UPPSALA
UNIVERSITET

IT mBM 26 013
Degree project 30 credits
07/2026

# Cross-Modality Image Translation from Routine Brightfield Slides to Synthetic cPOL Using Deep Learning

Joel Sundin

Master's Programme in Image Analysis & Machine Learning

UPPSALA
UNIVERSITET

# Cross-Modality Image Translation from Routine Brightfield Slides to Synthetic cPOL Using Deep Learning

Joel Sundin

## Abstract

Collagen fiber organization in the tumor microenvironment shapes disease progression, with fiber alignment and density identified as prognostic markers. Most pathology laboratories image collagen using picrosirius red (PSR)-stained brightfield (BF) microscopy, which reveals the spatial distribution of collagen but not fiber-level structure. Circularly polarized light (cPOL) microscopy addresses this through imaging collagen birefringence, but requires specialized optical hardware and strict acquisition protocols unavailable to most laboratories. Whether cPOL images can be synthesized computationally from routine BF acquisitions has not been established. Here we show that they can: a model trained on paired PSR-BF and cPOL whole-slide images reproduces dominant fiber orientation to within 2 degrees and birefringent color composition to within 2 percentage points of ground truth.

Adversarial regularization drives this result, a model trained with pixel-wise losses alone overestimates red-orange collagen by over 27 percentage points despite achieving the best pixel-level scores, showing that standard image similarity metrics are actively misleading in this setting. These results show that PSR-BF slides contain sufficient information for PSR-cPOL synthesis, enabling for retrospective collagen fiber analysis in existing archives without additional hardware or re-staining. Code is available at https://github.com/joelsundin/bf2cpol/.

**Faculty of Science and Technology**

**Uppsala University, Uppsala**

Supervisors: Francisco J. Peña (KTH), Tânia Costa (KI), Fahim Ebrahimi (USB)
Subject reader: Joakim Lindblad

Examiner: Nataša Sladoje

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

A tumor is best understood not only as a population of malignant epithelial cells, but as part of a larger tissue system in which the surrounding stroma and extracellular matrix (ECM) actively influence disease progression. Tumor development is not driven by epithelial cancer cells alone, but also by their interactions with a dynamic microenvironment where stromal cells and ECM components provide biochemical and mechanical cues. When communication is disrupted, tissue architecture can be altered in ways promoting tumor initiation and progression [1, 2]. Remodeling and stiffening of the ECM change the mechanical environment of the tissue and activate intracellular signaling pathways involved in cancer cell proliferation [3]. These effects are mediated through cell-ECM interactions, where receptors such as integrins allow cells to sense and respond to both biochemical composition and mechanical properties of their surroundings.

Fibrillar collagen is the most abundant ECM component. In tumors, fiber alignment, orientation, and density are extensively reorganized, influencing both tissue mechanics and cell signaling [4, 5]. These structural alterations have been identified as prognostic biomarkers [6], highlighting the importance of accurately imaging and quantifying collagen architecture.

Most institutions rely on standard brightfield (BF) microscopy, which when combined with picrosirius red (PSR) staining, reveals the spatial distribution of collagen within tissue sections. Since BF images optical density rather than birefringence, structural features such as fiber orientation and alignment are not directly accessible, and this information is routinely lost in standard histopathological workflows. Circularly polarized light (cPOL) imaging of PSR stained tissue has gained increasing attention as a tool to address this limitation. By exploiting the birefringent properties of fibrillar collagen, cPOL enables consistent visualization of fiber organization regardless of orientation, providing access to quantitative analysis of features such as fiber alignment, density, and spatial distribution. Ideally, cPOL imaging would be a routine part of histopathological assessment, but it requires specialized optical hardware and strict acquisition protocols, making it inaccessible to the majority of pathology laboratories worldwide.

This thesis investigates whether this gap can be closed computationally. The main research question of this thesis is: can a convolutional neural network learn a mapping from PSR-BF to PSR-cPOL images that recovers collagen structural information not directly visible in the BF input? By learning such a mapping, a convolutional neural network can potentially recover cPOL-specific information from images acquired with a standard BF microscope, making cPOL analysis accessible to

any laboratory that already performs routine PSR staining. This approach also enables for retrospective analysis: archives of PSR-BF slides exist for cases with known clinical outcomes, and synthetic cPOL images could be generated from this existing material to study collagen organization in relation to disease progression, without new hardware or re-staining.

