CELLDISC: BOOSTING VIRTUAL STAINING VIA CELL SEGMENTATION-GUIDED DISCRIMINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Virtual staining aims to generate histochemically stained images from unstained or differently stained tissue images, yet existing methods primarily rely on image-level objectives without explicitly modeling cellular structures that are fundamental to histopathological interpretation. Moreover, rigorous evaluation remains challenging because commonly used datasets often rely on consecutive tissue sections, resulting in imperfect cell-level correspondence between source and target images. To address these limitations, we propose **CellDisc**, a simple yet effective **cell segmentation-guided discrimination framework** that introduces dense cellular supervision into adversarial learning. Instead of conventional binary discrimination, CellDisc jointly models cell localization and directional flow fields, encouraging the preservation of cellular structure and spatial organization. As a plug-and-play discriminator enhancement, CellDisc can be integrated into diverse paired GAN baselines without modifying their generators. We further establish **three strongly paired same-section datasets** covering H&E-to-P63, H&E-to-P63&CK, and FFPE-to-H&E transformations, enabling unified reference-based evaluation of pixel-wise, perceptual, cell-level, and stain-specific fidelity. Extensive experiments across multiple baselines and staining tasks demonstrate the broad compatibility and effectiveness of CellDisc. When integrated with P2P-res9, CellDisc achieves competitive or superior performance compared with recent GAN- and diffusion-based approaches, particularly in perceptual and cell-level fidelity. Ablation studies further demonstrate the benefit of directional flow fields and their complementarity with perceptual supervision, while evaluations using independent cell segmentation models consistently support the improvements in cellular fidelity. Finally, downstream cancer classification demonstrates that CellDisc preserves diagnostically relevant information across different classifier architectures. Overall, our results establish explicit cellular modeling as an effective and compatible strategy for improving structural fidelity in virtual staining. Code and data will be made publicly available.

1 INTRODUCTION

Virtual staining aims to computationally transform label-free or routinely stained tissue images into their histochemically or immunohistochemically stained counterparts, with the potential to reduce staining cost, tissue consumption, and laboratory workload while extracting additional morphological or molecular information from existing specimens Bai et al. (2023). Recent advances in image-to-image translation have led to diverse virtual staining approaches, spanning GAN- and diffusion-based generative methods Latonen et al. (2024); Klöckner et al. (2025). Despite this progress, two fundamental aspects remain insufficiently emphasized: *what information should be explicitly modeled during learning*, and *how different models can be evaluated rigorously and objectively*.

First, as illustrated in Fig. 1(a), virtual staining is intrinsically associated with single-cell morphology and staining patterns. For H&E staining, diagnostically relevant appearance is closely associated with nuclear morphology, including nuclear shape, size, chromatin pattern, and spatial organization Hu et al. (2025). For immunohistochemistry (IHC), staining patterns are further associated with cell-level marker status and subcellular localization: nuclear markers appear within positive nuclei, whereas membrane markers are localized around appropriate cellular boundaries Klöckner et al. (2025). These observations suggest that cellular morphology, localization, and spatial organiza-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

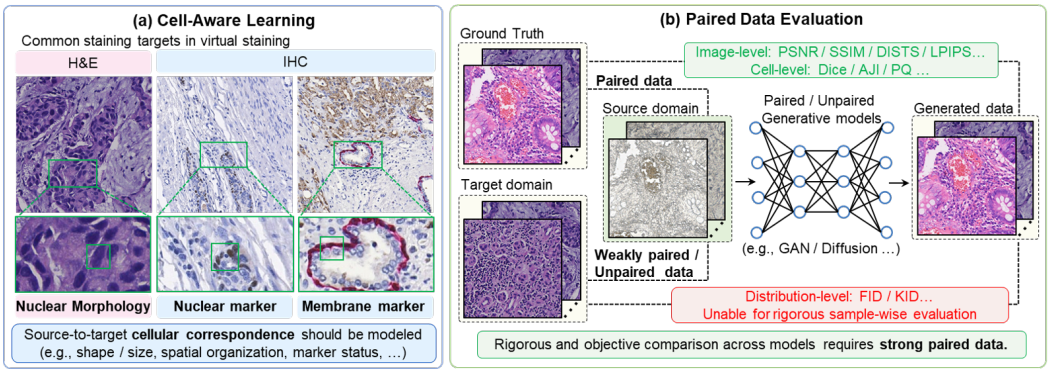

Figure 1: **Motivation. (a)** Cell-Aware Learning. Common staining targets are closely associated with cell-level morphology, marker status, and subcellular localization, highlighting the importance of **preserving cellular structure and spatial organization** during learning. **(b)** Paired Data Evaluation. Rigorous comparison across virtual staining models requires **strongly paired references**.

tion provide fundamental structural priors for virtual staining. Virtual staining should therefore preserve fine-grained cellular structures rather than merely reproduce global color or texture distributions. However, existing pathology-aware techniques, such as **pathology/foundation-model feature conditioning** Yang et al. (2025); Großkopf et al. (2025); Yellapragada et al. (2025), **protein-expression/pathological-semantic constraints** Chen et al. (2024; 2026), and **cross-image/feature correspondence mining** Guan et al. (2025); Peng et al. (2026); Andani et al. (2025), do not explicitly exploit dense cell-level structural information, potentially limiting their ability to preserve fine-grained cellular organization and reproduce cellular staining patterns.

Second, as illustrated in Fig. 1(b), rigorous evaluation of virtual staining raises a different but equally important consideration: paired reference data. This requirement is independent of whether a model is trained with paired or unpaired data, or whether its generator is based on GANs or diffusion models. **Widely adopted reference-based metrics (Appendix A.1)**, including PSNR and SSIM Hore & Ziou (2010), DISTS Ding et al. (2020), and LPIPS Zhang et al. (2018), compare a generated image with its corresponding real target and therefore rely on accurate spatial correspondence. **Cell-level evaluations**, such as Dice, AJI Kumar et al. (2017), and PQ Kirillov et al. (2019), also impose a stronger requirement because individual cellular structures must be compared at matched locations. However, **widely used public datasets (Appendix A.2)** such as BCI Liu et al. (2022), MIST Li et al. (2023), and IHC4BC Akbarnejad et al. (2025) provide only weakly matched references , limiting rigorous sample-wise evaluation Bian et al. (2024) and likely introducing metric-related pitfalls, as spatial misalignment can confound the measured discrepancies between generated and real images Reinke et al. (2024). Distribution-level metrics such as FID and KID characterize population-level similarity without requiring paired references, but cannot directly assess the faithful preservation of morphology or staining status. Consequently, strongly paired test data provide a common reference for more rigorous and objective comparison across heterogeneous virtual staining models.

Motivated by these observations, we introduce **CellDisc**, a simple yet effective framework for boosting virtual staining through **cell segmentation-guided discrimination**. CellDisc explicitly incorporates cellular information into adversarial learning to better preserve cellular morphology and spatial organization, while requiring minimal modification to existing paired GAN-based baselines. To enable rigorous reference-based evaluation, we further establish **three strongly paired same-section datasets**, covering FFPE-to-H&E, H&E-to-P63, and H&E-to-P63&CK transformations (Appendix A.2). The main contributions of this work are summarized as follows:

- We propose **CellDisc**, a cell segmentation-guided discrimination framework that introduces dense cellular supervision into adversarial learning by explicitly modeling cell localization and spatial organization. As a plug-and-play discriminator enhancement, CellDisc can be readily integrated into diverse paired GAN architectures without modifying the generator.
- We establish **three strongly paired same-section datasets** spanning FFPE-to-H&E, H&E-to-P63, and H&E-to-P63&CK transformations. Their strong spatial correspondence en-

ables unified reference-based evaluation across pixel-wise, perceptual, cell-level, and stain-specific dimensions.

- We conduct extensive experiments across multiple paired GAN architectures and three staining tasks, demonstrating the broad compatibility and effectiveness of CellDisc. Comparisons with recent GAN- and diffusion-based methods, independent cell segmentation models, and downstream cancer classification further support its improvements in cellular fidelity and preservation of diagnostically relevant information.

2 Related Work

2.1 Virtual Staining Models and Pathology-Aware Learning

Virtual staining has been widely formulated as an image-to-image translation problem. Early approaches were predominantly based on **generative adversarial networks (GANs)**, with paired methods largely following the conditional translation paradigm of Pix2Pix Isola et al. (2017). Subsequent variants introduced multi-scale Wang et al. (2018), pyramid Liu et al. (2022), and multi-branch Bian et al. (2024) architectures to improve high-resolution synthesis and structural preservation. Meanwhile, unpaired translation frameworks based on CycleGAN Zhu et al. (2017) and CUT Park et al. (2020) have also been widely explored, including StructuralCycleGAN Dubey et al. (2023), PC-StainGAN Liu et al. (2021), AI-FFPE Ozyoruk et al. (2022), SIM-GAN Guan et al. (2025), VirtualMultiplexer Pati et al. (2024), ASP Li et al. (2023), and PSPStain Chen et al. (2024), to reduce the dependence on accurately registered training pairs. Despite their reduced reliance on paired data training, unpaired models can underperform paired models when strongly paired data is available, particularly in structure-sensitive virtual staining tasks Hu et al. (2025).

More recently, **diffusion-based models** have emerged as an alternative generative paradigm for virtual staining. HistDiST Großkopf et al. (2025) employs latent diffusion with pathology-derived morphological conditioning, whereas D-VST Yang et al. (2025) adopts a diffusion Transformer with disentangled pathology and tone representations for cross-dye virtual staining. DiffVS Oh et al. (2026) further explores marker-wise conditioned diffusion for H&E-to-multiplex staining. Beyond pathology-specific models, recent general-purpose diffusion architectures such as JiT Li & He (2026) revisit pixel-space generation using simple image Transformers with direct clean-image prediction, providing a flexible generative backbone that can be adapted to virtual staining.

Beyond improving generative architectures, increasing attention has been devoted to incorporating **pathology-specific priors**. TDK-Stain Peng et al. (2024) introduces task-specific constraints on nuclei distribution and membrane staining for HER2 virtual staining. PSPStain Chen et al. (2024) and its extension PGVMS Chen et al. (2026) exploit protein-aware and pathological-semantic learning to improve molecular consistency. HistoPlexer Andani et al. (2025) predicts multiplex protein expression using customized pixel- and representation-level objectives. SIM-GAN Guan et al. (2025) and USIGAN Peng et al. (2026) further address imperfectly paired H&E–IHC data through semantic or feature correspondence modeling. Related studies have also incorporated pathology foundation-model representations or segmentation-derived priors to improve structural and semantic fidelity Großkopf et al. (2025); Yellapragada et al. (2025); Hu et al. (2025).

These approaches demonstrate the value of pathology-aware modeling, while most do not explicitly exploit cellular information. Among the few methods that incorporate cell-level cues, such as TDK-Stain and MCS-Stain, cellular information is primarily introduced on the generator side through semantic regularization, or architecture-dependent modules. A general discriminator-centric formulation that explicitly exploits cell-level information remains underexplored. In contrast, **CellDisc** introduces cell segmentation-guided discrimination, using dense cell-segmentation signals to enrich adversarial supervision while remaining readily compatible with diverse paired GAN baselines.

2.2 Paired Datasets and Reference-Based Evaluation in Virtual Staining

Several public datasets have therefore been developed to provide registered source–target image pairs. BCI Liu et al. (2022) contains registered H&E–HER2 image pairs, while MIST Li et al. (2023) extends this setting to multiple breast cancer biomarkers, including HER2, ER, PR, and Ki67. IHC4BC Akbarnejad et al. (2025) provides a similar but larger collection. Although registration can

substantially reduce global spatial mismatch at the WSI level, inherent section-to-section variations inevitably introduce local deformations, missing structures, and imperfect cellular correspondence at the patch level during training, motivating recent weakly paired methods that account for spatial inconsistencies Li et al. (2024); Ma et al. (2026).

Such incoherence also affects **reference-based evaluation** (Appendix A.1). PSNR measures pixel-wise intensity differences and is therefore highly sensitive to spatial misalignment. SSIM assesses local luminance, contrast, and structural similarity within sliding windows (typically 11×11), and likewise relies on local spatial correspondence Lin et al. (2025). Among perceptual metrics, DISTS Ding et al. (2020) measures structural and textural similarity across the input image and hierarchical representations, whereas LPIPS Zhang et al. (2018) computes perceptual distances from multi-level deep features. Although they are more tolerant to low-level appearance variations than PSNR/SSIM, substantial local misalignment can still introduce discrepancies unrelated to virtual staining quality. The requirement for accurate correspondence becomes even more stringent for cell-level evaluation: Dice measures overlap between corresponding cellular regions, whereas AJI Kumar et al. (2017) and PQ Kirillov et al. (2019) additionally account for individual cell instances.

