

Pathology-Guided Virtual Staining Metric for Evaluation and Training

Qiankai Wang¹, James E.D. Tweel¹, Parsin Haji Reza^{1*}†,
Anita Layton^{2,3,4,5†}

¹*Department of Systems Design Engineering, University of Waterloo,
200 University Ave W, Waterloo, N2L 3G1, Ontario, Canada.

²Department of Applied Mathematics, University of Waterloo, 200
University Ave W, Waterloo, N2L 3G1, Ontario, Canada.

³Cheriton School of Computer Science, University of Waterloo, 200
University Ave W, Waterloo, N2L 3G1, Ontario, Canada.

⁴Department of Biology, University of Waterloo, 200 University Ave W,
Waterloo, N2L 3G1, Ontario, Canada.

⁵School of Pharmacy, University of Waterloo, 10 Victoria St S A,
Kitchener, N2G 1C5, Ontario, Canada.

*Corresponding author(s). E-mail(s): parsin.hajireza@uwaterloo.ca;

Contributing authors: qiankai.wang@uwaterloo.ca;

james.tweel@uwaterloo.ca; anita.layton@uwaterloo.ca;

†These authors contributed equally to this work.

Abstract

Virtual staining has emerged as a powerful alternative to traditional histopathological staining techniques, enabling rapid, reagent-free image transformations. However, existing evaluation methods predominantly rely on full-reference image quality assessment (FR-IQA) metrics such as structural similarity, which are originally designed for natural images and often fail to capture pathology-relevant features. Expert pathology reviews have also been used, but they are inherently subjective and time-consuming.

In this study, we introduce PaPIS (Pathology-Aware Perceptual Image Similarity), a novel FR-IQA metric specifically tailored for virtual staining evaluation. PaPIS leverages deep learning-based features trained on cell morphology segmentation and incorporates Retinex-inspired feature decomposition to better reflect histological perceptual quality. Comparative experiments demonstrate

that PaPIS more accurately aligns with pathology-relevant visual cues and distinguishes subtle cellular structures that traditional and existing perceptual metrics tend to overlook. Furthermore, integrating PaPIS as a guiding loss function in a virtual staining model leads to improved histological fidelity.

This work highlights the critical need for pathology-aware evaluation frameworks to advance the development and clinical readiness of virtual staining technologies.

Keywords: Image Quality Assessment, Virtual Staining, Perceptual Similarity, Pathology-Aware Metrics

1 Introduction

In recent years, numerous virtual staining methods have been developed as an alternative to traditional histopathological staining techniques such as hematoxylin and eosin (H&E), immunohistochemistry (IHC), and Masson’s trichrome [1]. Conventional staining methods rely on chemical reagents, which can cause irreversible tissue alterations, limiting subsequent analyses and multi-modal imaging studies. In contrast, virtual staining preserves the tissue in its original state while computationally generating stained representations, enabling repeated analyses and facilitating the integration of different imaging modalities [1].

One of the key advantages of virtual staining is the elimination of chemical reagent consumption. Traditional staining protocols require costly chemical reagents, whereas virtual staining is purely computational, significantly reducing the expense of histopathological workflows. Additionally, virtual staining dramatically accelerates the staining process. While conventional staining can take tens of minutes to several hours, virtual staining can be completed within seconds or minutes, greatly improving the efficiency of pathology pipelines[2].

Despite these advantages, virtual staining currently lacks evaluation methods specifically tailored to the medical domain. Most assessments rely on full-reference image quality assessment (FR-IQA) metrics, which compare a processed image to a reference standard based on structural and perceptual similarities. While FR-IQA metrics are well-established in general image processing and have been widely applied in natural image quality assessment, their effectiveness in microscopic pathology imaging remains limited [3]. A fundamental limitation of traditional FR-IQA approaches is their focus on textural fidelity rather than histopathological relevance. These metrics are primarily designed for natural images captured by cameras, prioritizing structural and perceptual consistency without considering cellular morphology and tissue architecture, which are crucial for medical applications [3, 4]. Consequently, conventional FR-IQA metrics often fail to accurately assess the diagnostic quality of virtual staining images.

With advances in deep learning and feature engineering, perceptual evaluation metrics have increasingly enabled domain-specific comparisons [5, 6]. However, existing perceptual similarity metrics remain heavily biased toward natural image characteristics, making them inadequate for assessing medical imaging data. As a result,

many virtual staining images cannot be effectively evaluated using current methodologies, highlighting the need for a domain-specific, histopathology-aware evaluation framework [7].

To address this gap, this study proposes a histopathology-guided perceptual full-reference image assessment metric specifically designed for virtual staining evaluation. Additionally, an optimization framework for virtual staining models is developed based on the proposed evaluation metric. By integrating domain-specific histopathological knowledge into image quality assessment, this approach aims to provide a more clinically relevant evaluation of virtual staining results, ultimately improving their reliability and applicability in medical practice.

2 Literature Review

Virtual staining has emerged as a powerful approach for synthesizing H&E-equivalent images from label-free modalities, offering a reagent-free alternative to traditional histopathological staining. This label-free to stain transformation aims to replace physical dyes by generating stained-like outputs directly from inputs such as autofluorescence or photon-based signals. Ecclestone et al. [8] introduced the Photon Absorption Remote Sensing (PARS) system, which captures spectral and temporal photon absorption signatures to emulate H&E staining with enhanced cellular detail. Building on this direction, Rivenson et al. [9] and Wang et al. [10] demonstrated the efficacy of deep learning in mapping autofluorescence images to realistic H&E counterparts, broadening the applicability of virtual staining in clinical workflows.

The advent of generative deep learning models has further propelled the quality and fidelity of virtual staining. Extensions of the PARS framework by Tweel et al. [11] and Boktor et al. [12] integrated generative networks to process expanded spectral inputs and time-domain signals, improving visual realism and structural preservation. In parallel, diffusion-based models, such as those proposed by Saharia et al. [13], have shown strong potential for high-fidelity image-to-image translation in biomedical contexts. These advances build on foundational work in generative translation, notably Pix2Pix [14] and CycleGAN [15], which have been adapted to address the specific challenges of virtual staining tasks.

FR-IQA methods compare a processed image to a reference image to quantify differences in structure, perception, and statistical features. Wang et al. [16] introduced structural similarity index (SSIM), which evaluate pixel-wise differences and structural information to measure image fidelity. To further enhance structural evaluations, Wang et al. [17] developed multi-scale structural similarity (MS-SSIM), incorporating multi-resolution analysis to improve robustness. Beyond hand-crafted approaches, perceptual-based FR-IQA methods have been introduced to align more closely with human perception. Zhang et al. [5] proposed the Learned Perceptual Image Patch Similarity (LPIPS) metric, which utilizes deep neural networks to model perceptual judgments based on feature representations from trained vision models. Ding et al. [6] extended perceptual similarity analysis through Deep Image Structure and Texture Similarity (DISTS), integrating structural and texture information to improve quality assessment. Tian et al. [18] explored the impact of multiple reference images in

quality assessment, while Xian et al. [19] proposed structure-aware methods utilizing high-order statistical moments to capture intricate quality attributes.

