Intelligent Oncology 1 (2025) 105–119

Contents lists available at ScienceDirect

Intelligent Oncology

journal homepage: www.keaipublishing.com/en/intonc/

Review

# Virtual staining for pathology: Challenges, limitations and perspectives

Weiping Lin[a], Yihuang Hu[a], Runchen Zhu[b], Baoshun Wang[a], Liansheng Wang[a,*]

[a] Department of Computer Science, School of Informatics, Xiamen University, Xiamen Fujian 361005, China
[b] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen Fujian 361005, China

ARTICLE INFO

*Keywords:*
Virtual staining
Pathological image generation
Models and datasets
Model evaluation

ABSTRACT

In pathological examinations, tissue must first be stained to meet specific diagnostic requirements, a meticulous process demanding significant time and expertise from specialists. With advancements in deep learning, this staining process can now be achieved through computational methods known as virtual staining. This technique replicates the visual effects of traditional histological staining in pathological imaging, enhancing efficiency and reducing costs. Extensive research in virtual staining for pathology has already demonstrated its effectiveness in generating clinically relevant stained images across a variety of diagnostic scenarios. Unlike previous reviews that broadly cover the clinical applications of virtual staining, this paper focuses on the technical methodologies, encompassing current models, datasets, and evaluation methods. It highlights the unique challenges of virtual staining compared to traditional image translation, discusses limitations in existing work, and explores future perspectives. Adopting a macro perspective, we avoid overly intricate technical details to make the content accessible to clinical experts. Additionally, we provide a brief introduction to the purpose of virtual staining from a medical standpoint, which may inspire algorithm-focused researchers. This paper aims to promote a deeper understanding of interdisciplinary knowledge between algorithm developers and clinicians, fostering the integration of technical solutions and medical expertise in the development of virtual staining models. This collaboration seeks to create more efficient, generalized, and versatile virtual staining models for a wide range of clinical applications.

## 1. Introduction

Pathological examination is the gold standard for cancer diagnosis.[1] This process involves the analysis of tissues, cells, and organs to detect diseases, particularly cancers and other abnormalities. The primary objectives of pathological examination include diagnosing diseases, staging and grading cancers, estimating prognosis, monitoring treatment responses, and guiding treatment plans.[2] In pathological examination, the initial step involves staining the tissue of the lesion, a critical process that requires selecting appropriate staining methods based on specific diagnostic needs. For example, hematoxylin and eosin (H&E) staining is the most commonly used technique, where hematoxylin stains cell nuclei blue, and eosin stains the cytoplasm and extracellular matrix pink.[3] Other types of pathological staining include periodic acid–Schiff staining,[4] Masson's trichrome staining,[5] Giemsa staining,[6] immunohistochemistry (IHC),[7] and Ziehl–Neelsen staining.[8]

* Corresponding author.
*E-mail address:* lswang@xmu.edu.cn (L. Wang).

https://doi.org/10.1016/j.intonc.2025.03.005
Received 2 November 2024; Received in revised form 25 January 2025; Accepted 10 March 2025
Available online 18 March 2025
2950-2616/© 2025 The Author(s). Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Traditional pathological staining relied on specialized pathologists or clinicians in the laboratory. The process generally involves several key steps to prepare and stain tissue samples. The workflow includes tissue collection and processing (e.g., dehydration, clearing, and embedding), sectioning, deparaffinization, and staining. Each step is crucial for ensuring high-quality tissue samples and accurate staining results, which are both vital to reliable pathological diagnoses. Traditional pathological staining is often labor-intensive and time-consuming, typically requiring several days to weeks from tissue collection to microscopic examination. Furthermore, certain diagnoses require multiple staining techniques,[9] adding repetition and complexity to the process. Additionally, antibodies for some immunohistochemical stains can be prohibitively expensive. In summary, traditional pathological staining is a time-consuming and labor-intensive process that not only increases the workload of pathologists but also carries the risk of delaying patient treatment due to its cumbersome and lengthy procedures.

In recent years, the widespread adoption of digital pathology[10] and advancements in deep learning have greatly enhanced the performance of computational pathology across various diagnostic tasks. Notably, virtual staining methods are beginning to replace traditional staining processes in certain scenarios.[11] In essence, virtual staining is a deep learning-based method for image generation or transformation. It uses an image of tissue stained with one type of dye as input, and the model produces an output image as though the tissue were stained with a different dye. For example, a model designed for virtual staining from H&E to IHC requires a substantial dataset of paired images for training. Once trained, the model can take an H&E-stained image as input and generate the corresponding IHC-stained image. Virtual staining offers several advantages over traditional staining, such as increased efficiency, greater convenience, reduced risk of contamination, cost-effectiveness, and lower variability. However, current virtual staining models are not yet positioned to replace traditional staining entirely.[12] While promising results have been achieved in some scenarios, these models still face considerable limitations and require further development before they can be reliably applied in clinical practice.

Extensive research has been conducted on virtual staining, with several papers summarizing the findings.[13–15] However, these reviews typically focus on the clinical perspective, emphasizing the applications of virtual staining in various clinical tasks. In contrast, this paper reviews existing research from the perspective of technical theory and algorithms and emphasizes the implementation details, including existing virtual staining models, datasets, evaluation methods, the unique challenges of virtual staining, the limitations of existing models, and the future perspectives for developing novel models. Specifically, we introduce the workflow of virtual staining and several typical virtual staining methods. We also summarize the publicly available datasets and evaluation methods. Next, we analyze how virtual staining differs from traditional image translation. Finally, we summarize the limitations of current research and highlight potential future perspectives in this field. We hope this review will provide researchers in related fields with a comprehensive understanding of virtual staining, particularly for those focused on algorithm development. We aim to guide the development of efficient, accurate, and versatile virtual staining models. Furthermore, virtual staining is an interdisciplinary field that combines deep learning with medical knowledge. The development of models requires the active involvement of clinicians, pathology experts, and biological researchers. We aim to present the technical principles and details of virtual staining in an accessible manner to enable these professionals to engage in model design and contribute to the development of improved models based on interdisciplinary knowledge.

The rest of this paper is organized as follows. Section 2 introduces the workflow of virtual staining, including data processing, image registration, model training, and testing. Section 3 reviews classic virtual staining methods. Section 4 analyzes the unique challenges of virtual staining. Section 5 and Section 6 summarize the limitations of current work and explore future perspectives. Finally, Section 7 draws the conclusion.

## 2. The workflow of virtual staining

The workflow for virtual staining typically involves several key steps, including data processing, image registration, model training, and testing.

### 2.1. Data processing

Artifacts such as air bubbles, dust, or uneven staining can interfere with analysis and compromise the quality of pathology images. These artifacts are typically identified through visual inspection or automated algorithms. Manual or algorithmic techniques are then employed to correct them, which may involve filling in gaps, removing noise, or adjusting brightness and contrast in affected areas to enhance overall image quality. Extensive research has concentrated on identifying and eliminating these artifacts in pathology images,[16–18] resulting in notable improvements in image quality. Pathological tissues are often scanned into whole-slide images (WSIs), which have extremely high resolutions. At magnification $\times 40$, the width or height of a WSI can exceed 100,000 pixels, and few devices can directly process images of such size. Therefore, similar to most processing workflows for pathology analysis tasks, it is essential to first eliminate the background region from WSIs and then crop the remaining valid regions into fixed-size patches (e.g., $1024 \times 1024$ pixels). This process can be referred to as clustering-constrained-attention multiple-instance learning.[19]

After that, color normalization techniques are utilized to mitigate differences caused by various imaging equipment or settings, helping to standardize the appearance of stained images. Common color normalization methods can generally be categorized into two types: those based on image color distribution[20,21] and those that utilize deep learning models.[22–24]

### 2.2. Image registration

In supervised virtual staining methods, image registration plays a crucial role. It ensures that key anatomical structures in the unstained images are correctly positioned relative to their stained counterparts, which is essential for the model to learn the accurate

mapping of features. Image registration methods can generally be categorized into two types: rigid registration and nonrigid registration. Rigid registration involves applying transformations such as rotation and scaling. It is used when the images being aligned are assumed to be of the same shape and size but may be positioned differently. Nonrigid registration accommodates changes in shape, allowing for more complex transformations that account for biological variability or distortion in the tissue. It employs techniques such as spline-based warping or elastic deformations. Both types of methods have been applied in medical imaging.[25–27]

### 2.3. Model training and testing

Take pix2pix[28] as an example; during each training iteration, an unstained image is fed into the generator, which produces a corresponding stained image. The discriminator then evaluates both the generated image and the real stained image, assessing how realistic each one appears. In this process, the generator loss typically consists of two components: a pixel-wise loss (often L1 or L2 loss), which measures the difference between the generated image and the real image, and a generative adversarial network (GAN) loss, which encourages the generator to produce images that the discriminator cannot distinguish from real images. The discriminator loss measures the ability of the discriminator to correctly classify real and generated images. Training continues over multiple epochs, with the generator and discriminator improving iteratively. The generator aims to create more realistic images, while the discriminator becomes better at identifying fakes.