Consequently, misaligned references may conflate virtual staining errors with registration artifacts, confounding objective comparison. **Strongly paired data** therefore provide a more reliable basis for fine-grained structural and cellular evaluation. To this end, our datasets are constructed from the same tissue sections and aligned using WSI-level rigid and patch-level non-rigid registration, yielding strong source–target spatial correspondence (Appendix A.2), enabling unified reference-based evaluation across pixel-wise, perceptual, cell-level, and stain-specific dimensions.

3 METHOD

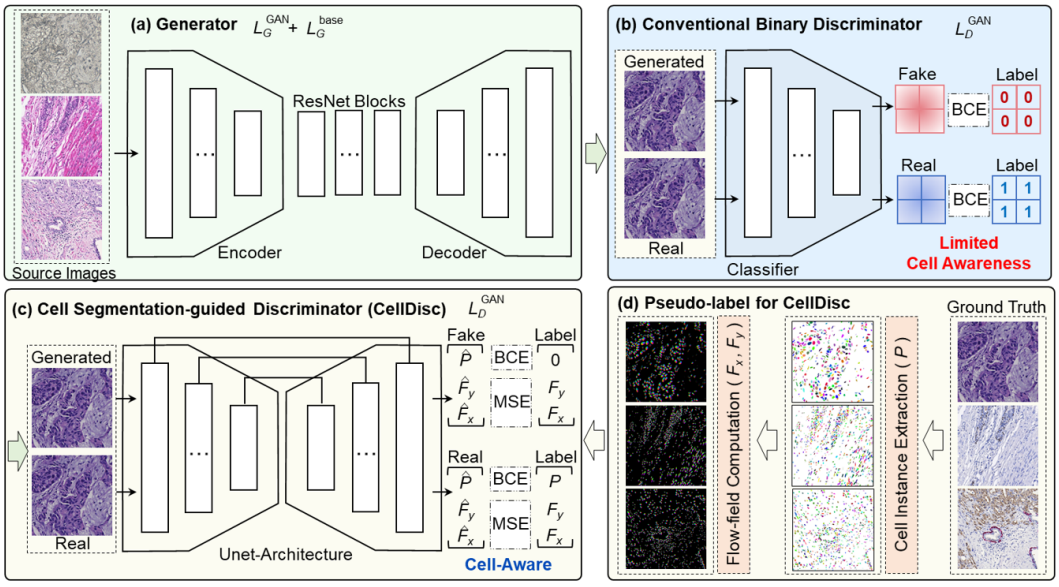

Figure 2: **Overview of CellDisc. (a)** Source-to-target image generation. **(b)** Conventional binary discrimination provides limited cell-level supervision. **(c)** CellDisc introduces cell segmentation-guided discrimination, where the predicted **cell probability map** $\hat{P}$ provides the adversarial real/fake contrast and the **flow-fields** $(\hat{F}_y, \hat{F}_x)$ encode cell-instance geometry and spatial organization. **(d) Pseudo-labels** for cell segmentation are derived from target images using CellposeSAM-based cell instance extraction and flow-field computation.

3.1 CELL SEGMENTATION-GUIDED DISCRIMINATION

As illustrated in Fig. 2(a), Virtual staining is commonly formulated as an image-to-image translation problem, where a generator G transforms a source image S into a fake target image $\hat{I} = G(S)$.

As illustrated in Fig. 2(b), existing GAN-based methods typically use a binary discriminator to distinguish $\hat{I}$ from the real image I, providing limited explicit guidance on cellular structures. As illustrated in Fig. 2(c), we propose **CellDisc**, which replaces binary discrimination with cell segmentation-guided discrimination. Specifically, a U-Net-style discriminator D_{cell} predicts $\hat{Y}$:

$$\hat{Y} = D_{\text{cell}}(I) = [\hat{P}, \hat{F}_y, \hat{F}_x]. \tag{1}$$

where P denotes the cell probability map, and (F_y, F_x) encode cell-instance spatial organization by representing, at each pixel, the direction toward the center of its corresponding cell.

We define discrimination labels for the real images I and the fake image $\hat{I}$ as

$$Y_{\text{real}} = [P, F_y, F_x], \qquad Y_{\text{fake}} = [0, F_y, F_x]. \tag{2}$$

Following CellposeSAM setting, CellDisc uses binary cross-entropy (BCE) for the cell probability map and mean squared error (MSE) for the directional flow fields:

$$\mathcal{L}_{\text{cell}} = \text{BCE}(\hat{P}, P) + \frac{1}{2} \sum_{i \in \{x,y\}} \text{MSE}(\hat{F}_i, F_i). \tag{3}$$

CellDisc is then optimized by:

$$\mathcal{L}_D^{\text{GAN}} = \mathbb{E}_{T_{\text{real}}} \left[\mathcal{L}_{\text{cell}}(\hat{Y}_{\text{real}}, Y_{\text{real}}) \right] + \mathbb{E}_{T_{\text{fake}}} \left[\mathcal{L}_{\text{cell}}(\hat{Y}_{\text{fake}}, Y_{\text{fake}}) \right]. \tag{4}$$

This asymmetric design assigns the adversarial real/fake contrast to the cell probability channel, while retaining the same flow fields for real and generated images to provide a shared representation of cellular geometry. During generator optimization, $\hat{Y}_{\text{fake}}$ is instead supervised toward Y_{real}, requiring T_{fake} to reproduce both cell localization and instance-level spatial organization:

$$\mathcal{L}_G^{\text{GAN}} = \mathbb{E}_{T_{\text{fake}}} \left[\mathcal{L}_{\text{cell}}(\hat{Y}_{\text{fake}}, Y_{\text{real}}) \right]. \tag{5}$$

The overall generator loss $\mathcal{L}_G$ is $\mathcal{L}_G^{\text{GAN}}$ together with the additional generator-side loss $\mathcal{L}_G^{\text{base}}$ of baselines. CellDisc therefore provides a dense cellular learning signal while modifying only the discriminator, making it readily compatible with various GAN-based virtual staining models.

3.2 Pseudo-label Generation

To enable CellDisc without manual annotation, we leverage CellposeSAM (Pachitariu et al., 2025) to automatically generate pseudo-labels from target images through cell instance extraction and flow-field computation, as illustrated in Fig. 2(d).

Cell Instance Extraction: Built upon the integration of Cellpose3.0 (Stringer & Pachitariu, 2025) and Segment Anything (SAM) (Kirillov et al., 2023), CellposeSAM exhibits strong generalization across diverse microscopy and pathology datasets without requiring task-specific retraining. Given a target image T_{real}, Let $\mathcal{C} = \{C_1, \ldots, C_N\}$ denote the resulting set of cell instances obtained by CellposeSAM. Pseudo-label of the cell probability map P is then defined as

$$P(x, y) = \begin{cases} 1, & (x, y) \in \bigcup_i C_i, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Flow-field Computation: For a pixel (x, y) belonging to the i-th cell with center (x_i^c, y_i^c), the displacement is

$$\Delta_x^{(i)} = x_i^c - x, \qquad \Delta_y^{(i)} = y_i^c - y. \tag{7}$$

The pseudo-label of (F_y, F_x) is then computed following the distance-diffusion formulation as

$$F_x(x, y) = \sum_i \Delta_x^{(i)} \exp\left(-d_i^2/\sigma^2\right), \qquad F_y(x, y) = \sum_i \Delta_y^{(i)} \exp\left(-d_i^2/\sigma^2\right). \tag{8}$$

where d_i is the distance from (x, y) to the center of the i-th cell and σ controls the diffusion radius.

4 EXPERIMENT

4.1 IMPLEMENTATION

Datasets. Detailed statistics are summarized in Table 1. We establish three paired virtual staining datasets covering distinct tissue types and staining transformations. **H&E-to-P63** consists of prostate cancer specimens and targets virtual IHC staining of basal cells with nuclear-localized P63 staining (brown). **H&E-to-P63&CK** is also derived from prostate cancer specimens and targets multiplex IHC generation, combining the same pattern of P63 staining with membranous-like CK(HMW) staining (red) in basal cells. **FFPE-to-H&E** consists of breast cancer specimens and aims to translate FFPE images into their H&E-stained counterparts. All datasets are constructed from paired WSIs acquired from the same tissue sections. Training and test sets are split at the patient level to prevent data leakage. Rigid registration is performed first between paired WSIs, followed by manual quality control to exclude poorly aligned regions. The aligned regions are then cropped into paired 1024×1024 patches. For the two H&E-to-IHC datasets, additional non-rigid registration using DeeperHistReg (Wodzinski et al., 2024) is applied to further improve local cellular correspondence (Appendix A.3).

Table 1: Summary of Datasets

Datasets	Train Pairs	Test Pairs
H&E-to-P63	10,856	5,398
H&E-to-P63&CK	10,171	4,281
FFPE-to-H&E	6,403	1,605

Training Details and Evaluation Metrics. All generative models are trained for 50 epochs with a batch size of 1. All experiments were conducted on GPUs with 80 GB of memory. We employ the widely used PSNR / SSIM for pixel-wise evaluation and DISTS / LPIPS for perceptual evaluation. We further use a general cell segmentation model **(CellposeSAM unless otherwise specified)** to evaluate cell-level Dice, AJI, and PQ between virtual and target images. To avoid circular evaluation, we also employ independent **Hover-net** Graham et al. (2019), **Cellvit** Hörst et al. (2024), and **Cellpose** Pachitariu & Stringer (2022) for evaluation on FFPE-to-H&E, as they are not applicable to IHC images. Following Hu et al. (2025); Kataria et al. (2025), we also perform stain-specific evaluation by decomposing images into corresponding stain channels and computing Dice scores: Dice-H and Dice-P63 for H&E-to-P63; Dice-H, Dice-P63, and Dice-CK for H&E-to-P63&CK; and Dice-H and Dice-E for FFPE-to-H&E. The **bold** and underlined values in tables indicate the **best** and second-best results, respectively.

Baselines for Evaluating CellDisc Compatibility. As described in Section 2.1, existing paired GANs largely follow the paradigm established by Pix2Pix Isola et al. (2017). We therefore integrate CellDisc with Pix2Pix (P2P) and its ResNet-9 variant (P2P-res9), PyramidPix2Pix Liu et al. (2022) (PyP2P) and its ResNet-9 variant (PyP2P-res9), and Pix2PixDualBranch Bian et al. (2024) (P2PDB). For all experiments, CellDisc adopts the commonly used U-Net256 architecture for its cell segmentation-guided discriminator. To examine the contribution of directional flow fields, we additionally evaluate CellDisc(w/o F), which retains only the cell-probability discrimination. Since CellDisc primarily imposes cellular structural constraints, we further investigate complementary perceptual supervision using the widely adopted VGG-based perceptual loss (denoted as $*$).

State-of-the-art (SOTA) Methods for Comparison. We compare P2P-res9 equipped with CellDisc and CellDisc* against recent diffusion-based methods, including D-VST Yang et al. (2025), DiffVS Oh et al. (2026), and JiT Li & He (2026), and GAN-based methods, including HistoPlexer Andani et al. (2025) and DGR Ma et al. (2026). We also include H&E-to-IHC-specific TDK-Stain Peng et al. (2024) and FFPE-to-H&E-specific MCS-Stain Hu et al. (2025) for their corresponding tasks.

Downstream Task Analysis. To assess whether virtual staining preserves diagnostically relevant information, we conduct a downstream cancer classification task on FFPE-to-H&E. The evaluation set contains 1,605 patches, including 661 cancer-positive and 944 cancer-negative samples annotated by an experienced pathologist. We perform five-fold cross-validation (train:val:test = 3:1:1). ResNet10, ResNet18, ResNet34, ResNet50, and EfficientNet classifiers are trained exclusively on real H&E images and evaluated on virtual H&E images under a *Real(train)-to-Virtual(test)* protocol. For SOTA comparison, we use P2P-res9, consistent with primary SOTA experiments. For ablation, we use P2PDB as the baseline for this analysis because the others exhibit performance saturation on this binary classification task, making differences among configurations less discriminative. All

classifiers are trained for 50 epochs with an initial learning rate of 1×10^{-4} and a CosineAnnealingLR scheduler. Accuracy (ACC) is reported as the mean $\pm$ standard deviation over five folds.

4.2 RESULTS

Compatibility of CellDisc. As shown in Fig. 3(a–c), integrating CellDisc generally improves the corresponding baselines across the three staining tasks and multiple evaluation dimensions, demonstrating its compatibility with different paired GAN baselines. The improvements are particularly consistent when combined with perceptual supervision (CellDisc*). Although CellDisc alone shows less consistent gains for P2PDB on FFPE-to-H&E, CellDisc* substantially improves its overall performance, indicating greater robustness across architectures and tasks. Importantly, improvements in Dice, AJI, and PQ are also observed using independent cell segmentation models, including Hover-net, Cellvit, and Cellpose (Fig. 3(d)), supporting the robustness of cell-level evaluation beyond CellposeSAM. Detailed results corresponding to Fig. 3 are provided in Appendix A.4.