Recent research has explored the application of FR-IQA in medical imaging. Breger et al. [3] demonstrated that standard FR-IQA metrics, originally designed for natural images, may not be directly applicable to medical imaging tasks, including MRI, CT, OCT, and digital pathology. Ohashi et al. [20] evaluated and adapted FR-IQA methods to better align with medical image quality requirements. Varga et al. [21] proposed optimized metric combinations to enhance prediction accuracy in medical imaging contexts. Additionally, Sujana et al. [22] investigated FR-IQA for structural MRI preprocessing, while Rodrigues et al. [23] examined perceptual quality assessment of medical images and videos.

3 Methods

Figure 1a illustrates the overall framework of the proposed perceptual similarity metric, PaPIS. This similarity metric assigns perceptual weights that reflect histological performance. To extract feature representations, image patches are first transformed into multi-channel embeddings using a pre-trained cell morphology segmentation model based on the work of Ignatov et al. [24]. Their model, which achieves state-of-the-art performance in nuclei segmentation and classification, adopts a dual-layer encoder-decoder architecture with an EfficientNet-B7-based encoder, as shown in Figure 1b.

Subsequently, intrinsic properties are extracted from the image features to obtain the reflection map R and the estimated illumination map L , using the Retinex algorithm [25]. The histology perceptual distance between these maps is then computed, inspired by prior works such as SSIM [16] and DISTS [6]. By calculating the distance between features derived from cell morphology, PaPIS offers a histology-aware quantification of image differences. This contrasts with traditional texture-based metrics such as SSIM and PSNR, which are primarily designed to align with human visual perception rather than pathological relevance.

To validate the practical utility of the proposed metric in pathological image analysis, we further apply it to evaluate the performance of a deep learning model that translates label-free histological images into H&E-stained representations.

3.1 Cell Morphology Level Feature Representation

Current state-of-the-art Full-Reference perceptual image similarity metrics, such as LPIPS [5] and DISTS [6], rely on feature extractors pre-trained on ImageNet to evaluate differences between generated and reference images. However, in the histology domain, pathologists focus on distinct perceptual attributes that cannot be adequately captured by models trained on natural image datasets.

To achieve superior performance in histological assessments, we utilize a feature extractor built upon the implementation of a state-of-the-art cell morphology segmentation task [24], which has demonstrated exceptional accuracy in nuclei segmentation and classification. In work by Ignatov et al. [24], cell morphology segmentation is

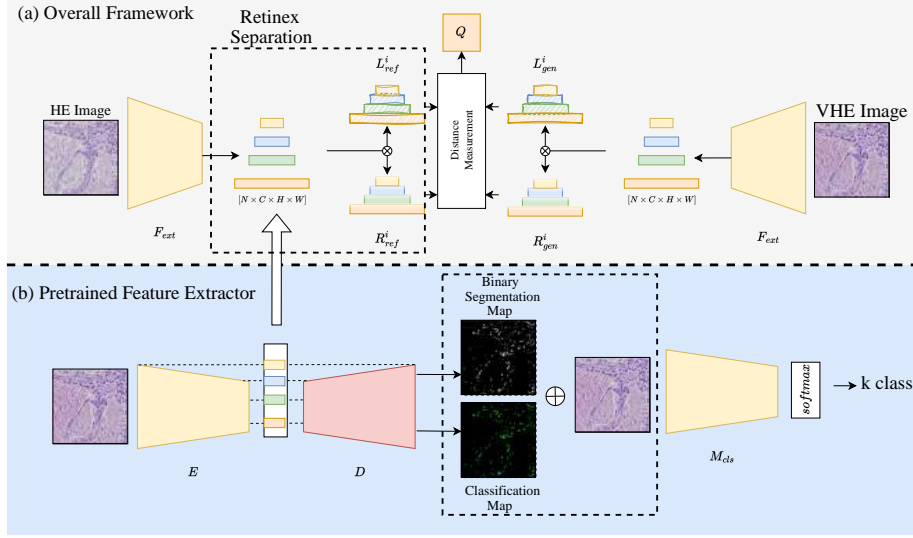


Fig. 1 (a) The figure illustrates the overall framework of the PaPIS. Image patches are fed into a pretrained feature extractor to obtain multi-channel feature representations. These features are decomposed into a set of estimated illumination maps L_a^i and reflectance maps R_a^i , where the superscript i denotes the feature layer and the subscript a indicates the image type (Reference or Generated images). The collection of decomposed components is subsequently used to compute perceptual similarity distances between image pairs. (b) This panel presents the internal architecture of the pretrained feature extractor. The feature extractor functions as the encoder component of a cell morphology segmentation model. The subsequent decoder generates both a binary segmentation map M_b , which preserves the spatial dimensions of the input, and a multi-layer classification map M_c . The fusion of M_b , M_c , and the original image patch is then fed into a classification model to produce the final classification label.

achieved using a dual-layer encoder-decoder architecture with an EfficientNet-B7-based encoder, the structure of which is shown in Figure 1b. The computed latent features utilized in our work, formally, can be written as:

$$f_{eff}(x) = \{\tilde{x}_j^{(i)}; i = 0, \dots, m; j = 1, \dots, n_i\} \quad (1)$$

where $x \in \mathbb{R}^{H \times W \times C}$ denotes the input image patch, $m = 4$ represents the number of selected convolutional blocks within the EfficientNet backbone, and n_i is the number of output channels (feature maps) at block i . The term $\tilde{x}_j^{(i)}$ refers to the j -th feature map from block i , normalized via channel-wise min-max normalization to ensure scale consistency across layers.

This formulation explicitly defines the multi-layer, multi-channel output of the encoder, which serves as the foundation for subsequent perceptual analysis. The architecture of the feature extractor is illustrated in Figure 1b.

To compare the perceptual focus of these distinct feature extractors, and obtain a unified representation from the multi-layer feature maps, we compute a reconstructed image I_{Rec} as follows:

$$I_{\text{Rec}} = \frac{1}{m} \sum_{i=0}^m f_{in} \left(\frac{1}{n} \sum_{j=0}^n \tilde{x}_j^{(i)} \right) \quad (2)$$

Here, $\tilde{x}_j^{(i)}$ denotes the j -th normalized feature map at layer i , as previously defined. The inner summation computes the average feature response across all channels n within each layer, while the function $f_{in}(\cdot)$ integrates these responses. Specifically, f_{in} denotes a linear interpolation function that maps the current image feature of size $\mathbb{R}^{w \times h}$ to the original spatial resolution $\mathbb{R}^{w_0 \times h_0}$, where w_0 and h_0 correspond to the width and height of the input image. The outer average aggregates information across all m selected layers, resulting in the final reconstructed representation I_{Rec} , which captures both local and hierarchical morphological features from the input image.

