K-Stain: Keypoint-Driven Correspondence for H&E-to-IHC Virtual Staining

Sicheng Yang[1] **Zhaohu Xing**[1] **Haipeng Zhou**[1] **Lei Zhu**[1,2*]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology

Abstract

Virtual staining offers a promising method for converting Hematoxylin and Eosin (H&E) images into Immunohistochemical (IHC) images, eliminating the need for costly chemical processes. However, existing methods often struggle to utilize spatial information effectively due to misalignment in tissue slices. To overcome this challenge, we leverage keypoints as robust indicators of spatial correspondence, enabling more precise alignment and integration of structural details in synthesized IHC images. We introduce K-Stain, a novel framework that employs keypoint-based spatial and semantic relationships to enhance synthesized IHC image fidelity. K-Stain comprises three main components: (1) a Hierarchical Spatial Keypoint Detector (HSKD) for identifying keypoints in stain images, (2) a Keypoint-aware Enhancement Generator (KEG) that integrates these keypoints during image generation, and (3) a Keypoint Guided Discriminator (KGD) that improves the discriminator's sensitivity to spatial details. Our approach leverages contextual information from adjacent slices, resulting in more accurate and visually consistent IHC images. Extensive experiments show that K-Stain outperforms state-of-the-art methods in quantitative metrics and visual quality.

1 Introduction

Histopathological stain is an essential process in clinical pathological analysis [1]. Hematoxylin and Eosin (H&E) and immunohistochemical (IHC) staining are two widely used staining methods. H&E staining is a basic and cost-effective technique for observing tissue morphology and structure, but it lacks specificity for specific proteins [2]. IHC stain uses specific antibodies to identify target proteins in tissues, making it crucial for cancer diagnosis. Although IHC staining offers high specificity, it is more complex and expensive [3]. These characteristics have motivated researchers to explore whether H&E images can be computationally converted into IHC-like counterparts.

Digital pathology is rapidly evolving, and numerous researchers are developing various virtual staining methods based on deep learning [4–14]. These methods enable the generation of one type of stain from another without the need for chemical staining procedures. However, many traditional generative models cannot adapt to a key characteristic of virtual staining. In practice, the paired images are obtained from two consecutive tissue sections that are stained separately. Although these sections are adjacent, they are not identical since cells and structures may appear in one section but not in the other, and their spatial arrangement inevitably shifts due to tissue slicing. Consequently, the dataset lacks truly aligned image pairs, which makes pixel-wise supervision (e.g., ℓ_1 or ℓ_2 loss) unreliable. This intrinsic misalignment poses a fundamental challenge for conventional generative frameworks that rely on strictly paired data [15]. To address this issue, there are generally three approaches: (a) Contrastive Learning, which constructs corresponding patch pairs to align semantically similar

[*] Lei Zhu (leizhu@ust.hk) is the corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

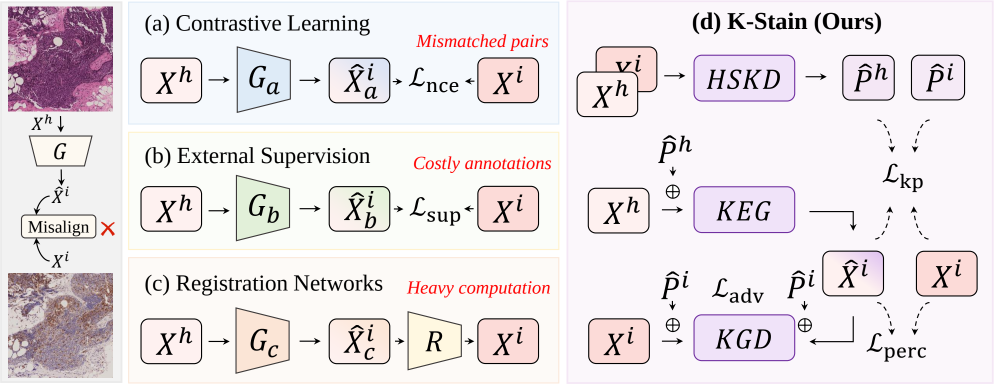

Figure 1: Illustration of different strategies for handling misalignment in virtual staining. (a) Contrastive Learning, (b) External Supervision, (c) Registration Networks, and (d) our proposed K-Stain framework. Here, X_h and X_i denote the H&E and IHC images, respectively; $\hat{X}_i$ is the generated IHC; $\hat{P}_h$ and $\hat{P}_i$ are the predicted keypoints; $\mathcal{L}_{\text{kp}}$, $\mathcal{L}_{\text{perc}}$, and $\mathcal{L}_{\text{adv}}$ represent the keypoint-guided reconstruction, perceptual, and adversarial losses, respectively.

patches [15–18] (Fig. 1a). (b) External Supervision Signals, such as segmentation maps [19], nuclei density graph [16] or protein prototype [20] (Fig. 1b). (c) An additional registration network, which calibrates the spatial position of generated images [21] (Fig. 1c).

However, contrastive learning methods only enforce semantic alignment, which prevents them from fully exploiting the fine-grained spatial correspondence available in paired data. This often results in mismatches for subtle but clinically important structures, such as nuclei boundaries or glandular morphology. External supervision signals can provide useful auxiliary constraints, yet they do not explicitly address spatial misalignment and their acquisition, such as segmentation maps, nuclei density graphs, or protein prototypes, is labor-intensive and costly, limiting practical scalability. Registration networks aim to resolve alignment by calibrating spatial positions, but they introduce considerable computational burden, particularly when processing high-resolution whole-slide images.

To address these limitations, we introduce **K-Stain**, a novel framework that explicitly integrates keypoint-based spatial cues into the virtual staining process (Fig. 1d). Keypoints act as compact and discriminative descriptors of structural landmarks, effectively capturing local correspondences between different stains. Leveraging these spatial anchors enables the network to align subtle tissue structures more accurately without relying on dense annotations or computationally expensive registration. By embedding keypoint relationships into the generation pipeline, K-Stain achieves improved spatial consistency and reduced computational cost.