Ablation of Structural and Perceptual Supervision. Comparing CellDisc(w/o F) with the full CellDisc shows that incorporating directional flow fields generally provides additional gains, supporting the benefit of modeling cellular spatial organization beyond cell localization. Moreover, the generally stronger performance of the $*$ variants indicate that VGG-based perceptual supervision complements the cellular structural constraints of CellDisc. Together, these results explain the stronger and more robust performance of CellDisc*. Qualitative comparisons further illustrate the improved preservation of cellular structures and stain patterns (Appendix A.5).

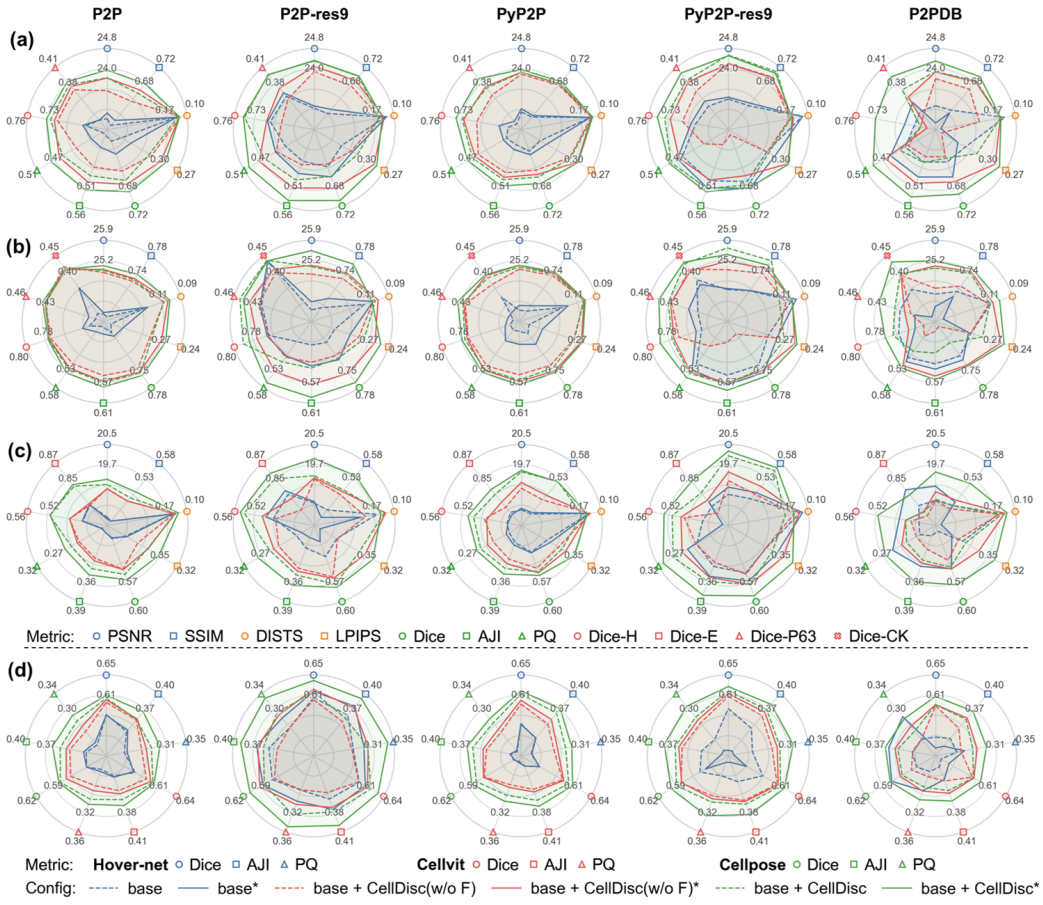

Figure 3: **Compatibility of CellDisc. (a–c)** Metrics on H&E-to-P63, H&E-to-P63&CK, and FFPE-to-H&E, respectively. **(d)** Independent cell segmentation metrics on FFPE-to-H&E.

Ablation on Downstream Task. As shown in Table 2, incorporating CellDisc consistently improves downstream cancer classification over the P2PDB baseline across all five classifier architectures.

The full CellDisc achieves the highest ACC for four of the five classifiers, while removing the directional flow fields generally reduces performance, supporting the contribution of cellular spatial information. Notably, perceptual supervision alone (P2PDB*) substantially degrades classification performance, whereas its combination with CellDisc maintains strong performance. These results suggest that perceptual similarity alone does not necessarily translate into diagnostic fidelity, while explicitly modeling cellular structure helps preserve diagnostically relevant information.

Table 2: Downstream cancer classification (ACC) of different configurations on FFPE-to-H&E.

Config	ResNet10	ResNet18	ResNet34	ResNet50	EfficientNet
P2PDB	0.837 ± 0.025	0.831 ± 0.022	0.822 ± 0.031	0.832 ± 0.027	0.852 ± 0.025
P2PDB*	0.622 ± 0.036	0.485 ± 0.111	0.570 ± 0.056	0.490 ± 0.078	0.568 ± 0.083
P2PDB + CellDisc(w/o F)	0.852 ± 0.020	0.860 ± 0.036	0.872 ± 0.041	0.854 ± 0.031	0.858 ± 0.023
P2PDB + CellDisc(w/o F)*	0.865 ± 0.012	0.870 ± 0.021	0.865 ± 0.048	0.860 ± 0.024	0.854 ± 0.023
P2PDB + CellDisc	**0.874 ± 0.024**	**0.879 ± 0.027**	**0.874 ± 0.034**	0.860 ± 0.027	**0.874 ± 0.015**
P2PDB + CellDisc*	0.866 ± 0.010	0.872 ± 0.020	0.873 ± 0.034	**0.865 ± 0.024**	0.870 ± 0.015
Real images	0.905 ± 0.013	0.900 ± 0.011	0.900 ± 0.012	0.895 ± 0.013	0.898 ± 0.012

Comparison with SOTA Methods. Results on the three datasets are shown in Tables 3–5. On H&E-to-P63 (Table 3), the CellDisc variants achieve the best results on eight of nine metrics, with particularly strong perceptual and cell-level performance. On H&E-to-P63&CK (Table 4), CellDisc* achieves the best DISTS, LPIPS, Dice, AJI, and PQ, while JiT performs better on pixel-wise and stain-specific metrics. On FFPE-to-H&E (Table 5), CellDisc* achieves the best PSNR, DISTS, LPIPS, Dice, and AJI, and matches the task-specific MCS-Stain in PQ. Overall, equipping P2P-res9 with CellDisc yields competitive performance across different staining transformations and evaluation dimensions, demonstrating that its effectiveness is not restricted to a specific task. Results using independent cell segmentation models are reported in Table 6. CellDisc* achieves the best or second-best performance across all Dice, AJI, and PQ, further supporting the robustness of the observed improvements beyond CellposeSAM-based evaluation. Qualitative comparisons are further provided in Appendix A.6, where CellDisc shows more consistent preservation of cellular structures and stain distributions, while several competing methods exhibit over-staining, blurred cellular details, or staining distortions across the three tasks.

Table 3: Comparison with SOTA Methods on H&E-to-P63.

Methods	Pixel-wise Metrics		Perceptual Metrics		Cell Segmentation Metrics			Stain-specific Metrics	
	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	Dice↑	AJI↑	PQ↑	Dice-H↑	Dice-P63↑
Diffusion-based									
D-VST (NeurIPS, 2025)	19.729	0.596	0.180	0.343	0.660	0.474	0.410	0.708	0.372
DiffVS (AAAI, 2026)	17.460	0.517	0.267	0.436	0.448	0.257	0.106	0.575	0.323
JiT (CVPR, 2026)	22.524	0.577	0.258	0.413	0.510	0.322	0.185	0.638	0.343
GAN-based									
TDK-Stain (MICCAI, 2024)	22.364	0.595	**0.137**	0.340	0.650	0.475	0.422	0.698	0.335
HistoPlexer (Nat. Mach. Intell., 2025)	20.753	0.475	0.315	0.477	0.141	0.071	0.022	0.561	0.242
DGR (Nat. Commun., 2026)	18.468	0.322	0.203	0.444	0.401	0.240	0.091	0.498	0.114
P2P-res9 + CellDisc	**24.538**	**0.707**	0.164	0.318	0.675	0.517	0.479	0.746	**0.397**
P2P-res9 + CellDisc*	24.516	0.704	0.159	**0.286**	**0.680**	**0.524**	**0.483**	**0.747**	0.396

Table 4: Comparison with SOTA Methods on H&E-to-P63&CK.

Methods	Pixel-wise Metrics		Perceptual Metrics		Cell Segmentation Metrics			Stain-specific Metrics		
	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	Dice↑	AJI↑	PQ↑	Dice-H↑	Dice-P63↑	Dice-CK↑
Diffusion-based										
D-VST (NeurIPS, 2025)	20.932	0.668	0.169	0.321	0.706	0.510	0.450	0.738	0.380	0.314
DiffVS (AAAI, 2026)	18.470	0.658	0.252	0.389	0.585	0.364	0.275	0.694	0.377	0.427
JiT (CVPR, 2026)	**25.865**	**0.769**	0.170	0.309	0.746	0.562	0.530	**0.792**	**0.463**	**0.472**
GAN-based										
TDK-Stain (MICCAI, 2024)	23.580	0.669	0.115	0.314	0.718	0.522	0.469	0.738	0.390	0.389
HistoPlexer (Nat. Mach. Intell., 2025)	21.258	0.527	0.369	0.482	0.030	0.017	0.009	0.538	0.283	0.308
DGR (Nat. Commun., 2026)	20.080	0.534	0.185	0.381	0.486	0.288	0.163	0.556	0.227	0.072
P2P-res9 + CellDisc	25.657	0.766	0.121	0.272	0.745	0.575	0.537	0.787	0.454	0.437
P2P-res9 + CellDisc*	25.447	0.756	**0.106**	**0.255**	**0.760**	**0.589**	**0.549**	0.780	0.448	0.435

Comparison with SOTA on Downstream Task. Under the same downstream evaluation protocol (Table 7), the CellDisc variants demonstrate strong and stable downstream performance across all five classifier architectures, generally outperforming the other general-purpose GAN- and diffusion-based methods. Notably, CellDisc remains competitive with MCS-Stain, which is specifically de-

Table 5: Comparison with SOTA Methods on FFPE-to-H&E.

Methods	Pixel-wise Metrics		Perceptual Metrics		Cell Segmentation Metrics			Stain-specific Metrics	
	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	Dice↑	AJI↑	PQ↑	Dice-H↑	Dice-E↑
Diffusion-based									
D-VST (NeurIPS, 2025)	15.612	0.365	0.152	0.407	0.455	0.252	0.152	0.370	0.790
DiffVS (AAAI, 2026)	18.404	0.462	0.268	0.445	0.462	0.269	0.180	0.460	0.830
JiT (CVPR, 2026)	19.467	0.512	0.205	0.388	0.523	0.323	0.225	0.502	0.851
GAN-based									
HistoPlexer (Nat. Mach. Intell., 2025)	17.006	0.330	0.350	0.524	0.323	0.168	0.080	0.376	0.804
DGR (Nat. Commun., 2026)	13.444	0.175	0.309	0.556	0.091	0.041	0.003	0.112	0.627
MCS-Stain (IEEE TMI, 2026)	19.995	**0.578**	0.159	0.339	0.592	0.381	**0.304**	**0.554**	**0.864**
P2P-res9 + CellDisc	20.044	0.559	0.178	0.365	0.573	0.361	0.281	0.541	0.853
P2P-res9 + CellDisc*	**20.223**	0.567	**0.148**	**0.330**	**0.593**	**0.383**	**0.304**	0.548	0.859

Table 6: Comparison with SOTA using independent H&E cell segmentation on FFPE-to-H&E.

Methods	Hover-net			Cellvit			Cellpose		
	Dice	AJI	PQ	Dice	AJI	PQ	Dice	AJI	PQ
Diffusion-based									
D-VST (NeurIPS, 2025)	0.511	0.275	0.184	0.501	0.273	0.186	0.473	0.262	0.170
DiffVS (AAAI, 2026)	0.501	0.260	0.187	0.532	0.300	0.226	0.494	0.273	0.190
JiT (CVPR, 2026)	0.582	0.321	0.244	0.588	0.356	0.276	0.547	0.329	0.240
GAN-based									
HistoPlexer (Nat. Mach. Intell., 2025)	0.460	0.226	0.113	0.444	0.219	0.111	0.461	0.217	0.102
DGR (Nat. Commun., 2026)	0.119	0.053	0.005	0.116	0.050	0.004	0.110	0.047	0.003
MCS-Stain (IEEE TMI, 2026)	**0.641**	0.390	**0.339**	**0.636**	**0.405**	0.348	0.607	0.387	0.324
P2P-res9 + CellDisc	0.617	0.352	0.296	0.620	0.381	0.324	0.588	0.357	0.291
P2P-res9 + CellDisc*	0.638	**0.392**	**0.339**	0.633	0.402	**0.351**	**0.610**	**0.390**	**0.328**

signed for FFPE-to-H&E virtual staining, and achieves comparable ACC across multiple classifiers. Together with the image- and cell-level evaluations, these results suggest that the structural fidelity promoted by CellDisc translates into the preservation of diagnostically relevant information.