Evaluating generative models in this setting is not straightforward, standard image similarity metrics do not capture whether biologically relevant information is preserved, motivating task-specific evaluation.

This thesis makes the following contributions:

1. A paired image-to-image translation pipeline for synthesizing PSR-cPOL WSIs from PSR-BF.

2. A preprocessing pipeline addressing modality-specific alignment errors introduced during WSI stitching, using raw field-of-view tiles to ensure spatial consistency.

3. A multi-level evaluation pipeline, combining perceptual, and distributional metrics with collagen-specific downstream analysis using CTFIRE-based fiber extraction and color-based birefringent collagen segmentation.

4. An empirical demonstration that pixel-wise metrics are insufficient and might be actively misleading for evaluating cross-modality image translation.

Code for this work is publicly available on GitHub.

# Chapter 2

# Background

The methods developed in this thesis build on established imaging modalities and deep learning architectures. The following sections introduce BF imaging, cPOL microscopy, optical properties of collagen related to cPOL imaging, and the U-Net architecture on which the implementations in this work are based.

## 2.1 Brightfield Microscopy and PSR Staining

Brightfield (BF) microscopy is a standard imaging modality in histopathology, where thin tissue sections are placed on glass slides and transilluminated by a light source. The tissue is commonly stained to enhance contrast and enable assessment of morphology. In BF imaging, transmitted light passes through the sample, and image contrast arises from differences in light absorption and scattering within the tissue. As stains are applied, the recorded image also reflects variations in stain uptake, allowing specific tissue components, such as collagen in picrosirius red (PSR)-stained sections, to be highlighted.

Whole-slide images (WSIs) are typically acquired by scanning a sample in a tiled manner, where high-resolution image patches are captured sequentially across the slide and subsequently stitched together to form a gigapixel-scale representation, enabling visualization of tissue structure both at cellular level, and large-scale spatial levels.

In the context of collagen analysis, PSR staining is commonly applied due to its selectivity for fibrillar collagen. PSR-BF images allows for assessment of the spatial distribution of collagen within the sample, but since BF microscopy records intensity-based contrast, structural features such as fiber orientation and alignment are not directly accessible.

## 2.2 Polarized Light and cPOL Imaging

Fibrillar collagen is highly anisotropic, meaning that its interaction with light depends on the orientation of the collagen fibers. As polarized light passes through an anisotropic material, it is split into two components that propagate at different velocities and are polarized orthogonally to each other; this is referred to as birefringence [7].

Birefringence arises from differences in refractive index along different axes of the material, resulting in a phase shift between the two orthogonal light components. When these components are recombined at an analyzer, their interference produces

intensity variations that depend on fiber orientation, allowing for visualization of anisotropic structure such as collagen.

PSR is an elongated dye molecule that enhances the natural birefringence of collagen. When bound to collagen fibers, PSR aligns parallel to the fibrillar structure, amplifying the birefringent signal. Under POL microscopy, PSR-stained collagen appears bright and highly contrasted against the surrounding tissue, which remains comparatively dark [7].

In linearly polarized imaging systems, the detected signal intensity depends on the orientation of the collagen fibers relative to the polarization axis, resulting in reduced visibility at specific angles. This directional dependence can be mitigated by acquiring multiple images at different sample orientations and combining them, or more effectively by imaging using circularly polarized light.

Circularly polarized microscopy employs an arrangement of polarizing elements that allows birefringent structures to be visualized independently of their orientation, enabling for consistent detection of collagen fibers across all angles.

## 2.3 Motivation for Cross-Modality Translation

Brightfield images of PSR-stained tissue are widely available, while cPOL imaging is rarely acquired outside specialized research settings. This disparity in availability motivates cross-modality image translation as a way to recover polarization-based collagen information from routinely available inputs.

The central challenge is not only to generate visually plausible cPOL-like images, but to determine, and evaluate in quantifiable terms, whether the generated representations preserve collagen features meaningful for downstream analysis, such as fiber orientation, alignment, and birefringent color composition. Figure 2.1 illustrates this modality gap on the same tissue section: the BF image captures the spatial distribution of collagen as a uniform red signal, while the cPOL image reveals fiber-level structure through birefringence.