Specifically, K-Stain consists of three components. First, we introduce a Hierarchical Spatial Keypoint Detector (HSKD) for efficient keypoint prediction in HE and IHC images, capturing both spatial and semantic relationships. Second, we propose a Keypoint-aware Enhancement Generator (KEG) that leverages keypoint information to guide the generation process, improving spatial consistency and enhancing the quality of synthesized IHC images. Third, we design a Keypoint Guided Discriminator (KGD) that incorporates keypoints to distinguish real from synthetic images, enabling the discriminator to focus simultaneously on semantic content and spatial details. Extensive experiments on two datasets demonstrate that K-Stain significantly enhances virtual staining quality and achieves superior performance compared with state-of-the-art methods.

2 Related Work

Deep learning for virtual staining. A variety of deep learning approaches have been developed to enable virtual staining between H&E and IHC images. Early models, such as CycleGAN-like frameworks [22–24] and Pix2Pix-like models [25, 26], demonstrated the feasibility of stain conversion, yet their dependence on pixel-level supervision made them susceptible to slice misalignment. To alleviate this issue, contrastive learning objectives were introduced to enhance semantic alignment across patches [15, 18]; however, inaccurate correspondences often led to the loss of clinically relevant fine structures. To preserve structural fidelity, Dubey *et al.* [27] proposed SC-GAN, which integrates edge-based priors, decoder-side attention, and a structural loss to maintain contextual integrity, though its reliance on handcrafted features restricts generalization. Zeng *et al.* [28] further developed a semi-supervised PR-staining method that combines patch-level labels derived from registration with a classifier to enforce pathological consistency, but the requirement for precise registration remains a key limitation. More recently, diffusion-based frameworks have emerged as a promising alternative. Shen *et al.* [29] introduced StainDiff, a probabilistic diffusion framework for virtual staining, while Jewsbury *et al.* [30] proposed StainFuser, a conditional latent diffusion model trained on large-scale pathology data. Efficiently capturing fine-grained structural correspondence under misaligned conditions remains an open challenge.

Keypoint detection in medical imaging. Keypoint representations provide a compact means of encoding anatomical landmarks and establishing spatial correspondence, and have been widely adopted in medical image analysis. In segmentation tasks [31, 32], keypoints serve as localized anchors that facilitate instance separation by grouping predefined points into structured objects, while also enriching features with long-range dependencies through keypoint-augmented fusion layers. In registration tasks [33, 34], sparse yet distinctive keypoints act as stable geometric constraints, enabling accurate estimation of dense deformation fields while alleviating the memory burden by obviating the need for additional registration networks. In the context of virtual staining, adjacent tissue slices exhibit strong local similarity, and paired H&E and IHC images naturally contain matching points that preserve structural semantics. Although slight pixel shifts or deformations may occur between slices, their glandular layout and cellular organization remain largely consistent within local regions. K-Stain leverages this property by detecting and matching keypoints to capture robust structural correspondences under small misalignments.

3 Methodology

3.1 Overview

Our proposed framework, K-Stain, addresses the misalignment challenge in virtual staining by incorporating spatial correspondence through keypoints. As illustrated in Fig. 2, K-Stain consists of three main modules: (1) a Hierarchical Spatial Keypoint Detector (HSKD) that adaptively learns to predict spatially and semantically consistent keypoints between H&E and IHC images in an end-to-end manner, (2) a Keypoint-aware Enhancement Generator (KEG) that embeds the detected keypoints into dense feature maps to guide the generation process, thereby enhancing structural fidelity, and (3) a Keypoint Guided Discriminator (KGD) that leverages keypoint-derived structural priors to enforce consistent adversarial supervision. These components are jointly optimized under a combination of keypoint-guided reconstruction loss, perceptual loss, and adversarial loss, which together enforce spatial alignment, semantic consistency, and realistic appearance in the synthesized IHC images. This adaptive and modular design enables K-Stain to effectively enhance both the quality and structural consistency of virtual staining compared with conventional generative approaches.

3.2 Hierarchical Spatial Keypoint Detector

We introduce the Hierarchical Spatial Keypoint Detector (HSKD) to predict keypoints from paired H&E and IHC images, as shown in Fig. 2(a). Given the input images X^h and X^i, we first process them with a series of Convolutional Neural Network (CNN) layers to extract hierarchical feature representations. These feature maps are subsequently embedded into a sequence of tokens, denoted as $\{z_1, z_2, \ldots, z_N\}$, where N is the number of tokens. To enhance keypoint localization by capturing long-range dependencies, the token sequence is processed by a Transformer block [35]. Specifically, the multi-head self-attention (MHSA) mechanism computes pairwise interactions among tokens. Given the query, key, and value projections:

$$Q = ZW^Q, \quad K = ZW^K, \quad V = ZW^V, \tag{1}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{2}$$