Table 7: Downstream cancer classification (ACC) of Comparison with SOTA on FFPE-to-H&E.

Methods	ResNet10	ResNet18	ResNet34	ResNet50	EfficientNet
Diffusion-based					
D-VST (NeurIPS, 2025)	0.795 ± 0.052	0.823 ± 0.052	0.821 ± 0.032	0.836 ± 0.023	0.827 ± 0.034
DiffVS (AAAI, 2026)	0.733 ± 0.136	0.811 ± 0.053	0.770 ± 0.079	0.803 ± 0.046	0.710 ± 0.051
JiT (CVPR, 2026)	0.840 ± 0.035	0.851 ± 0.035	0.820 ± 0.063	0.822 ± 0.032	0.840 ± 0.050
GAN-based					
DGR (Nat. Commun., 2026)	0.687 ± 0.058	0.617 ± 0.058	0.630 ± 0.074	0.597 ± 0.093	0.592 ± 0.095
HistoPlexer (Nat. Mach. Intell., 2025)	0.844 ± 0.030	0.760 ± 0.030	0.755 ± 0.066	0.682 ± 0.106	0.760 ± 0.123
MCS-Stain (IEEE TMI, 2026)	**0.872 ± 0.034**	**0.887 ± 0.018**	**0.882 ± 0.029**	**0.844 ± 0.019**	**0.874 ± 0.034**
P2P-res9 + CellDisc	0.854 ± 0.034	0.874 ± 0.021	0.862 ± 0.038	**0.844 ± 0.020**	0.846 ± 0.027
P2P-res9 + CellDisc*	0.856 ± 0.027	0.865 ± 0.025	0.864 ± 0.030	0.825 ± 0.017	0.861 ± 0.037
Real images	0.905 ± 0.013	0.900 ± 0.011	0.900 ± 0.012	0.895 ± 0.013	0.898 ± 0.012

5 CONCLUSION

In this work, we present CellDisc, a simple yet effective cell segmentation-guided discrimination framework that explicitly incorporates cellular localization and spatial organization into adversarial learning for virtual staining. We further establish three strongly paired same-section datasets covering H&E-to-P63, H&E-to-P63&CK, and FFPE-to-H&E transformations, enabling comprehensive evaluation of pixel-wise, perceptual, cell-level, and stain-specific fidelity. Extensive experiments across multiple paired GAN baselines and staining tasks demonstrate the broad compatibility and effectiveness of CellDisc, while ablation studies confirm the benefit of directional flow fields and the complementarity of perceptual supervision. Comparisons with recent GAN- and diffusion-based methods demonstrate competitive or superior performance, particularly in perceptual and cell-level fidelity. These findings are further supported by independent cell segmentation evaluation and downstream cancer classification, where CellDisc preserves strong performance across multiple classifiers. Overall, our results highlight explicit cellular modeling as an effective strategy for improving structural fidelity while preserving diagnostically relevant information in virtual staining.

AI USE STATEMENT

Generative AI tools were used solely for language editing and polishing to improve the clarity, grammar, and readability of this manuscript. They were not used to generate research ideas, develop methodologies, conduct experiments, analyze data, or draw scientific conclusions. All AI-assisted edits were reviewed and verified by the authors, who take full responsibility for this work.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of CellDisc, including its architecture, training objectives, and integration with paired GAN baselines, in the main paper. Dataset construction, registration, preprocessing, training and inference settings, evaluation protocols, and baseline implementations are also described in the corresponding Methods and Implementation sections. Additional experimental results and analyses are provided in the Appendix. The source code and paired datasets will be made publicly available upon publication to facilitate reproducibility.

REFERENCES

Amir Akbarnejad, Nilanjan Ray, Penny J Barnes, and Gilbert Bigras. Toward accurate deep learning-based prediction of ki67, er, pr, and her2 status from h&e-stained breast cancer images. *Applied Immunohistochemistry & Molecular Morphology*, 33(3):131–141, 2025.

Sonali Andani, Boqi Chen, Joanna Ficek-Pascual, Simon Heinke, Ruben Casanova, Bernard Friedrich Hild, Bettina Sobottka, Bernd Bodenmiller, Viktor H Koelzer, et al. Histopathology-based protein multiplex generation using deep learning. *Nature Machine Intelligence*, 7(8):1292–1307, 2025.

Bijie Bai, Xilin Yang, Yuzhu Li, Yijie Zhang, Nir Pillar, and Aydogan Ozcan. Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications*, 12(1):57, 2023.

Chang Bian, Beth Phillips, Tim Cootes, and Martin Fergie. Hemit: H&e to multiplex-immunohistochemistry image translation with dual-branch pix2pix generator. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 184–197. Springer, 2024.

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*, pp. 384–394. Springer, 2024.

Fuqiang Chen, Ranran Zhang, Wanming Hu, Deboch Eyob Abera, Yue Peng, Boyun Zheng, Yiwen Sun, Jing Cai, and Wenjian Qin. Pgvms: A prompt-guided unified framework for virtual multiplex ihc staining with pathological semantic learning. *IEEE Transactions on Medical Imaging*, 2026.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2567–2581, 2020.

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*, pp. 447–456. Springer, 2023.

Simon Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir Rajpoot. Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. *Medical image analysis*, 58:101563, 2019.

Erik Großkopf, Valay Bundele, Mehran Hosseinzadeh, and Hendrik PA Lensch. Histdist: Histopathological diffusion-based stain transfer. In *DAGM German Conference on Pattern Recognition*, pp. 28–40. Springer, 2025.

Xianchao Guan, Zheng Zhang, Yifeng Wang, Yueheng Li, and Yongbing Zhang. Supervised information mining from weakly paired images for breast ihc virtual staining. *IEEE Transactions on Medical Imaging*, 2025.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Fabian Hörst, Moritz Rempe, Lukas Heine, Constantin Seibold, Julius Keyl, Giulia Baldini, Selma Ugurel, Jens Siveke, Barbara Grünwald, Jan Egger, et al. Cellvit: Vision transformers for precise cell segmentation and classification. *Medical Image Analysis*, 94:103143, 2024.

Yihuang Hu, Zhicheng Du, Weiping Lin, Shurong Yang, Lequan Yu, Guojun Zhang, and Liansheng Wang. Mcs-stain: Boosting ffpe-to-he virtual staining with multiple cell semantics. *IEEE Transactions on Medical Imaging*, 2025.

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*, pp. 1125–1134, 2017.

Tushar Kataria, Shikha Dubey, Mary Bronner, Jolanta Jedrzkiewicz, Ben J Brintz, Shireen Y Elhabian, and Beatrice S Knudsen. Building trust in virtual immunohistochemistry: Automated assessment of image quality. *arXiv preprint arXiv:2511.04615*, 2025.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9404–9413, 2019.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Pascal Klöckner, José Teixeira, Diana Montezuma, João Fraga, Hugo M Horlings, Jaime S Cardoso, and Sara P Oliveira. H&e to ihc virtual staining methods in breast cancer: an overview and benchmarking. *Npj digital medicine*, 8(1):384, 2025.

Neeraj Kumar, Ruchika Verma, Sanuj Sharma, Surabhi Bhargava, Abhishek Vahadane, and Amit Sethi. A dataset and a technique for generalized nuclear segmentation for computational pathology. *IEEE transactions on medical imaging*, 36(7):1550–1560, 2017.

Leena Latonen, Sonja Koivukoski, Umair Khan, and Pekka Ruusuvuori. Virtual staining for histology by deep learning. *Trends in Biotechnology*, 2024.

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*, pp. 632–641. Springer, 2023.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36115–36125, 2026.

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.

Weiping Lin, Yihuang Hu, Runchen Zhu, Baoshun Wang, and Liansheng Wang. Virtual staining for pathology: Challenges, limitations and perspectives. *Intelligent Oncology*, 1(2):105–119, 2025.

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*, pp. 1815–1824, 2022.

Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired stain transfer using pathology-consistent constrained generative adversarial networks. *IEEE Transactions on Medical Imaging*, 40(8):1977–1989, 2021.

Jiabo Ma, Wenqiang Li, Jinbang Li, Ziyi Liu, Linshan Wu, Fengtao Zhou, Li Liang, Ronald Cheong Kin Chan, Terence TW Wong, and Hao Chen. Generative ai for misalignment-resistant virtual staining to accelerate histopathology workflows. *Nature Communications*, 2026.

Hyun-Jic Oh, Junsik Kim, Zhiyi Shi, Yichen Wu, Yu-An Chen, Peter K Sorger, Hanspeter Pfister, and Won-Ki Jeong. Virtual multiplex staining for histological images using a marker-wise conditioned diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 8168–8176, 2026.

Kutsev Bengisu Ozyoruk, Sermet Can, Berkan Darbaz, Kayhan Başak, Derya Demir, Guliz Irem Gokceler, Gurdeniz Serin, Uguray Payam Hacisalihoglu, Emirhan Kurtuluş, Ming Y Lu, et al. A deep-learning model for transforming the style of tissue images from cryosectioned to formalin-fixed and paraffin-embedded. *Nature Biomedical Engineering*, 6(12):1407–1419, 2022.

Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature methods*, 19(12):1634–1641, 2022.

Marius Pachitariu, Michael Rariden, and Carsen Stringer. Cellpose-sam: superhuman generalization for cellular segmentation. *bioRxiv*, pp. 2025–04, 2025.

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*, pp. 319–345. Springer, 2020.

Pushpak Pati, Sofia Karkampouna, Francesco Bonollo, Eva Compérat, Martina Radić, Martin Spahn, Adriano Martinelli, Martin Wartenberg, Marianna Kruithof-de Julio, and Marianna Rapsomaniki. Accelerating histopathology workflows with generative ai-based virtually multiplexed tumour profiling. *Nature Machine Intelligence*, 6(9):1077–1093, 2024.

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*, pp. 3–13. Springer, 2024.

Yue Peng, Bing Xiong, Fuqiang Chen, Deboch Eybo Abera, RanRan Zhang, Wanming Hu, Jing Cai, and Wenjian Qin. Usigan: Unbalanced self-information feature transport for weakly paired image ihc virtual staining. *IEEE Transactions on Image Processing*, 2026.

Annika Reinke, Minu D Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A Emre Kavur, Tim Rädsch, Carole H Sudre, Laura Acion, Michela Antonelli, et al. Understanding metric-related pitfalls in image analysis validation. *Nature methods*, 21(2):182–194, 2024.

Carsen Stringer and Marius Pachitariu. Cellpose3: one-click image restoration for improved cellular segmentation. *Nature methods*, 22(3):592–599, 2025.

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*, pp. 8798–8807, 2018.

Marek Wodzinski, Niccolo Marini, Manfredo Atzori, and Henning Müller. Deeperhistreg: robust whole slide images registration framework. *arXiv preprint arXiv:2404.14434*, 2024.

Shurong Yang, Dong Wei, Yihuang Hu, Qiong Peng, Hong Liu, Yawen Huang, Xian Wu, Yefeng Zheng, and Liansheng Wang. D-vst: diffusion transformer for pathology-correct tone-controllable cross-dye virtual staining of whole slide images. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Srikar Yellapragada, Alexandros Graikos, Zilinghan Li, Kostas Triaridis, Varun Belagali, Saarthak Kapse, Tarak Nath Nandi, Ravi K Madduri, Prateek Prasanna, Tahsin Kurc, et al. Pixcell: A generative foundation model for digital histopathology images. *ArXiv*, pp. arXiv–2506, 2025.

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*, pp. 586–595, 2018.

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*, pp. 2223–2232, 2017.

A Appendix

A.1 Reference-Based Metrics and Paired Evaluation

Reference-based image quality metrics quantify the discrepancy between a generated image $\hat{I}$ and its corresponding real target I. Although these metrics can mathematically be computed between arbitrary image pairs, their use for sample-wise virtual staining evaluation implicitly assumes that $\hat{I}$ and I depict the same underlying tissue structures. Otherwise, the measured discrepancy may reflect not only virtual staining errors, but also spatial misalignment or biological differences between the compared images. Below, we summarize the reference-based metrics used in this work and discuss their dependence on source–target correspondence.

