5**

Table 6: Quantitative results of stain transfer on MITOS & ATYPIA14. Each image from A-domain is converted into H-domain, and both SSIM and PSNR are calculated between converted image and corresponding ground truth image in H-domain.

Figure 3 and Figure 4 present some typical samples obtained when applying SGCD on Camelyon17. Figure 3 illustrates the cyclic architecture ($S \to T \to S$ and $T \to S \to T$) used for training the diffusion model, while Figure 4 demonstrates the results generated by the diffusion model under different conditional constraints. The images transformed from S to T are used to train the target classifier. The transformed images in Figure 5 and quantitative results in Table 6 reveal that the stain normalization method suffers from a loss of detail information and distortion due to its normalization process, resulting in lower SSIM and PSNR scores. In contrast, StainGAN, which is based on a GAN architecture, generates images of a higher quality and greater accuracy, thus outperforming the stain normalization methods. Nonetheless, among all the considered methods, the proposed SGCD method, which incorporates cyclic and conditional constraints, and leverages the image generation capabilities

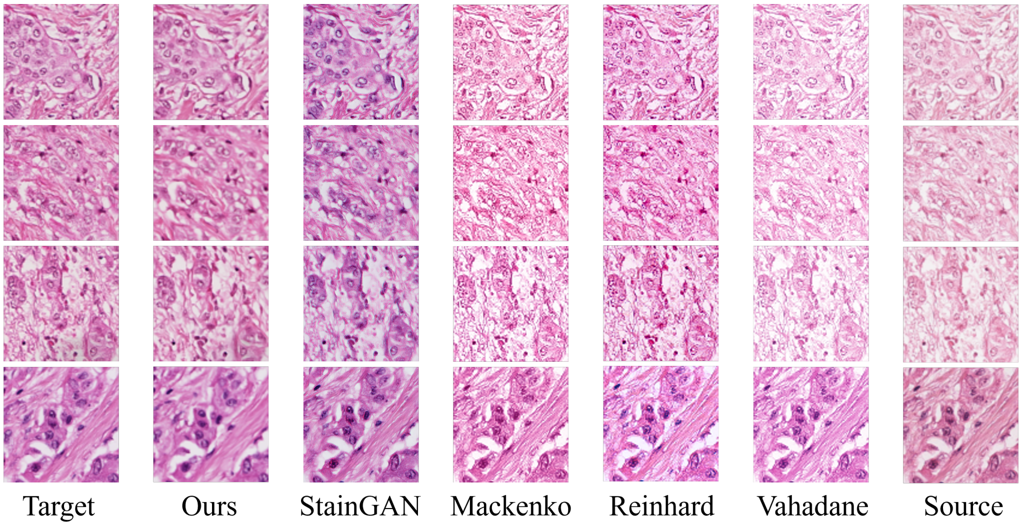


Figure 5: Visualization results of different conversion methods on MITOS & ATYPIA14 dataset. Paired images from domains A and H are used, with H-domain serving as ground truth. Rightmost column shows source images from A-domain, and leftmost column shows corresponding target images from H-domain.

of diffusion models, achieves the best performance on this task. Moreover, Figure 6 illustrates the UMAP embeddings of color statistics from different domains wherein the embeddings demonstrate that the transformed images closely match the target domain distribution, a critical factor for effective downstream task performance.