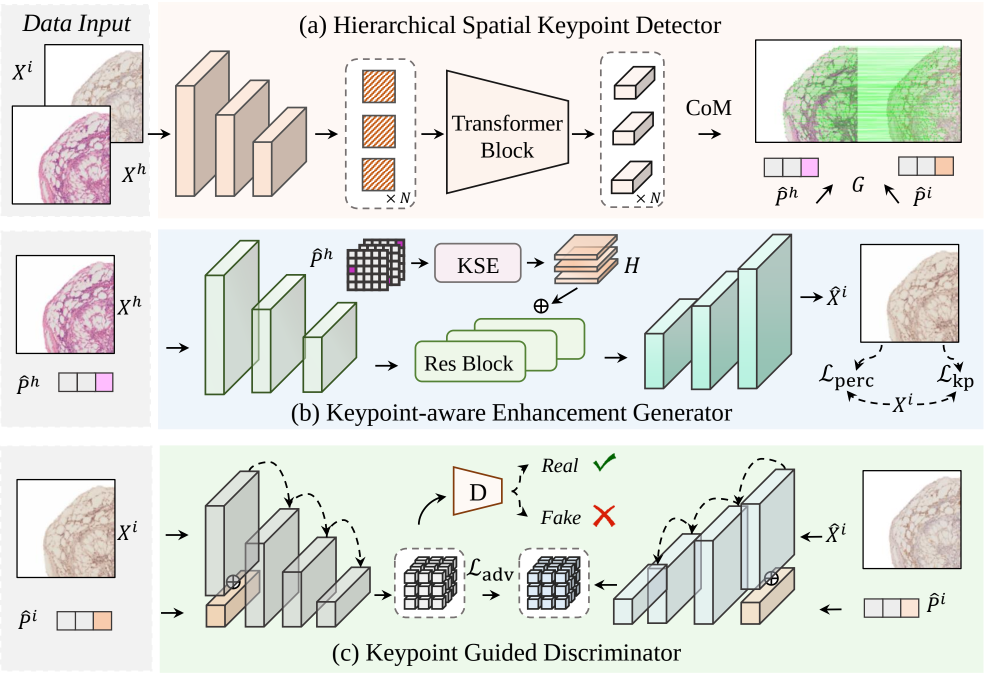
Data Input
X^i
X^h
(a) Hierarchical Spatial Keypoint Detector
×N
Transformer Block
×N
CoM
$\hat{P}^h$
G
$\hat{P}^i$
X^h
$\hat{P}^h$
$\hat{P}^h$
KSE
H
Res Block
$\hat{X}^i$
$\mathcal{L}_{\text{perc}}$
$\mathcal{L}_{\text{kp}}$
X^i
(b) Keypoint-aware Enhancement Generator
X^i
$\hat{P}^i$
D
Real ✓
Fake ✗
$\mathcal{L}_{\text{adv}}$
$\hat{X}^i$
$\hat{P}^i$
(c) Keypoint Guided Discriminator

Figure 2: The proposed K-Stain framework integrates keypoint-based spatial correspondence to address misalignment in virtual staining. It consists of (a) a Hierarchical Spatial Keypoint Detector (HSKD) that predicts consistent keypoints and estimates affine transformation, (b) a Keypoint-aware Enhancement Generator (KEG) that embeds keypoints into dense feature maps for IHC synthesis, and (c) a Keypoint Guided Discriminator (KGD) that enforces consistent adversarial supervision.

where $Z \in \mathbb{R}^{N \times d}$ represents the token embeddings, $W^Q, W^K, W^V \in \mathbb{R}^{d \times d}$ are learnable projection matrices, and d is the embedding dimension. The attention-enhanced features are then updated as

$$Z' = \text{MHSA}(Z). \tag{3}$$

Finally, the enhanced feature maps are passed through a Center of Mass (CoM) operator [36], which converts the feature activations into spatial coordinates. This yields two sets of predicted keypoints for the HE and IHC images, denoted as $\hat{P}^h, \hat{P}^i \in \mathbb{R}^{N \times 2}$, where each row corresponds to the (x, y)-coordinate of a keypoint.

Based on the predicted keypoints $\hat{P}^h$ and $\hat{P}^i$, we directly estimate an affine transformation that aligns X^h to X^i. The optimal affine matrix A is obtained via a least-squares fit:

$$A = \arg\min_{M} \left\| M\tilde{P}^h - \tilde{P}^i \right\|_F^2, \tag{4}$$

where $\tilde{P}^h$ and $\tilde{P}^i$ denote the homogeneous coordinates of $\hat{P}^h$ and $\hat{P}^i$, and $\|\cdot\|_F$ is the Frobenius norm. To warp the HE image, we construct a differentiable sampling grid $\mathcal{G}$ in normalized coordinates using the inverse affine matrix A^{-1}:

$$\mathcal{G}(p) = A^{-1}p, \tag{5}$$

where p denotes a pixel coordinate in the target (IHC) space. The warped HE image is then obtained by bilinear sampling:

$$X^h_{\text{w}}(p) = X^h\big(\mathcal{G}(p)\big). \tag{6}$$

The HSKD module jointly learns keypoint consistency and affine transformation, suppressing unreliable correspondences by optimizing spatially consistent keypoint pairs instead of raw pixels. Being fully differentiable, it enables adaptive keypoint learning without requiring a separate registration network. The resulting keypoints and alignment matrix are subsequently utilized by KEG (Sec. 3.3) and KGD (Sec. 3.4) for generation and discrimination.

3.3 Keypoint-aware Enhancement Generator

We propose the Keypoint-aware Enhancement Generator (KEG), which leverages spatial information from predicted keypoints to guide the synthesis of IHC images. As shown in Fig. 2(b), the detected keypoints $\hat{P}^h$ are first converted into the image domain by the Keypoint Spatial Embedding (KSE) module (see Fig. 3). Instead of performing registration directly in the feature space, we project keypoints back into the image domain through the KSE module. This mapping preserves spatial interpretability and avoids high-dimensional feature-space registration, which would otherwise incur significant computational overhead and hinder end-to-end optimization. KSE projects each $\hat{p}_i \in \hat{P}^h$ onto a 2D grid using a localized kernel $\phi(\cdot)$. In our experiments, we adopt a Gaussian kernel:

$$h_i(x,y) = \phi\big((x,y), \hat{p}_i; \sigma\big), \tag{7}$$

where σ controls the spatial influence. The Gaussian kernel is chosen because it provides smooth and spatially localized activations, effectively modeling the gradual structural variations in tissue and avoiding sharp discontinuities that may arise from alternative kernels [37]. Each heatmap h_i is then passed through a 1×1 convolution to obtain $\tilde{h}_i$, and all embeddings are concatenated:

$$H = \mathrm{Concat}(\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_N), \tag{8}$$

producing a dense tensor H that encodes the structural layout indicated by keypoints.

In parallel, the H&E image X^h is encoded into multi-scale feature maps through a downsampling path:

$$F_{\text{down}} = \mathrm{Down}(X^h), \tag{9}$$

which reduces spatial resolution while preserving semantic context. The ResBlock pathway then integrates structural priors by concatenating F_{down} with H:

$$F_{\text{res}} = \mathrm{ResBlock}\big(\mathrm{Concat}(F_{\text{down}}, H)\big), \tag{10}$$

where the residual blocks [38] refine the fused representation by maintaining stable gradient flow and enhancing feature discrimination. This design allows the model to capture both appearance information and keypoint-guided structural information. Finally, an upsampling path reconstructs the virtual IHC image:

$$\hat{X}^i = \mathrm{Up}(F_{\text{res}}). \tag{11}$$

By embedding sparse keypoints into dense feature maps and integrating them with residual appearance features, KEG leverages the spatial correspondence encoded by keypoints to ensure alignment and improve the fidelity of virtual staining.