Peak Signal-to-Noise Ratio (PSNR). PSNR is derived from the pixel-wise mean squared error (MSE). For two images I and $\hat{I}$ with N pixels (including channels), the MSE is defined as

$$\mathrm{MSE}(I, \hat{I}) = \frac{1}{N} \sum_{i=1}^{N} \left(I_i - \hat{I}_i \right)^2, \tag{9}$$

and PSNR is computed as

$$\mathrm{PSNR}(I, \hat{I}) = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}(I, \hat{I})} \right), \tag{10}$$

where L denotes the maximum possible image intensity. Because PSNR directly compares intensities at corresponding pixel locations, even a small spatial displacement of an otherwise correctly reproduced nucleus or tissue boundary can substantially increase the MSE. Therefore, accurate spatial correspondence is particularly important for interpreting PSNR in virtual staining.

Structural Similarity Index Measure (SSIM). SSIM compares local luminance, contrast, and structural information between two images. For local image regions x and y, SSIM is defined as

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x \mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{11}$$

where μ_x and μ_y denote local means, σ_x^2 and σ_y^2 denote local variances, and σ_{xy} denotes local covariance. C_1 and C_2 are stabilization constants. In our evaluation, these local statistics are computed using an 11×11 Gaussian window. The final SSIM is obtained by averaging the local measurements across the image.

Unlike PSNR, SSIM tolerates moderate intensity variations, but its local comparison still assumes that corresponding structures fall within matched spatial neighborhoods. For example, displacement of a nucleus beyond or within the local window changes the estimated means, variances, and cross-covariance even when its morphology is otherwise faithfully generated. Consequently, local registration errors can directly affect the measured structural similarity.

Deep Image Structure and Texture Similarity (DISTS). DISTS evaluates perceptual similarity using hierarchical deep features while separating structure- and texture-related comparisons Ding et al. (2020). Let $\phi_l(I) \in \mathbb{R}^{C_l \times H_l \times W_l}$ denote the feature map at layer l. For channel c, the similarity of feature means is defined as

$$l_{lc} = \frac{2\mu_{lc}^{I} \mu_{lc}^{\hat{I}} + c_1}{(\mu_{lc}^{I})^2 + (\mu_{lc}^{\hat{I}})^2 + c_1}, \tag{12}$$

and the structure-related similarity is

$$s_{lc} = \frac{2\sigma_{lc}^{I,\hat{I}} + c_2}{(\sigma_{lc}^{I})^2 + (\sigma_{lc}^{\hat{I}})^2 + c_2}, \tag{13}$$

where μ, σ^2, and $\sigma^{I,\hat{I}}$ denote the corresponding feature statistics. DISTS aggregates these similarities across layers and channels as

$$\mathrm{DISTS}(I, \hat{I}) = 1 - \sum_{l,c} \left(\alpha_{lc} l_{lc} + \beta_{lc} s_{lc}\right), \tag{14}$$

where α_{lc} and β_{lc} are learned aggregation weights.

Because DISTS operates on multi-level feature representations rather than raw pixels, it is generally more tolerant to low-level appearance variation and small geometric perturbations than PSNR or SSIM. Nevertheless, it still compares a generated image against a specific reference image. Structural differences arising from imperfect registration or different tissue sections therefore contribute to the measured distance and may be indistinguishable from actual virtual staining errors.

Learned Perceptual Image Patch Similarity (LPIPS). LPIPS measures perceptual distance using normalized deep features extracted from a pretrained network Zhang et al. (2018). Let $\hat{\phi}_l(I)$ denote channel-normalized features at layer l. LPIPS can be expressed as

$$\mathrm{LPIPS}(I, \hat{I}) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \left\| \mathbf{w}_l \odot \left(\hat{\phi}_l(I)_{hw} - \hat{\phi}_l(\hat{I})_{hw} \right) \right\|_2^2, \tag{15}$$

where $\mathbf{w}_l$ contains learned channel-wise weights and $\odot$ denotes element-wise multiplication. Although deep representations make LPIPS less sensitive than pixel-wise metrics to small intensity changes, features are still compared at corresponding spatial locations. Thus, displacement, deformation, or absence of matched cellular structures can increase LPIPS even when such differences originate from imperfect pairing rather than the virtual staining model itself.

Implications for virtual staining evaluation. These formulations illustrate that reference-based metrics evaluate different aspects of fidelity but all rely, to varying degrees, on correspondence between the generated image and its real reference. This requirement is strongest for pixel-wise metrics such as PSNR and remains important for local structural metrics such as SSIM. DISTS and LPIPS provide greater tolerance to low-level appearance and modest geometric variations, but they are not invariant to structural mismatch.

This distinction is particularly important in histopathology. When source and target images originate from adjacent or consecutive tissue sections, individual nuclei, glands, or other microscopic structures may be displaced, deformed, newly present, or absent altogether. In such cases, the measured distance between a virtual image and its nominal target reflects a mixture of model-induced error and inherent source–target mismatch. Consequently, accurately paired test data do not merely facilitate paired model training; they provide a more reliable reference for objective and comparable sample-wise evaluation across different virtual staining methods.

A.2 Visual Assessment of Dataset Pairing Quality

Figure A1 shows source–target overlays for our three datasets. The close alignment of tissue structures and cellular morphology across all three tasks visually confirms the high pairing quality used for paired training and reference-based evaluation.

Figs. A2–A5 qualitatively compare the pairing quality of different H&E–IHC datasets using representative image pairs and CellposeSAM segmentations. Our datasets are constructed from the same tissue sections and aligned using WSI-level rigid and patch-level non-rigid registration, resulting in close tissue- and cell-level correspondence (Fig. A2). In contrast, BCI, MIST, and IHC4BC are based on consecutive or adjacent sections, where intrinsic section-to-section differences lead to local structural and cellular mismatches despite registration and pair selection (Figs. A3–A5). These examples illustrate the distinction between strongly and weakly paired datasets.

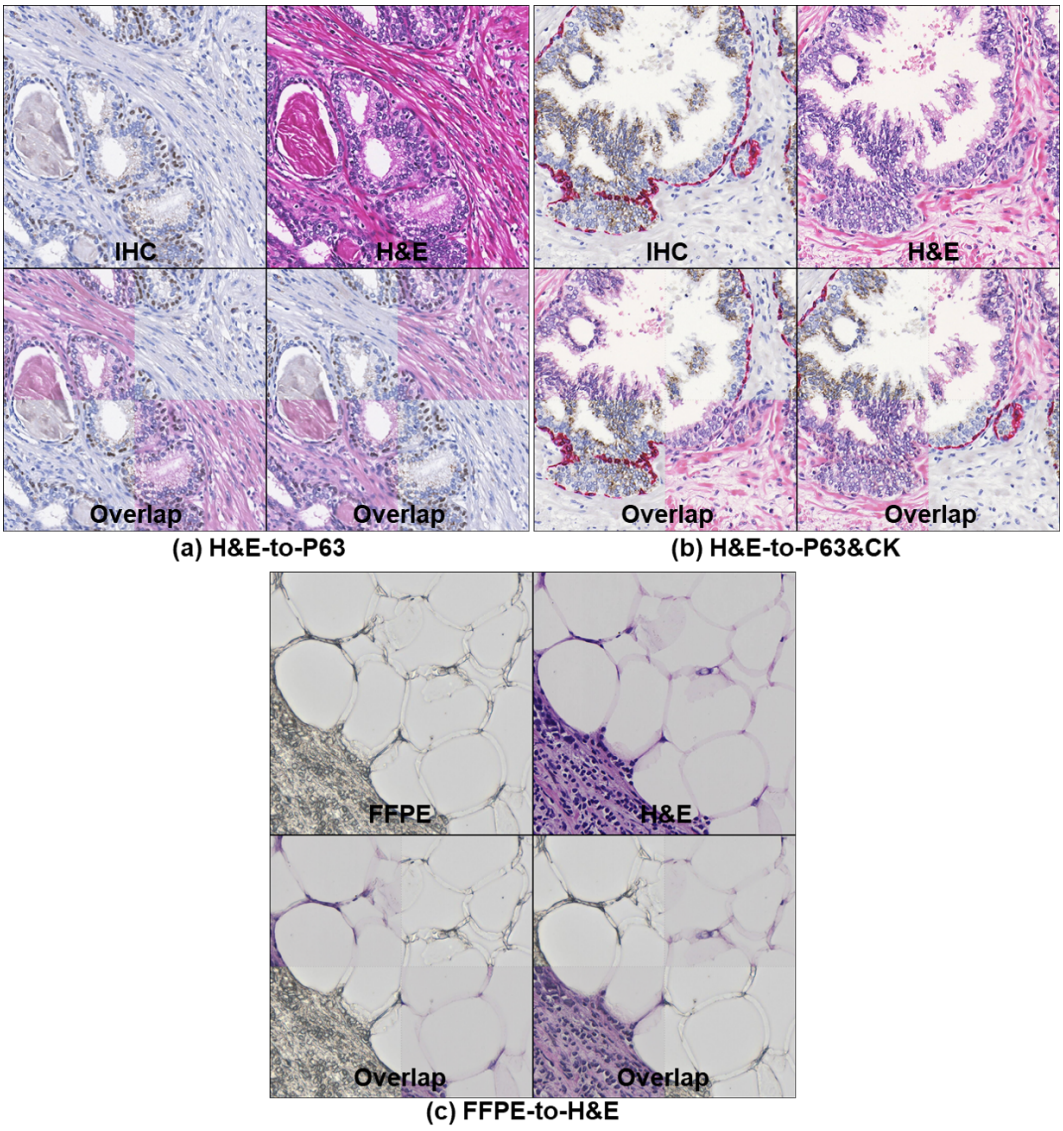

Figure A1: Representative paired images from **our three datasets**. Source and target images and their **overlays** are shown for H&E-to-P63 (left), H&E-to-P63&CK (right), and FFPE-to-H&E (bottom), illustrating the spatial correspondence of tissue and cellular structures within each pair.

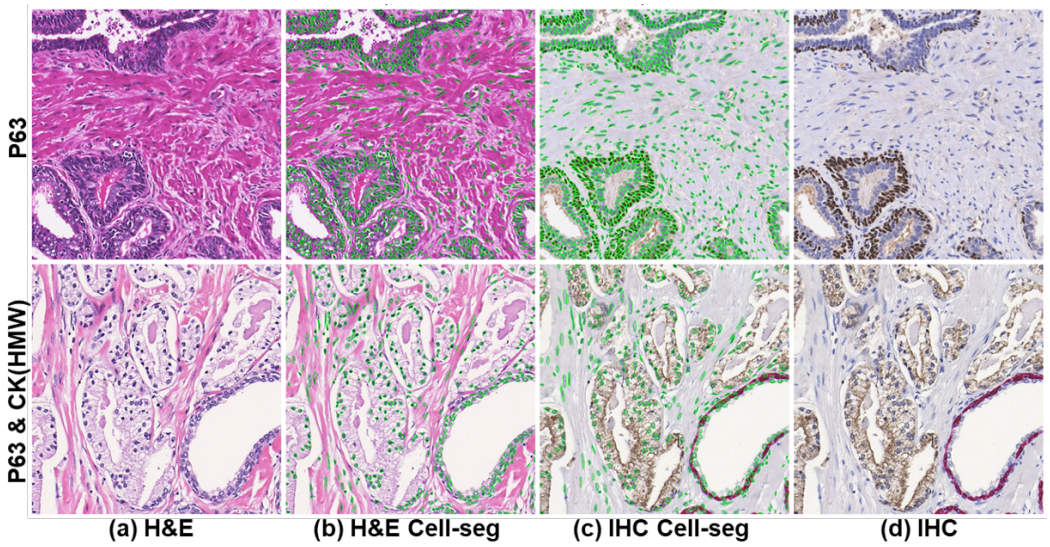

Figure A2: Representative H&E–IHC pairs from **our two H&E-to-IHC datasets** and their CellposeSAM segmentation results. The H&E and IHC images are derived from the same tissue section and aligned through WSI-level rigid registration followed by patch-level non-rigid registration. CellposeSAM segmentations (green) highlight the close correspondence of cellular structures between the paired H&E and IHC images.