Figure 2 illustrates the visualization of extracted features, comparing the feature maps obtained from models pre-trained on the cell morphology histological task (F_{histo}) with those from models pre-trained on natural image datasets (F_{natural}). Specifically, Figure 2b shows that the deeper layers of F_{histo} focus more strongly on nuclear regions, exhibiting activation patterns that correspond to the locations of cell nuclei. In contrast, Figure 2a demonstrates that F_{natural} primarily captures low-level texture patterns such as edges and gradients, which are characteristic of natural images and reflect the visual features typically emphasized by models trained to match human perception.

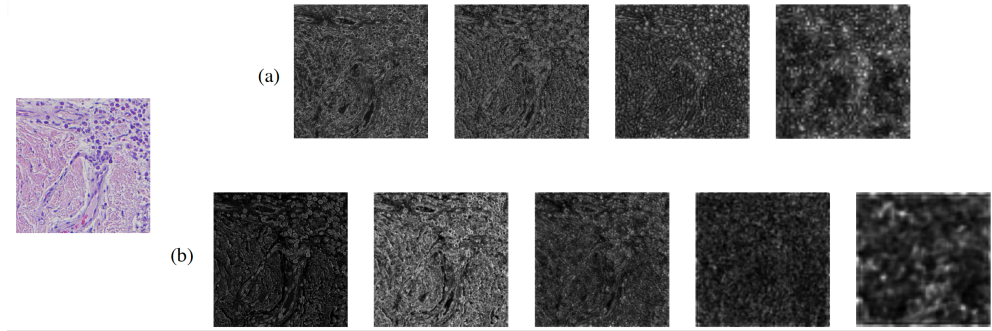


Fig. 2 (a) The first set of feature maps was visualized from a VGG16 model pretrained on the ImageNet dataset. These feature maps correspond to channels 3, 8, 15, 22, and 29, as used in the works of LPIPS [5] and DISTS [6]. (b) The second set of feature maps was extracted from the encoder output of a pretrained segmentation model, with dimensions of R^{192} , R^{288} , R^{480} , and R^{1344} , corresponding to channels at different encoder stages.

3.2 Retinex Properties on Features

In order to enhance the contrast of histological features, we apply the Multi-Scale Retinex (MSR) algorithm [25] to decompose the extracted features into two distinct components: the estimated illumination map and the reflectance map, which correspond to the low-frequency and high-frequency components within the extracted

feature space, respectively. The MSR algorithm models the decomposition of an input feature map based on the following relationship:

$$I(x, y) = R(x, y) \cdot L(x, y) \quad (3)$$

where $I(x, y)$ denotes the observed feature intensity at pixel location (x, y) , $R(x, y)$ is the reflectance component capturing intrinsic feature structures (e.g., edges and nuclei boundaries), and $L(x, y)$ represents the illumination component modeling smooth variations in local intensity due to shading or uneven activation.

To obtain a multi-scale estimation, we compute the illumination component at each scale i as:

$$L_i(x, y) = F_i(x, y) * G_{\sigma_i}(x, y) \quad (4)$$

where $F_i(x, y)$ denotes the extracted feature map at scale i , $G_{\sigma_i}(x, y)$ is a 2D Gaussian filter with standard deviation σ_i , and $*$ denotes the 2D convolution operation. The choice of σ_i controls the spatial frequency captured at each scale.

The reflectance component at each scale is then derived by logarithmic subtraction:

$$R_i(x, y) = \log F_i(x, y) - \log L_i(x, y) \quad (5)$$

where the logarithmic transformation serves to suppress multiplicative illumination effects and enhance the structural contrast in the feature map. The resulting $R_i(x, y)$ highlights fine-grained morphological features relevant to histological analysis.

3.3 PaPIS: Pathological Perceptual Distance

PaPIS integrates both low-frequency and high-frequency perceptual distances to evaluate histopathological image similarity. The decomposition of feature maps into illumination and reflectance components is based on the Retinex theory [25], which models an image as the product of its intrinsic reflectance and smooth illumination.

Inspired by SSIM [16] and DISTS [6], we define the high-frequency distance by comparing reflectance features extracted from Retinex decomposition. The high-frequency distance, denoted as $D_{\text{high}}(x, y, \alpha, \beta)$, is formulated as:

$$D_{\text{high}}(x, y, \alpha, \beta) = \alpha_{ij} \frac{2\mu_{R\tilde{x}_j}^{(i)}\mu_{R\tilde{y}_j}^{(i)} + c_1}{(\mu_{R\tilde{x}_j}^{(i)})^2 + (\mu_{R\tilde{y}_j}^{(i)})^2 + c_1} + \beta_{ij} \frac{2\sigma_{R\tilde{x}_j}^{(i)}\sigma_{R\tilde{y}_j}^{(i)} + c_2}{(\sigma_{R\tilde{x}_j}^{(i)})^2 + (\sigma_{R\tilde{y}_j}^{(i)})^2 + c_2} \quad (6)$$

In this formulation, the first term captures the similarity of mean values (μ) of the reflected features, representing brightness consistency across spatial locations and feature channels. The second term evaluates the similarity of standard deviations (σ), which reflects structural consistency in the high-frequency domain. The reflection statistics $\mu_{R\tilde{x}_j}^{(i)}$, $\mu_{R\tilde{y}_j}^{(i)}$, $\sigma_{R\tilde{x}_j}^{(i)}$, and $\sigma_{R\tilde{y}_j}^{(i)}$ denote the mean and standard deviation of

reflectance features \tilde{x} and \tilde{y} at layer i and channel j . Constants c_1 and c_2 are included to ensure numerical stability. The weights α_{ij} and β_{ij} are randomized such that $\sum_{i,j} \alpha_{ij} + \beta_{ij} = 1$, ensuring a balanced contribution between brightness and structure.

The low-frequency distance, D_{low} , is computed as the mean squared error (MSE) between illumination maps derived from each feature channel via the Retinex algorithm [25].

Finally, the complete PaPIS metric is defined as:

$$\begin{aligned} PaPIS(x, y, \lambda, \alpha, \beta) = & \lambda \sum_{i=0}^m \sum_{j=0}^{n_i} \text{MSE}(L(\tilde{x}_j^{(i)}), L(\tilde{y}_j^{(i)})) \\ & + D_{\text{high}}(x, y, \alpha, \beta) \end{aligned} \quad (7)$$

Here, λ is a hyperparameter that balances the contributions of the low-frequency and high-frequency components in the final distance score.

3.4 PaPIS Metric-Guided Loss for Enhancing Virtual Staining Models

Since PaPIS is utilized as a quality evaluation metric in generative tasks, incorporating it as an optimization objective during training could further improve the performance of virtual staining. Building on the unpaired training paradigm of the CycleGAN model, we propose a PaPIS-guided framework for virtual staining, designed to transform images from the PARS modality into H&E-stained representations. The complete pipeline is illustrated in Figure 3.