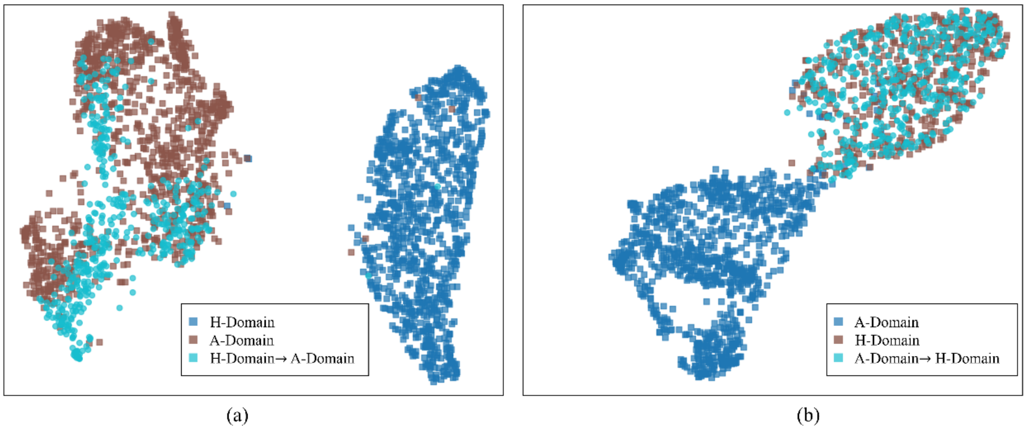


Figure 6: UMAP embeddings of color statistics across domains. (a) Embeddings for A-domain, H-domain, and the images converted from H-domain to A-domain. (b) Embeddings for H-domain, A-domain, and the images converted from A-domain to H-domain.

A.5 Validation of the Two-step Conversion Process

Experiments were conducted on the paired data in MITOS & ATYPIA14 to validate the effectiveness of SGCD in adapting a diffusion model trained on domain A to generate images resembling domain A from domain B images. To evaluate the sensitivity of SGCD to different hyperparameter settings, k was varied. The performance of SGCD was measured by computing PSNR and SSIM metrics between the generated images and their ground truth equivalents in both domain S and domain T. The results are visualized in Figure 7, where the stain guidance process was stopped at step 100. It can be seen that as k increases, the quality of the transformed results improves accordingly. Specifically,