3.4 Keypoint Guided Discriminator

The Keypoint Guided Discriminator (KGD) integrates spatial keypoint information into the adversarial training process. Specifically, the keypoint set is first embedded into dense heatmaps through the Keypoint-to-Structural Embedding (KSE) module, yielding a structural tensor H. As shown in Fig. 2(c), the discriminator then receives both the IHC image (X^i or $\hat{X}^i$) and the corresponding structural tensor H. These tensors are fused with image features at multiple scales within the discriminator backbone, ensuring that structural correspondence is explicitly encoded during discrimination. Formally, the discriminator is defined as

$$D(X, H) \rightarrow [0, 1], \tag{12}$$

where X denotes either a real IHC image X^i or a generated image $\hat{X}^i$, and $H = \mathrm{KSE}(P)$ represents the structural embedding derived from its associated keypoints.

We adopt a conditional non-saturating GAN objective [39]. The discriminator aims to assign high scores to real pairs and low scores to generated pairs:

$$\mathcal{L}_D = -\,\mathbb{E}_{(X^i,P^i)}\big[\log D\big(X^i, H\big)\big] - \mathbb{E}_{(\hat{X}^i,\hat{P}^i)}\Big[\log\big(1 - D\big(\hat{X}^i, H\big)\big)\Big], \tag{13}$$

$$\mathcal{L}_G = -\,\mathbb{E}_{(\hat{X}^i,\hat{P}^i)}\Big[\log D\big(\hat{X}^i, H\big)\Big], \tag{14}$$

Here, the generator G corresponds to the Keypoint-aware Enhancement Generator (KEG) introduced in Sec. 3.3. By introducing keypoint information, the discriminator evaluates not only global appearance realism but also the consistency of local structures around keypoint-defined regions. This design enforces that generated IHC images $\hat{X}^i$ preserve semantic and spatial information represented by keypoints, thereby providing more reliable and structure-aware adversarial feedback.

3.5 Optimization Objective

The training objective of our framework combines keypoint-guided reconstruction loss, perceptual loss, and adversarial loss. These complementary objectives jointly enforce spatial alignment, semantic consistency, and visual realism in the generated IHC images.

Keypoint-guided reconstruction loss. To explicitly enforce spatial correspondence, we employ the affine transformation estimated in the HSKD module. Based on the predicted keypoints, HSKD yields an affine matrix A, which is used to warp the H&E image into the IHC coordinate space, resulting in X^h_{w}. The warped image is then passed through the generator $G(\cdot)$, i.e., the Keypoint-aware Enhancement Generator (KEG), to synthesize the virtual IHC image. We compute an ℓ_1 loss between the generated image and the ground truth IHC:

$$\mathcal{L}_{\text{kp}} = \big\| G\big(X^h_{\mathrm{w}}\big) - X^i \big\|_1. \tag{15}$$

Compared with a direct pixel-wise ℓ_1 loss, $\mathcal{L}_{\text{kp}}$ explicitly leverages the spatial correspondence established by HSKD, thereby reducing misalignment artifacts and improving semantic consistency.

Perceptual loss. To further enhance visual quality, we employ a perceptual loss φ_{perc} based on a pretrained feature extractor [40]. Specifically, we adopt the widely used VGG16 network [41] to compute high-level feature representations. The perceptual loss is defined as the ℓ_2 distance between the features of the generated IHC image and those of the ground truth:

$$\mathcal{L}_{\text{perc}} = \big\| \varphi(\hat{X}^i) - \varphi(X^i) \big\|_2^2, \tag{16}$$

where $\varphi(\cdot)$ denotes the feature representation extracted by the fixed VGG16 network [41]. This encourages the generator to preserve semantic structures and stain-specific texture details that may not be captured by pixel-level supervision alone.

Adversarial loss. Finally, we introduce an adversarial loss [39] with the Keypoint Guided Discriminator (KGD). The discriminator is conditioned on both the synthesized IHC image and the structural tensor derived from the detected keypoints. By jointly evaluating the image appearance and the encoded spatial information, KGD enforces consistency between visual realism and the spatial relationships. This adversarial training objective drives the generator to produce IHC images that are not only indistinguishable from real ones, but also spatially aligned with the tissue morphology.

Overall objective. The overall generator loss is formulated as

$$\mathcal{L} = \mathcal{L}_{\text{adv}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{kp}}\mathcal{L}_{\text{kp}}, \tag{17}$$

where λ_{perc} and λ_{kp} are trade-off hyperparameters, set to 1 and 5, respectively, as these values yield the best performance in our ablation study (Sec. 4.3).

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We use two publicly available datasets to evaluate K-Stain, namely the Breast Cancer Immunohistochemical (BCI) dataset [42] and the H&E to IHC image Translation (HIT) dataset [18]. The BCI dataset has 3,896 image pairs for training and 977 image pairs for testing. For the HIT dataset, we focus on its CD3 section, which includes 1,652 image pairs for training and 155 image pairs for testing. We follow the official train-test split provided by the respective datasets.

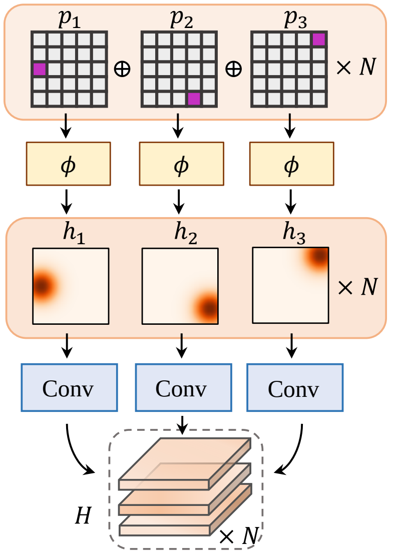

Figure 3: Illustration of the Keypoint Spatial Embedding (KSE) module, which encodes keypoint information into feature maps for virtual staining.