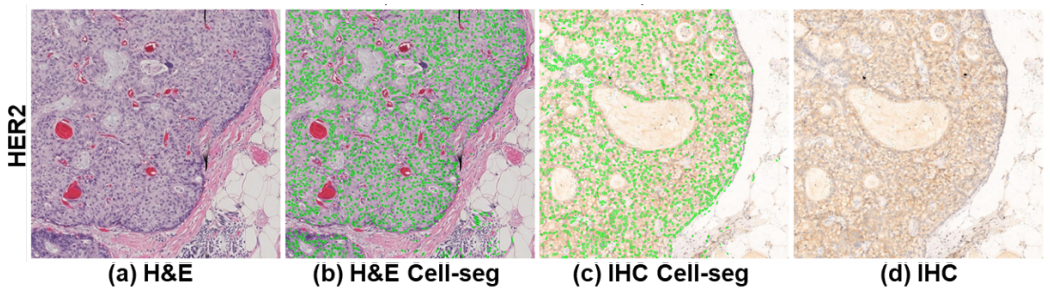

Figure A3: Representative H&E–IHC pairs from **BCI** Liu et al. (2022) dataset and their CellposeSAM segmentations results. Despite registration and pair selection, the use of consecutive tissue sections results in local structural differences between paired H&E and IHC images. CellposeSAM segmentations (green) further illustrate imperfect cell-level correspondence between the two modalities.

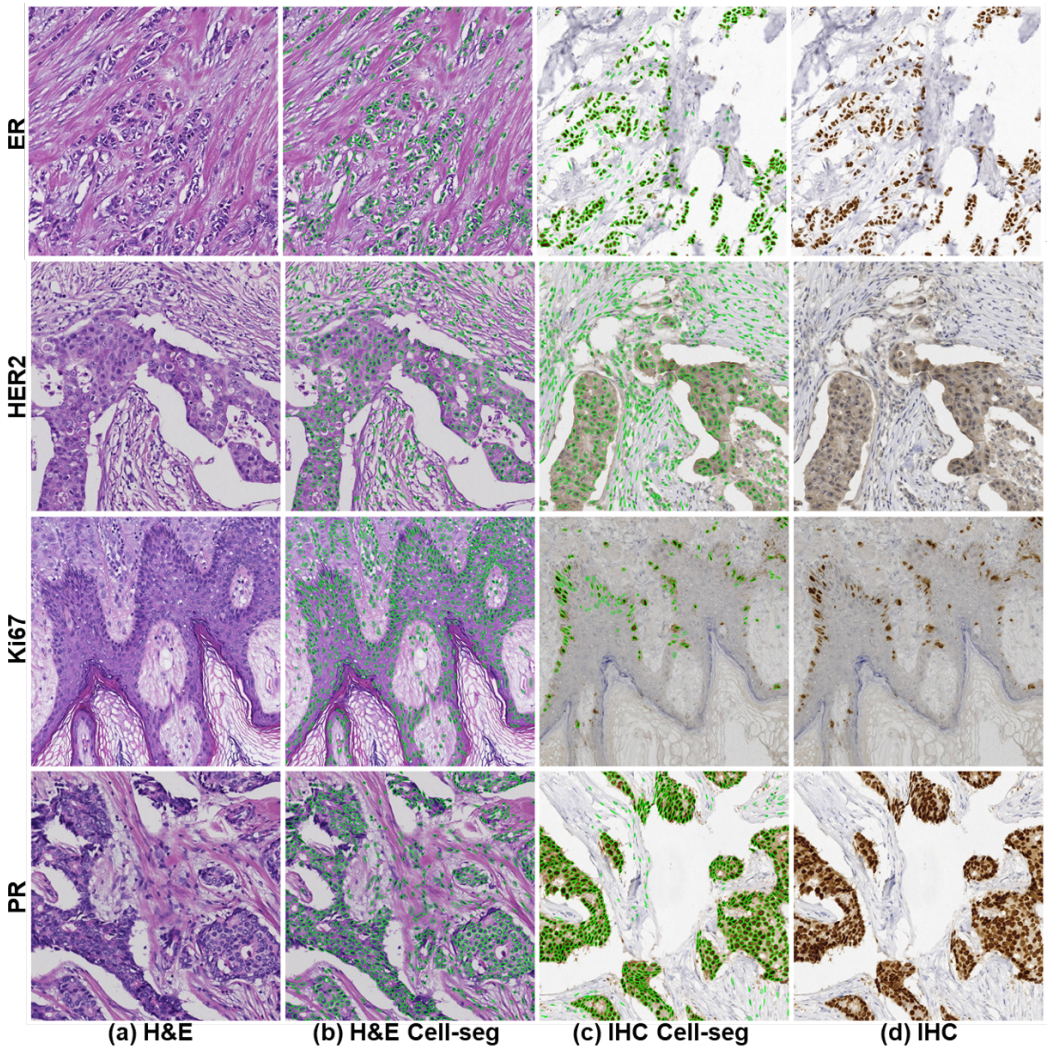

Figure A4: Representative H&E–IHC pairs from **MIST** Li et al. (2023) dataset and their CellposeSAM segmentation results. The paired images are derived from consecutive tissue sections, resulting in local differences in tissue morphology and cellular composition that remain after registration. CellposeSAM segmentations (green) visualize the resulting imperfect correspondence of individual cellular structures.

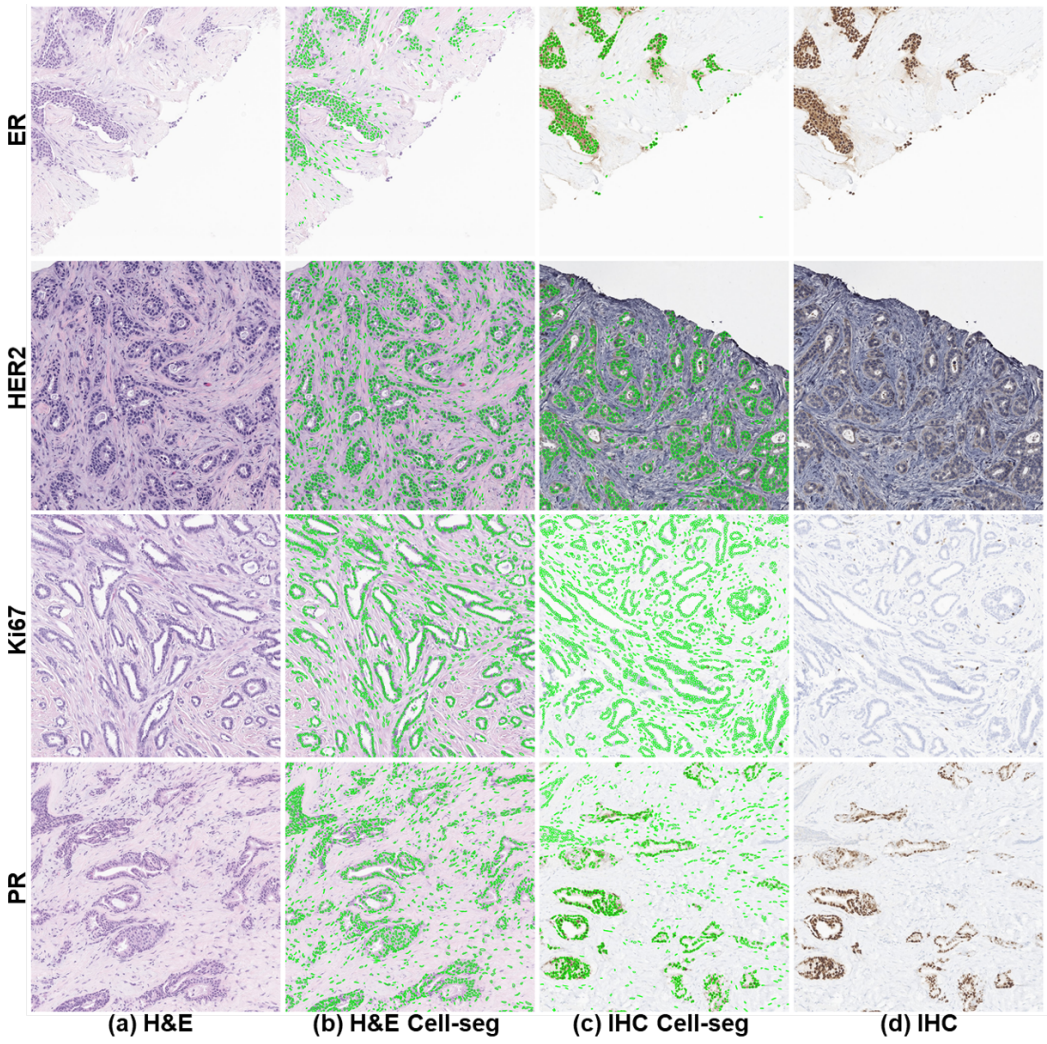

Figure A5: Representative H&E–IHC pairs from **IHC4BC** Akbarnejad et al. (2025) dataset and their CellposeSAM segmentation results. Although registration reduces global spatial mismatch, the use of consecutive tissue sections introduces local structural differences between H&E and IHC images. CellposeSAM segmentations (green) highlight the resulting imperfect cell-level correspondence.

A.3 Ablation of H&E–IHC registration

To assess the effect of additional non-rigid H&E–IHC registration on our H&E-to-IHC datasets, we compare five paired GAN baselines trained on H&E–IHC pairs before and after Deeper-HistReg Wodzinski et al. (2024). As shown in Tables A1 and A2, registration consistently improves performance across all baselines and nearly all evaluation metrics on both datasets, with particularly pronounced gains in pixel- and cell-level metrics. These results support the benefit of improved spatial correspondence for paired virtual staining.

Table A1: Ablation of H&E–IHC registration on H&E-to-P63.

Model	Data	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	Dice↑	AJI↑	PQ↑
P2P	w/o Reg	21.098	0.523	0.146	0.364	0.595	0.399	0.283
	w/ Reg	22.408	0.591	0.142	0.353	0.608	0.419	0.361
P2P-res9	w/o Reg	21.090	0.523	0.143	0.350	0.607	0.420	0.298
	w/ Reg	22.896	0.621	0.125	0.321	0.668	0.499	0.448
PyP2P	w/o Reg	20.702	0.497	0.153	0.376	0.558	0.364	0.245
	w/ Reg	22.160	0.570	0.143	0.360	0.589	0.397	0.332
PyP2P-res9	w/o Reg	20.954	0.509	0.151	0.364	0.584	0.393	0.273
	w/ Reg	22.544	0.600	0.132	0.339	0.639	0.456	0.399
P2PDB	w/o Reg	20.937	0.517	0.148	0.355	0.589	0.403	0.282
	w/ Reg	22.639	0.609	0.139	0.335	0.630	0.458	0.403

Table A2: Ablation of H&E–IHC registration on H&E-to-P63&CK.

Model	Data	PSNR↑	SSIM↑	DISTS↓	LPIPS↓	Dice↑	AJI↑	PQ↑
P2P	w/o Reg	20.717	0.516	0.141	0.368	0.512	0.317	0.162
	w/ Reg	23.629	0.662	0.133	0.335	0.674	0.471	0.416
P2P-res9	w/o Reg	20.513	0.506	0.140	0.362	0.502	0.312	0.154
	w/ Reg	24.302	0.704	0.107	0.293	0.738	0.559	0.511
PyP2P	w/o Reg	20.620	0.500	0.164	0.389	0.476	0.280	0.128
	w/ Reg	23.509	0.651	0.136	0.338	0.673	0.463	0.405
PyP2P-res9	w/o Reg	20.397	0.492	0.152	0.376	0.486	0.296	0.141
	w/ Reg	23.643	0.669	0.121	0.317	0.692	0.498	0.448
P2PDB	w/o Reg	20.196	0.495	0.153	0.376	0.446	0.270	0.126
	w/ Reg	24.157	0.702	0.117	0.300	0.720	0.536	0.494

A.4 Detailed Results of CellDisc Compatibility

Tables A3–A6 provide the complete results of the CellDisc compatibility experiments. Across different paired GAN baselines and staining tasks, integrating CellDisc generally improves pixel-wise, cell-level, and stain-specific fidelity. For example, on H&E-to-P63, P2P + CellDisc* improves PSNR from 22.408 to 23.999 and Dice/AJI/PQ from 0.608/0.419/0.361 to 0.672/0.511/0.469. Similar gains are observed for PyP2P-res9, where Dice/AJI/PQ increase from 0.639/0.456/0.399 to 0.706/0.543/0.501. On H&E-to-P63&CK, PyP2P-res9 + CellDisc* improves PSNR from 23.643 to 25.568 and Dice/AJI/PQ from 0.692/0.498/0.448 to 0.773/0.603/0.564. The improvements also extend to the more challenging FFPE-to-H&E task: for P2P-res9, CellDisc* increases PSNR from 18.577 to 20.223 and PQ from 0.226 to 0.304. Although CellDisc alone is less consistent for some baseline–task combinations, particularly P2PDB on FFPE-to-H&E, combining CellDisc with perceptual supervision generally yields stronger and more consistent performance. Overall, these detailed results further support the compatibility of CellDisc across different paired GAN baselines and staining transformations.

Table A3: Compatibility of CellDisc on H&E-to-P63.