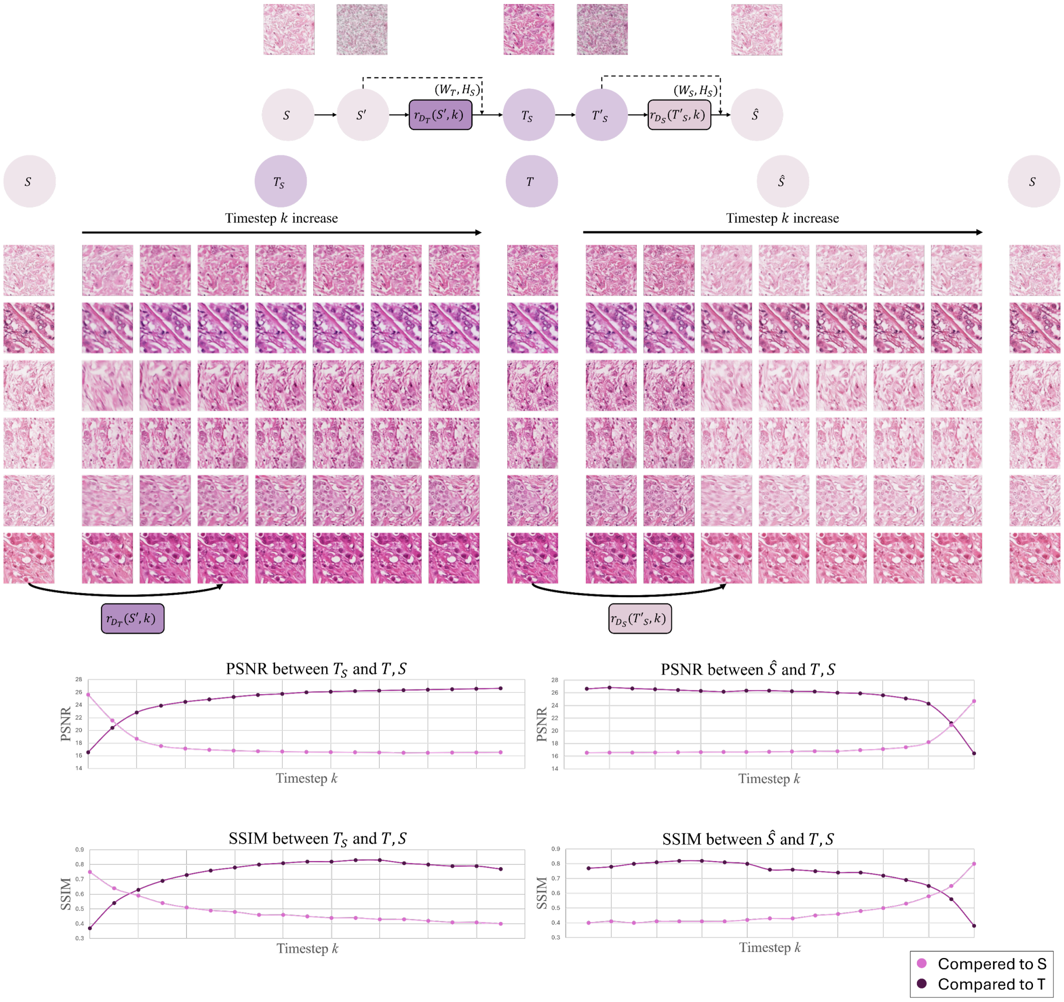


Figure 7: Quantitative results and visualization results of the two-stage conversion process on MITOS & ATYPIA14.

the transformed T_S becomes more similar to the actual T, and the transformed $\hat{S}$ becomes more similar to the actual S, as evidenced by the consistent increase in the PSNR and SSIM metrics. However, an excessively large k may lead to a loss of original image features, resulting in a decrease in the SSIM value after conversion. Therefore, $k = 600$ was employed in our experiments to achieve optimal performance. The same experiment was also conducted on Camelyon17. As Camelyon17 lacked a paired image, only the visualization results of the reconstructed and converted images are shown in Figure 8 and Figure 9, respectively. Overall, the proposed SGCD improves the ability of the diffusion model to perform bidirectional translation between S and T, making it a powerful tool for downstream task fine-tuning.

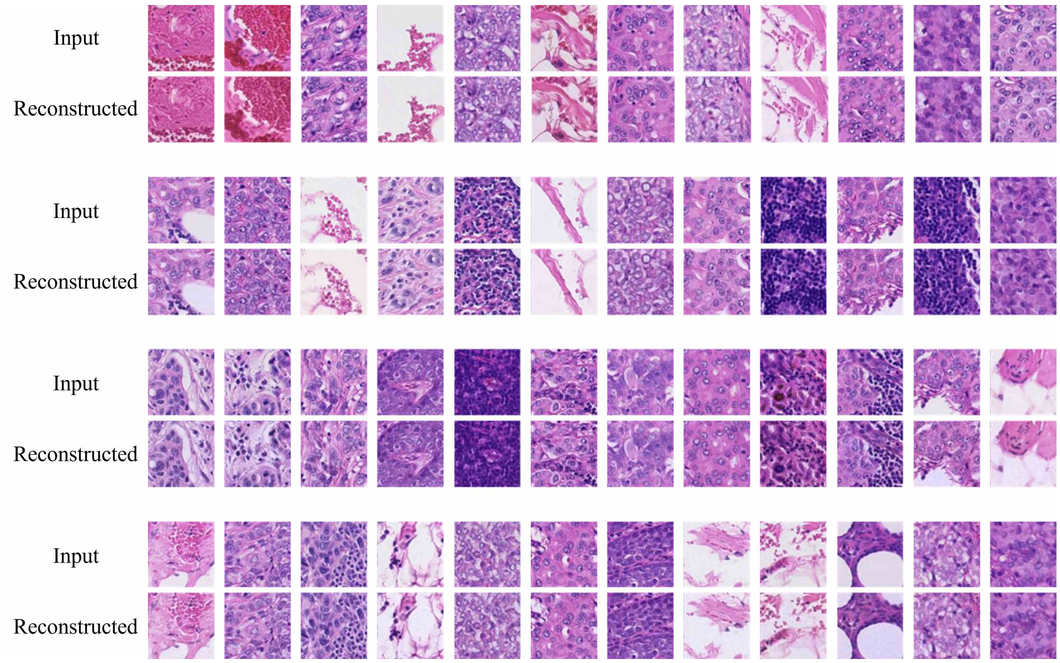


Figure 8: The visualization results of input and reconstructed images using SGCD on the Camelyon17 dataset.