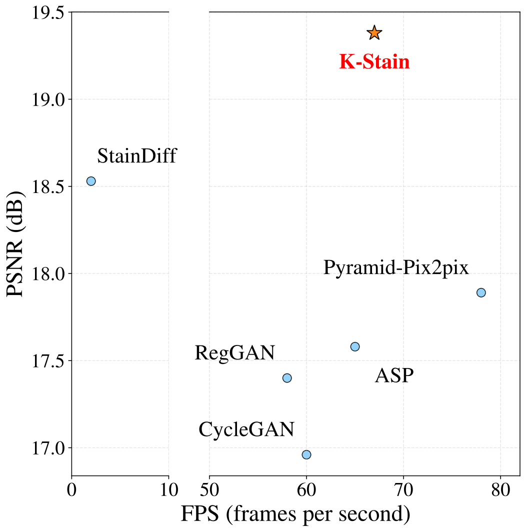

Figure 4: Comparison of model performance in terms of PSNR versus inference speed (FPS). The proposed K-Stain achieves the best trade-off between performance and efficiency compared to GAN- and diffusion-based baselines.

Table 1: Quantitative comparison. The **best** and the second-best results are highlighted.

Methods	BCI			HIT			Avg		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
CycleGAN	0.4739	16.68	0.2918	0.3838	17.24	0.2437	0.4289	16.96	0.2678
Pix2pix	0.4636	16.64	0.2947	0.3594	16.96	0.2478	0.4115	16.80	0.2713
Pix2pixHD	0.4685	16.84	0.2934	0.3622	17.47	0.2337	0.4153	17.16	0.2636
GcGAN	0.4720	16.98	0.2924	0.3922	17.90	0.2373	0.4327	17.44	0.2648
CUT	0.4752	17.05	0.2911	0.3594	17.37	0.2468	0.4173	17.21	0.2690
RegGAN	0.4701	16.99	0.2920	0.3720	17.82	0.2351	0.4210	17.40	0.2635
Pyramid-Pix2pix	0.4881	17.90	0.2977	0.3933	17.89	0.2373	0.4407	17.89	0.2675
ASP	0.4992	17.50	0.2949	0.3963	17.66	0.2463	0.4478	17.58	0.2706
StainFuser	0.5072	18.12	0.2750	0.3904	17.62	0.2486	0.4487	17.86	0.2538
StainDiff	0.5113	18.43	0.2809	0.3932	18.66	0.2142	0.4515	18.53	0.2471
K-Stain	**0.5268**	**19.82**	**0.2665**	**0.4162**	**18.93**	**0.2061**	**0.4720**	**19.38**	**0.2361**

Implementation Details. To ensure fair comparisons, we keep the discriminator architecture consistent with CycleGAN [23]. Specifically, we utilize 6 residual blocks in the generator. Our model is implemented in PyTorch 2.1.0 with CUDA 12.1 support. For data preprocessing, we perform random cropping to a resolution of 512×512. The batch size is set to 4 per GPU for each dataset. All experiments are conducted on a single NVIDIA A6000 GPU. We employ the Adam optimizer with an initial learning rate of 2×10^{-4} and train the network for 100 epochs.

Evaluation Metrics. We adopt the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index (SSIM), and the Learned Perceptual Image Patch Similarity (LPIPS) [43] to quantitatively evaluate different methods. In addition to reconstruction quality, we also measure the inference efficiency of each method using Frames Per Second (FPS), which reflects the number of images the model can process per second during inference.

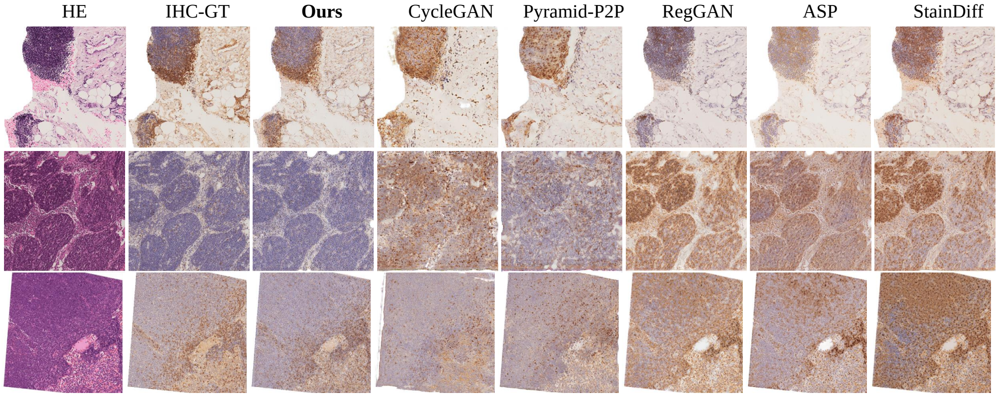

Figure 5: Visual comparisons of proposed K-Stain and other methods.

4.2 Comparison with SOTA Methods

We compare the proposed method with two representative categories of virtual staining approaches: (a) GAN-based methods, including CycleGAN [23], Pix2pix [44], Pix2pixHD [25], GcGAN [45], CUT [46], RegGAN [47], Pyramid-Pix2pix [42], and ASP [15]; and (b) diffusion-based methods, represented by StainFuser [30] and StainDiff [29].

Quantitative Comparisons. Table 1 reports the quantitative results of different methods on the BCI and HIT datasets in terms of SSIM, PSNR, and LPIPS. Compared with all competing approaches, our proposed K-Stain consistently achieves the best performance across both datasets. On BCI, K-Stain improves SSIM and PSNR by +1.55% and +1.39 dB over the second-best method, while also reducing LPIPS by –0.0085. On HIT, K-Stain attains clear gains as well, outperforming the next best competitor in SSIM by +1.99% and in PSNR by +0.27 dB, along with the lowest LPIPS (0.2061). Averaged across datasets, K-Stain reaches 0.4720 in SSIM, 19.38 dB in PSNR, and 0.2361 in LPIPS, surpassing the second-best results by +2.05%, +0.85 dB, and –0.011, respectively. These results highlight the superiority of K-Stain over existing virtual staining methods.

