Y-diff: Structure-Texture Decoupled Diffusion Distillation for H&E-to-pCLE Translation

Haodong Wang[1*], Yan Wen[1*], Hongen Liao[1,2], Fang Chen[1†], and Tianqi Huang[1†]

[1] School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] School of Biomedical Engineering, Tsinghua University, Beijing, China
{hdw2agon, yan.wen, liaohongen, chen-fang, huangtianqi}@sjtu.edu.cn

Abstract. Probe-based Confocal Laser Endomicroscopy (pCLE) enables real-time *in vivo* optical biopsies. However, acquiring large-scale, high-quality pCLE images is prohibitively expensive and technically complex, causing severe training data scarcity. Furthermore, when applied to pCLE image generation, existing cross-modal translation models consistently introduce artifacts and cellular structure distortions. To address this, we propose Y-diff, a novel generative framework translating widely accessible H&E-stained pathology images into the scarce pCLE modality with high fidelity, providing robust data support for computational pathology. Y-diff innovatively introduces a knowledge distillation mechanism into this task paradigm, efficiently decoupling the learning of pCLE-domain textures from that of H&E pathology staining domain spatial structures. Specifically, a Teacher model captures distinct pCLE optical textures via color encoding, while a Student model extracts and distills this knowledge under strict H&E spatial constraints. Extensive experiments demonstrate that Y-diff consistently outperforms baselines, while significantly alleviating conventional translation artifacts and structural distortions. What's more, the synthesized data further expands the scale of this rare modality, reducing mean absolute percentage error to 13.84% in downstream cell counting tasks. Source code is available at: `https://github.com/Hdw2agon/Y_diff`.

Keywords: Cross-Modality Translation · Diffusion Model · Knowledge Distillation · pCLE

1 Introduction

Probe-based Confocal Laser Endomicroscopy (pCLE) enables real-time, *in vivo* cellular-level optical biopsies, significantly enhancing the efficiency of clinical pathological diagnosis [22, 38]. However, limited by inherent optical physical mechanisms, pCLE images universally suffer from defects such as low contrast and blurred cellular boundaries. This not only makes the clinical interpretation of these images heavily reliant on senior experts [1, 30], but also escalates

*These authors contributed equally. †Corresponding authors.

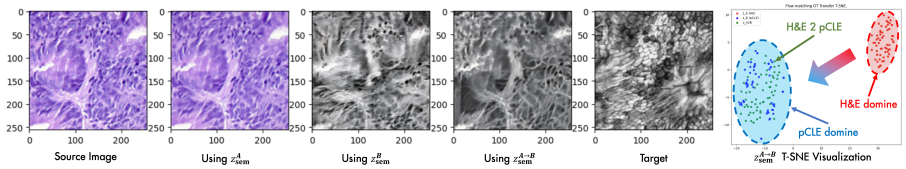

Fig. 1: $z_{\mathbf{sem}}$ replacing simulation results. This figure illustrates the synthesized results obtained by merely substituting different z_{sem} encodings given the source H&E and target pCLE images, while freezing all network parameters. t-SNE visualization demonstrates that our mapped semantics align with the target pCLE distribution.

the cost of acquiring and annotating large-scale, high-quality data [25, 36]. Although deep learning holds the promise of breaking the bottleneck of manual interpretation [4, 45, 48], its performance is severely hampered by data scarcity. Consequently, existing works are predominantly confined to small-sample image classification tasks [2, 6, 13, 42], rendering them inadequate for complex downstream pathological analyses such as cell detection or segmentation.

Translating widely accessible H&E pathology images into the pCLE modality via unsupervised image translation presents a promising avenue to address the scarcity of high-quality data [5]. However, lacking paired supervision signals, existing mainstream methods are consistently plagued by artifact generation and cellular structure distortions. Cycle-consistency-based generic generative models [5, 26, 50] struggle to establish robust mappings between target texture synthesis and source-domain structure preservation. Consequently, they are highly susceptible to irreversible structural deformations, spurious artifacts, or the loss of cellular topological features caused by training distribution deviations. Concurrently, virtual staining methods tailored for pathology images [17, 20, 43], despite their superior performance in pixel-level weakly paired tasks, struggle to generalize to the complex structural and color modeling required in such extremely unpaired scenarios

To tackle the prevalent artifacts and cellular structure distortions in existing cross-modal translation, we propose Y-diff, a novel structure-texture decoupled diffusion distillation framework. Deviating from conventional methods that implicitly map images as monolithic entities, Y-diff innovatively and efficiently decouples pCLE-domain optical textures from H&E-domain spatial structures. Specifically, we design a Teacher-Student dual-branch denoising architecture: the Teacher model concentrates on learning distinct pCLE optical textures, utilizing Flow matching [19] to achieve optimal transport of global semantics encoded by a cross-modal DiffAE [31]; concurrently, under strict H&E spatial priors, the Student model distills the Teacher's color knowledge to precisely synthesize images possessing both high-fidelity pCLE textures and original H&E topological structures. In summary, the main contributions of this work are as follows:

– **Proposing Y-diff, a novel cross-modal pathology image generation framework**: By designing a Teacher-Student-based diffusion distilla-

tion mechanism, we achieve the efficient decoupling of optical textures from spatial structures in unpaired H&E-to-pCLE translation.
– **Significantly alleviating artifacts and cellular structure distortions**: To bridge the substantial modality gap, we transfer semantic features via Flow matching and impose strict spatial constraints. This approach achieves effective structural alignment with the source H&E images while precisely synthesizing complex pCLE textures.
– **Demonstrating downstream clinical value**: Extensive experiments show that Y-diff outperforms state-of-the-art baselines in synthesis quality. Serving as a robust data augmentation strategy, the synthesized pseudo-pCLE data substantially enhances deep learning model accuracy in downstream tasks such as real pCLE cell detection, broadening the application scope of computational pathology.

2 Related Work

Cross-modal translation aims to synthesize images visually matching the target modality while preserving source-domain structural and semantic information. For cellular-level translation tasks, early studies [5] predominantly rely on cycle consistency to model color-structure relationships. Despite yielding visual improvements, these methods frequently suffer from anatomical distortions and semantic misalignment. Subsequently, approaches incorporating contrastive learning or explicit constraints [11,17,27,47] achieved substantial advancements in visual fidelity. However, confronted with massive domain gaps, patch-based contrastive mapping becomes highly unreliable, and acquiring accurate explicit constraints is rendered fundamentally unfeasible. Consequently, they still struggle to circumvent artifact generation and cellular structure distortions. Recently, the rapid proliferation of diffusion models [9, 35, 49] has advanced translation tasks like FFPE-to-H&E toward higher precision and resolution [7, 10, 43]. Nevertheless, as elucidated by Yang et al. [43], performance degrades significantly when severe discrepancies exist between source and target modalities, coupled with a lack of reliable pixel-level or semantic correlations. This challenge is further exacerbated by the inherent domain gap in H&E-to-pCLE translation. Since early attempts by Gu et al. [5], this specific task has remained largely underexplored due to inherent difficulty [1] and data bottlenecks [2]. The failure of existing generative paradigms in this cross-modal translation fundamentally exposes their inability to establish a smooth, minimal-cost mapping path between two markedly divergent, independent probability distributions.