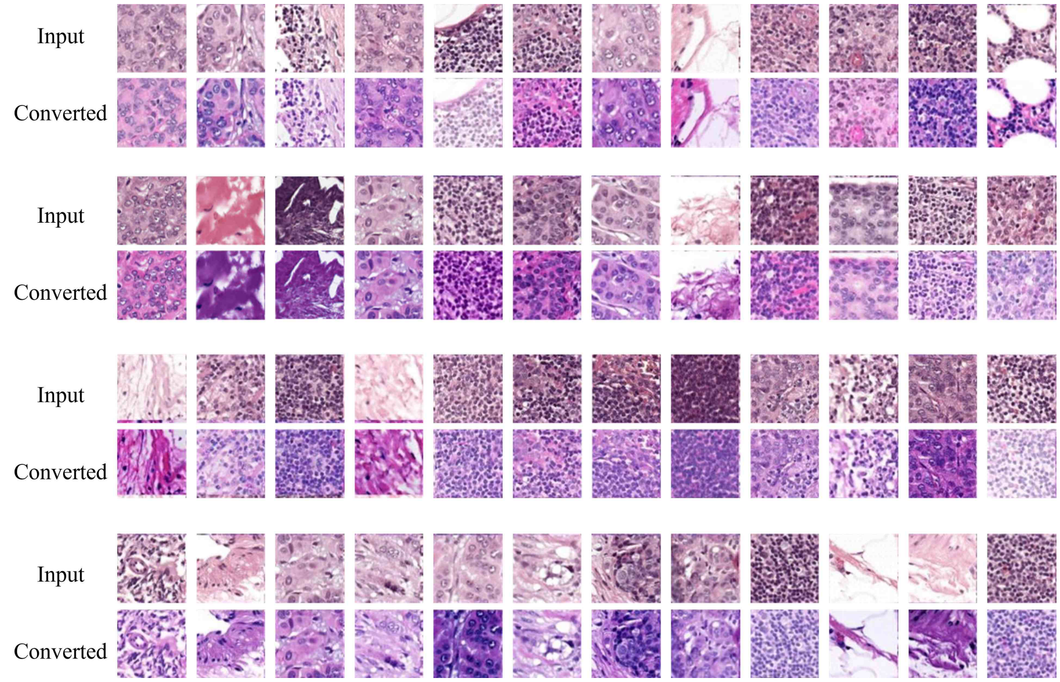


Figure 9: The visualization results of input and converted images using SGCD on the Camelyon17 dataset.

A.6 Performance with the new image Ts

This section presents supplementary experimental evidence on Camelyon16 dataset to quantify the intrinsic value of our generated synthetic images (T_S) and clarify their role in our overall domain adaptation pipeline.Our method's core contribution is the ability to produce high-quality, labeled T_S images that are both stylistically consistent with the target domain and class-consistent. We argue that even before the full domain adaptation (DA) process, these generated images are a powerful resource that establishes a strong baseline.The full adaptation pipeline then refines the model further by incorporating unlabeled target images (T) via a feature alignment strategy (e.g., MMD) to optimize performance on the true target domain distribution.Table 7 shows the results that quantify the impact of these two distinct steps.

Method	AUC (%)
No Adaptation (Source Only)	75.9
Training with only T_S	92.6
Training with T_S and T	95.8

Table 7: Quantitative analysis of the contribution of synthetic images (T_S).

The results clearly demonstrate that training a classifier on only our generated T_S images achieves a strong AUC of 92.6%, representing a substantial gain over the Source Only baseline (75.9%). This highlights the primary benefit of our method: generating class-consistent, labeled images that can be used directly for Domain Adaptation.The further performance gain to 95.8%, achieved by additionally incorporating the unlabeled target domain images (T), confirms that our approach provides an excellent, high-performance starting point with T_S, which subsequent feature alignment steps can leverage to achieve maximum performance.

A.7 Ablation Study of Components

Dual Diffusion Model	SGC Loss	FT of Diffusion Models	AUC (%)
V	V	V	95.8
-	V	V	92.8
V	-	V	89.4
-	-	-	86.8

Table 8: Component-wise ablation study on the Camelyon16 dataset.