Once the model is trained, the generator can produce virtual stained images from unseen unstained images. Then, the generated images are compared to real images using qualitative and quantitative metrics. Meanwhile, expert pathologists may also evaluate the quality and diagnostic relevance of the generated images.

## 3. Overview of current work

### 3.1. Typical virtual staining methods

#### 3.1.1. Base models

Most virtual staining models are built upon traditional image translation frameworks, which are modified and adapted to meet specific clinical staining requirements. These image translation models serve as the base model or foundational architecture for learning the complex mapping between different staining domains (e.g., H&E to IHC). Widely utilized models include GAN,[29] CycleGAN,[30] CUT,[31] pix2pix,[28] pix2pixHD,[32] and diffusion models.[33]

#### 3.1.2. Supervised versus unsupervised methods

According to whether paired images are utilized, these methods can be categorized into two primary types: unsupervised methods and supervised methods (Table 1). Supervised methods rely on directly paired data, enabling superior feature preservation by ensuring a precise mapping between input and output domains. However, these methods necessitate complex image registration pipelines, which can greatly increase computational complexity and introduce potential registration errors. In contrast, unsupervised approaches eliminate the need for image registration by learning mappings from unpaired data. This flexibility makes them more practical in scenarios in which paired data are scarce or difficult to obtain. Nonetheless, unsupervised approaches typically require longer training times and may achieve lower levels of feature preservation compared to supervised methods.

**Table 1**
Typical virtual staining methods.

| Method | Architecture | Type | Task | Quantitative metric | EE | DT |
|---|---|---|---|---|---|---|
| ASP[34] | Not specified | Unsupervised | H&E to IHC | SSIM, PHV[35], FID, KID | × | × |
| PC-stain GAN[35] | CycleGAN | Unsupervised | H&E to IHC | PSNR, SSIM, MS-SSIM, PCC | × | × |
| BFF-GAN[36] | CycleGAN | Unsupervised | H&E to IHC | PSNR, SSIM, CSS, FID | √ | √ |
| VirtualMultiplexer[37] | CUT | Unsupervised | H&E to IHC | FID, CSS | √ | √ |
| AI-FFPE[38] | GAN | Unsupervised | H&E to H&E | FID | √ | √ |
| Bao et al.[39] | CycleGAN | Unsupervised | MxIF to H&E | Euclidian distance | × | √ |
| confusion-GAN[40] | CycleGAN | Unsupervised | H&E to IHC | FID, LPIPS, SSIM | × | √ |
| MIM[41] | Swin-transformer + MAE | Unsupervised | FFPE to H&E | PSNR, SSIM | × | × |
| Pyramid Pix2pix[42] | pix2pix | Supervised | H&E to IHC | PSNR, SSIM | × | √ |
| Rana et al.[43] | pix2pix | Supervised | FFPE to H&E | SSIM, PCC | × | √ |
| Rana et al.[44] | pix2pix | Supervised | FFPE to H&E | PSNR, SSIM, PCC | × | √ |
| Hu et al.[45] | pix2pixHD | Supervised | FFPE to H&E | PSNR, SSIM, MS-SSIM, PCC | × | × |
| Peng et al.[12] | pix2pix | Supervised | H&E to IHC | SSIM, FID, DISTS | × | √ |
| VIMs[46] | LDM | Supervised | H&E to IHC | SSIM, MSE, FID | √ | √ |
| PST-Diff[47] | DDIM | Supervised | H&E to IHC | PSNR, SSIM, LC | × | √ |
| StainDiffuser[48] | FFT-based U-Net | Supervised | H&E to IHC | PSNR, SSIM, FID | × | × |

Abbreviations: 1. EE: expert evaluation 2. DT: downstream task evaluation 3. H&E: hematoxylin and eosin 4. HC: immunohistochemistry 5. MxIF: multiplexed immunofluorescence 6. FFPE: formalin fixed paraffin embedded 7. SSIM: structural similarity index 8. PHV: perceptual hash value 9. FID: Fréchet inception distance 10. KID: kernel inception distance 11. PSNR: peak signal-to-noise ratio 12. MS-SSIM: multiscale-structural similarity index 13. PCC: Pearson correlation coefficient 14. CSS: contrast structural similarity 15. LPIPS: learned perceptual image patch similarity 16. DISTS: deep image structure and texture similarity 17. MSE: mean squared error 18. LC: luminance & contrast similarity.

#### 3.1.3. Summary

Supervised methods rely on paired datasets, where corresponding images from different staining (e.g., H&E and IHC) are available. The primary advantage of supervised methods lies in their precision and reliability, especially when high-quality paired data are provided. However, their reliance on large-scale, well-aligned paired datasets makes data preparation labor-intensive and costly, limiting their scalability and applicability in scenarios wherein paired data are scarce or unavailable. Unsupervised methods do not require paired datasets and instead utilize unpaired datasets to learn mappings between staining domains. The key advantage of unsupervised methods is their flexibility and reduced dependency on paired data, which simplifies dataset collection and extends their applicability to a broader range of tasks. However, these methods often have lower accuracy, and they may produce artifacts or fail to capture subtle details critical to clinical applications.

In summary, supervised and unsupervised methods each have their advantages and limitations. Supervised methods excel in accuracy and detail, while unsupervised methods offer greater flexibility at the cost of precision. When paired datasets are available, supervised methods should be prioritized to achieve more reliable and clinically applicable results.

### 3.2. Virtual staining datasets

#### 3.2.1. Paired versus unpaired datasets

Virtual staining datasets can be broadly categorized into paired and unpaired datasets, each with distinct characteristics and acquisition processes.

**Paired datasets:** These datasets consist of corresponding image pairs, where the same tissue sample is processed using two different staining protocols, such as H&E and IHC, and imaged under precisely aligned conditions. The acquisition process for paired datasets is highly resource-intensive and involves the following steps: (1) cutting adjacent or sequential tissue sections to ensure comparable morphology, (2) applying different staining procedures to each section, (3) imaging the tissue sections with strict alignment, and (4) postprocessing to match image pairs at the pixel level. Such meticulous preparation enables paired datasets to be ideal for supervised learning, as they provide direct mappings between staining domains. However, the challenges of tissue section alignment, staining protocol standardization, and imaging reproducibility significantly increase the difficulty and cost of constructing paired datasets.

**Unpaired datasets:** These datasets include images from different staining protocols without requiring direct correspondence between them. The acquisition process is more straightforward, as tissue samples can be independently stained and imaged without concern for alignment. While this reduces the time and effort needed for data collection, the lack of explicit pixel-level alignment complicates model training, often resulting in suboptimal performance.

#### 3.2.2. Public paired datasets

In this subsection, we provide a summary of publicly available datasets for virtual staining. It is noteworthy that acquiring unpaired datasets is relatively straightforward. For instance, to construct a dataset for virtual staining from H&E to IHC, one can simply select H&E-stained and IHC-stained slides from the same organ. Consequently, our focus is exclusively on summarizing publicly available paired datasets. More details can be found in Table 2.

#### 3.2.3. Challenges in existing paired datasets

Despite their contributions, current datasets still face several challenges. First, variability in staining protocols, including differences in reagents, imaging systems, and institutional practices, introduces inconsistencies that hinder model generalizability across datasets or institutions. Second, many patch-level paired datasets lack rigorous alignment, leading to noise during training and potentially degrading model performance. Third, there is an insufficient sample size and limited diversity in current datasets, which are predominantly focused on H&E to IHC tasks for a narrow range of tissue types and biomarkers. These limitations highlight the need for more diverse, standardized, and well-aligned datasets to strengthen the robustness and scalability of virtual staining models.