Modeling the **Optimal Transport** (OT) trajectory for textures and structures constitutes the core challenge in cross-modal translation tasks. Cellular morphology varies significantly across diverse tissues, whereas textures strictly depend on the imaging modality. Consequently, directly establishing stable OT paths within a highly entangled pixel space presents formidable difficulties. In the task of OT path construction, early methods, typified by Flow matching [19], directly regress OT paths by constructing continuous velocity fields, achieving

substantial breakthroughs in both training and inference efficiency. However, in complex cellular scenarios, standard Flow matching directly fits the evolution trajectory on the global pixel manifold. Lacking local structural constraints, it struggles to preserve intricate cellular topological structures. Recently, the Schrödinger Bridge (SB) framework [3,37] revealed that theoretically established OT paths mathematically satisfy cycle consistency by nature, elegantly simplifying traditional multi-generator architectures [26, 41, 50]. Nevertheless, mapping both source and target domain distributions to a shared Gaussian noise space incurs exorbitant training overhead. Subsequent SB optimization works [7,15,34] attempt to construct unified diffusion networks to mitigate training complexity. Although these methods excel in texture modeling, they inevitably introduce cellular distortions or hallucinatory artifacts in pCLE synthesis. These issues stem from the models attempting to simultaneously preserve cellular structures and execute optical texture translations across a massive domain gap within a highly entangled pixel space. Inspired by [43], we argue that efficiently decoupling the representations of textures and structures, and independently constructing OT paths, constitutes a more reliable paradigm for translation to pCLE tasks. Furthermore, to safely inject the decoupled pCLE-domain textures into the generation process while preserving H&E-domain structural priors, we introduce a Teacher-Student architecture [39, 44]. By specifically applying a distillation mechanism to the transfer and recombination of decoupled features, our approach ingeniously circumvents the entanglement trap of the pixel space, achieving high-fidelity modality conversion.

3 Method

To alleviate artifacts and cellular distortions in H&E-to-pCLE cross-modal translation, we present the overall framework of Y-diff (Fig. 2). The diffusion pipeline of Y-diff comprises a shared spatial forward-noising branch alongside two independent denoising branches: a Teacher and a Student. To achieve the efficient decoupling of texture and structural representations, we physically isolate the cross-modal translation process into two independent information flows:

Texture Flow. A pre-trained DiffAE [31] disentangles the color and optical features of the images into low-dimensional global semantic encodings. Subsequently, an independent Flow matching module [19] learns the optimal transport mapping at the optical texture level between the H&E and pCLE domains.

Structure Flow. The microscopic cellular topology of the images is isolated into pure spatial noise and grayscale features, which are transmitted under the structural guidance of a Structure-Preserving Network (SPNet).

During the inference and distillation in stage 2 (Fig. 2), the Student model recombines the structure flow preserved from the H&E domain with the pCLE texture flow mapped via Flow matching. This recombination is optimized through direct feature alignment and knowledge distillation against the generation outputs of the Oracle Teacher model.

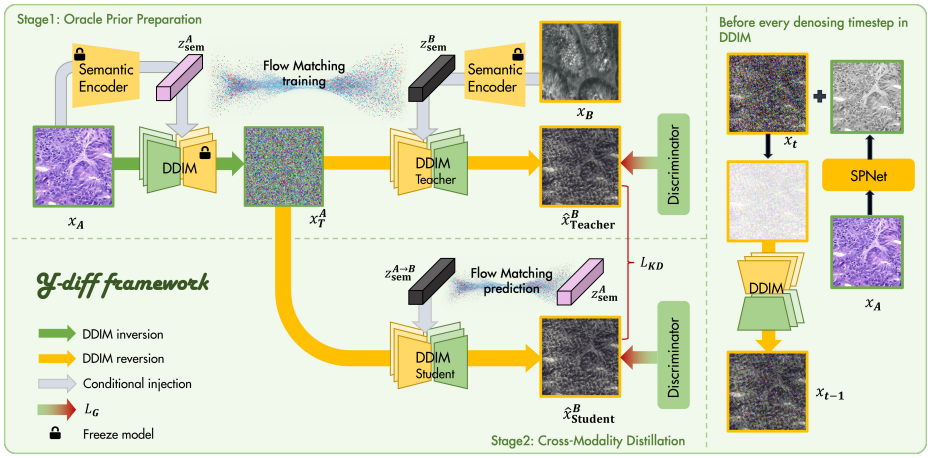

Fig. 2: Overview of Y-diff framework. The pre-trained DiffAE semantic encoder extracts the low-dimensional global semantic codes of H&E and pCLE domain images; by learning the optimal transport velocity field between these two semantic distributions, the Flow Matching module deterministically maps the semantic codes of the H&E domain to the pCLE domain semantics during Student model training; The Student model distills the optical texture generative priors from the Oracle Teacher model, and, guided by SPNet, learns the cellular topological structures.

3.1 Preliminaries : Diffusion-based Semantic Extraction

The semantic encoder (Enc) proposed in DiffAE [31] encodes an input image x into a highly compressed global semantic representation $z_{\text{sem}} = Enc(x)$. This representation captures rich color information of x. When employing z_{sem} as a conditional injection to guide the DDIM [35], the forward noising and reverse denoising processes for the representation x_t of image x at timestep t within the diffusion pipeline can be formulated as:

$$x_{t+1} = \sqrt{\alpha_{t+1}} f_\theta(x_t, t, z_{\text{sem}}) + \sqrt{1-\alpha_{t+1}}\,\epsilon_\theta(x_t, t, z_{\text{sem}}), \tag{1}$$

$$x_{t-1} = \sqrt{\alpha_{t-1}} f_\theta(x_t, t, z_{\text{sem}}) + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(x_t, t, z_{\text{sem}}). \tag{2}$$

Here $f_\theta(\cdot)$ is the denoising prediction network, and $\epsilon_\theta(\cdot)$ is the predicted noise. where $f_\theta(\mathbf{x}_t, t, z_{\text{sem}}) = \frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t, z_{\text{sem}})}{\sqrt{\alpha_t}}$. α_t denotes the cumulative noise schedule coefficient. This design ensures that the forward noising process and reverse denoising processes are guided by the semantic representation. Compared to conventional diffusion models [9, 35], DiffAE [31] enables high-quality unimodal image encoding and reconstruction.