Baselines	Config	Pixel-wise Metrics		Perceptual Metrics		Cell Segmentation Metrics			Stain-specific Metrics	
		PSNR↑	SSIM↑	DISTS↓	LPIPS↓	Dice↑	AJI↑	PQ↑	Dice-H↑	Dice-P63↑
P2P	base	22.408	0.591	0.142	0.353	0.608	0.419	0.361	0.689	0.314
	base*	22.524	0.595	0.140	0.343	0.615	0.430	0.369	0.693	0.314
	base + CellDisc(w/o F)	23.822	0.673	0.137	0.305	0.646	0.481	0.439	0.722	0.369
	base + CellDisc(w/o F)*	23.904	0.676	0.137	0.292	0.654	0.491	0.448	0.726	0.375
	base + CellDisc	23.893	0.677	**0.132**	0.303	0.659	0.497	0.456	0.730	0.379
	base + CellDisc*	**23.999**	**0.682**	0.133	**0.291**	**0.672**	**0.511**	**0.469**	**0.735**	**0.384**
P2P-res9	base	22.896	0.621	0.125	0.321	0.668	0.499	0.448	0.715	0.358
	base*	22.965	0.627	**0.123**	0.308	**0.681**	0.515	0.461	0.719	0.364
	base + CellDisc(w/o F)	24.229	0.694	0.168	0.321	0.576	0.419	0.388	0.703	0.353
	base + CellDisc(w/o F)*	24.209	0.692	0.154	0.287	0.657	0.498	0.457	0.720	0.372
	base + CellDisc	**24.538**	**0.707**	0.164	0.318	0.675	0.517	0.479	0.746	**0.397**
	base + CellDisc*	24.516	0.704	0.159	**0.286**	0.680	**0.524**	**0.483**	**0.747**	0.396
PyP2P	base	22.160	0.570	0.143	0.360	0.589	0.397	0.332	0.680	0.307
	base*	22.369	0.583	0.139	0.345	0.614	0.425	0.363	0.690	0.312
	base + CellDisc(w/o F)	23.204	0.632	**0.126**	0.311	0.648	0.469	0.406	0.716	0.360
	base + CellDisc(w/o F)*	23.682	0.660	0.130	0.295	0.674	0.502	0.453	0.731	0.375
	base + CellDisc	23.691	0.653	0.135	0.315	0.666	0.488	0.444	0.733	0.373
	base + CellDisc*	**23.979**	**0.676**	0.129	**0.289**	**0.688**	**0.520**	**0.475**	**0.740**	**0.389**
PyP2P-res9	base	22.544	0.600	0.132	0.339	0.639	0.456	0.399	0.702	0.340
	base*	22.631	0.611	**0.129**	0.321	0.659	0.483	0.427	0.708	0.349
	base + CellDisc(w/o F)	23.917	0.672	0.152	0.323	0.634	0.464	0.425	0.718	0.367
	base + CellDisc(w/o F)*	24.118	0.685	0.138	0.286	0.681	0.515	0.473	0.732	0.384
	base + CellDisc	24.322	**0.696**	0.156	0.313	0.658	0.498	0.458	0.739	0.387
	base + CellDisc*	**24.341**	**0.696**	0.143	**0.282**	**0.706**	**0.543**	**0.501**	**0.747**	**0.400**
P2PDB	base	22.639	0.609	**0.139**	0.335	0.630	0.458	0.403	0.689	0.330
	base*	21.997	0.602	0.360	0.347	0.659	0.491	0.445	0.700	0.346
	base + CellDisc(w/o F)	23.657	0.663	0.227	0.374	0.619	0.446	0.395	0.681	0.323
	base + CellDisc(w/o F)*	23.907	0.681	0.155	0.286	0.669	0.505	0.459	0.674	0.356
	base + CellDisc	23.932	0.683	0.156	0.316	0.622	0.458	0.419	0.682	0.369
	base + CellDisc*	**24.321**	**0.695**	0.153	**0.279**	**0.693**	**0.535**	**0.493**	**0.734**	**0.393**

Table A4: Compatibility of CellDisc on H&E-to-P63&CK.

Baselines	Config	Pixel-wise Metrics		Perceptual Metrics		Cell Segmentation Metrics			Stain-specific Metrics		
		PSNR↑	SSIM↑	DISTS↓	LPIPS↓	Dice↑	AJI↑	PQ↑	Dice-H↑	Dice-P63↑	Dice-CK↑
P2P	base	23.629	0.662	0.133	0.335	0.674	0.471	0.416	0.727	0.346	0.289
	base*	23.786	0.673	0.126	0.317	0.696	0.497	0.446	0.735	0.357	0.361
	base + CellDisc(w/o F)	24.965	0.735	0.106	0.275	0.738	0.558	0.517	0.769	0.431	**0.419**
	base + CellDisc(w/o F)*	25.095	0.740	0.106	0.265	0.749	0.571	0.528	0.771	0.431	0.410
	base + CellDisc	25.039	0.738	0.107	0.274	0.744	0.568	0.526	0.772	0.429	0.415
	base + CellDisc*	**25.096**	**0.742**	**0.105**	**0.263**	**0.752**	**0.577**	**0.534**	**0.773**	**0.436**	0.412
P2P-res9	base	24.302	0.704	0.107	0.293	0.738	0.559	0.511	0.758	0.416	0.434
	base*	24.337	0.705	**0.101**	0.281	0.753	0.577	0.528	0.760	0.419	0.434
	base + CellDisc(w/o F)	24.949	0.740	0.114	0.280	0.675	0.493	0.461	0.759	0.423	0.403
	base + CellDisc(w/o F)*	25.222	0.750	0.106	0.258	0.749	0.574	0.534	0.771	0.437	0.411
	base + CellDisc	**25.657**	**0.766**	0.121	0.272	0.745	0.575	0.537	**0.787**	**0.454**	**0.437**
	base + CellDisc*	25.447	0.756	0.106	**0.255**	**0.760**	**0.589**	**0.549**	0.780	0.448	0.435
PyP2P	base	23.509	0.651	0.136	0.338	0.673	0.463	0.405	0.722	0.343	0.335
	base*	23.552	0.661	0.132	0.322	0.679	0.477	0.421	0.727	0.352	0.294
	base + CellDisc(w/o F)	24.861	0.723	0.111	0.286	0.746	0.559	0.510	0.768	0.424	0.370
	base + CellDisc(w/o F)*	24.959	0.729	0.107	0.270	0.752	0.571	0.523	0.769	0.426	0.394
	base + CellDisc	24.930	0.727	0.112	0.283	0.754	0.573	0.525	0.772	0.428	0.388
	base + CellDisc*	**25.083**	**0.737**	**0.104**	**0.265**	**0.759**	**0.582**	**0.536**	**0.775**	**0.437**	**0.397**
PyP2P-res9	base	23.643	0.669	0.121	0.317	0.692	0.498	0.448	0.740	0.382	0.383
	base*	23.906	0.688	0.110	0.296	0.727	0.540	0.489	0.746	0.398	0.370
	base + CellDisc(w/o F)	24.807	0.729	0.110	0.285	0.719	0.533	0.492	0.765	0.426	0.412
	base + CellDisc(w/o F)*	25.095	0.741	**0.104**	0.263	0.754	0.577	0.533	0.770	0.429	0.407
	base + CellDisc	25.051	0.738	0.107	0.277	0.725	0.544	0.505	0.772	0.433	0.428
	base + CellDisc*	**25.568**	**0.759**	0.111	**0.255**	**0.773**	**0.603**	**0.564**	**0.785**	**0.448**	**0.432**
P2PDB	base	24.157	0.702	0.117	0.300	0.720	0.536	0.494	0.752	0.392	0.356
	base*	23.747	0.692	0.175	0.303	0.728	0.546	0.504	0.734	0.367	0.279
	base + CellDisc(w/o F)	24.347	0.713	0.120	0.300	0.660	0.471	0.434	0.724	0.374	0.394
	base + CellDisc(w/o F)*	25.071	0.743	0.115	0.260	0.738	0.560	0.516	0.741	0.384	0.398
	base + CellDisc	25.000	0.740	0.118	0.285	0.692	0.514	0.476	0.758	0.401	0.400
	base + CellDisc*	**25.237**	**0.748**	**0.107**	**0.256**	**0.741**	**0.567**	**0.522**	**0.767**	**0.428**	**0.432**

Table A5: Compatibility of CellDisc on FFPE-to-H&E.

Baselines	Config	Pixel-wise Metrics		Perceptual Metrics		Cell Segmentation Metrics			Stain-specific Metrics	
		PSNR↑	SSIM↑	DISTS↓	LPIPS↓	Dice↑	AJI↑	PQ↑	Dice-H↑	Dice-E↑
P2P	base	17.988	0.426	0.140	0.397	0.517	0.299	0.181	0.460	0.827
	base*	18.044	0.430	0.138	0.391	0.518	0.303	0.187	0.461	0.827
	base + CellDisc(w/o F)	18.811	0.469	**0.136**	0.377	0.544	0.325	0.218	0.484	0.827
	base + CellDisc(w/o F)*	19.031	0.483	0.143	0.364	0.552	0.333	0.231	0.483	0.828
	base + CellDisc	19.432	**0.525**	0.166	0.371	0.550	0.340	0.247	**0.519**	0.846
	base + CellDisc*	**19.489**	0.524	0.147	**0.350**	**0.557**	**0.349**	**0.255**	0.518	**0.848**
P2P-res9	base	18.577	0.472	0.127	0.375	0.553	0.337	0.226	0.489	0.838
	base*	18.849	0.491	**0.118**	0.357	0.567	0.353	0.248	0.503	0.841
	base + CellDisc(w/o F)	19.101	0.487	0.147	0.376	0.555	0.330	0.228	0.485	0.822
	base + CellDisc(w/o F)*	19.429	0.510	0.130	0.346	0.572	0.354	0.260	0.509	0.833
	base + CellDisc	20.044	0.559	0.178	0.365	0.573	0.361	0.281	0.541	0.853
	base + CellDisc*	**20.223**	**0.567**	0.148	**0.330**	**0.593**	**0.383**	**0.304**	**0.548**	**0.859**
PyP2P	base	17.713	0.404	0.150	0.409	0.493	0.281	0.163	0.442	0.819
	base*	17.642	0.404	0.150	0.407	0.491	0.282	0.165	0.441	0.817
	base + CellDisc(w/o F)	18.798	0.466	0.144	0.388	0.546	0.324	0.217	0.482	0.825
	base + CellDisc(w/o F)*	18.787	0.465	0.140	0.371	0.547	0.328	0.222	0.478	0.830
	base + CellDisc	18.950	0.475	0.138	0.371	0.555	0.340	0.236	0.502	0.838
	base + CellDisc*	**19.148**	**0.494**	**0.128**	**0.352**	**0.565**	**0.350**	**0.249**	**0.511**	**0.843**
PyP2P-res9	base	18.246	0.438	0.158	0.402	0.524	0.307	0.193	0.463	0.826
	base*	18.298	0.416	0.214	0.431	0.498	0.291	0.181	0.427	0.830
	base + CellDisc(w/o F)	19.159	0.485	0.198	0.401	0.560	0.340	0.235	0.495	0.819
	base + CellDisc(w/o F)*	19.208	0.494	**0.128**	0.354	0.562	0.345	0.247	0.501	0.832
	base + CellDisc	19.286	0.499	0.133	0.363	0.573	0.350	0.255	0.517	0.839
	base + CellDisc*	**19.938**	**0.542**	0.145	**0.339**	**0.578**	**0.367**	**0.283**	**0.531**	**0.851**
P2PDB	base	18.351	0.452	0.159	0.394	0.480	0.288	0.189	0.446	0.826
	base*	18.887	0.462	0.360	0.413	0.546	0.337	0.252	0.481	**0.842**
	base + CellDisc(w/o F)	18.334	0.442	0.154	0.401	0.531	0.309	0.197	0.460	0.816
	base + CellDisc(w/o F)*	18.678	0.468	0.136	0.366	0.546	0.332	0.231	0.470	0.820
	base + CellDisc	18.390	0.452	0.144	0.389	0.542	0.327	0.220	0.470	0.820
	base + CellDisc*	**19.428**	**0.519**	**0.127**	**0.338**	**0.573**	**0.362**	**0.270**	**0.519**	0.837

Table A6: Compatibility of CellDisc using independent cell segmentation models on FFPE-to-H&E.