3.4.1 PARS images Acquisition

The pixel-by-pixel registration of scanned whole slide images (WSIs) and the enhancement process are based on the work by Tweel et al. [26]. Training patches are extracted using an automated script specifically designed for large-scale whole slide images. This script is capable of extracting the required patches within 15 seconds, significantly improving efficiency. The training dataset is constructed in paired format rather than in a random order, following the methodology outlined in the work of Tweel et al. [11].

3.4.2 PaPIS-guided Virtual Staining CycleGAN

Building upon the standard CycleGAN framework [15], we introduce an auxiliary perceptual loss term based on the proposed PaPIS metric to encourage the generated images to better align with histological properties of the target domain. This PaPIS-based loss is defined as:

$$\mathcal{L}_{\text{papis}}(x, G(x)) = 1 - \text{PaPIS}(x, G(x), \lambda, \alpha, \beta) \quad (8)$$

where x is the input unstained image and $G(x)$ is the corresponding virtual H&E image generated by the generator G . The PaPIS score quantifies the pathological perceptual similarity between the input and generated image; thus, minimizing $\mathcal{L}_{\text{papis}}$ encourages perceptual closeness in both low- and high-frequency histological features.

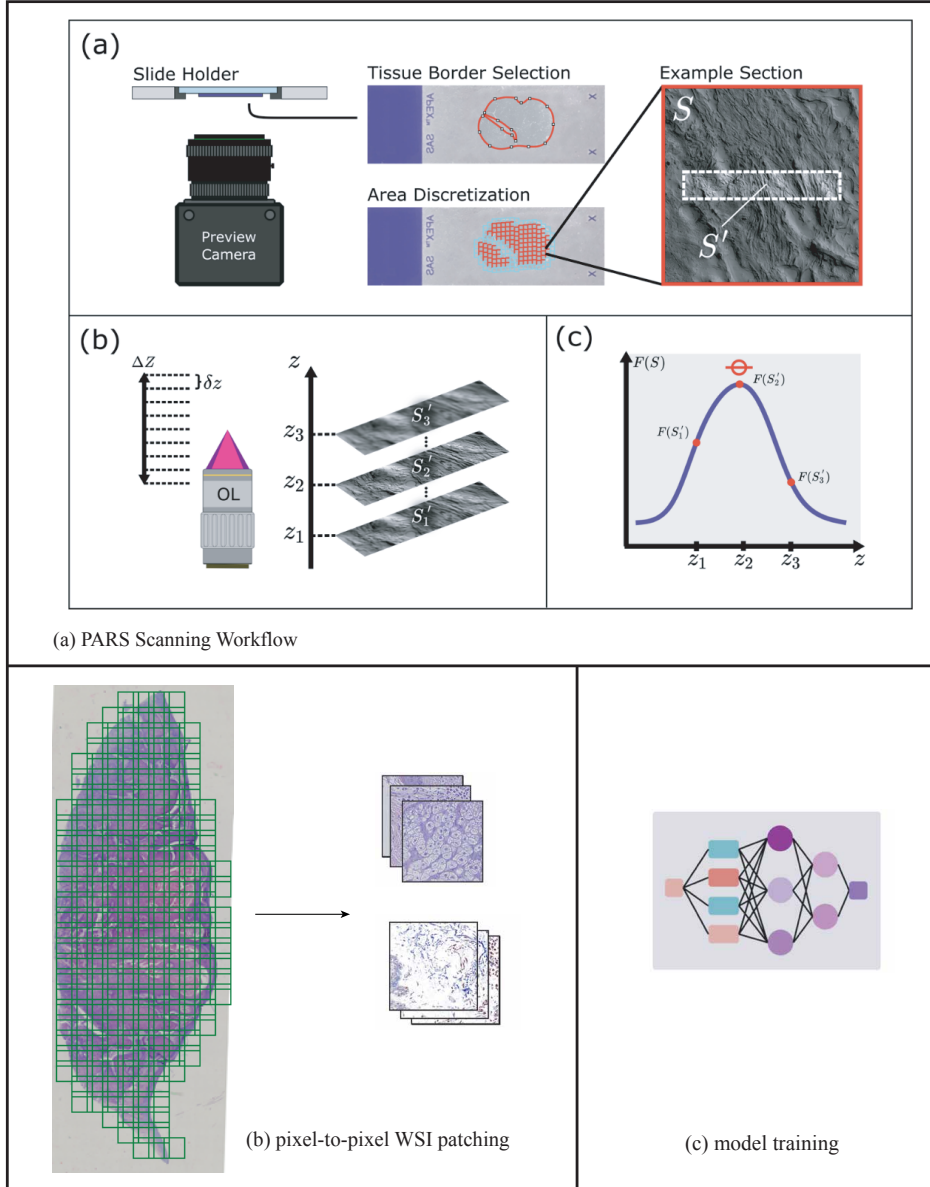


Fig. 3 Pipeline for the PaPIS-guided virtual staining. Part (a) is adapted from [26]. The complete pipeline consists of slide scanning, patch extraction, and model training, forming an end-to-end framework for virtual staining.

The overall training loss for the modified CycleGAN is formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cycle}} + \lambda_2 \mathcal{L}_{\text{papis}} + \lambda_3 \mathcal{L}_g + \lambda_4 \mathcal{L}_d \quad (9)$$

Here, $\mathcal{L}_{\text{cycle}}$ denotes the cycle consistency loss, \mathcal{L}_g is the generator adversarial loss, and \mathcal{L}_d is the discriminator loss. The weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters that control the contribution of each loss component. Figure 4 illustrates the modified CycleGAN architecture with PaPIS-guided perceptual supervision.

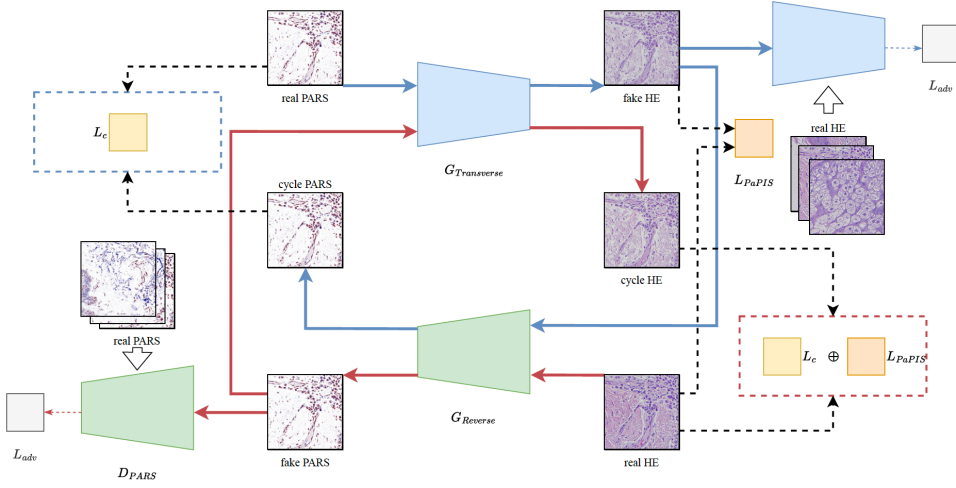


Fig. 4 The architecture of the PaPIS-guided virtual staining model. Built upon the CycleGAN[15] framework, the model incorporates additional loss terms to measure the distance between the generated (fake) H&E images and the real H&E images (ground truth), as well as the distance between cycle-consistent H&E images and the real H&E images.