3.2 Flow Matching for Semantic Translation

Since semantic information such as color and texture is extensively encoded and compressed into z_{sem} by the semantic encoder of DiffAE [31], as illustrated in

Fig. 1, retaining the intermediate state x_t derived from the source H&E image and freezing all network parameters allows the H&E image to translate to the pCLE domain merely by substituting z_{sem}. To obtain a reliable z_{sem} of the pCLE domain during inference, Y-diff introduces a continuous normalizing flow parameterized by an Ordinary Differential Equation (ODE) to achieve deterministic and smooth transport of global semantics, formulated as:

$$\frac{\mathrm{d}z_t}{\mathrm{d}t} = v_\phi(z_t, t) \tag{3}$$

Here z_t is the time-dependent semantic code and v_ϕ is the learned vector field parameterized by ϕ.

To construct an OT path between the source semantic distribution $p(z^A_{\text{sem}})$ and the target distribution $p(z^B_{\text{sem}})$, we employ Flow matching to optimize the neural network vector field v_ϕ. Given a minibatch OT-coupled semantic pair $(z^A_{\text{sem}}, z^B_{\text{sem}})$, the target vector field is intuitively defined as $u_t(z_t|z^A_{\text{sem}}, z^B_{\text{sem}}) = z^B_{\text{sem}} - z^A_{\text{sem}}$. The optimization objective for Flow matching is as follows:

$$\mathcal{L}_{FM} = \mathbb{E}_{t\sim U[0,1], z^A_{\text{sem}}, z^B_{\text{sem}}}[\|v_\phi(z_t, t) - (z^B_{\text{sem}} - z^A_{\text{sem}})\|_2^2] \tag{4}$$

Here $t \sim U[0, 1]$ denotes the continuous time variable uniformly sampled from the interval $[0, 1]$; z^A_{sem} and z^B_{sem} represent semantic encodings sampled from H&E (A) and pCLE (B) marginal, paired via minibatch optimal transport; z_t indicates the intermediate semantic state at time t, constructed via the linear interpolation path $z_t = (1 - t)z^A_{\text{sem}} + tz^B_{\text{sem}}$; and $\|\cdot\|_2^2$ computes the squared L_2 norm.

In the subsequent prediction stage, the target-domain semantics can be deterministically synthesized by solving the ODE from the initial state z^A_{sem}:

$$z^{A\to B}_{\text{sem}} = z^A_{\text{sem}} + \int_0^1 v_\phi(z_t, t)\mathrm{d}t \tag{5}$$

Here $z^{A\to B}_{\text{sem}}$ denotes the transported semantic code from H&E (A) to pCLE (B). As shown by the t-SNE in Fig. 1, the flow matching module effectively translates H&E semantics to the pCLE domain, providing a reliable texture prior that circumvents conventional mode collapse.

3.3 Network Optimization and Adversarial Supervision

SPNet-Guided Spatial Constraints. The reverse diffusion process must effectively preserve the cellular topology of the H&E domain while simultaneously learning pCLE textures. At each denoising timestep t, SPNet(S_ψ) operates as a time-conditioned pixel-wise MLP that extracts grayscale morphological features from x_A and transforms them through a hierarchical feature space. The timestep embedding enables adaptive feature modulation via affine transformations at each hidden layer, allowing the network to generate structure-aware

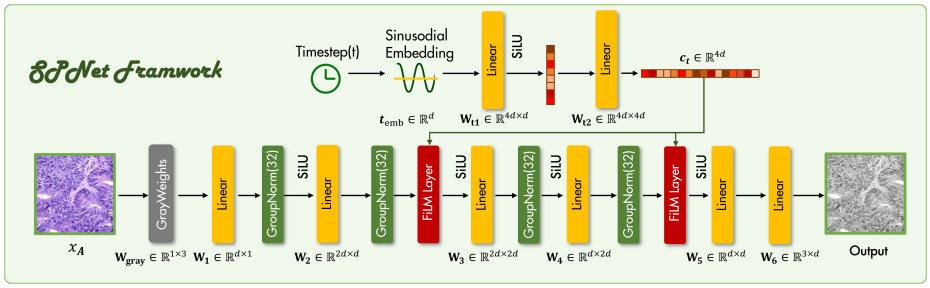

Fig. 3: Detail Structure of SPNet. SPNet is designed as a time-conditioned pixel-wise MLP. The input RGB image x_A is first converted to grayscale to reduce the influence of color channels on the network output. Temporal embedding c_t is injected via FiLM layers [29], enabling the output to adapt to the time-aware denoising process of the DDIM. This network operates at the pixel level, preserving image structural information while maintaining time awareness.

outputs conditioned on the diffusion stage. These spatial priors are injected into the diffusion latent state x_t^A via explicit modulation:

$$x_t^A \leftarrow x_t^A + \lambda_{\text{map}} \cdot \sigma(\text{S}_\psi(x_A, t)) \tag{6}$$

Here $\sigma(\cdot)$ is the sigmoid function, λ_{map} is a weighting coefficient controlling the strength of spatial modulation.

Knowledge Distillation. To stabilize the optimization process of the Student model, we employ a frozen Oracle Teacher model. Conditioned on the real semantics z_{sem}^B, the Teacher generates a pseudo-ground-truth (Pseudo-GT) image $\hat{\mathbf{x}}_{\text{Teacher}}^B$. Concurrently, conditioned on the semantics $z_{\text{sem}}^{A\to B}$ translated via Flow matching, the Student model generates the translated output $\hat{\mathbf{x}}_{\text{Student}}^B$. We enforce strict alignment between the Student and Teacher models via a comprehensive Knowledge Distillation (KD) loss including pixel, perceptual, and semantic feature spaces:

$$\begin{aligned}\mathcal{L}_{KD} = \lambda_{\text{L1}} \|\hat{x}_{\text{Student}}^B - \hat{x}_{\text{Teacher}}^B\|_1 \\ + \lambda_{\text{LPIPS}} \text{LPIPS}(\hat{x}_{\text{Student}}^B, \hat{x}_{\text{Teacher}}^B) \\ + \lambda_{\text{CLIP}} \big(1 - \text{CosSim}(\text{CLIP}(\hat{x}_{\text{Student}}^B), \text{CLIP}(\hat{x}_{\text{Teacher}}^B))\big)\end{aligned} \tag{7}$$

Here $\|\cdot\|_1$ denotes the L_1 distance, LPIPS is a perceptual similarity metric, CosSim denotes cosine similarity, and CLIP$(\cdot)$ denotes feature extraction by the CLIP encoder, which robustly preserves high-level semantics despite domain shiftswhich robustly preserves high-level semantics despite domain shifts (see Supp. for ablation). The weights λ_{L1}, λ_{LPIPS}, and λ_{CLIP} balance the pixel, perceptual, and semantic terms, respectively.