Model	Config	Hover-net			Cellvit			Cellpose		
		Dice↑	AJI↑	PQ↑	Dice↑	AJI↑	PQ↑	Dice↑	AJI↑	PQ↑
P2P	base	0.576	0.324	0.219	0.569	0.325	0.225	0.543	0.313	0.207
	base*	0.575	0.327	0.225	0.566	0.327	0.230	0.543	0.317	0.213
	base + CellDisc(w/o F)	0.598	0.348	0.256	0.589	0.348	0.261	0.567	0.337	0.242
	base + CellDisc(w/o F)*	0.603	0.350	0.266	0.597	0.353	0.272	0.574	0.342	0.254
	base + CellDisc	0.606	0.349	0.277	0.601	0.362	0.291	0.583	0.351	0.269
	base + CellDisc*	**0.611**	**0.365**	**0.295**	**0.603**	**0.371**	**0.302**	**0.585**	**0.362**	**0.283**
P2P-res9	base	0.607	0.358	0.265	0.599	0.361	0.271	0.576	0.349	0.252
	base*	0.618	0.376	0.288	0.610	0.377	0.295	0.590	0.367	0.276
	base + CellDisc(w/o F)	0.603	0.345	0.262	0.598	0.351	0.270	0.573	0.337	0.252
	base + CellDisc(w/o F)*	0.622	0.373	0.297	0.614	0.374	0.302	0.595	0.365	0.285
	base + CellDisc	0.617	0.352	0.296	0.620	0.381	0.324	0.588	0.357	0.291
	base + CellDisc*	**0.638**	**0.392**	**0.339**	**0.633**	**0.402**	**0.351**	**0.610**	**0.390**	**0.328**
PyP2P	base	0.560	0.310	0.202	0.551	0.310	0.207	0.524	0.299	0.190
	base*	0.558	0.311	0.204	0.550	0.310	0.210	0.523	0.300	0.192
	base + CellDisc(w/o F)	0.595	0.340	0.251	0.592	0.346	0.259	0.569	0.334	0.240
	base + CellDisc(w/o F)*	0.602	0.349	0.259	0.594	0.350	0.265	0.570	0.339	0.248
	base + CellDisc	0.613	0.362	0.276	0.603	0.363	0.282	0.581	0.354	0.264
	base + CellDisc*	**0.619**	**0.371**	**0.289**	**0.610**	**0.372**	**0.295**	**0.588**	**0.362**	**0.275**
PyP2P-res9	base	0.587	0.331	0.232	0.579	0.334	0.239	0.556	0.323	0.221
	base*	0.511	0.284	0.192	0.550	0.304	0.210	0.548	0.294	0.193
	base + CellDisc(w/o F)	0.609	0.357	0.274	0.605	0.363	0.284	0.585	0.351	0.264
	base + CellDisc(w/o F)*	0.614	0.364	0.285	0.607	0.365	0.290	0.584	0.356	0.272
	base + CellDisc	0.619	0.369	0.291	0.611	0.369	0.297	0.592	0.361	0.279
	base + CellDisc*	**0.627**	**0.380**	**0.319**	**0.618**	**0.385**	**0.327**	**0.600**	**0.377**	**0.307**
P2PDB	base	0.535	0.310	0.226	0.527	0.311	0.230	0.542	0.318	0.222
	base*	0.515	0.294	0.242	0.541	0.331	0.272	0.580	0.352	**0.280**
	base + CellDisc(w/o F)	0.592	0.331	0.235	0.583	0.333	0.241	0.550	0.317	0.222
	base + CellDisc(w/o F)*	0.594	0.352	0.269	0.586	0.352	0.273	0.567	0.343	0.255
	base + CellDisc	0.594	0.349	0.260	0.585	0.348	0.264	0.564	0.339	0.243
	base + CellDisc*	**0.609**	**0.366**	**0.289**	**0.598**	**0.364**	**0.292**	**0.582**	**0.357**	0.275

A.5 Qualitative Ablation of Structural and Perceptual Supervision

Figures A6 and A7 provide qualitative comparisons across the virtual staining tasks without and with VGG-based perceptual supervision, respectively. Across different paired GAN baselines, incorporating CellDisc generally produces virtual images with cellular structures and stain patterns that more closely correspond to the real targets. As highlighted by the green boxes, the CellposeSAM segmentation maps further reveal improved preservation of cell instances, particularly in regions where the baseline models introduce, miss, or poorly delineate cellular structures. Similar improvements remain visible under perceptual supervision, indicating that the cell-level structural constraints introduced by CellDisc are complementary to image-level perceptual supervision.

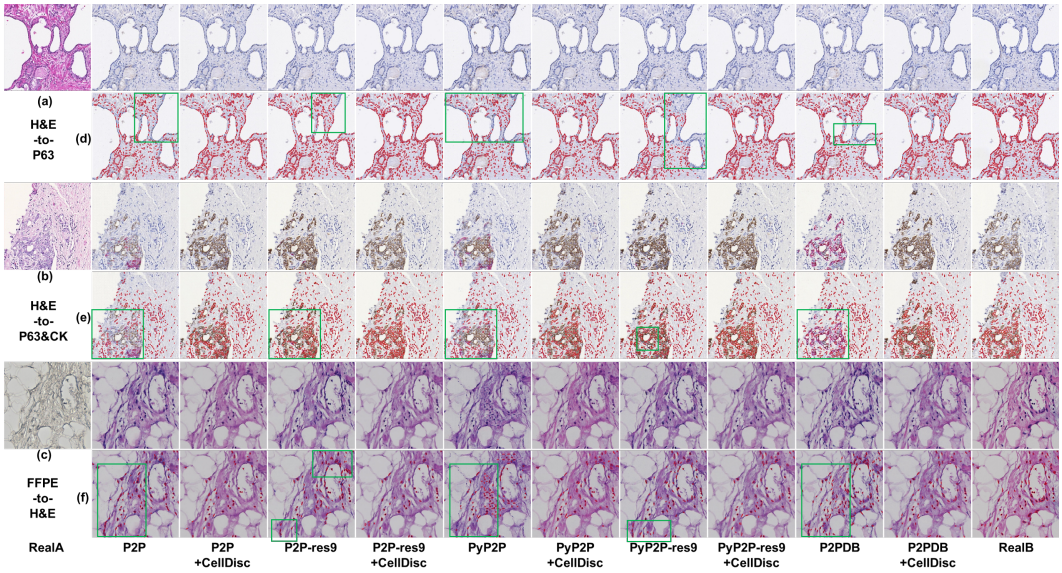

Figure A6: **Qualitative comparison without perceptual supervision. (a–c)** Real source (RealA), real target (RealB), and virtual staining results generated by different paired GAN baselines with and without CellDisc. **(d–f)** Corresponding CellposeSAM segmentation results. Green boxes highlight representative regions where CellDisc better preserves cell instances and stain patterns.

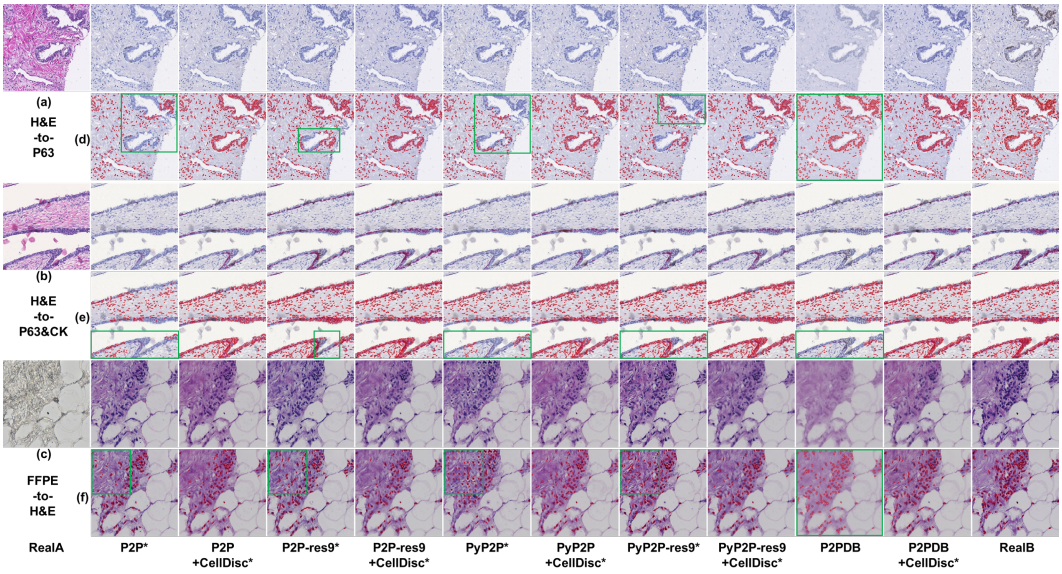

Figure A7: **Qualitative comparison with perceptual supervision. (a–c)** Real source (RealA), real target (RealB), and virtual staining results generated by different paired GAN baselines with VGG-based perceptual supervision (Base*) and their CellDisc-enhanced variants (CellDisc*). **(d–f)** Corresponding CellposeSAM segmentation results. Green boxes highlight representative regions where CellDisc better preserves cell instances and stain patterns under perceptual supervision.

A.6 Qualitative Comparison with SOTA Methods

Figure A8 provides qualitative comparisons with SOTA methods across the three virtual staining tasks. On H&E-to-P63, D-VST and DiffVS tend to introduce excessive staining in non-target regions, whereas HistoPlexer and DGR frequently fail to reproduce the target P63 staining patterns. DiffVS and JiT additionally produce relatively blurred cellular details. On H&E-to-P63&CK, several competing methods, including D-VST, DiffVS, JiT, HistoPlexer, and DGR, exhibit varying degrees of over-staining, resulting in stain distributions that deviate from the corresponding real targets. On FFPE-to-H&E, D-VST occasionally introduces additional nucleus-like structures and shows locally inconsistent coloration, while DiffVS and JiT again exhibit relatively blurred cellular details. HistoPlexer shows noticeable staining distortions and occasionally misses nuclei present in the real H&E images, whereas DGR exhibits severe visual distortions, suggesting unsuccessful convergence on this task. In contrast, P2P-res9 equipped with either CellDisc or CellDisc* better preserves cellular structures and stain distributions across the three distinct staining transformations, in agreement with the quantitative comparisons reported in the main text.

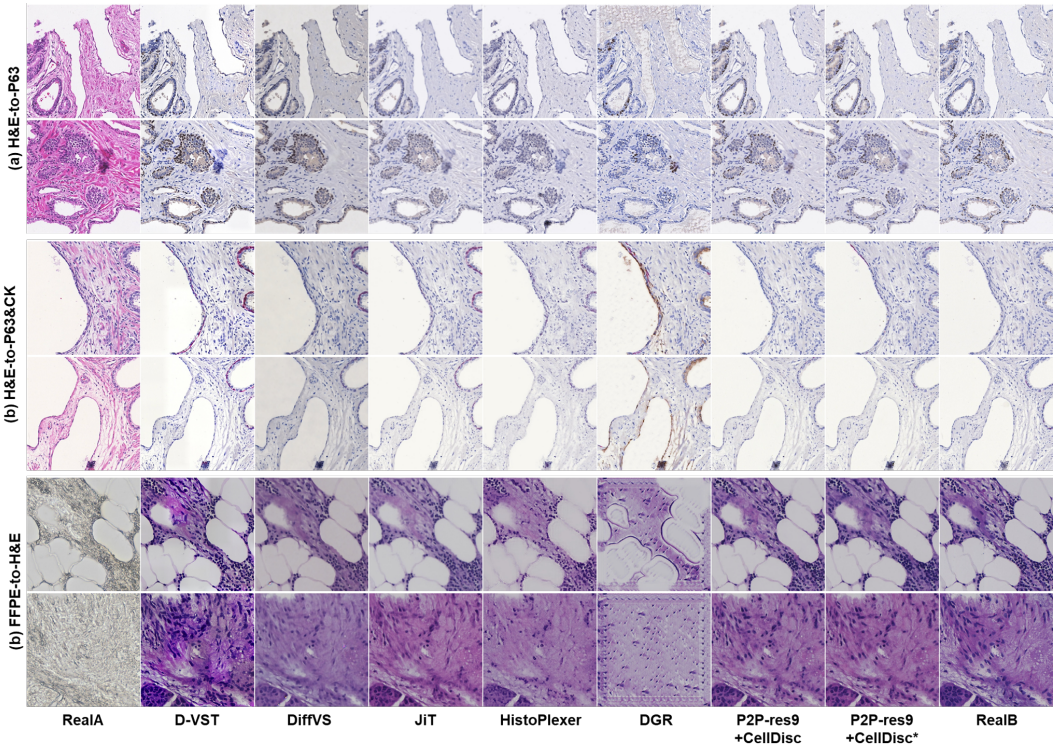

Figure A8: **Qualitative comparison with SOTA methods.** Representative results on **(a)** H&E-to-P63, **(b)** H&E-to-P63&CK, and **(c)** FFPE-to-H&E virtual staining. For each task, the real source (RealA), real target (RealB), and virtual staining results generated by the compared methods are shown for visual comparison. CellDisc produces virtual images with cellular structures and stain patterns that more closely resemble the corresponding real targets across the three staining tasks.