This section provides a component-wise ablation study to precisely quantify the individual contributions of our key architectural elements: the dual diffusion model, the Stain-Guided Consistency (SGC) loss, and the fine-tuning (FT) of the diffusion models. This analysis confirms the source of performance gains stems from the synergistic effect of these targeted components, rather than diffusion in general. The results, measured on the Camelyon16 dataset, are summarized in Table 8

The results underscore the following key findings:

- Dual Diffusion's Crucial Role: Comparing the full model ('V V V') to the single-diffusion approach ('- V V') shows a significant performance contribution from the dual diffusion model (95.8 vs. 92.8). This indicates that the cyclic nature and the bidirectional generative constraints are essential for achieving the highest performance.
- Impact of SGC Loss: The introduction of the SGC loss provides a substantial boost to the method's effectiveness (comparing 'V V V' to 'V - V': 95.8 vs. 89.4). This confirms the value of targeted stain guidance in aligning features during the adaptation process.
- Synergistic Gains: The performance difference between the full model (95.8) and the baseline without any of our proposed components (86.8) demonstrates that the gains are primarily derived from the synergistic effect of both cycle consistency and targeted stain guidance, rather than solely from the general properties of diffusion models.

A.8 Robustness to Non-Stain Domain Degradations

Augmentation Method	No Adaptation (AUC (%))	With SGCD (AUC (%))
Blur	61.4	86.3
Noise	59.3	82.9
Blur + Noise	62.1	83.5

Table 9: Robustness comparison against non-stain related domain degradations.

While our core loss function is stain-guided, the bidirectional generative constraints inherent in our dual-diffusion framework provide a degree of robustness against other common types of domain shift, including structural variations and image artifacts.

To demonstrate this broader applicability, we conducted supplementary experiments on the Camelyon16 dataset where common image imperfections (blur and noise) were simulated through data augmentation. The results in Table 9 compare the baseline performance (No Adaptation) against our SGCD method under these corrupted conditions. As shown, the SGCD method significantly improves performance even when input images are corrupted with common artifacts like blur and noise. This suggests that the dual-diffusion framework, while optimized for stain variations, possesses a broader adaptability to structural or artifactual variations frequently encountered in real-world medical imaging. This aligns with findings in related workGao et al. (2023) exploring diffusion-driven adaptation to test-time corruption.

A.9 Ablation Study of Diffusion Timesteps

We conducted a dedicated ablation study on the key hyperparameters k and k_G (as defined in Eq. 5 and Eq.7 of the manuscript) to evaluate proposed SGCD's robustness to their variation. The AUC(%) results, measured on the Camelyon16 dataset, are presented in Table10. The experimental results demonstrate that our proposed method consistently achieves superior performance compared to existing methods across a wide range of k values. This suggests that while these hyperparameters influence peak performance, our method's overall effectiveness is robust to reasonable variations.

k \ k_G	10	100	150
200	88.6	94.5	85.6
400	93.6	94.2	89.1
600	94.1	95.8	90.9
800	85.9	94.7	90.4
1000	91.5	94.9	83.9

Table 10: Ablation study on the hyperparameters k and k_G.

A.10 Quantifying Semantic Preservation (Class Consistency)

Method	Class Consistency Ratio
No Adaptation	0.66
With SGCD	0.85

Table 11: Quantitative analysis of Class Consistency.

Semantic preservation is critical for clinical decision-making. We address this through collaborative training where the target classifier actively guides the diffusion model, ensuring generated images retain semantic information consistent with the source.

To quantify this, we measured the Class Consistency Ratio on the Camelyon16 dataset, comparing the class labels of original images with their transformed counterparts. This metric demonstrates that our method significantly improves the preservation of class-level semantic information during domain translation, thereby establishing a necessary foundation for clinical trust.

A.11 Robustness to Limited Target Domain Data

We acknowledge that handling data scarcity is critical in real-world applications. To quantify our method's ability to adapt with minimal target domain samples, we conducted a supplementary experiment on Camelyon16 using varying percentages of available target data.

The results, shown in Table12, demonstrate the effectiveness of our Stain-Guided Consistency Diffusion (SGCD) even with heavily restricted data access. The results indicate that our method shows promising adaptability even with only 1% of target domain data (AUC 87.5), significantly outperforming the Source-Only baseline (75.9). This confirms our method's capability to generalize effectively in challenging, data-scarce scenarios.

Target Data %	Source-only	1%	10%	50%	100%
AUC (%)	75.9	87.5	89.4	93.5	95.8

Table 12: Robustness to Limited Target Domain Data.