**Table 2**
Paired datasets for virtual staining.

| Dataset | Organ | Task | Antibody | WSI No. | Patch No. | Patch size |
|---|---|---|---|---|---|---|
| BCI[42] | Breast | H&E to IHC | HER2 | 51 pairs | 4870 pairs | 1024 × 1024 |
| MIST[34] | Breast | H&E to IHC | HER2 | 64 pairs | 5642 pairs | 1024 × 1024 |
| | | | ER | 56 pairs | 5153 pairs | |
| | | | Ki67 | 56 pairs | 5161 pairs | |
| | | | PR | 56 pairs | 5139 pairs | |
| IHC4BC[49] | Breast | H&E to IHC | HER2 | 52 pairs | 17,995 pairs | 1000 × 1000 |
| | | | ER | 59 pairs | 27,395 pairs | |
| | | | Ki67 | 60 pairs | 18,745 pairs | |
| | | | PR | 60 pairs | 21,971 pairs | |
| RegH2I[12] | Breast | H&E to IHC | 4B5 | Not specified | 2592 pairs | 1024 × 1024 |
| HCI[40] | Liver | H&E to IHC | GPC3 | 100 pairs | Approximately 1.4 million pairs | 256 × 256 |

**Table 3**
Image quality evaluation metrics.

| Evaluation metric | Expected value | Pixel alignment | Feature type | With medical knowledge |
|---|---|---|---|---|
| PSNR | ↑ | √ | Original images | × |
| PPC | ↑ | √ | Original images | × |
| SSIM | ↑ | √ | Original images | × |
| MS-SSIM | ↑ | √ | Original images | × |
| FID | ↓ | × | Deep features | × |
| DISTS | ↓ | × | Deep features | × |

### 3.3. Conventional image quality metrics

In this subsection, we provide an overview of the commonly used evaluation metrics for virtual staining, covering their calculation methods, ideal values, alignment requirements, feature types, and incorporation of medical knowledge. Sections 3.3.1–3.3.6 offer detailed explanations for each metric, while Table 3 presents a concise summary.

#### 3.3.1. Peak signal-to-noise ratio (PSNR)

PSNR is a widely used metric for assessing the quality of reconstructed images, particularly in the context of lossy compression.[50] PSNR is calculated using the following formula:

$$\mathrm{PSNR} = 10\cdot\log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right)$$

where $\mathrm{MAX}_I$ is the maximum possible pixel value of the image (typically 255 for 8-bit images), and mean squared error is defined by:

$$\mathrm{MSE} = \frac{1}{m^2}\sum_{i=0}^{m-1}\sum_{j=0}^{m-1}[I(i,j) - K(i,j)]^2$$

here, $I(i, j)$ and $K(i, j)$ denote the pixel values of the original and generated images, respectively, and both the height and width of the image are $m$.

#### 3.3.2. Pearson correlation coefficient (PCC)

The PCC is commonly used to measure the similarity between two images by comparing their pixel intensities.[51] It is useful for evaluating the quality of generated images against real images. The formula for PCC is as follows:

$$\mathrm{PCC} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where $X_i$ and $Y_i$ represent the pixel values of the real image and the generated image, respectively, $\bar{X}$ and $\bar{Y}$ are the mean values of $X$ and $Y$, and $n$ is the total number of pixels in the image.

#### 3.3.3. Structural similarity index (SSIM)

SSIM is also used to measure the similarity between two images. In existing research,[52] the luminance, contrast, and structure comparison measures are defined as follows:

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}$$

where $C_1$,$C_2$, and $C_3$ are small constants. The signals $\mathbf{x} = \{x_i | i = 1, 2, \cdots, N\}$ and $\mathbf{y} = \{y_i | i = 1, 2, \cdots, N\}$ are two discrete nonnegative signals that have been aligned with each other. Additionally, $\mu_x$ denotes the mean of $\mathbf{x}$, $\sigma_x^2$ represents the variance of $\mathbf{x}$, and $\sigma_{xy}$ indicates the covariance between $\mathbf{x}$ and $\mathbf{y}$. The general form of SSIM is defined as follows:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha}\cdot[c(\mathbf{x},\mathbf{y})]^{\beta}\cdot[s(\mathbf{x},\mathbf{y})]^{\gamma}$$

where $\alpha$, $\beta$, and $\gamma$ are parameters to define the relative importance of the three components. Specifically, when we set $\alpha = \beta = \gamma = 1$, the resulting SSIM is formulated as

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

#### 3.3.4. Multiscale SSIM (MS-SSIM)

In a previous study,[53] a MS-SSIM method for image quality assessment was proposed. In this approach, the original image is designated as Scale 1, while the highest scale is referred to as Scale $M$, which is obtained after $M - 1$ iterations. The overall MS-SSIM

evaluation is obtained by combining the measurements at different scales using

$$\text{MS-SSIM}(\mathbf{x}, \mathbf{y}) = [l_M(\mathbf{x}, \mathbf{y})]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(\mathbf{x}, \mathbf{y})]^{\beta_j} [s_j(\mathbf{x}, \mathbf{y})]^{\gamma_j}$$

Similar to SSIM, the exponents $\alpha_M$, $\beta_j$, and $\gamma_j$ are used to adjust the relative importance of different components.

#### 3.3.5. Fréchet inception distance (FID)

FID is a commonly used metric for evaluating the quality of images from generative models, such asGANs.[54] It measures the similarity between the distribution of real images and generated images by comparing the statistics (mean and covariance) of their feature representations extracted from a pretrained Inception-v3 model. The FID score is calculated as follows:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2})$$

where $\mu_r$ and $\mu_g$ are the mean feature vectors of the real and generated images, respectively, $\Sigma_r$ and $\Sigma_g$ are the covariance matrices of the real and generated images, respectively, and Tr denotes the trace of the matrix.

#### 3.3.6. Deep image structure and texture similarity (DISTS)

The DISTS evaluates image quality by measuring the similarity between reference and generated images based on their structure and texture features.[55] The formula for DISTS is as follows:

$$\text{DISTS}(x, y) = \alpha \cdot \sum_l (1 - S(f^l(x), f^l(y))) + (1 - \alpha) \cdot \sum_l T(f^l(x), f^l(y))$$

where $f^l(x)$ and $f^l(y)$ represent the feature maps extracted from layer $l$ of a pretrained deep neural network (usually a visual geometry group model). $S(f^l(x), f^l(y))$ measures the "structural similarity" between the features of the two images at each layer. $T(f^l(x), f^l(y))$ captures the "texture similarity" between the features. $\alpha$ is a weighting factor that balances the contributions of the structure and texture terms and is typically set between zero and one.

#### 3.3.7. Summary

While conventional metrics such as PSNR, SSIM, and FID are widely used for assessing image quality, their relevance to virtual staining remains limited. These metrics primarily concentrate on pixel-level accuracy or global visual similarity, which do not necessarily align with the diagnostic requirements in pathology. These metrics can be considered necessary but not sufficient conditions for evaluating the quality of virtual staining. High scores on these metrics typically indicate that the virtual staining results closely resemble traditional staining at the pixel level or in terms of structural similarity. However, the sufficiency is limited because they fail to capture the diagnostic relevance of the generated images. Even if a virtual staining model achieves high PSNR and SSIM, this does not guarantee diagnostically critical features. Moreover, these metrics can be influenced by nondiagnostic factors, such as uniform background consistency or noncritical regions of the image, which may artificially inflate scores without contributing to the model's clinical utility.

### 3.4. Pathology-oriented evaluation approach

#### 3.4.1. Content consistency evaluation

Content consistency evaluation refers to assessing the extent to which the various entities in the generated images, such as cells, glands, tissues, and task-specific staining (e.g., antibody-marked regions), match the morphology, quantity, and distribution seen in the real images. In the context of pathological images, this evaluation ensures that the generated content mirrors the key features and diagnostic information present in the original. If the generated image maintains a high degree of consistency with the real image in these aspects, it can be considered to have conveyed the same diagnostic information, thus achieving an ideal virtual staining outcome.

There have been related studies that have adopted this approach to evaluate generated images, such as the research by Peng et al.[12] They aimed to achieve virtual staining from H&E to IHC for HER2 scoring. The key features of IHC images are the distribution of cells and the intensity of cell membrane staining. Therefore, they evaluated the distribution of nuclei and the cell membrane staining regions in the generated images. Similarly, Martell et al.[56] analyzed the consistency of cellular attributes between generated and real images, such as internuclear distance, cross-sectional nuclear area, eccentricity, compactness, and cell count. The evaluation method, which was based on content consistency, can flexibly design metrics and calculation methods according to the pathological entities focused on specific diagnostic tasks to assess whether the generated images can accurately convey diagnostic information.