Adversarial Supervision. Inspired by general diffusion bridge frameworks [15, 37], it is imperative to ensure the alignment between the generated distribution q_θ and the target domain distribution p: $D_{\text{KL}}(q_\theta||p) = 0$. Since direct

computation of the KL divergence in high-dimensional spaces is intractable, we train a discriminator D to differentiate between real target-domain images and generated images, thereby propelling the generated distribution to converge towards the target distribution. Consequently, the adversarial training objective is formulated as:

$$\mathcal{L}_G = \mathbb{E}_{x_B \sim p_B}[\log D(x_B)] + \mathbb{E}_{x_A \sim p_A}[\log(1 - D(\hat{x}_B))] \tag{8}$$

where $\mathbb{E}$ denotes the mathematical expectation; p_B and p_A represent the marginal data distributions of the real pCLE and H&E domains, respectively. $x_B \sim p_B$ denotes a real sample drawn from the pCLE domain distribution, whereas $x_A \sim p_A$ signifies a conditional sample drawn from the H&E domain, which is utilized to generate the cross-modal Pseudo-pCLE $\hat{x}_B$. $D(\cdot)$ outputs the probability that the input image belongs to the real target-domain data manifold. Ultimately, the overall loss function for optimizing the Y-diff Student model is computed as the weighted sum of the distillation loss and the adversarial loss:

$$\mathcal{L}_{\text{Student}} = \lambda_{\text{KD}} \mathcal{L}_{\text{KD}} + \lambda_G \mathcal{L}_G \tag{9}$$

Here λ_{KD} and λ_G control the relative contributions of the distillation and adversarial objectives. For the complete loss function of the Y-diff Teacher model, please refer to the supplementary material.

3.4 Training and Sampling

To guarantee the stable convergence of decoupled features and optimal synthesis quality, the training strategy of Y-diff is formulated as two sequential optimization phases:

Oracle Prior Preparation. Guided by the ground-truth semantics of the pCLE domain and the structural constraints of SPNet, the Oracle Teacher model is independently optimized to approximate the real pCLE data manifold. Concurrently, the Flow matching module is trained in parallel to solve the OT Ordinary Differential Equation (ODE) between the semantics of the two domains.

Cross-Modality Distillation. With the Teacher model and Flow matching module frozen, the Student model is trained conditioned on the cross-modal semantics translated via Flow matching. Given that the Teacher now yields stable Pseudo-GT targets, the Student learns the decoupled recombination by fusing source-domain structures with target-domain textures through knowledge distillation. This paradigm effectively circumvents the gradient conflicts and objective misalignment prevalent in conventional end-to-end training.

During the inference and sampling phase, the Teacher model prior is discarded. Instead, the Student model operates in tandem with the Flow matching module to achieve reliable translation into the pCLE modality.

4 Experiment

4.1 Setting

Datasets. We evaluate the generalization capability and synthesis quality of our model across multiple datasets:

Clinical pCLE Dataset (CCD) is an in-house dataset constructed specifically for the H&E-to-pCLE translation task. It comprises 296 *ex vivo* H&E-stained slide images and 296 *in vivo* pCLE video frames acquired from the colon. Although roughly matched at the anatomical region level, inherent spatiotemporal acquisition discrepancies and strict clinical de-identification protocols preclude the establishment of any pixel-level spatial correspondence between the two modalities. The dataset is partitioned into 240 training and 56 testing images per modality. All images were resized to 256×256. The dataset is available at: `https://github.com/Hdw2agon/Y_diff`.

NuInsSeg [21] is a public pathology staining cell segmentation dataset. To validate the viability of the Y-diff-synthesized data for downstream tasks, we utilized a total of 132 H&E-stained test samples from tissues including the jejunum and cerebellum. All images were resized to 256×256.

MIST [17] is a public immunohistochemistry (IHC) staining dataset. To verify the robustness of Y-diff on weakly paired data, we employed the ER-004 subset, which contains 4,000 training samples and 1,000 testing samples obtained by registering serial tissue sections. Following the experimental setup of [7], we conduct model training and analysis on this dataset.

Baselines. We comprehensively evaluate Y-diff against state-of-the-art methods encompassing diverse generative paradigms. These baselines include the classic cycle-consistency model CycleGAN [50] and its recent one-step diffusion-enhanced variant CycleGAN-Turbo [28]; the contrastive-learning-based cellular-level translation framework PPT [47]; the Schrödinger Bridge-based generative model UNSB [15]; and recent cutting-edge medical modality translation and virtual staining architectures, namely SynDiff [26], PyramidPix2Pix [20], and D-VST [43]. To assess their generalizability to unpaired pCLE scenarios, all baseline models utilize their official implementations. While preserving their core optimization strategies and objective weights, we adapted resolution-dependent configurations strictly following official guidelines, training all baselines to full convergence to ensure a rigorous and fair evaluation on the CCD dataset.

Quantitative Evaluation. Due to inherent spatial misalignments and pixel-level discrepancies across domains, traditional metrics such as SSIM and PSNR offer limited reference value in unpaired translation tasks [7,18,40], while FID [8] suffers from statistical instability on small-scale datasets [14]. Following the quantitative evaluation protocol established by He et al [7] , we adopt Structure Similarity (SS), Luminance and Contrast (LC), and Histogram Correlation (Hist) to independently and accurately quantify structural preservation and style transfer performance. Additionally, LPIPS [46] is utilized to assess the overall perceptual image quality. For MIST evaluation, we report A-SSIM/A-PSNR for grayscale structural consistency with input H&E images, B-SSIM/B-PSNR for

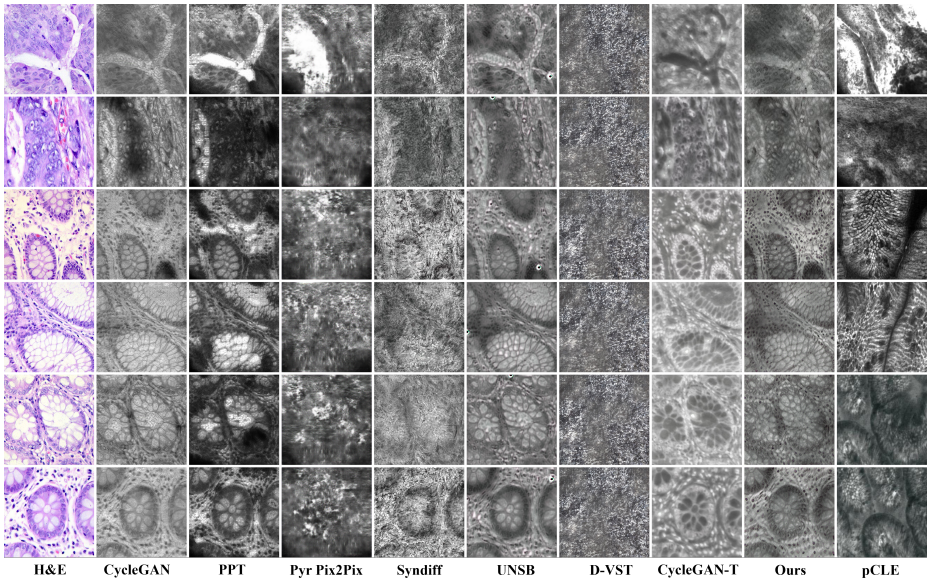

Fig. 4: Qualitative results on CCD datasets

fidelity to real target images, and FID for target-domain distributional similarity in addition.