A.12 Fine-Grained Pathological Fidelity

Histopathology relies on subtle details. To provide quantitative validation that our method preserves diagnostically meaningful structures, we measured the pixel-level overlap of tumor nuclei regions before and after image translation. We used a semantic segmentation model trained on the target domain for evaluation consistency.The results in Table13 compare segmentation performance on original target images with that on our translated images (T_S). These high metrics confirm that our method is highly effective at preserving fine-grained pathological structures. The marginal performance drop provides strong quantitative evidence that our approach maintains the critical pixel-level details essential for accurate pathological interpretation.

	Original	Translated
IoU	0.9661	0.9124
Dice Score	0.9827	0.9542

Table 13: Quantitative validation of Fine-Grained Pathological Fidelity.

A.13 Robustness to Rare Cohorts (Positive Class Performance)

Our datasets are inherently class-imbalanced, with tumor regions often representing rare cohorts. To explicitly address performance on the most challenging, clinically relevant rare samples, we compared our full method against a Source-Only baseline in the 1% target data setting. The results demonstrate a severe performance degradation in the Source-Only baseline for the rare positive class (Recall: 0.212). In stark contrast, our full SGCD method achieves a robust Recall of 0.819 and a high F1-score of 0.861 for the same rare class. This evidence confirms that our domain adaptation strategy provides a crucial and decisive benefit in accurately identifying challenging, clinically relevant rare samples.

Metric	No Adaptation (Source Only)	SGCD with 1% Target Data
Precision	0.985	0.835
Recall	0.967	0.917
F1-score	0.976	0.874

Table 14: Comparison of performance metrics for the **Negative/Majority Class** on the Camelyon16 dataset with only 1% target domain data.

A.14 Reference image selection

Our method does not rely on a specific, fixed reference image for domain translation. Instead, we dynamically sample images from the target domain during training to provide stain information for

Metric	No Adaptation (Source Only)	SGCD with 1% Target Data
Precision	0.104	0.908
Recall	0.212	0.819
F1-score	0.140	0.861

Table 15: Comparison of performance metrics for the **Positive/Rare Class** on the Camelyon16 dataset with only 1% target domain data.

the Stain-Guided Consistency (SGC) loss. This inherent design makes the adaptation process robust by accounting for the natural variations in stain matrices within the target domain.

Furthermore, this dynamic process leads to an implicit benefit: occasionally generated images with slight stylistic deviations from the target mean act as a form of on-the-fly data augmentation for the target classifier. This strengthens the model's generalization capability against minor distribution shifts Chang et al. (2021).

However, while beneficial, this randomness is also the source of potential failure cases, as shown in Figure 10. When the dynamic reference image leads to an overly aggressive style shift or excessive distortion of fine-grained pathological structures, the resulting synthetic image may become diagnostically ambiguous, leading to classifier errors.

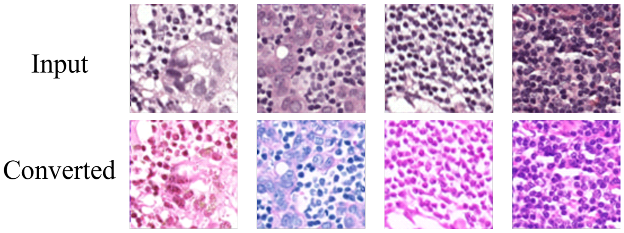


Figure 10: The fail cases of input and converted images using SGCD on the Camelyon17 dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have explained the contribution and scope of the paper in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of relying on pre-trained models and hyperparameter selection are discussed in the main paper (Section 5) and Appendix (Section A).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This thesis does not contain theoretical studies.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all experimental details in the main paper (Section 5) and in the Appendix (Section A).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not include the code in this submission, but in the future, we will organize the experimental code and documentation, and released on GitHub.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details are in the main paper (Section 5) and in the Appendix (Section A).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We took the average of the results of three independent repeated experiments as the results of the main experiment, but did not analyze and discuss the differences in detail.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The details are in the main paper (Section 5).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully comply with the NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not pose significant risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have added appropriate citation notes to the existing papers, models, and datasets mentioned in this paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.