#### 3.4.2. Expert evaluation

Expert evaluation is widely used in pathology virtual staining, relying on the subjective judgment of experienced pathologists. In this approach, experts compare generated virtual stained images with real histological images to assess quality, diagnostic accuracy, and the ability to reflect pathological features.

In a previous study,[36] pathologists were presented with 100 real and 100 generated images and tasked with distinguishing the real ones. These findings demonstrate that virtual staining models can, to some extent, deceive even experienced pathologists. In another

study, professional pathologists were tasked with visually inspecting generated images to assess their quality and accuracy. This process involved evaluating the overall image quality, identifying false positive and false negative cells, and ranking the models based on their accuracy.[46]

#### 3.4.3. Downstream task-based evaluation

Downstream task-based evaluation is also a critical approach to evaluating the practical effectiveness of virtual staining models. Unlike conventional image quality assessments, this method focuses on assessing how well the generated images perform in actual clinical or diagnostic tasks.

Previous research has utilized this type of evaluation method. In the research by Peng et al., generated IHC images were used for HER2 scoring-related classification tasks, achieving an accuracy nearly identical to that of real images.[12] Similarly, Dubey et al. employed a U-Net model trained on gland segmentation tasks to evaluate both paired and unpaired test sets.[46] Additionally, He et al. used ImageJ for stain estimation on both the generated and real images in the test set.[47] The generated and real images can be presented to doctors for diagnosis to evaluate their diagnostic accuracy when using the generated images.

#### 3.4.4. Summary

Pathology-oriented evaluation methods provide more diagnostic relevance compared to conventional metrics. However, due to certain limitations, these methods have not yet been widely adopted. Content consistency evaluation requires detailed manual or semi-automated annotations of regions of interest, which can be time-consuming and resource-intensive. Additionally, weighting the importance of different regions introduces complexity in metric design. Expert evaluation is inherently subjective and time-consuming, making it challenging to scale for large datasets. Inter-expert variability can also affect the reproducibility of results. Downstream task-based evaluation requires additional labeled datasets for task-specific training and validation, increasing data preparation complexity. In summary, these methods lack generalizability, as each method requires task-specific designs tailored to particular diagnostic scenarios, limiting their broader applicability. Additionally, their implementation often involves considerable additional effort. These requirements render them resource-intensive and less scalable for large-scale studies.

## 4. Challenges of virtual staining compared to image translation

Virtual staining may appear to be a straightforward adaptation of image translation methods from the domain of natural images to pathological images.[57] In reality, there are considerable differences between them. The characteristics of pathological images make virtual staining inherently more challenging than common image translation tasks. In this paper, we focus on three challenges: the gigapixel resolution in WSIs, the severe homogenization of image features, and the invisibility of certain features.

### 4.1. The gigapixel of WSI

Pathological tissues are often scanned into WSIs, which have extremely high resolutions.[10] At magnification $\times 40$, the width or height of an image can exceed 100,000 pixels. Additionally, virtual staining is a complex task that depends on large-scale image generation models with numerous parameters. Training these models requires substantial computational resources, rendering it nearly impossible for existing devices to directly process images of this size. Thus, virtual staining of WSI generally entails cropping the WSI into fixed-size patches. Model training and testing are performed at the patch level, and the resulting stained patches are then stitched together to create the final staining outcome at the slide level. However, this approach introduces three major challenges.

First, inconsistencies in staining style may occur because each sample is stained independently. For example, some patches might exhibit deeper staining, while others appear lighter. When these inconsistent stained patches are combined to create a slide-level image, a pronounced block effect becomes evident, leading to unnatural transitions between patches. Figure 1 illustrates an example of this issue. The block effect not only affects the diagnostic accuracy of pathologists but also suggests that the model may be exploiting spurious correlations unrelated to the diagnosis during the virtual staining process, such as differences in background staining intensity.[58]

Additionally, cropping a WSI into patches neglects the spatial relationships between them, leading to a loss of the contextual information that is intrinsic to the WSI.[59] This, in turn, diminishes the diagnostic information that can be gathered from the entire slide. If a virtual staining model is developed solely based on patch-level images without considering the contextual information of WSI, the resulting model may be suboptimal.

Finally, strict registration between the source and target images in the training dataset is essential for supervised virtual staining methods. While numerous image registration techniques can achieve satisfactory results at the slide level, attaining perfect registration remains nearly impossible. The considerable size differences between patch-level and slide-level images mean that even minor registration errors at the slide level can be greatly magnified at the patch level. Consequently, achieving high-quality registration at the patch level presents a considerable challenge.

The virtual staining task for ultra-high-resolution WSI requires careful consideration of block effects, contextual information, and patch-level registration, making it inherently more complex than traditional image translation tasks.

### 4.2. Homogenization of image features

Most existing virtual staining methods are constructed on image generation models. These models were initially developed for natural image translation. However, there are considerable differences between pathological images and natural images. In natural

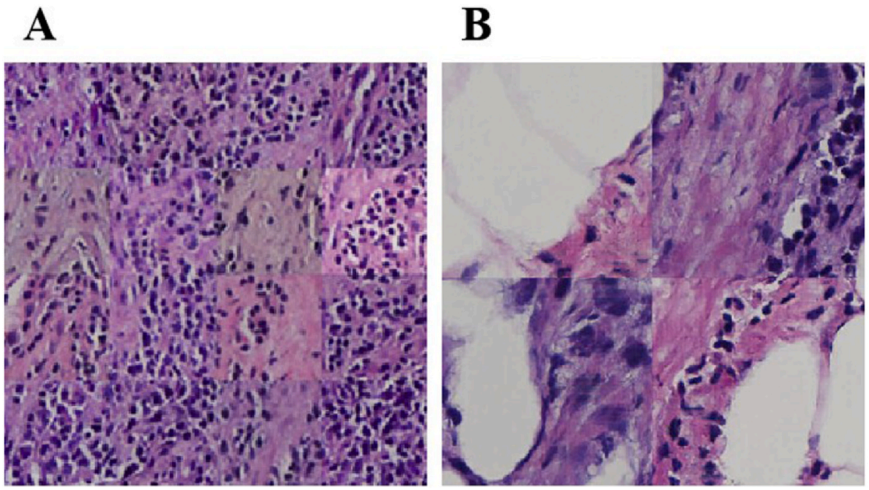



**Figure 1.** Illustration of block effects in virtual staining.
(A) shows a virtual stained image composed of 4 × 4 patches; (B) shows a virtual stained image composed of 2 × 2 patches.

images, distinct objects are usually visible, with clear differences in features and boundaries between different entities. However, pathological images display considerable homogeneity in their features. The entities present in these images consist of various cells and tissues, exhibiting minimal visual differences among different cell types. Additionally, the boundaries between different tissues are often difficult to discern. Figure 2 shows an example of the differences between image translation and virtual staining. Furthermore, even if two pathological images seem highly similar, subtle differences between them can convey entirely distinct diagnostic information.

This homogeneity complicates these models' abilities to establish a mapping function between source and target images, particularly when training unsupervised virtual staining models with unpaired images. In contrast to natural images, where entities are