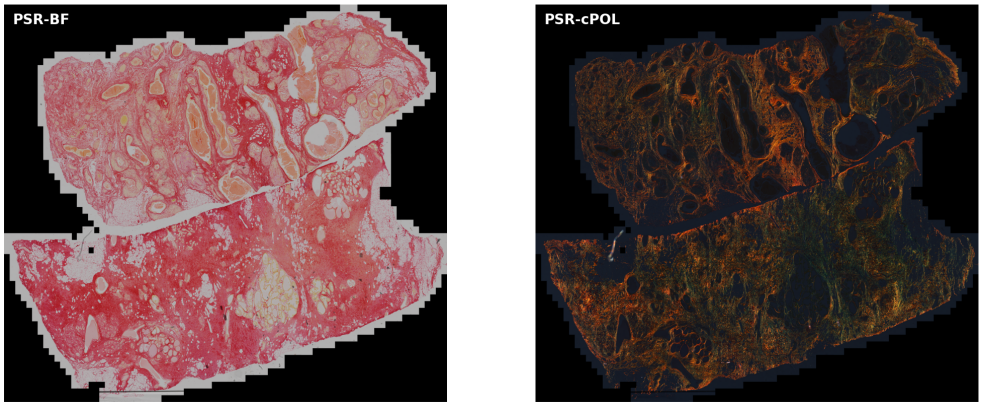



Figure 2.1: The same PSR-stained tissue section imaged in BF (left) and cPOL (right), shown at the lowest pyramid level for overview purposes. BF reveals the spatial distribution of collagen as a uniform red signal, while cPOL exploits collagen birefringence to reveal fiber-level structure.

## 2.4 The U-net Architecture

The models implemented in this thesis are derived from the U-net architectural family. The original U-Net is therefore briefly described before detailing the specific modifications used in each implementation.

The U-Net is a fully convolutional encoder-decoder network, developed primarily for biomedical image segmentation. Its architecture consists of a contracting path leading into an expansive path. The encoder follows the typical structure of a CNN. At each resolution level, two successive $3 \times 3$ unpadded convolutions are applied, each followed by a ReLU activation, before the feature maps are downsampled using $2 \times 2$ max-pooling with stride 2. After each downsampling step, the number of feature channels is doubled, shifting the representation from high-resolution spatial detail towards a lower-resolution, richer feature encoding.

The expansive path (i.e., decoder) restores the spatial resolution of the data through a sequence of upsampling operations. At each step, the feature map is upsampled using a $2 \times 2$ up-convolution, halving the number of feature channels. The upsampled feature map is then concatenated with the corresponding feature map from the contracting path. This skip-connection mechanism is one of the defining features of the U-Net architecture, as these concatenations allow for high-resolution spatial information, that might have been lost during the contraction path, to be reintroduced during reconstruction. After each concatenation, two additional $3 \times 3$ convolutions followed by ReLU activations are applied. Finally, a $1 \times 1$ convolution maps the resulting feature representation to the desired number of output classes. In the original formulation, the network contains 23 convolutional layers [8].

# Chapter 3

# Related Work

Histology images acquired under different modalities often reveal complementary aspects of the same tissue, and information present in one modality may not be directly visible in another. This has motivated deep learning-based image-to-image translation in computational pathology, where models are trained to synthesize missing or alternative imaging modalities from more accessible inputs. Existing approaches have primarily been based on convolutional encoder-decoder architectures and conditional generative adversarial networks (GANs), demonstrating that modality-specific structural information can be recovered from a range of histological image types.

U-Net based architectures [8] has been widely adopted for cross-modality translation in biomedical imaging and histopathology [6, 9–11]. Keikhosravi et al. [6] demonstrated that a modified U-Net trained with a combined mean squared error (MSE) and structural similarity index measure (SSIM) objective can synthesize second harmonic generation (SHG) images directly from hematoxylin and eosin (H&E) brightfield inputs, enabling quantitative assessment of collagen fiber without requiring specialized SHG imaging systems.

Rather than the standard convolutional blocks and pooling operations of the original U-Net, Keikhosravi et al. [6] proposed an encoder-decoder network with residual blocks and strided convolutions for downsampling. In the decoder, the final upsampling stage is replaced by a pixel shuffle layer, which reconstructs the full-resolution output from lower-resolution, higher-channel feature maps. To constrain the output to a fixed range ([0,1]), the reconstructed image is passed through a sigmoid activation function.