4 Experiments

4.1 Dataset Acquisition

4.1.1 Modalities of Images Acquisition

All scanning and post-processing procedures are detailed in Section 3.4.1. The dataset consists of five pairs of whole slide images (WSIs) with dimensions of $18,202 \times 48,800$, $9,200 \times 36,200$, $59,600 \times 61,418$, $17,664 \times 17,876$ and $35,344 \times 37,957$, respectively. Four of the scanned WSIs are divided into 1024×1024 patches, resulting in a total of 10,086 patches. These patches are then randomly flipped and resized to 256×256 for the training process. The remaining WSI pair is designated for the comparative experiment. A filtering process is first applied to extract histological regions, followed by randomly cropping into 206 synchronized 1024×1024 patches for subsequent analysis.

4.1.2 Virtual Stained Image Generate

The dataset utilized in this study was generated using a virtual staining model based on the CycleGAN architecture [15]. The loss function parameters were set as follows: $\lambda_1 = 2.0$, $\lambda_2 = 1.0$, $\lambda_3 = 1.0$, and $\lambda_4 = 1.0$. Training was conducted with a batch size of 1. The Adam optimizer was used for both the generator and discriminator networks, with a learning rate of 0.001 and β parameters of (0.5, 0.999). An accumulative gradient strategy with an accumulation count of 2 was applied. A linear learning rate decay strategy was employed using the `LinearLrInterval` scheduler, with updates every 1000 iterations. The learning rate decayed linearly from 0.001 to 0 between the 10,000th and 50,000th iterations.

4.2 Comparison with Previous Metrics

For the comparison of image pairs, we calculate the following metrics: PSNR, SSIM [16], MS-SSIM [17], LPIPS [5], and DISTS [6]. Based on the relationship between these metrics and PaPIS scores, we define four distinct categories:

1. **AH (Aligned High)**: Points exhibiting both high PaPIS scores and high values across the compared metrics, indicating strong agreement between PaPIS and traditional measures.
2. **AL (Aligned Low)**: Points with both low PaPIS scores and low metric values, reflecting consistent low similarity across both evaluation methods.
3. **PD (PaPIS-Dominant)**: Points where PaPIS scores are high despite low traditional metric values, suggesting cases where PaPIS captures relevant pathological features overlooked by conventional metrics.
4. **TD (Traditionally-Dominant)**: Points with low PaPIS scores but high traditional metric values, highlighting instances where traditional metrics indicate high similarity, yet PaPIS identifies deficiencies in pathology-relevant features.

We categorize the comparison metrics into two groups: traditional handcrafted metrics and perceptual feature-based perceptual metrics.

4.2.1 PaPIS and Handcrafted Metrics

In this section, we compare our proposed metric with traditional handcrafted metrics, specifically SSIM [16]. A case-by-case scatter plot is presented in Figure 5, while additional image-to-image examples are provided in Figure 6 for further comparison.

Among the AH and AL points, both selected data points exhibit strong alignment between PaPIS and traditional metrics in assessing image quality. In contrast, the PD and TD points highlight the differing sensitivities of PaPIS and traditional approaches. PD points correspond to cases where nuclear structures and spatial alignment are well-preserved, yet fine textures and subtle color variations result in lower scores from traditional metrics, revealing their limitations in capturing histological relevance. Conversely, TD points receive high scores from traditional metrics due to uniform textures or consistent coloration but may lack proper histological structures or exhibit misalignment in cellular organization, leading to lower PaPIS scores. This

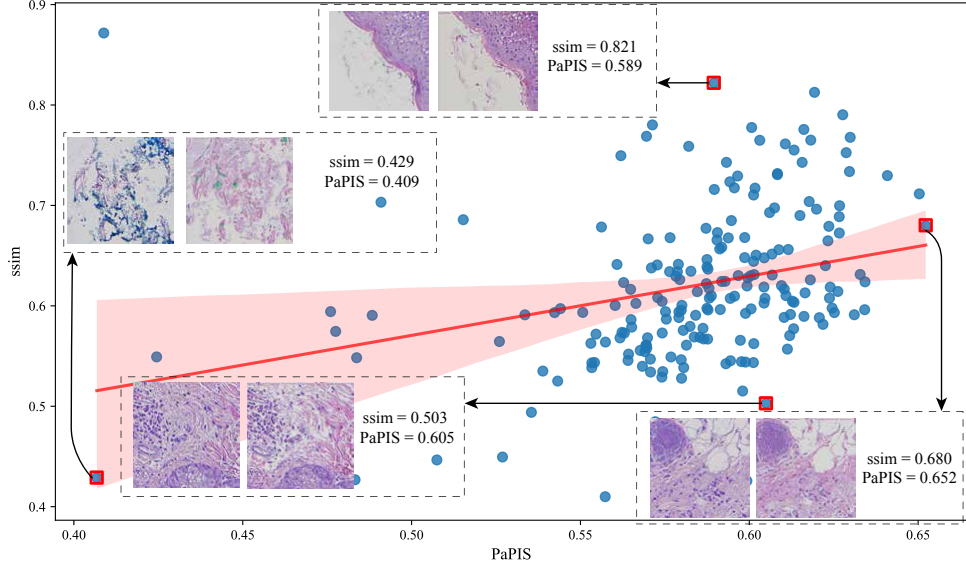


Fig. 5 Scatter plot comparing SSIM and PaPIS across different image pairs. Selected data points are labeled as AH (Aligned High), AL (Aligned Low), PD (PaPIS-Dominant), and TD (Traditionally-Dominant), representing specific cases for analysis.

contrast underscores the ability of PaPIS to better capture pathologically significant features that traditional methods may overlook.

4.2.2 PaPIS and Other Feature-Based Perceptual Metrics

In this section, PaPIS is compared with other perceptual-based similarity metrics, specifically LPIPS [5] and DISTS [6]. A case-by-case scatter plot is presented in Figure 7 and 8. At the AH and AL points, both metrics consistently identify images as either high or low in overall quality. However, at the PD points, the images tend to be concentrated in regions with high cellular density, whereas at the TD points, they are primarily located at tissue edges or in areas with sparse cellular distribution. Notably, neither perceptual metric provides explicit information regarding morphological differences in cellular structures. Nevertheless, compared to handcrafted metrics, traditional perceptual metrics exhibit higher consistency with PaPIS, suggesting an improved alignment in assessing histologically relevant features.