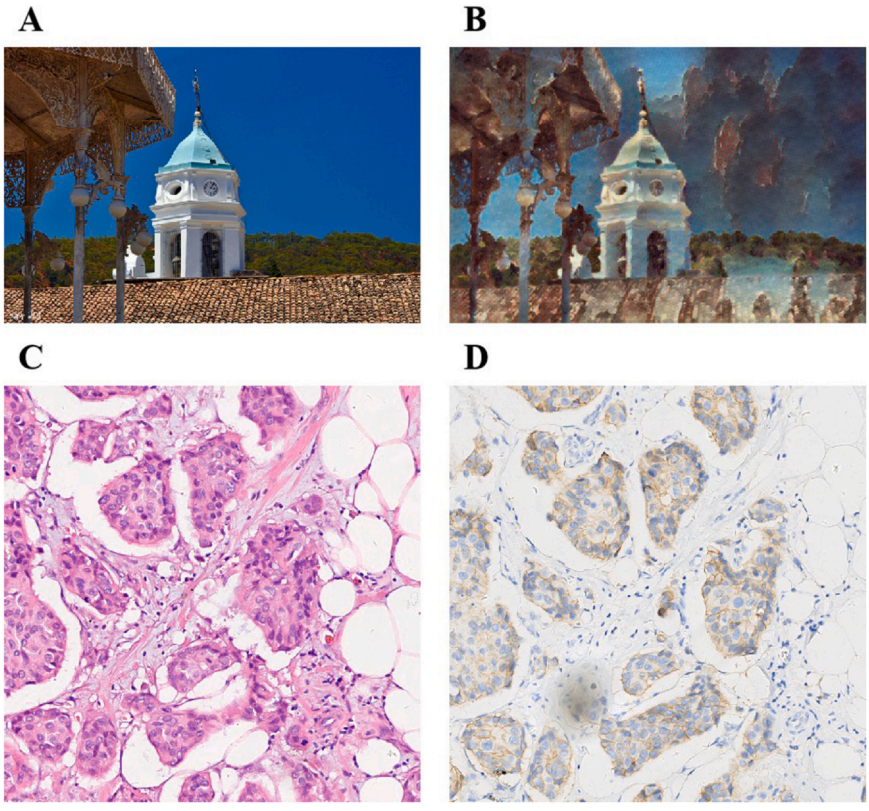



**Figure 2.** Differences between image translation and virtual staining.

discernible, the model can easily align regions representing the same entity across both images. When working with unpaired natural images, even though it is impossible to identify corresponding entities based solely on spatial coordinates, the distinctly defined entities are easily differentiated from the background regions of the images. Consequently, the model can more effectively create an entity dictionary between images of different styles and learn the mapping function from this dictionary.

Such feature homogeneity greatly restricts the application of unsupervised image translation methods in virtual staining. For supervised image translation methods, this homogeneity requires improved registration between source and target images in the training data. Consequently, feature homogeneity presents a substantial challenge that should be addressed when transitioning from traditional image translation to virtual staining.

### 4.3. Invisibility of image features

The invisibility of certain features in pathological images is another key factor that makes virtual staining more challenging than traditional image translation. As noted in the previous section, entities in natural images possess clear contours, making their features visually identifiable. As shown in Figure 2A, elements such as the blue sky and buildings are distinctly visible. In traditional image translation tasks,[30] the emphasis is primarily on applying style transformations to these visible entities, ensuring that the generated images (Figure 2B) contain the same entities represented in the target style. However, in addition to directly applying style transformations to the entities within the images, pathological virtual staining aims to highlight elements of diagnostic value in the generated images, i.e., the invisible features, according to clinical diagnostic guidelines.

We illustrate this concept using an example of virtual staining from H&E to IHC for HER2 scoring in breast cancer, explaining what these invisible features are and why they hold significance. H&E staining shows tissues in shades of blue and pink. Hematoxylin stains nuclei blue, while eosin stains the cytoplasm and extracellular matrix pink. IHC employs specific antibodies to mark particular proteins within tissues. Generally, the background is white, while the cell nuclei are stained light blue. Regions containing antigens exhibit a higher contrast color, such as brown or red. In IHC staining images, the cancer cell membranes in breast cancer slides display varying staining statuses (brown), such as weak incomplete staining, moderate complete staining, and strong complete staining. The HER2 scoring guidelines for breast cancer classify patients into four risk levels based on the number of cells exhibiting different staining intensities. However, even experienced pathologists cannot directly distinguish between cells with different staining statuses in H&E-stained images. Thus, virtual staining involves two tasks: (1) performing style transformations, which entails transforming the pink and blue elements of H&E staining into the blue and white tones of IHC staining (blue regions in Figure 2D), and (2) highlighting the intensity of the brown staining on the cell membranes (brown regions in Figure 2D). If a virtual staining model achieves perfect style transformation (Task 1) but fails to indicate the brown staining intensity (Task 2), it would still be considered unsuccessful. The brown intensity of the cell membrane is invisible in H&E images, yet it contains the most crucial diagnostic information. In general, the successful identification and highlighting of these diagnostic elements is more important and challenging than mere style transformation.

## 5. Limitations in current practices

### 5.1. Limited generalizability

Generalizability refers to a model's ability to perform effectively on unseen data that is not included in its training set.[60] A model is considered to generalize well if it can accurately capture the underlying patterns in the training data and successfully apply that knowledge to make predictions on new or previously unseen samples. The ultimate goal of any pathological analysis model is to assist clinicians in decision-making. Once the model has been trained, it is used for unknown samples, and its generalizability determines whether it can be successfully deployed in real clinical settings.

However, we conducted some preliminary experiments and found that some existing virtual staining models lack strong generalization capabilities. While these models perform impressively when trained and validated on datasets with the same distribution, i.e., successfully completing style transformations and highlighting elements of diagnostic value, they show a marked performance decline when tested on samples from a different dataset. In such cases, they fail to identify diagnostic elements and even fail to achieve satisfactory style transformations. Figure 3 illustrates an example of poor generalization performance in virtual staining from H&E to IHC.

### 5.2. Limited robustness

Robustness refers to a model's ability to maintain its performance when faced with varying conditions, including noise, outliers, or changes in the input data distribution. A robust model is not easily affected by small perturbations or unexpected variations in the data, allowing it to produce reliable results even in less-than-ideal circumstances. It is important to note that generalizability, which we discussed in the previous section, is a component of robustness. Here, we focus on the model's resilience to perturbations.

In this subsection, we also conducted some preliminary experiments to assess the stability of models when faced with minor perturbations. Specifically, we selected an H&E image and created a perturbed image by adjusting its hue and saturation within a defined range. We then input these images into a well-trained virtual staining model to obtain the corresponding IHC-stained image for each. Because these H&E images differ only in hue and saturation, the diagnostic information, such as the distribution of cells and tissues, should remain entirely consistent. Ideally, the virtual staining model should produce similar staining results for these images. However, in practice, we observed notable differences in the generated images, including variations in style and the identification of

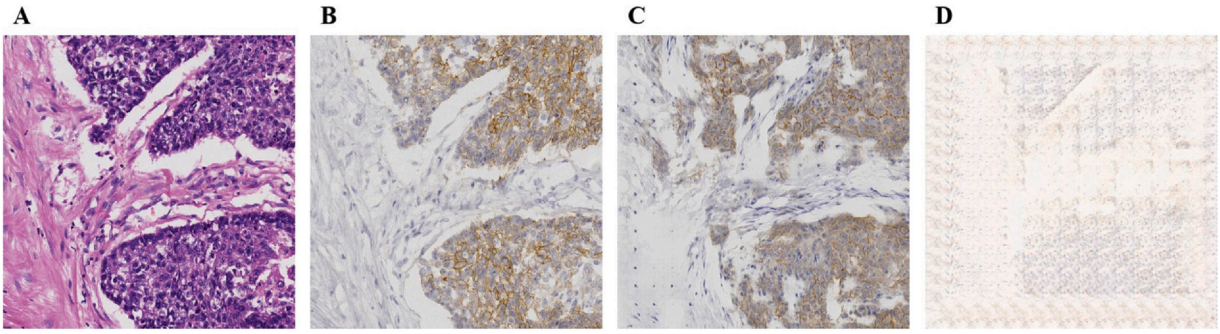



**Figure 3.** Visualization of H&E to IHC staining. A, H&E; B, real IHC; C, generated IHC; D, IHC generated by models trained on another dataset.

diagnostic elements. Figure 4 illustrates an example of this case. Such results not only conflict with the intuitive understanding of pathologists but also indicate that the model may acquire spurious knowledge related to color variability, thereby undermining the credibility of the virtual staining model.

### *5.3. Limited training data*

Pathological virtual staining is essentially a more complex image translation task. Training an image translation model typically requires a large number of image pairs, which helps capture the diversity of the images, including variations in style, color, texture, and content. The inclusion of more samples enhances the model's ability to learn the complex mapping function between the two domains. Adequate training data also help prevent model overfitting, resulting in improved generalizability and robustness.