Implementation Details. All experiments were conducted on one RTX PRO6000 GPU. Teacher and Student models were trained for 30 and 10 epochs, respectively, using Adam optimization and initialized with pre-trained DiffAE weights [31]. To reduce DDIM training cost, we adopt truncated diffusion with 1,000 total timesteps, 50 sampling steps, and a 0.5 truncation ratio, reducing both forward and reverse processes to 25 steps from the intermediate state x_t. The Flow matching MLP [19] was trained with minibatch OT coupling to construct probability paths between source and target latent vectors. During inference, the Dopri5 ODE solver was adopted for a single ODE integration pass. Detailed hyperparameters are provided in the supplementary material.

4.2 Main Result

Qualitative Results. As illustrated in Fig. 4, the modality discrepancy between H&E and pCLE within the CCD dataset poses a formidable challenge to existing paradigms. Specifically, although CycleGAN [50] partially captures texture and structural mappings, it introduces spurious shadows into local cellular structures and generates scale-like hallucinations alongside faint streak artifacts in cavity regions. The contrastive-learning-based PPT [47] exhibits severe structural degradation, manifesting extensive dark artifactual regions where the cellular topology is entirely obliterated, accompanied by streak noise akin to that of CycleGAN [50]. More critically, while excelling in standard paired settings,

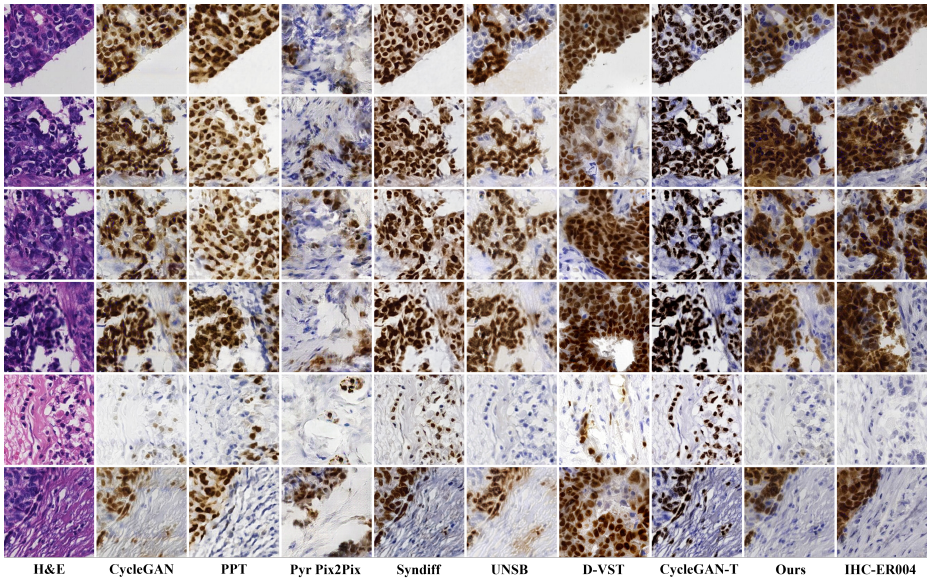

Fig. 5: Qualitative results on the MIST dataset

the virtual staining models PyramidPix2Pix [20] and D-VST [43], along with the modality translation models SynDiff [26] and CycleGAN-Turbo [28], suffer from catastrophic distortions, precipitating an almost total collapse of cellular topological structures. While the overall performance of the Schrödinger Bridge-based UNSB [15] approximates that of CycleGAN [50], it degrades when attempting to preserve extremely fine cellular details, causing minute structures to be obscured by dark patches. Furthermore, conspicuous black-and-white speckle artifacts emerge across the generated images, indicating a failure of local mapping during the construction of the Schrödinger Bridge. In stark contrast, our Y-diff significantly alleviates the aforementioned severe cellular distortions and cluttered artifacts. It preserves most intricate cellular topologies while synthesizing optical textures of the target pCLE modality with high fidelity.

Quantitative Results. As demonstrated in Table 1(a), CycleGAN [50] achieves the second-best scores in SS and LPIPS, maintaining fundamental structural fidelity. However, its performance is severely constrained by extensive artifacts and erroneous structural mappings, precluding further improvements. Benefiting from the effective attention mechanism within its stage-wise training paradigm, D-VST [43] attains the highest LC score. Nevertheless, struggling to effectively model textures and structures within an entangled pixel space, D-VST [43], alongside PyramidPix2Pix [20], SynDiff [26], and CycleGAN-Turbo [28], suffers from catastrophic structural collapse. The contrastive-learning-based PPT [47], leveraging patch-mapping logic, achieves the second-best performance in Hist. Yet, its local constraints fail to maintain global consistency, resulting in a severe loss of cellular topology and a low overall perceptual score. Although the

Table 1: Quantitative comparison on the CCD and MIST datasets. **Best** and second best results are highlighted. ↑ indicates higher is better, and ↓ indicates lower is better.

(a) CCD dataset.

Metric	CycleGAN	PyrPix2Pix	PPT	SynDiff	D-VST	UNSB	CycleGAN-T	Ours
SS ↑	0.614	0.114	0.217	0.226	0.103	0.488	0.316	**0.689**
LC ↑	0.908	0.926	0.925	0.927	**0.953**	0.920	0.887	0.939
Hist ↑	0.731	0.823	0.854	0.815	0.853	0.798	0.672	**0.868**
LPIPS ↓	0.511	0.974	0.907	0.529	0.811	0.655	0.777	**0.495**

(b) MIST dataset.