4.3 PaPIS-guided Virtual Staining

For the PaPIS-guided virtual staining process, the model is trained following the parameters outlined in Section 4.1.2, ensuring a fair comparison of performance. The weight coefficient λ_2 is set to 2.0, introducing the \mathcal{L}_{PaPIS} loss term to balance the overall loss function. As shown in Figure 9, the PaPIS-guided model enhances the consistency of cellular morphology, including cell position, shape, and size. Additionally,

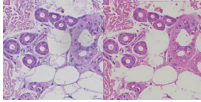
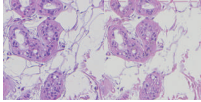
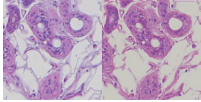
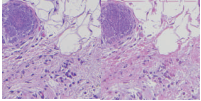
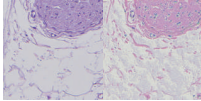
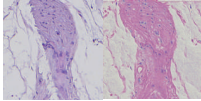
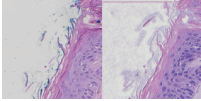
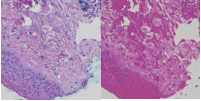
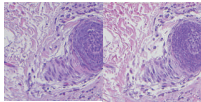
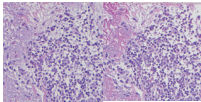
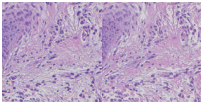
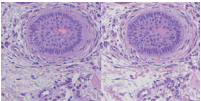
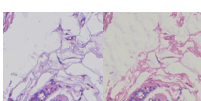
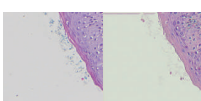
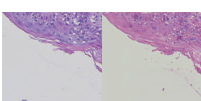
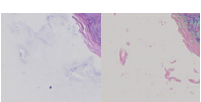
AH	 PaPIS = 0.650 SSIM = 0.712	 PaPIS = 0.641 SSIM = 0.730	 PaPIS = 0.630 SSIM = 0.734	 PaPIS = 0.652 SSIM = 0.680
AL	 PaPIS = 0.476 SSIM = 0.594	 PaPIS = 0.484 SSIM = 0.548	 PaPIS = 0.478 SSIM = 0.575	 PaPIS = 0.483 SSIM = 0.427
PD	 PaPIS = 0.611 SSIM = 0.557	 PaPIS = 0.634 SSIM = 0.596	 PaPIS = 0.631 SSIM = 0.591	 PaPIS = 0.613 SSIM = 0.577
TD	 PaPIS = 0.562 SSIM = 0.750	 PaPIS = 0.590 SSIM = 0.822	 PaPIS = 0.619 SSIM = 0.813	 PaPIS = 0.498 SSIM = 0.878

Fig. 6 Case-wise comparison of PaPIS and SSIM values across different histological image samples.

the model effectively mitigates certain image detail losses, leading to a more faithful reconstruction of fine structures. In terms of overall perceptual quality, the generated images exhibit improved resemblance to the ground truth. The PaPIS-guided virtual staining model demonstrates enhanced performance in both histological similarity and high-frequency texture preservation, optimizing pathological feature retention in the generated images.

4.3.1 Regional Sensitivity Analysis via Heatmap Visualization

To further assess the spatial sensitivity of PaPIS compared to traditional metrics, we conducted a region-wise similarity analysis on a representative whole slide image. The image was divided into non-overlapping 1024×1024 patches, and similarity scores were computed using both SSIM and PaPIS. These values were then visualized as spatial heatmaps, as shown in Fig. 10. Warmer colors indicate higher similarity.

Observations.

From this analysis, we derive the following observations:

1. **Texture Sensitivity vs. Morphological Awareness:** SSIM shows a sharp contrast between background and tissue regions, primarily responding to local

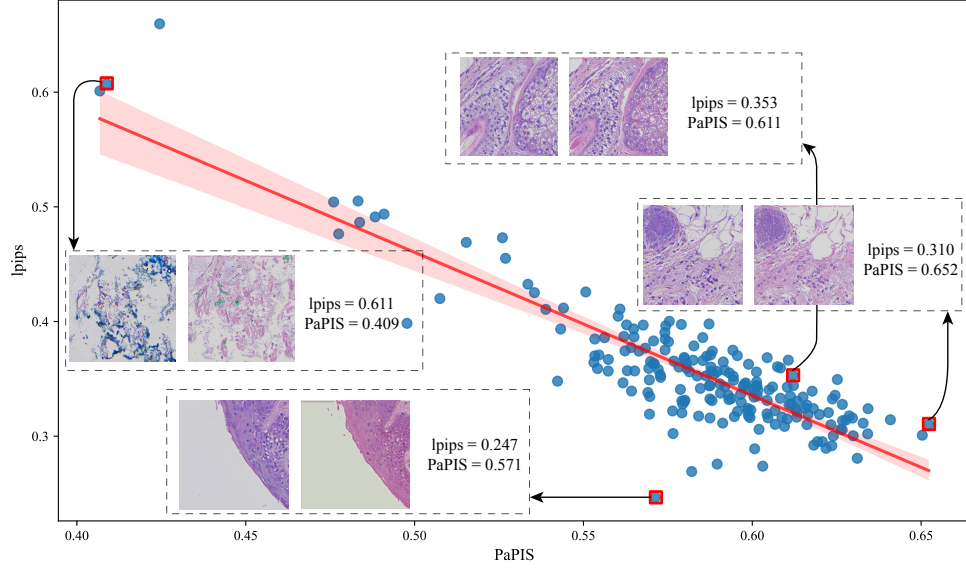


Fig. 7 Scatter plot comparing LPIPS and PaPIS across different image pairs. Each data point represents an image pair evaluated by both metrics, illustrating variations in their assessments. This comparison highlights differences in perceptual and pathology-aware similarity measurements between the two approaches.

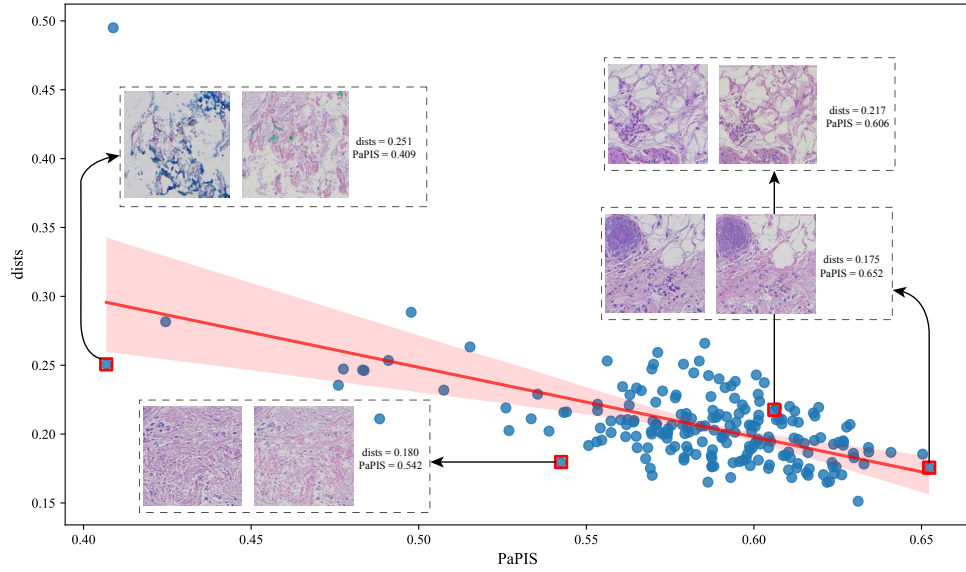


Fig. 8 Scatter plot comparing DISTIS and PaPIS across different image pairs.