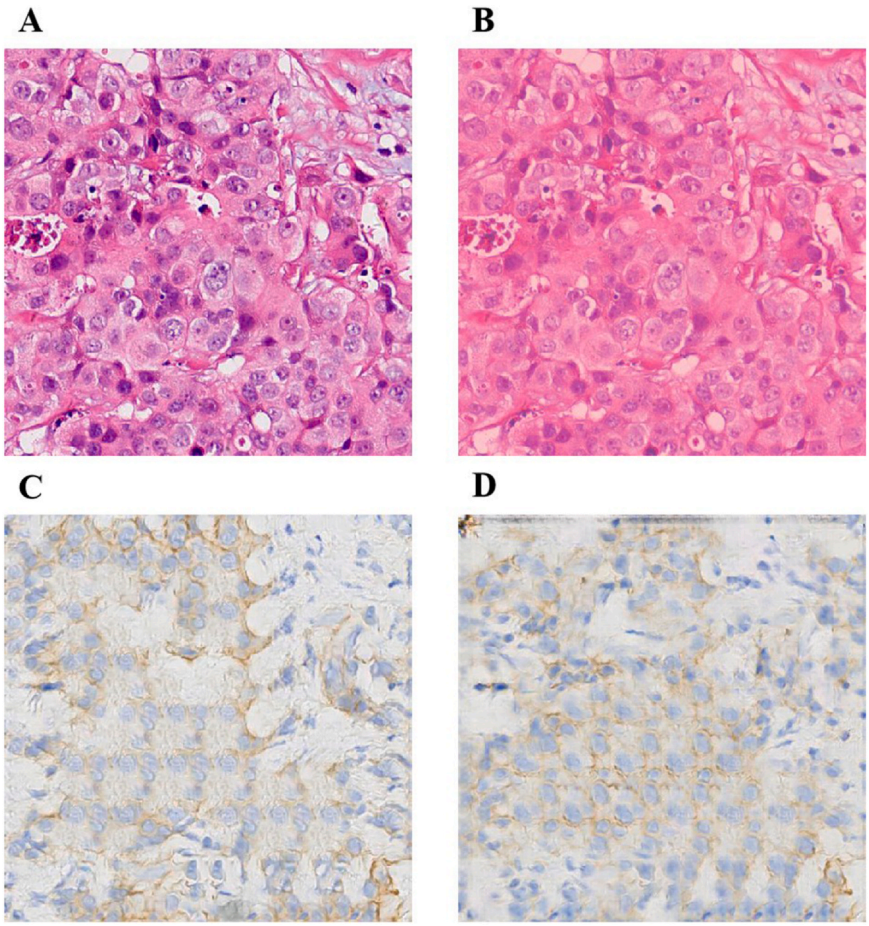



**Figure 4.** Visualization of inconsistent staining. A, H&E; B, a perturbed image of H&E; C, generated IHC from H&E; D, generated IHC from perturbed image of H&E.

However, most existing virtual staining models utilize a relatively insufficient number of training samples. As shown in Table 2, the majority of datasets contain only a few thousand pairs of source and target images. This is part of the reason why some existing models exhibit poor generalizability and robustness.

### 5.4. Limited theoretical models

As discussed in Section 4, there are considerable differences between pathological images and natural images, making pathological virtual staining more challenging than typical image translation tasks. However, most current virtual staining methods are based on classical image translation approaches, such as pix2pix. Even though some research has made improvements to traditional image translation models based on the characteristics of pathological images, resulting in better virtual staining performance, there is still considerable room for advancement. So far, aside from a few methods that adapt and improve upon transformer models,[41] GANs,[34–38,40] and diffusion models,[46–48] there have been few entirely novel model architectures specifically designed to address the challenges of virtual staining.

### 5.5. Limited evaluation methods

Similarly, most current evaluation metrics for virtual staining are derived from natural images. While these metrics can effectively assess the quality of the generated images, they are not entirely suitable for evaluating virtual staining performance.

Image translation emphasizes pixel-level accuracy, requiring that the positions of various entities in the source and target images be absolutely precise. In other words, the pixels between the two images must correspond one-to-one, which has led to the development of numerous pixel-level and local region-level evaluation metrics. However, pathological virtual staining prioritizes the staining effectiveness of regions with diagnostic value in the generated images. It emphasizes whether the resulting images accurately convey the diagnostic information rather than achieving perfect pixel alignment. Consider a scenario wherein the regions containing diagnostic information in a pathological image are quite small, and the virtual staining model achieves perfect pixel-level style transformations but fails to highlight the diagnostic information. From the perspective of pixel-level metrics, this would be considered an excellent virtual staining model. However, from a clinical diagnostic standpoint, it remains a failure.

Currently, most virtual staining evaluations still rely on metrics from natural images, which poses the risk of inaccurately assessing virtual staining performance.

## 6. Future perspectives

In response to the challenges and limitations, we provide our preliminary insights on addressing these issues in this section.

### 6.1. Improving representation learning

Pathological virtual staining is a complex task that encompasses style transformations between different images and the extraction of visually invisible diagnostic information. During this process, variations in staining intensity can introduce spurious correlations, potentially impairing the virtual model's ability to learn an intrinsic mapping function between the two types of images. The block effect also suggests that the model may learn spurious correlations related to variations in staining intensity. Such spurious mapping relationships typically exist only in specific datasets, but they reduce the model's generalization and robustness.

Therefore, we suggest promoting the feature representation learning of virtual staining models. For instance, introducing certain perturbations and learning strategies could guide the model to learn more robust feature representations. Alternatively, some research has incorporated causal inference into pathological image analysis, yielding promising results.[61,62] From a causal inference perspective, identifying causal relationships within datasets and addressing spurious correlations may enable models to acquire genuine medical knowledge.

### 6.2. Leveraging contextual information

Due to the gigapixel WSIs, the source WSI is typically cropped into patches for virtual staining and then stitched back together to reconstruct the target WSI. As described in Section 4.1, this process can result in a loss of contextual information during virtual staining, leading to suboptimal patch-level outputs and slide-level block effects. These issues can be addressed by incorporating contextual information from the WSI.

When contextual information is incorporated, each patch gains awareness of the surrounding regions. Interacting with neighboring areas helps capture diagnostic-relevant information from a broader field of view, enabling a more accurate mapping function. Moreover, the staining style of each generated patch becomes more consistent with its surroundings, which helps alleviate the block effect.

A simple method for utilizing WSI contextual information is multiresolution joint analysis. For instance, virtual staining is performed at two magnification levels: high-magnification patches capture image details, while low-magnification patches cover a larger field of view at the same location to provide contextual information. This approach has achieved outstanding performance in WSI segmentation and could also be considered for application in virtual staining.

### 6.3. Integrating medical knowledge

While virtual staining is a technical procedure through deep learning methods, it remains inherently connected to medical knowledge. Therefore, we recommend integrating medical knowledge into existing virtual staining models.

A simple and effective method is to predefine regions that are relevant to diagnosis and guide the model to focus more on these regions. This can be achieved by assigning greater weight to these regions in the loss function. Identifying diagnostic-relevant regions is essentially a straightforward application of integrating medical knowledge. One of the main purposes of H&E staining is to observe the morphology and distribution of cells, so regions associated with cells should be emphasized. By using pretrained cell segmentation models, the approximate distribution of cells in H&E images can be obtained. For example, in the FFPE-to-H&E virtual staining process, better performance was achieved by constraining the semantic consistency of various cell types between the generated and real images.[45] IHC staining detects and localizes specific protein expressions based on the principle of antibody binding to target antigens. Therefore, we can identify antibody-marked regions as they constitute the diagnostic-relevant regions. In most cases, antibody-marked regions typically appear brown or red, while other regions are white or blue. These antibody-marked regions can be determined using simple color analysis methods. For instance, Peng et al. focused on H&E to IHC virtual staining for HER2 scoring, with emphasis placed on increasing the staining accuracy of nuclear and membrane regions. Similarly, in Masson's trichrome staining, collagen fibers are usually stained blue or green. When performing collagen-related analyses, diagnostic-relevant regions can also be identified through similar methods. Integrating medical knowledge into virtual staining models can help achieve clinically relevant outcomes, especially for specific tasks.

### 6.4. Innovating theoretical methods

As outlined in Section 4, pathological virtual staining presents greater challenges than typical image translation tasks, requiring the achievement of additional tasks. Nevertheless, most current virtual staining methods continue to depend on classical image translation models. Undoubtedly, these models have produced encouraging results in virtual staining. However, we believe that model architectures and loss functions that are meticulously designed from scratch and based on the characteristics of pathology images and the objectives of virtual staining can achieve more accurate virtual staining than making limited improvements to traditional image translation models. Virtual staining does not aim for absolute pixel-level alignment; instead, it focuses more on preserving diagnostic information. The classic image translation model pix2pix aims to model pixel-level mappings between the source and target domains. However, in virtual staining, we can consider scaling this mapping to the cellular or tissue level and develop entirely new model architectures based on this concept. For instance, a "cell2cell" model could be developed to achieve cellular style transfer between two different stained images. Similarly, a "tissue2tissue" model could learn the mapping function between the source and target domains at the tissue level.