Method	SS ↑	LC ↑	Hist ↑	LPIPS ↓	FID ↓	B-SSIM ↑	B-PSNR ↑	A-SSIM ↑	A-PSNR ↑
CycleGAN	0.7875	0.8938	0.6958	0.5025	59.8261	0.2198	13.6596	0.7198	12.9434
PPT	0.5420	0.8965	0.6873	0.4838	72.7841	0.1944	13.6642	0.6018	11.3111
PyrPix2Pix	0.0034	0.8966	0.6965	0.5367	108.3851	0.1754	13.3405	0.1015	9.6877
SynDiff	0.6626	0.8597	0.7020	0.4985	88.2607	0.1789	12.7212	0.7315	**14.1787**
UNSB	0.6543	0.8902	0.6908	0.5672	98.6930	0.2104	13.6935	0.6571	12.5585
D-VST	0.4293	0.8546	0.6778	0.5198	74.1360	0.1897	12.8145	0.4237	10.1237
CycleGAN-T	0.7812	0.8725	0.7123	0.5718	75.1487	0.1874	11.8567	0.7292	13.9730
Ours	**0.8164**	**0.9359**	**0.7938**	**0.4485**	**41.4010**	**0.2225**	**14.7326**	**0.7379**	13.9396

Table 2: Ablation study on key components of Y-diff. The *w/o Student* variant directly utilizes the target domain's true semantic code, serving as an oracle upper bound. **Best** and second best results are highlighted.

Method	SS ↑	LC ↑	Hist ↑	LPIPS ↓
w/o Flow matching	0.5397	0.9339	0.8646	0.4810
w/o SPNet	0.6090	0.9092	0.7814	0.4915
w/o Discriminator	0.6039	0.9199	0.7802	0.4992
Ours (Full)	0.6892	0.9387	0.8684	0.4948
w/o Student (Oracle)	**0.7015**	**0.9664**	**0.8807**	**0.4460**

Schrödinger Bridge-based UNSB [15] surpasses CycleGAN in LC and Hist, exacerbated unnatural black-and-white speckle artifacts precipitate a drastic decline in its SS and LPIPS metrics. In stark contrast, Y-diff shows a clear advantage. Aside from securing the second-best score in LC, it achieves the best performance on SS, Hist, and LPIPS, supporting our method's effectiveness.

Due to the limited space of the main text, we provide the qualitative results (Fig. 4) and quantitative comparison (Table 1(b)) on the MIST dataset in the main paper, while detailed experimental settings and analyses are provided in the supplementary material. The results show that Y-diff also achieves competitive performance in the virtual staining scenario.

Ablation Study. Table 2 presents the ablation study results concerning the core components of Y-diff. Bypassing the Student model to synthesize images directly utilizing the ground-truth global semantics of the target pCLE domain establishes the theoretical upper bound of our generative architecture. Upon ablating the Flow matching module, the network relies solely on the semantics of the source H&E domain. This severe domain misalignment introduces conflicting generative priors, resulting in a pronounced degradation in the SS metric. Furthermore, the removal of SPNet deprives the Student model of explicit

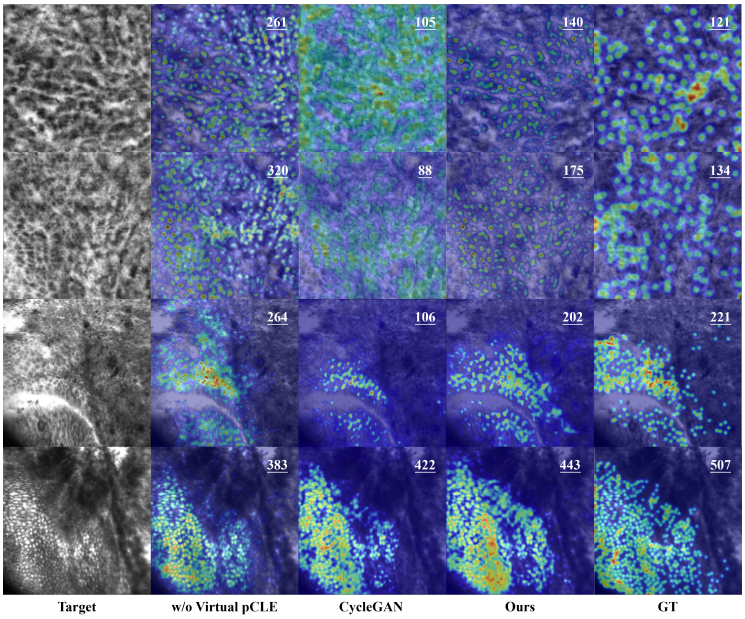

Fig. 6: Qualitative results on downstream verification. This figure presents the Class Activation Map (CAM) overlays and the corresponding predicted cell counts from U-Net models trained under diverse strategies. Specifically, *w/o Virtual pCLE* denotes that without injection of pseudo-pCLE images for training set augmentation; instead, the U-Net was trained exclusively on the manually annotated dataset, relying solely on conventional data augmentation techniques such as image rotation and color inversion.

structural constraints, compelling it to blindly distill all entangled information from the Teacher. This precipitates a significant distributional deviation between the synthesized outputs and the authentic pCLE modality. Finally, without the adversarial supervision of the Discriminator, the Student model struggles to capture high-frequency perceptual details and fails to converge to the target distribution relying merely on the remaining losses, triggering a comprehensive decline across all quantitative metrics. Detailed hyperparameter ablations are provided in the supplementary material.

4.3 Downstream Verification and Clinical Evaluation

We further employ a cell counting task as a clinical downstream application to validate the clinical significance of our algorithm. In practical clinical diagnostics, a prominent anatomical characteristic of the tumor microenvironment is an abnormally elevated cell density [23, 24], a feature that is particularly conspicuous under pCLE examination [16]. Consequently, cell counting serves as an effective modality for assessing the nature of examined regions in pCLE imaging. In this study, we adopt the widely utilized U-Net [4, 33] as the downstream

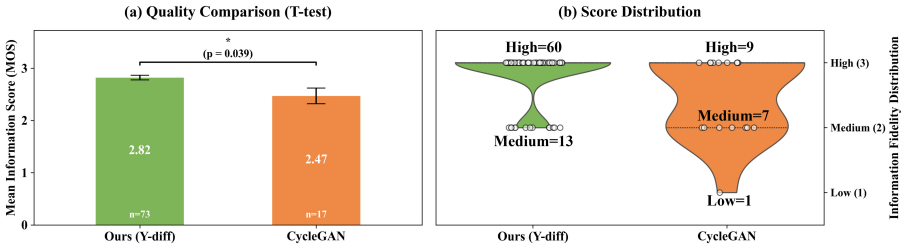

Fig. 7: Clinical Evaluation result. "High", "Medium", and "Low" denote a high degree of cellular structural fidelity, partial structural degradation with fundamentally correct overall topologies, and severe structural collapse rendering the image clinically unusable, respectively, mapped to scores of 3, 2, and 1. (a) Presents quantitative clinical evaluation results, where Y-diff attains a higher Mean Opinion Score (MOS). Welch's T-test shows this improvement is statistically significant ($p < 0.05$, denoted by $*$). (b) Visualizes score distributions across different models, highlighting Y-diff's pronounced advantage in achieving high structural fidelity.