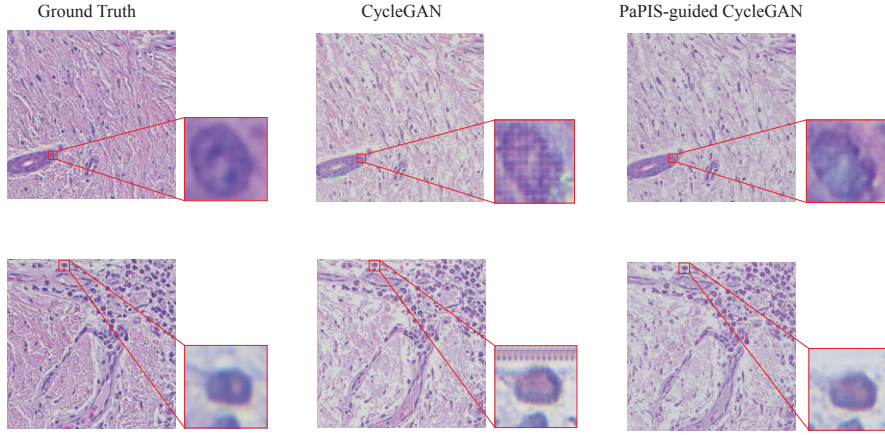


Fig. 9 Comparison of virtual staining results, showing ground truth images, CycleGAN-generated images, and PaPIS-guided CycleGAN-generated images. The figure illustrates differences in staining quality and structural preservation across the methods.

texture presence. In contrast, PaPIS produces smoother gradients at tissue boundaries and highlights internal structure variations, suggesting better alignment with morphological integrity.

2. **Detection of Staining Artifacts:** In certain regions with high SSIM but visually suboptimal staining, PaPIS assigns noticeably lower scores, demonstrating its capacity to penalize virtual staining artifacts even when conventional metrics fail to do so.
3. **Intra-Tissue Structural Differentiation:** PaPIS reveals greater variance within tissue interiors than SSIM, which tends to yield flat responses. This indicates that PaPIS is more sensitive to histologically relevant features such as nuclear organization or glandular structure.
4. **Sensitivity to Functional Regions:** In regions containing critical micro-anatomical structures (e.g., ducts, tumor nests), SSIM maintains high scores despite morphological degradation. In contrast, PaPIS shows significant score drops, reflecting its better alignment with pathology-relevant structures.
5. **Complementary Applications:** The divergence between SSIM and PaPIS highlights potential complementary use cases: SSIM may remain useful for detecting broad texture-based distortions, while PaPIS provides superior guidance for histology-aware evaluation and diagnostic quality control.

These insights underscore the importance of incorporating domain-specific perceptual criteria into similarity metrics. PaPIS enables fine-grained, region-aware evaluation that aligns more closely with expert interpretation and diagnostic relevance,

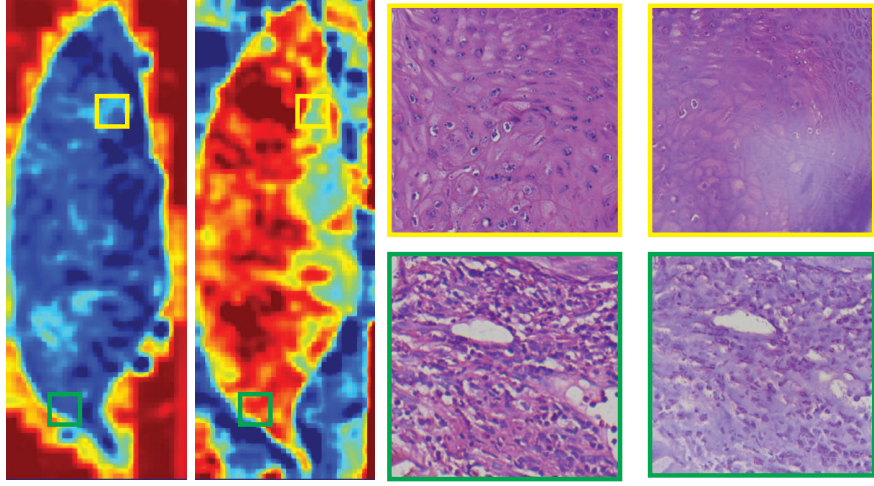


Fig. 10 Patch-wise similarity heatmaps of a representative WSI. **Left:** SSIM heatmap; **Middle:** PaPIS heatmap; **Right:** Magnified views of selected regions. Warmer colors indicate higher similarity. Compared to SSIM, which often assigns high similarity scores in low-texture or background areas, PaPIS demonstrates greater sensitivity to morphological and contextual cues.

offering a distinct advantage over traditional full-reference IQA methods in virtual staining workflows.

5 Discussions and Conclusions

In this study, we introduced PaPIS, a pathology-guided FR-IQA metric tailored for virtual staining evaluation. By leveraging a cell morphology-aware feature extractor and Retinex-based feature decomposition, PaPIS captures both high-frequency nuclear structures and low-frequency illumination characteristics—features highly relevant to histological interpretation. Our experimental results demonstrate that PaPIS achieves superior performance in identifying pathology-relevant differences across image pairs, outperforming traditional handcrafted metrics (e.g., SSIM, PSNR) and perceptual feature-based metrics (e.g., LPIPS, DISTs). Furthermore, the integration of PaPIS as a perceptual loss into a CycleGAN-based virtual staining model improved histological fidelity in generated outputs, suggesting that PaPIS can be used not only as an evaluation tool but also as a training objective.

A notable limitation of the current study is the absence of subjective validation by expert pathologists. While PaPIS shows promising quantitative alignment with structural and morphological features important to histopathology, its clinical validity remains to be established through direct correlation with expert perception and diagnostic utility. This limitation stems from the logistical complexity and resource demands of conducting large-scale reader studies. Nevertheless, we acknowledge this as a critical direction for future work. We plan to design a reader study in collaboration with board-certified pathologists, involving pairwise comparisons of virtual staining outputs across different models and correlation analysis between PaPIS scores and

expert ratings. This will help quantify the interpretability and clinical relevance of the proposed metric.