Additionally, we also recommend concentrating on the development of unsupervised virtual staining methods, which can address the pressing challenge of insufficient training data in this field. For example, pathological entities at the cellular or tissue level could be identified within unpaired source and target domains to obtain corresponding entity prototypes. These prototypes could then be matched across the source and target domains using specific strategies, enabling domain pairing in a particular semantic space.

Most existing image generation models use L1 loss as the optimization objective, which is focused on achieving pixel-level alignment. However, when the proportion of diagnostic-relevant regions is small, these loss functions may not guarantee accurate staining in those regions. Therefore, we recommend designing loss functions and evaluation metrics that are closely related to the characteristics of pathology images or the objectives of virtual staining. For example, as discussed in Section 6.3, we could incorporate medical knowledge to calculate generation errors at the cellular or tissue level. These innovations may allow the generated images to capture diagnostic information rather than simply achieve perfect pixel alignment.

### 6.5. Cross-modal learning

Cross-modal learning is a subfield of machine learning that focuses on understanding and utilizing data from different modalities, such as text, images, and audio. The primary goal is to enhance the learning process by leveraging the complementary information available in multiple modalities.

The invisibility of certain features in source images is a critical factor contributing to the inability of virtual staining to accurately highlight diagnostically relevant regions. Even experienced pathologists may find it challenging to find such information solely from source images. Therefore, leveraging supplementary data from other modalities (e.g., medical histories and genetic sequences) to capture features that are difficult to discern from source images holds promise for realizing better virtual staining.

Currently, several studies have incorporated cross-modal learning into virtual staining, demonstrating improved results.[46–48,63–65] We believe this represents a promising avenue for exploration, as obtaining additional information through cross-modal learning to achieve more precise virtual staining has substantial potential.

### 6.6. Developing foundational models

In computational pathology, numerous foundational models have emerged, pretrained on extensive datasets of pathological images, demonstrating exceptional performance across a variety of downstream tasks.[66,67] However, foundational models specifically designed for virtual staining are still scarce.

There are numerous types of virtual staining. Even within IHC staining, different antibodies can produce varying staining results. Consequently, clinical scenarios may involve hundreds or even thousands of virtual staining tasks. Customizing a model for each specific virtual staining task is not only time-consuming and labor-intensive but also often impractical.

Therefore, we recommend developing a foundational model for virtual staining. Taking the H&E -to IHC virtual staining as an example the paired dataset for a single antibody between H&E and IHC contains a maximum of only 210 WSI pairs (Table 2). However, we can decompose the H&E to IHC virtual staining into two tasks, cellular or tissue style transfer and staining of antibody-marked regions, corresponding to the image background and diagnostic-relevant regions. Even for different antibodies, the generation of background regions tends to be similar. Therefore, we can first ignore the antibody type and use all H&E to IHC paired datasets to pretrain a foundational model focused on achieving the first goal. Subsequently, we can fine-tune the model using the dataset for a specific antibody to achieve the second goal.

Additionally, obtaining unpaired datasets is relatively easy. Therefore, we can acquire a vast amount of unpaired images, covering various types of virtual staining tasks. Then, we can use existing or the novel unsupervised virtual staining models described in Section 6.4 to train a foundation model. This model could be easily fine-tuned or directly applied to address various staining tasks. By simply inputting source domain images and prompts of the target domain, the model can generate the desired stained images. Ideally, this would provide a versatile “all-in-one” solution for a wide range of virtual staining applications.

## 7. Conclusions

In this paper, we present the workflow of pathological virtual staining and review several representative technical models, publicly available datasets, and evaluation methods. We analyze the additional challenges that pathological virtual staining faces compared to traditional image translation. We also summarize some limitations of existing research and offer several future perspectives of virtual staining.

Unlike previous reviews that focused on the clinical applications of virtual staining, this paper presents an overview of virtual staining from the perspective of algorithmic models. However, we adopt a macro perspective on the algorithms without delving into overly complex technical details, making it more accessible for clinical experts. Additionally, we also provide a brief introduction to the objective of virtual staining from a medical viewpoint, which may inspire experts focusing on algorithms. This paper aims to promote a deeper understanding of interdisciplinary knowledge between algorithm developers and clinicians, facilitating a better understanding of each other’s needs. This collaboration seeks to create more efficient, generalized, and versatile virtual staining models.

## CRediT authorship contribution statement

**Weiping Lin:** Writing – review & editing, Writing – original draft, Visualization, Validation, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yihuang Hu:** Investigation, Formal analysis, Data curation, Conceptualization. **Runchen Zhu:** Writing – review & editing, Writing – original draft, Conceptualization. **Baoshun Wang:** Visualization, Methodology. **Liansheng Wang:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Formal analysis, Conceptualization.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62371409, Fujian Provincial Natural Science Foundation of China under Grant 2023J01005.

## References

1. Cai Z, Song H, Fingerhut A, et al. A greater lymph node yield is required during pathological examination in microsatellite instability-high gastric cancer. *BMC Cancer.* 2021;21(1):319.
2. Musumeci G. Past, present and future: overview on histology and histopathology. *J Histol Histopathol.* 2014;1(5):1–3.
3. Titford M. The long history of hematoxylin. *Biotech Histochem.* 2005;80(2):73–78.
4. Pollitt CC. Basement membrane pathology: a feature of acute equine laminitis. *Equine Vet J.* 1996;28(1):38–46.
5. Foot NC. The Masson trichrome staining methods in routine laboratory use. *Stain Technol.* 1933;8(3):101–110.
6. Perry P, Wolff S. New Giemsa method for the differential staining of sister chromatids. *Nature.* 1974;251(5471):156–158.
7. Ramos-Vara JA, Miller MA. When tissue antigens and antibodies get along: revisiting the technical aspects of immunohistochemistry–the red, brown, and blue technique. *Vet Pathol.* 2014;51(1):42–87.
8. Van Deun A, Hossain MA, Gumusboga M, Rieder HL. Ziehl-Neelsen staining: theory and practice. *Int J Tuberc Lung Dis.* 2008;12(1):108–110.
9. Harms PW, Frankel TL, Moutafi M, et al. Multiplex Immunohistochemistry and Immunofluorescence: A Practical Update for Pathologists. *Mod Pathol.* 2023;36(7):100197.
10. Madabhushi A, Lee G. Image analysis and machine learning in digital pathology: Challenges and opportunities. *Med Image Anal.* 2016;33:170–175.
11. Rivenson Y, de Haan K, Wallace WD, Ozcan A. Emerging Advances to Transform Histopathology Using Virtual Staining. *BME Front.* 2020;2020:9647163.