Table 3: Quantitative results of the downstream cell counting task on the CCD dataset. *w/o Virtual pCLE* denotes that without injection of pseudo-pCLE images for training set augmentation. Present the statistical results of Mean Absolute Error(MAE), Root Mean Square Error(RMSE), and Mean Absolute Percentage Error(MAPE) for each method. **Best** and second best results are highlighted.

Method	MAE ↓	RMSE ↓	MAPE (%) ↓
w/o Virtual pCLE	83.29	103.29	55.77
CycleGAN	63.71	73.17	30.86
Ours (Y-diff)	**28.43**	**34.30**	**13.84**

validation model to rigorously verify the synthesis quality of the virtual pCLE data through controlled experiments.

We partition 65 images from the CCD, manually annotated for cells by pathology experts, into a fixed training set of 35 images and a testing set of 30 images based on approximate anatomical locations, and use the H&E ground-truth images from NuInsSeg [21] to generate pseudo-pCLE images for augmenting the training set while retaining their original segmentation labels. Referring to the qualitative and quantitative performance of the various baselines, this experiment selects the strongest-performing CycleGAN [50] and a method trained exclusively on the small-scale manually annotated data as our baselines. Please refer to the supplementary material for detailed experimental settings.

As shown in Fig. 6 and Table 3, the injection of pseudo-pCLE data generated by Y-diff effectively enhances the cell recognition capability of the U-Net network in this scenario, reducing its cell counting Mean Absolute Percentage Error (MAPE) from 55.77% to 13.84%. Referring to the clinical evaluation guidelines for virtual staining mentioned by Huang et al. [12], we define two criteria for clinical assessment: **(1) Morphological Fidelity**: Assesses the strict preser-

vation of cellular topologies and geometric contours. **(2) Texture Realism**: Evaluates the alignment of the synthesized images with the target modality's inherent optical properties.

To ensure the statistical significance of the evaluation results, we convened an expert panel comprising three pathologists of varying seniority (both junior and senior). Each expert independently conducted blind scoring on 30 pairs of generated images. Specifically, after each expert selects the more similar images based on texture realism, they then conduct further evaluation according to morphological fidelity; As shown in Fig. 7, the blind evaluation results indicate that the selection rate of images generated by Y-diff is significantly higher than that of other baseline method. Expert feedback reveals that the pseudo-pCLE synthesized by Y-diff exhibits a pronounced advantage in maintaining local cellular morphological features. Quantitative statistical results substantiate the extreme significance of this subjective preference.

5 Limitation

DDIM [35] improves structural reliability but its iterative sampling increases computation, slows inference relative to GAN-based methods [50], and limits translation to 256×256. Future work will explore Latent Diffusion Models [32] for real-time high-resolution generation and incorporate physical imaging priors to improve robustness under variable fluorescence intensity.

6 Conclusion

In this paper, We propose Y-diff for *ex vivo* H&E to *in vivo* pCLE translation. By synergizing Flow matching and Teacher-Student distillation, Y-diff decouples the learning of optical textures and spatial structures, effectively mitigating severe artifacts prevalent in current models. Extensive benchmark and clinical evaluations demonstrate that Y-diff achieves high-fidelity synthesis efficiently topological preservation, offering a reliable data augmentation solution for computational pathology.

Acknowledgements

This work was supported by National Key Research and Development Program of China (2025YFC2426300), the Science and Technology Commission of Shanghai Municipality (Nos.24511104100, 25ZR1402225, 24ZR1439800), National Nature Science Foundation of China Grants (82572314, 62477031, 62403307, 62271246), the Open Research Fund of The State Key Laboratory of Multimodal Artificial Intelligence Systems.

References

1. Ayyad, M., Gala, D., Albandak, M., Goyal, R.M., Abboud, Y., Al-Khazraji, A., Hajifathalian, K.: Probe-based confocal laser endomicroscopy: progress, challenges, and emerging applications. Surgical Endoscopy **39**, 7958–7972 (2025)
2. Cho, H., Moon, D., Heo, S.M., Chu, J., Bae, H., Choi, S., Lee, Y., Kim, D., Jo, Y., Kim, K., et al.: Artificial intelligence-based real-time histopathology of gastric cancer using confocal laser endomicroscopy. NPJ Precision Oncology **8**(1), 131 (2024)
3. De Bortoli, V., Korshunova, I., Mnih, A., Doucet, A.: Schrodinger bridge flow for unpaired data translation. In: Adv. Neural Inform. Process. Syst. vol. 37, pp. 103384–103441 (2024)
4. Falk, T., Mai, D., Bensch, R., Çiçek, Ö., Abdulkadir, A., Marrakchi, Y., Böhm, A., Deubner, J., Jäckel, Z., Seiwald, K., et al.: U-Net: deep learning for cell counting, detection, and morphometry. Nature methods **16**(1), 67–70 (2019)
5. Gu, Y., Vyas, K., Yang, J., Yang, G.Z.: Transfer recurrent feature learning for endomicroscopy image recognition. IEEE Trans. Med. Imaging **38**(3), 791–801 (2018)
6. Guleria, S., Shah, T.U., Pulido, J.V., Fasullo, M., Ehsan, L., Lippman, R., Sali, R., Mutha, P., Cheng, L., Brown, D.E., et al.: Deep learning systems detect dysplasia with human-like accuracy using histopathology and probe-based confocal laser endomicroscopy. Scientific reports **11**(1), 5086 (2021)
7. He, Y., Liu, Z., Qi, M., Ding, S., Zhang, P., Song, F., Ma, C., Wu, H., Cai, R., Feng, Y., et al.: PST-Diff: achieving high-consistency stain transfer by diffusion models with pathological and structural constraints. IEEE Trans. Med. Imaging **43**(10), 3634–3647 (2024)
8. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Adv. Neural Inform. Process. Syst. vol. 30. Curran Associates, Inc. (2017)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Adv. Neural Inform. Process. Syst. vol. 33, pp. 6840–6851 (2020)
10. Hu, Y., Du, Z., Lin, W., Yang, S., Yu, L., Zhang, G., Wang, L.: MCS-Stain: Boosting FFPE-to-HE virtual staining with multiple cell semantics. IEEE Trans. Med. Imaging **45**(4), 1475–1487 (2026)
11. Hu, Y., Peng, Q., Du, Z., Zhang, G., Wu, H., Liu, J., Chen, H., Wang, L.: Boosting FFPE-to-HE virtual staining with cell semantics from pretrained segmentation model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 67–76. Springer (2024)
12. Huang, L., Li, Y., Pillar, N., Keidar Haran, T., Wallace, W.D., Ozcan, A.: A robust and scalable framework for hallucination detection in virtual tissue staining and digital pathology. Nature Biomedical Engineering **9**(12), 2196–2214 (2025)
13. Iacucci, M., Jeffery, L., Acharjee, A., Grisan, E., Buda, A., Nardone, O.M., Smith, S.C., Labarile, N., Zardo, D., Ungar, B., et al.: Computer-aided imaging analysis of probe-based confocal laser endomicroscopy with molecular labeling and gene expression identifies markers of response to biological therapy in ibd patients: the endo-omics study. Inflammatory bowel diseases **29**(9), 1409–1420 (2023)
14. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking FID: Towards a better evaluation metric for image generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9307–9315 (2024)
15. Kim, B., Kwon, G., Kim, K., Ye, J.C.: Unpaired image-to-image translation via neural Schrödinger bridge. In: Int. Conf. Learn. Represent. vol. 2024, pp. 19312–19331 (2024)