Beyond the current experiments based on PARS imaging, we also anticipate the applicability of PaPIS to other label-free virtual staining modalities such as autofluorescence microscopy, quantitative phase imaging, and hyperspectral imaging. Since PaPIS operates in a feature space that captures cell-level morphology rather than being tied to specific pixel distributions, it is inherently modality-agnostic and can generalize across structurally diverse imaging sources, as long as the target virtual stain preserves histological structures.

Moreover, the modular structure of PaPIS enables generalization across tissue types and staining protocols. While the current implementation uses a feature encoder trained on H&E-based nuclei segmentation, the framework can be extended by incorporating encoders trained on different tissue domains or fine-tuned using transfer learning. This flexibility opens opportunities for PaPIS to serve as a generalized perceptual metric across various virtual staining scenarios, including different stain types (e.g., Masson’s trichrome, IHC) and across species or clinical contexts.

Despite the limitations in subjective validation, the strong quantitative performance and architectural flexibility of PaPIS position it as a valuable tool for automated quality assessment in computational pathology. It offers an interpretable, pathology-aware alternative to conventional IQA metrics and has the potential to guide model selection, training optimization, and quality control in large-scale virtual staining workflows. In future work, we aim to further explore PaPIS’s clinical relevance, extend its applicability to broader imaging contexts, and integrate it with human-in-the-loop systems for semi-automated quality assurance.

Acknowledgements. The authors used OpenAI’s ChatGPT (version GPT-4, accessed May 2025) to assist in improving the English language and clarity of the manuscript during the revision stage. All intellectual content was written and verified by the authors.

References

- [1] Bai, B., Yang, X., Li, Y., Zhang, Y., Pillar, N., Ozcan, A.: Deep learning-enabled virtual histological staining of biological samples. *Light: Science & Applications* **12**(1), 57 (2023)
- [2] Latonen, L., Koivukoski, S., Khan, U., Ruusuvuori, P.: Virtual staining for histology by deep learning. *Trends in Biotechnology* (2024)
- [3] Breger, A., Biguri, A., Landman, M.S., Selby, I., Amberg, N., Brunner, E., Gröhl, J., Hatamikia, S., Karner, C., Ning, L., et al.: A study of why we need to reassess full reference image quality assessment with medical images. *arXiv preprint arXiv:2405.19097* (2024)
- [4] Pambrun, J.-F., Noumeir, R.: Limitations of the ssim quality metric in the context of diagnostic imaging. In: 2015 IEEE International Conference on Image Processing (ICIP), pp. 2960–2963 (2015). IEEE

- [5] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: 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)
- [6] Ding, K., Ma, K., Wang, S., Simoncelli, E.P.: Image quality assessment: Unifying structure and texture similarity. *IEEE transactions on pattern analysis and machine intelligence* **44**(5), 2567–2581 (2020)
- [7] Breger, A., Karner, C., Selby, I., Gröhl, J., Dittmer, S., Lilley, E., Babar, J., Beckford, J., Else, T.R., Sadler, T.J., et al.: A study on the adequacy of common iqa measures for medical images. *arXiv preprint arXiv:2405.19224* (2024)
- [8] Ecclestone, B.R., Simmons, J.A.T., Tweel, J.E., Kaur, C., Hajiahmadi, A., Reza, P.H.: Photon absorption remote sensing (pars): A comprehensive approach to label-free absorption microscopy across biological scales. *arXiv preprint arXiv:2403.04229* (2024)
- [9] Rivenson, Y., Wang, H., Wei, Z., Zhang, Y., Gunaydin, H., Ozcan, A.: Deep learning-based virtual histology staining using auto-fluorescence of label-free tissue. *arXiv preprint arXiv:1803.11293* (2018)
- [10] Wang, J., Xiong, B., Zhou, Y., Cao, X., Ma, Z.: A value mapping virtual staining framework for large-scale histological imaging. *arXiv preprint arXiv:2501.03592* (2025)
- [11] Tweel, J.E., Ecclestone, B.R., Boktor, M., Simmons, J.A.T., Fieguth, P., Reza, P.H.: Virtual histology with photon absorption remote sensing using a cycle-consistent generative adversarial network with weakly registered pairs. *arXiv preprint arXiv:2306.08583* (2023)
- [12] Boktor, M., Tweel, J.E., Ecclestone, B.R., Ye, J.A., Fieguth, P., Haji Reza, P.: Multi-channel feature extraction for virtual histological staining of photon absorption remote sensing images. *Scientific Reports* **14**(1), 2009 (2024)
- [13] Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10 (2022)
- [14] Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134 (2017)
- [15] Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232 (2017)

- [16] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)
- [17] Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, vol. 2, pp. 1398–1402 (2003). Ieee
- [18] Tian, Y., Chen, B., Wang, S., Kwong, S.: Towards thousands to one reference: Can we trust the reference image for quality assessment? *IEEE Transactions on Multimedia* **26**, 3278–3290 (2023)
- [19] Xian, W., Zhou, M., Fang, B., Xiang, T., Jia, W., Chen, B.: Perceptual quality analysis in deep domains using structure separation and high-order moments. *IEEE Transactions on Multimedia* **26**, 2219–2234 (2023)
- [20] Ohashi, K., Nagatani, Y., Yoshigoe, M., Iwai, K., Tsuchiya, K., Hino, A., Kida, Y., Yamazaki, A., Ishida, T.: Applicability evaluation of full-reference image quality assessment methods for computed tomography images. *Journal of Digital Imaging* **36**(6), 2623–2634 (2023)
- [21] Varga, D.: Full-reference image quality assessment based on an optimal linear combination of quality measures selected by simulated annealing. *Journal of Imaging* **8**(8), 224 (2022)
- [22] Sujana, D.S., Augustine, D.P., Grace, D.S.R.: Full reference image quality assessment (fr-iqa) of pre-processed structural magnetic resonance images. In: *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, vol. 1, pp. 1–5 (2024). IEEE
- [23] Rodrigues, R., Lévêque, L., Gutiérrez, J., Jebbari, H., Outtas, M., Zhang, L., Chetouani, A., Al-Juboori, S., Martini, M.G., Pinheiro, A.M.: Objective quality assessment of medical images and videos: Review and challenges. *Multimedia Tools and Applications*, 1–34 (2024)
- [24] Ignatov, A., Yates, J., Boeva, V.: Histopathological image classification with cell morphology aware deep neural networks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6913–6925 (2024)
- [25] Jobson, D.J., Rahman, Z.-u., Woodell, G.A.: A multiscale retinex for bridging the gap between color images and the human observation of scenes. *IEEE Transactions on Image processing* **6**(7), 965–976 (1997)
- [26] Tweel, J.E., Ecclestone, B.R., Boktor, M., Dinakaran, D., Mackey, J.R., Reza, P.H.: Automated whole slide imaging for label-free histology using photon absorption remote sensing microscopy. *IEEE Transactions on Biomedical Engineering* **71**(6), 1901–1912 (2024)