Visual Comparisons. Moreover, we provide visual comparisons between K-Stain and state-of-the-art methods on the BCI and HIT datasets. As shown in Figure 5, our approach produces synthesized IHC images that better preserve fine-grained cellular morphology and large-scale tissue architecture, avoiding the structural distortions or artifacts often observed in other methods. In addition, the staining appearance generated by our framework is more consistent with the ground truth, exhibiting natural color distribution and clearer boundary delineation. These visual comparisons further demonstrate the effectiveness of our model in achieving both structural fidelity and staining realism.

Efficiency Comparisons. As shown in Fig. 4, the proposed K-Stain achieves the most favorable balance between accuracy and inference efficiency compared with GAN- and diffusion-based baselines. Specifically, K-Stain attains the highest PSNR of 19.38 dB while sustaining a fast inference speed of 67 FPS. In contrast, diffusion-based StainDiff achieves a relatively high PSNR of 18.53 dB but suffers from extremely low efficiency (2 FPS). GAN-based approaches such as CycleGAN, RegGAN, ASP, and Pyramid-Pix2pix provide higher inference speeds (58–78 FPS) but remain clearly inferior in terms of PSNR (16.96–17.89 dB). These results highlight that K-Stain consistently achieves the best trade-off between accuracy and efficiency , outperforming both GAN and diffusion counterparts.

4.3 Ablation Study

Module Ablation. We conduct an ablation study to assess the contributions of HSKD, KEG, and KGD, where the baseline ("Basic") is a GAN [39] implemented with residual convolutional blocks. As summarized in Table 2, introducing KEG into the baseline ("M1") increases SSIM by approximately +2% and PSNR by +1.56 dB, accompanied by a notable reduction in LPIPS. Similarly, substituting the baseline discriminator with KGD ("M2") brings a +0.7% improvement in SSIM and +1.11 dB in PSNR, while also slightly lowering LPIPS. When both KEG and KGD are jointly

Table 2: Ablation study for different modules on BCI and HIT dataset.

Methods	KEG	KGD	SSIM ↑	PSNR ↑	LPIPS ↓
Basic			0.4289	16.96	0.2678
M1	✓		0.4465	18.52	0.2470
M2		✓	0.4360	17.80	0.2535
Ours w/o HSKD	✓	✓	0.4380	17.95	0.2620
Ours	✓	✓	**0.4720**	**19.38**	**0.2361**

Table 3: Ablation study on the trade-off hyperparameters λ_{perc} and λ_{kp} in the loss function.

λ_{perc}	λ_{kp}	SSIM ↑	PSNR ↑	LPIPS ↓
1	1	0.4685	19.15	0.2395
1	5	**0.4720**	**19.38**	**0.2361**
1	10	0.4698	19.20	0.2384
5	1	0.4662	19.02	0.2412
10	1	0.4650	18.95	0.2420

Table 4: Ablation study on the Gaussian kernel parameter σ (smoothing strength).

σ	SSIM ↑	PSNR ↑	LPIPS ↓
0.1	0.4655	18.95	0.2432
0.5	0.4695	19.20	0.2385
1.0	**0.4720**	**19.38**	**0.2361**
5.0	0.4678	19.02	0.2403
10.0	0.4542	17.90	0.2840

integrated, the model achieves the most significant gains. In contrast, removing the HSKD module ("Ours w/o HSKD") results in a substantial degradation: SSIM decreases by –3.4%, PSNR drops by –1.43 dB, and LPIPS increases by +0.026 relative to the full model. These findings collectively demonstrate that each module contributes positively, while their combination produces the most pronounced performance improvements (see Appendix B for additional ablation results).

Hyperparameter Ablation. We conduct ablation experiments to investigate how different hyperparameter settings influence the performance of K-Stain. As shown in Table 3, balancing the perceptual and keypoint losses is essential. Setting $\lambda_{\text{perc}} = 1$ and $\lambda_{\text{kp}} = 5$ yields the best results, improving SSIM by +0.35% and PSNR by +0.23 dB compared to the equal-weight case ($\lambda_{\text{perc}} = \lambda_{\text{kp}} = 1$), while also reducing LPIPS. Excessively increasing λ_{kp} to 10 or λ_{perc} to 5 or 10 leads to consistent performance drops, indicating that an imbalanced loss trade-off can weaken structural guidance. Regarding the Gaussian kernel parameter (Table 4), a moderate smoothing strength ($\sigma = 1.0$) achieves the best overall results. Smaller values (e.g., $\sigma = 0.1, 0.5$) slightly reduce structural fidelity, while overly large values (e.g., $\sigma = 10.0$) cause a pronounced degradation, with SSIM dropping by –1.78% and PSNR by –1.48 dB, and LPIPS increasing by +0.048. Based on these observations, we adopt $\lambda_{\text{perc}} = 1$, $\lambda_{\text{kp}} = 5$, and $\sigma = 1.0$ as the default hyperparameter configuration in our framework. For additional hyperparameter ablations, please refer to Appendix A.

5 Conclusion

In this work, we present **K-Stain**, a framework that addresses intrinsic misalignment in H&E-to-IHC virtual staining. By leveraging spatial correspondence via keypoints, K-Stain integrates three key modules: the Hierarchical Spatial Keypoint Detector (HSKD), Keypoint-aware Enhancement Generator (KEG), and Keypoint Guided Discriminator (KGD), which together align fine morphological structures while maintaining global staining fidelity. Experiments on two public datasets show that K-Stain surpasses GAN and diffusion-based baselines.

Despite its strong results, keypoint reliability may decline in regions lacking clear landmarks, and the method still depends on paired H&E–IHC data for training. Future work will explore multimodal extensions, domain robustness via self-supervision or adaptation, and applications to downstream tasks such as biomarker quantification and diagnostic assistance.