16. Li, C., Zhang, Z., Ge, J., Liu, M., Wang, H., Zhu, L., Yao, H., Zhu, Y., Hu, X., Cai, Z., et al.: An in vivo pilot study of probe-based confocal laser endomicroscopy for detecting of pancreatic ductal adenocarcinoma intraoperatively. International Journal of Surgery **112**(3), 6686–6695 (2026)
17. Li, F., Hu, Z., Chen, W., Kak, A.: 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)
18. Lin, W., Hu, Y., Zhu, R., Wang, B., Wang, L.: Virtual staining for pathology: Challenges, limitations and perspectives. Intelligent Oncology **1**(2), 105–119 (2025)
19. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: Int. Conf. Learn. Represent. (2022), `https://arxiv.org/abs/2210.02747`, accessed 2026.6.23
20. Liu, S., Zhu, C., Xu, F., Jia, X., Shi, Z., Jin, M.: BCI: Breast cancer immunohistochemical image generation through pyramid pix2pix. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 1815–1824 (2022)
21. Mahbod, A., Polak, C., Feldmann, K., Khan, R., Gelles, K., Dorffner, G., Woitek, R., Hatamikia, S., Ellinger, I.: Nuinsseg: a fully annotated dataset for nuclei instance segmentation in H&E-stained histological images. Scientific Data **11**(1), 295 (2024)
22. Martirosyan, N.L., Georges, J., Eschbacher, J.M., Belykh, E., Carotenuto, A., Spetzler, R.F., Nakaji, P., Preul, M.C.: Confocal scanning microscopy provides rapid, detailed intraoperative histological assessment of brain neoplasms: experience with 106 cases. Clinical neurology and neurosurgery **169**, 21–28 (2018)
23. Mohammadi, H., Sahai, E.: Mechanisms and impact of altered tumour mechanics. Nature cell biology **20**(7), 766–774 (2018)
24. Nia, H.T., Munn, L.L., Jain, R.K.: Physical traits of cancer. Science **370**(6516), eaaz0868 (2020)
25. Oakden-Rayner, L.: Exploring large-scale public medical image datasets. Academic radiology **27**(1), 106–112 (2020)
26. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Cukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Trans. Med. Imaging **42**(12), 3524–3539 (2023)
27. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: Eur. Conf. Comput. Vis. pp. 319–345. Springer (2020)
28. Parmar, G., Park, T., Narasimhan, S., Zhu, J.Y.: One-step image translation with text-to-image models. arXiv preprint arXiv:2403.12036 (2024)
29. Perez, E., Strub, F., de Vries, H., Dumoulin, V., Courville, A.: FiLM: Visual reasoning with a general conditioning layer. In: AAAI. AAAI Press (2018)
30. Pilonis, N.D., Januszewicz, W., di Pietro, M.: Confocal laser endomicroscopy in gastro-intestinal endoscopy: technical aspects and clinical applications. Translational gastroenterology and hepatology **7**, 7 (2022)
31. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10619–10629 (2022)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022)

33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
34. Shi, Y., De Bortoli, V., Campbell, A., Doucet, A.: Diffusion schrödinger bridge matching. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 62183–62223 (2023)
35. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: Int. Conf. Learn. Represent. (2020), `https://arxiv.org/abs/2010.02502`, accessed 2026.6.23
36. Spasic, I., Nenadic, G.: Clinical text data in machine learning: systematic review. JMIR medical informatics **8**(3), e17984 (2020)
37. Su, X., Song, J., Meng, C., Ermon, S.: Dual diffusion implicit bridges for image-to-image translation. In: Int. Conf. Learn. Represent. (2022), `https://arxiv.org/abs/2203.08382`, accessed 2026.6.23
38. Villard, A., Breuskin, I., Casiraghi, O., Asmandar, S., Laplace-Builhe, C., Abbaci, M., Plana, A.M.: Confocal laser endomicroscopy and confocal microscopy for head and neck cancer imaging: Recent updates and future perspectives. Oral Oncology **127**, 105826 (2022)
39. Wang, L., Yoon, K.J.: Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3048–3068 (2021)
40. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
41. Xu, S., Ma, Z., Huang, Y., Lee, H., Chai, J.: Cyclenet: Rethinking cycle consistency in text-guided diffusion for image manipulation. In: Adv. Neural Inform. Process. Syst. vol. 36, pp. 10359–10384 (2023)
42. Xue, P., Guo, J., Che, Z., Liu, G., Ji, R., Ma, T., Chen, X., Zhao, Y., Ma, M., Shi, D., et al.: An automatic severity grading system using confocal laser endomicroscopy to evaluate inflammatory activity of ulcerative colitis: a prospective study. Scientific Reports **15**(1), 38008 (2025)
43. Yang, S., Wei, D., Hu, Y., Peng, Q., Liu, H., Huang, Y., Wu, X., Zheng, Y., Wang, L.: D-VST: Diffusion transformer for pathology-correct tone-controllable cross-dye virtual staining of whole slide images. In: Adv. Neural Inform. Process. Syst. vol. 38, pp. 3128–3162. Curran Associates, Inc. (2025)
44. Yim, J., Joo, D., Bae, J., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4133–4141 (2017)
45. Ying, X., Guo, H., Ma, K., Wu, J., Weng, Z., Zheng, Y.: X2CT-GAN: reconstructing ct from biplanar x-rays with generative adversarial networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10619–10628 (2019)
46. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 586–595 (2018)
47. Zhang, W., Hui, T.H., Tse, P.Y., Hill, F., Lau, C., Li, X.: High-resolution medical image translation via patch alignment-based bidirectional contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 178–188. Springer (2024)
48. Zhao, J., Li, X., Yang, F., Zhai, Q., Luo, A., Zhao, Y., Cheng, H., Fu, H.: MExD: An expert-infused diffusion model for whole-slide image classification. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20789–20799 (2025)

49. Zhou, L., Lou, A., Khanna, S., Ermon, S.: Denoising diffusion bridge models. In: Int. Conf. Learn. Represent. (2023), https://arxiv.org/abs/2309.16948, accessed 2026.6.23
50. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Int. Conf. Comput. Vis. pp. 2223–2232 (2017)