12. Peng Q, Lin W, Hu Y, et al. Advancing H&E-to-IHC Virtual Staining with Task-Specific Domain Knowledge for HER2 Scoring. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2024:3-13.
13. Bai B, Yang X, Li Y, Zhang Y, Pillar N, Ozcan A. Deep learning-enabled virtual histological staining of biological samples. *Light Sci Appl.* 2023;12(1):57.
14. Latonen L, Koivukoski S, Khan U, Ruusuvuori P. Virtual staining for histology by deep learning. *Trends Biotechnol.* 2024;42(9):1177–1191.
15. Salvi M, Michielli N, Salamone L, et al. Computational synthesis of histological stains: a step toward virtual enhanced digital pathology. *Int J Imaging Syst Technol.* 2024;34(5):e23165.
16. Shakhawat H, Hossain S, Kabir A, et al. Review of artifact detection methods for automated analysis and diagnosis in digital pathology. In: Mostefaoui GK, Islam SMR, Tariq F, eds. *Artificial Intelligence for Disease Diagnosis and Prognosis in Smart Healthcare*.1st ed.CRC Press; 2023:177-202.
17. Foucart A, Debeir O, Decaestecker C. Artifact identification in digital pathology from weak and noisy supervision with deep residual networks. In: *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*. IEEE; 2018:1-6.
18. Kanwal N, López-Pérez M, Kiraz U, Zuiverloon TCM, Molina R, Engan K. Are you sure it's an artifact? Artifact detection and uncertainty quantification in histological images. *Comput Med Imaging Graph.* 2024;112:102321.
19. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng.* 2021;5(6):555–570.
20. Macenko M, Niethammer M, Marron JS, et al. A method for normalizing histology slides for quantitative analysis. In: *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*. IEEE; 2009:1107-1110.
21. Vahadane A, Peng T, Sethi A, et al. Structure-Preserving Color Normalization and Sparse Stain Separation for Histological Images. *IEEE Trans Med Imaging.* 2016;35(8):1962–1971.
22. Jeong J, Kim KD, Nam Y, Cho CE, Go H, Kim N. Stain normalization using score-based diffusion model through stain separation and overlapped moving window patch strategies. *Comput Biol Med.* 2023;152:106335.
23. Jewsbury R, Wang R, Bhalerao A, Rajpoot N, Vu QD. StainFuser: Controlling Diffusion for Faster Neural Style Transfer in Multi-Gigapixel Histology Images. *arXiv preprint arXiv:240309302*. Published online 2024.
24. Shen Y, Ke J. Staindiff: Transfer stain styles of histology images with denoising diffusion probabilistic models and self-ensemble. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2023:549-559.
25. Ourselin S, Roche A, Prima S, Ayache N. Block matching: A general framework to improve robustness of rigid registration of medical images. In: *International Conference on Medical Image Computing And Computer-Assisted Intervention*. Springer; 2000:557-566.
26. Crum WR, Hartkens T, Hill DL. Non-rigid image registration: theory and practice. *Br J Radiol.* 2004:S140–S153 77 Spec No 2.
27. Gatenbee CD, Baker AM, Prabhakaran S, et al. Virtual alignment of pathology image series for multi-gigapixel whole slide images. *Nat Commun.* 2023;14(1):4502.
28. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*; 2017:1125-1134.
29. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. Proceedings of *Advances in Neural Information Processing Systems*; 2014;27:2672–80.
30. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: *Proceedings of the IEEE International Conference on Computer Vision.*; 2017:2223-2232.
31. Park T, Efros AA, Zhang R, Zhu JY. Contrastive learning for unpaired image-to-image translation. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*. Springer; 2020:319-345.
32. Wang TC, Liu MY, Zhu JY, Tao A, Kautz J, Catanzaro B. High-resolution image synthesis and semantic manipulation with conditional gans. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*; 2018:8798-8807.
33. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst.* 2020;33:6840–6851.
34. 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*. Springer; 2023:632-641.
35. Liu S, Zhang B, Liu Y, et al. Unpaired Stain Transfer Using Pathology-Consistent Constrained Generative Adversarial Networks. *IEEE Trans Med Imaging.* 2021;40(8):1977–1989.
36. Sun K, Chen Z, Wang G, Liu J, Ye X, Jiang YG. Bi-directional feature fusion generative adversarial network for ultra-high resolution pathological image virtual re-staining. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*; 2023:3904-3913.
37. Pati P, Karkampouna S, Bonollo F, et al. Accelerating histopathology workflows with generative AI-based virtually multiplexed tumour profiling. *Nat Mach Intell.* 2024;6(9):1077–1093.
38. Ozyoruk KB, Can S, Darbaz B, et al. A deep-learning model for transforming the style of tissue images from cryosectioned to formalin-fixed and paraffin-embedded. *Nat Biomed Eng.* 2022;6(12):1407–1419.
39. Bao S, Lee HH, Yang Q, et al. Alleviating tiling effect by random walk sliding window in high-resolution histological whole slide image synthesis. *Proc Mach Learn Res.* 2024;227:1406–1422.
40. Li J, Dong J, Huang S, et al. Virtual immunohistochemistry staining for histological images assisted by weakly-supervised learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*; 2024:11259-11268.
41. Ma J, Chen H. Efficient Supervised Pretraining of Swin-Transformer for Virtual Staining of Microscopy Images. *IEEE Trans Med Imaging.* 2024;43(4):1388–1399.
42. Liu S, Zhu C, Xu F, Jia X, Shi Z, Jin M. Bci: Breast cancer immunohistochemical image generation through pyramid pix2pix. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*; 2022:1815-1824.
43. Rana A, Yauney G, Lowe A, Shah P. Computational histological staining and destaining of prostate core biopsy RGB images with generative adversarial neural networks. In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE; 2018:828-834.
44. Rana A, Lowe A, Lithgow M, et al. Use of Deep Learning to Develop and Analyze Computational Hematoxylin and Eosin Staining of Prostate Core Biopsy Images for Tumor Diagnosis. *JAMA Netw Open.* 2020;3(5):e205111.
45. Hu Y, Peng Q, Du Z, et al. Boosting FFPE-to-HE Virtual Staining with Cell Semantics from Pretrained Segmentation Model. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer; 2024:67-76.
46. Dubey S, Chong Y, Knudsen B, Elhabian SY. *VIMS: virtual immunohistochemistry multiplex staining via text-to-stain diffusion trained on uniplex stains. International Workshop on Machine Learning in Medical Imaging.* Springer; 2024:143–155.
47. He Y, Liu Z, Qi M, et al. PST-Diff: Achieving High-Consistency Stain Transfer by Diffusion Models With Pathological and Structural Constraints. *IEEE Trans Med Imaging.* 2024;43(10):3634–3647.
48. Kataria T, Knudsen B, Elhabian SY. StainDiffuser: multitask dual diffusion model for virtual staining. *arXiv preprint arXiv:240311340*. Published online 2024.
49. Akbarnejad A, Ray N, Barnes PJ, Bigras G. Predicting ki67, er, pr, and her2 statuses from h&e-stained breast cancer images. *arXiv preprint arXiv:230801982*. Published online 2023.
50. Wang Z, Bovik AC. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process Mag.* 2009;26(1):98–117.
51. Starovoitov VV, Eldarova EE, Iskakov KT. Comparative analysis of the SSIM index and the Pearson coefficient as a criterion for image similarity. *Eurasian J Math Comput Appl.* 2020;8(1):76–90.
52. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process.* 2004;13(4):600–612.
53. Wang Z, Simoncelli EP, Bovik AC. Multiscale structural similarity for image quality assessment. In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol 2. IEEE; 2003:1398-1402.
54. 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. *Adv Neural Inf Process Syst.* 2017:30.
55. Ding K, Ma K, Wang S, Simoncelli EP. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Trans Pattern Anal Mach Intell.* 2022;44(5):2567–2581.

56. Martell MT, Haven NJM, Cikaluk BD, et al. Deep learning-enabled realistic virtual histology with ultraviolet photoacoustic remote sensing microscopy. *Nat Commun.* 2023;14(1):5967.
57. Rivenson Y, Wang H, Wei Z, et al. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nat Biomed Eng.* 2019;3(6):466–477.
58. Campanella G, Hanna MG, Geneslaw L, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat Med.* 2019;25(8):1301–1309.
59. Yao J, Zhu X, Huang J. Deep multi-instance learning for survival prediction from whole slide images. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*. Springer; 2019:496-504.
60. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning (still) requires rethinking generalization. *Commun ACM.* 2021;64(3):107–115.
61. Lin T, Yu Z, Hu H, Xu Y, Chen CW. Interventional bag multi-instance learning on whole-slide pathological images. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.; 2023:19830-19839.
62. Lin W, Zhuang Z, Yu L, Wang L. Boosting multiple instance learning models for whole slide image classification: A model-agnostic framework based on counterfactual inference. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol 38.; 2024:3477-3485.
63. Min S, Oh HJ, Jeong WK. Co-synthesis of Histopathology Nuclei Image-Label Pairs using a Context-Conditioned Joint Diffusion Model. In: *European Conference on Computer* Vision. Springer; 2024:146-162.
64. Winter D, Triltsch N, Rosati M, et al. Mask-guided cross-image attention for zero-shot in-silico histopathologic image generation with a diffusion model. *arXiv preprint arXiv:240711664*. Published online 2024.
65. Ho MM, Dubey S, Chong Y, Knudsen B, Tasdizen T. F2FLDM: Latent Diffusion Models with Histopathology Pre-Trained Embeddings for Unpaired Frozen Section to FFPE Translation. *arXiv preprint arXiv:240412650. Published online* 2024.
66. Chen RJ, Ding T, Lu MY, et al. Towards a general-purpose foundation model for computational pathology. *Nat Med.* 2024;30(3):850–862.
67. Lu MY, Chen B, Williamson DFK, et al. A visual-language foundation model for computational pathology. *Nat Med.* 2024;30(3):863–874.