Acknowledgments and Disclosure of Funding

This work is supported by the Guangdong Science and Technology Department (2024ZDZX2004) and the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628).

References

[1] Marian Boktor, Benjamin R Ecclestone, Vlad Pekar, Deepak Dinakaran, John R Mackey, Paul Fieguth, and Parsin Haji Reza. Virtual histological staining of label-free total absorption photoacoustic remote sensing (ta-pars). *Scientific Reports*, 12(1):10296, 2022.

[2] Muhammad Zeeshan Asaf, Babar Rao, Muhammad Usman Akram, Sajid Gul Khawaja, Samavia Khan, Thu Minh Truong, Palveen Sekhon, Irfan J Khan, and Muhammad Shahmir Abbasi. Dual contrastive learning based image-to-image translation of unstained skin tissue into virtually stained h&e images. *Scientific Reports*, 14(1):2335, 2024.

[3] Fabienne Anglade, Danny A Milner Jr, and Jane E Brock. Can pathology diagnostic services for cancer be stratified and serve global health? *Cancer*, 126:2431–2438, 2020.

[4] Zhaohu Xing, Lequan Yu, Liang Wan, Tong Han, and Lei Zhu. Nestedformer: Nested modality-aware transformer for brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 140–150. Springer, 2022.

[5] Zhaohu Xing, Lei Zhu, Lequan Yu, Zhiheng Xing, and Liang Wan. Hybrid masked image modeling for 3d medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2024.

[6] Zhaohu Xing, Lihao Liu, Yijun Yang, Hongqiu Wang, Tian Ye, Sixiang Chen, Wenxue Li, Guang Liu, and Lei Zhu. Detect any mirrors: Boosting learning reliability on large-scale unlabeled data with an iterative data engine. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25476–25486, 2025.

[7] Zhaohu Xing, Sicheng Yang, Sixiang Chen, Tian Ye, Yijun Yang, Jing Qin, and Lei Zhu. Cross-conditioned diffusion model for medical image to image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–211. Springer, 2024.

[8] Zhaohu Xing, Tian Ye, Yijun Yang, Du Cai, Baowen Gai, Xiao-Jian Wu, Feng Gao, and Lei Zhu. Segmamba-v2: Long-range sequential modeling mamba for general 3d medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

[9] Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 578–588. Springer, 2024.

[10] Hongqiu Wang, Guang Yang, Shichen Zhang, Jing Qin, Yike Guo, Bo Xu, Yueming Jin, and Lei Zhu. Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. *IEEE Transactions on Medical Imaging*, 2024.

[11] Hongqiu Wang, Jian Chen, Shichen Zhang, Yuan He, Jinfeng Xu, Mengwan Wu, Jinlan He, Wenjun Liao, and Xiangde Luo. Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. *IEEE Transactions on Medical Imaging*, 2024.

[12] Hongqiu Wang, Yixian Chen, Wu Chen, Huihui Xu, Haoyu Zhao, Bin Sheng, Huazhu Fu, Guang Yang, and Lei Zhu. Serp-mamba: Advancing high-resolution retinal vessel segmentation with selective state-space model. *IEEE Transactions on Medical Imaging*, 2025.

[13] Hongqiu Wang, Shichen Zhang, Jian Chen, Yuan He, Jinfeng Xu, Hui Huang, Jianghong Xiao, Lu Li, Wenjun Liao, Shaoting Zhang, et al. Versatile source-free active domain adaptation for multi-center and multi-rater medical image segmentation. *Information Fusion*, page 103586, 2025.

[14] Hongqiu Wang, Wu Chen, Xiangde Luo, Zhaohu Xing, Lihao Liu, Jing Qin, Shaozhi Wu, and Lei Zhu. Toward fair and accurate cross-domain medical image segmentation: A vlm-driven active domain adaptation paradigm. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24102–24112, 2025.

[15] Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive supervised patchnce loss for learning h&e-to-ihc stain translation with inconsistent groundtruth image pairs. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 632–641. Springer, 2023.

[16] Fuqiang Chen, Ranran Zhang, Boyun Zheng, Yiwen Sun, Jiahui He, and Wenjian Qin. Pathological semantics-preserving learning for h&e-to-ihc virtual staining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 384–394. Springer, 2024.

[17] Yueheng Li, Xianchao Guan, Yifeng Wang, and Yongbing Zhang. Exploiting supervision information in weakly paired images for ihc virtual staining. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 113–122. Springer, 2024.

[18] Wei Zhang, Tik Ho Hui, Pui Ying Tse, Fraser Hill, Condon Lau, and Xinyue Li. High-resolution medical image translation via patch alignment-based bidirectional contrastive learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 178–188. Springer, 2024.

[19] Yihuang Hu, Qiong Peng, Zhicheng Du, Guojun Zhang, Huisi Wu, Jingxin Liu, Hao Chen, and Liansheng Wang. Boosting ffpe-to-he virtual staining with cell semantics from pretrained segmentation model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 67–76. Springer, 2024.

[20] Qiong Peng, Weiping Lin, Yihuang Hu, Ailisi Bao, Chenyu Lian, Weiwei Wei, Meng Yue, Jingxin Liu, Lequan Yu, and Liansheng Wang. Advancing h&e-to-ihc virtual staining with task-specific domain knowledge for her2 scoring. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 3–13. Springer, 2024.

[21] Yuzhu Li, Nir Pillar, Jingxi Li, Tairan Liu, Di Wu, Songyu Sun, Guangdong Ma, Kevin de Haan, Luzhe Huang, Yijie Zhang, et al. Virtual histological staining of unlabeled autopsy tissue. *Nature Communications*, 15(1):1684, 2024.

[22] 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.

[23] 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.

[24] Ruijie Wang, Sicheng Yang, Qiling Li, and Dexing Zhong. Cytogan: Unpaired staining transfer by structure preservation for cytopathology image analysis. *Computers in Biology and Medicine*, 180:108942, 2024.

[25] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[26] 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.

[27] Shikha Dubey, Tushar Kataria, Beatrice Knudsen, and Shireen Y Elhabian. Structural cycle gan for virtual immunohistochemistry staining of gland markers in the colon. In *International Workshop on Machine Learning in Medical Imaging*, pages 447–456. Springer, 2023.

[28] Bowei Zeng, Yiyang Lin, Yifeng Wang, Yang Chen, Jiuyang Dong, Xi Li, and Yongbing Zhang. Semi-supervised pr virtual staining for breast histopathological images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 232–241. Springer, 2022.

[29] 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. Springer, 2023.

[30] Robert Jewsbury, Ruoyu Wang, Abhir Bhalerao, Nasir Rajpoot, and Quoc Dang Vu. Stainfuser: Controlling diffusion for faster neural style transfer in multi-gigapixel histology images. *arXiv preprint arXiv:2403.09302*, 2024.

[31] Jingru Yi, Pengxiang Wu, Qiaoying Huang, Hui Qu, Bo Liu, Daniel J Hoeppner, and Dimitris N Metaxas. Multi-scale cell instance segmentation with keypoint graph based bounding boxes. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 369–377. Springer, 2019.

[32] Zhangsihao Yang, Mengwei Ren, Kaize Ding, Guido Gerig, and Yalin Wang. Keypoint-augmented self-supervised learning for medical image segmentation with limited annotation. *Advances in Neural Information Processing Systems*, 36, 2024.

[33] Lasse Hansen and Mattias P Heinrich. Graphregnet: Deep graph regularisation networks on sparse keypoints for dense registration of 3d lung cts. *IEEE Transactions on Medical Imaging*, 40(9):2246–2257, 2021.

[34] Alan Q Wang, M Yu Evan, Adrian V Dalca, and Mert R Sabuncu. A robust and interpretable deep learning framework for multi-modal registration via keypoints. *Medical image analysis*, 90:102962, 2023.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[36] Tianyu Ma, Ajay Gupta, and Mert R Sabuncu. Volumetric landmark detection with a multi-scale shift equivariant neural network. In *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*, pages 981–985. IEEE, 2020.

[37] Moo K Chung. Gaussian kernel smoothing. *arXiv preprint arXiv:2007.09539*, 2020.

[38] 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.

[39] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[40] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[42] Shengjie Liu, Chuang Zhu, Feng Xu, Xinyu Jia, Zhongyue Shi, and Mulan Jin. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1815–1824, 2022.

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[44] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[45] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2427–2436, 2019.

[46] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020.

[47] Lingke Kong, Chenyu Lian, Detian Huang, Yanle Hu, Qichao Zhou, et al. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems*, 34:1964–1978, 2021.

A Additional Ablation on the Number of Keypoints

We further investigate how the number of detected keypoints N influences the performance of K-Stain. As presented in Table 5, using a small number of keypoints (e.g., $N = 32$) provides limited structural anchors, which weakens the spatial guidance and leads to lower SSIM (0.4605) and PSNR (18.92 dB), together with a higher LPIPS (0.2440). Increasing the number of keypoints to $N = 64$ improves structural fidelity, yielding moderate gains in both SSIM and PSNR. The best performance is achieved when $N = 128$, where the model reaches the highest SSIM (0.4720) and PSNR (19.38 dB), along with the lowest LPIPS (0.2361). However, further increasing N to 256 introduces redundancy and noise in the structural representation, causing a slight performance drop and additional computational cost. These results indicate that a moderate number of keypoints offers the best balance between capturing sufficient structural details and maintaining model efficiency. Therefore, we adopt $N = 128$ as the default setting in all experiments.

Table 5: Ablation study on the number of keypoints N.

N	SSIM ↑	PSNR ↑	LPIPS ↓
32	0.4605	18.92	0.2440
64	0.4689	19.21	0.2392
128	**0.4720**	**19.38**	**0.2361**
256	0.4701	19.15	0.2375

B Additional Ablation on KSE within KGD

To verify the contribution of the keypoint-to-structural embedding (KSE) used in the Keypoint Guided Discriminator (KGD), we compare KGD with and without the KSE tensor while keeping all other settings identical. The results are summarized in Table 6.

Table 6: Ablation on the KSE tensor fed into KGD.

Method	SSIM↑	PSNR↑	LPIPS↓
KGD w/o KSE	0.4622	18.95	0.2418
KGD w/ KSE (ours)	**0.4720**	**19.38**	**0.2361**

As shown in Table 6, introducing KSE enables KGD to utilize explicit keypoint information during discrimination, improving structural alignment and demonstrating that keypoints enhance the discriminator's spatial perception.

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: In the abstract and introduction, we clearly outline the main contributions of this work: (1) We propose K-Stain, a novel virtual staining framework that incorporates keypoint-guided spatial correspondence to address the intrinsic misalignment in H&E-to-IHC translation. (2) We design a Hierarchical Spatial Keypoint Detector (HSKD) that adaptively predicts spatially consistent keypoints in paired H&E and IHC images. (3) We develop a Keypoint-aware Enhancement Generator (KEG) that embeds keypoint information into dense feature maps to improve staining quality. (4) We introduce a Keypoint Guided Discriminator (KGD) that integrates structural priors derived from keypoints to enforce spatial correspondence during adversarial supervision. (5) We conduct extensive experiments on the BCI and HIT datasets, demonstrating that K-Stain achieves superior performance and efficiency compared with both GAN- and diffusion-based baselines.

   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: We discuss the limitations of the work in Section 5.

   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 paper does not contain formal theorems or proofs, as it is primarily an applied method

   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 disclose all implementation details in Section 4.1.

   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: [Yes]

Justification: We will provide the data downloading link.

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: We disclose all implementation details in Section 4.1.

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:

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: In Section 4.1, we provide the information about the computer resources we use.

   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: The research does not raise ethical concerns related to privacy, bias, or misuse.

   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: [Yes]

Justification: We discuss the positive societal impacts in Section 5. As a research work in the field of medical image, we believe this paper will not have negative impacts on society.

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 poses no such 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 cited all the models used in the paper and stated their version.

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 paper does not employ large language models (LLMs) in any capacity.

    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.