Their implementation was trained on over one million paired BF-SHG image patches, derived from annotated regions of 278 normal and 211 breast cancer cases, along with 1196 pancreatic tissue microarray cores. The authors report a range of pixel-wise evaluation metrics, including a mean squared error (MSE) of $0.002$, peak signal-to-noise ratio (PSNR) of $32.11$, and structural similarity index (SSIM) of $0.66$.

In addition to pixel-level evaluation, Keikhosravi et al. assessed downstream collagen analysis in CurveAlign[12]. They reported no significant difference in fiber orientations and alignment between synthesized and real SHG images on an independent pancreatic TMA slide, and further demonstrated that synthesized SHG images could be used to identify similar TACS-3 regions and yield similar relative angle measurements to real SHG images in a breast cancer TMA core.

While producing visually realistic images, the authors observed that the synthesized SHG outputs exhibited smoother fiber structures compared to ground-truth

SHG microscopy. This effect is likely related to the use of pixel-wise and structural similarity loss functions, which tend to prioritise overall intensity and structural agreement while underrepresenting high-frequency details. Although smoothing may reduce noise, it also risks suppressing fine collagen morphology, a limitation that is central to the evaluation approach taken in this thesis.

Rivenson et al. [9] proposed a U-Net-based architecture combined with a discriminator network, forming a generative adversarial network (GAN). Their implementation demonstrated strong performance in virtually staining autofluorescence images of tissue, producing results comparable to standard histological staining. Their approach supports the generation of multiple stain types and is primarily evaluated through blinded assessment by expert pathologists, showing no major diagnostic discordance between virtual and real stains.

The authors note that a by-product of training on large datasets of whole-slide images is the emergence of a consistent average stain appearance, leading not only to virtual staining but also to implicit stain normalization. This is highly relevant also in our work, as a model trained on paired PSR-BF and cPOL images may similarly learn aspects of the cPOL imaging protocol, such as contrast characteristics and color distribution, alongside the target birefringent signal.

De Haan et al. [13] use a similar GAN-based approach for virtual staining of label-free (unstained) autofluorescence images, generating aligned hematoxylin and eosin (H&E) and special stains including Masson's Trichrome, periodic acid-Schiff, and Jones silver stain. These outputs are subsequently used to train supervised stain-to-stain transformation networks, ensuring spatial alignment of the training data. Performance is evaluated through a blinded study in which renal pathologists assess diagnoses made using H&E alone, H&E with generated stains, and H&E with histochemically stained sections, showing that generated stains improve diagnostic accuracy over H&E alone, with performance comparable to real histochemical stains. The authors further evaluate the perceptual quality of the generated stains through expert scoring of stain quality, nuclear detail, cytoplasmic detail, and extracellular fibrosis, showing that stain transformation outputs achieve quality ratings comparable to histochemically stained tissue across all evaluated metrics.

More recently, diffusion-based approaches have emerged as an alternative to GAN-based methods for image generation and translation, and have been explored for their potential to improve perceptual quality and controllability. Zhang et al. [14] employ a conditional diffusion model to generate pixel super-resolved, virtually stained images from label-free autofluorescence inputs. In this approach, low-resolution autofluorescence images are translated into H&E-stained brightfield images using a Brownian bridge diffusion process, with engineered sampling strategies introduced to reduce variance in the model output. Their model is trained on paired autofluorescence–H&E image patches and evaluated on held-out test images. The implementation outperforms cGAN-based approaches in super-resolution settings ($2\times$–$5\times$), while achieving comparable performance in the $1\times$ case, where no resolution enhancement is performed. In this setting, both models achieve similar average SSIM values ($\sim 0.69$), while the cGAN slightly outperforms the diffusion model in PSNR ($18.75$ vs. $18.50$), both methods yield comparable Learned Perceptual Image Patch Similarity (LPIPS) values ($\sim 0.27$).

In a broader foundation-model setting, PixCell [15] extends diffusion-based histopathology generation to large-scale pretraining on 30.8 million H&E image patches from 69,184 whole-slide images, evaluated both for synthetic H&E generation and virtual staining. For H&E-to-IHC translation, the authors adapt the pretrained model

to HER2 stain generation and compare it to the CycleGAN-based MIST baseline. PixCell with lightweight LoRA adaptation improves perceptual image quality, reducing Fréchet Inception Distance (FID) from $54.28$ to $52.32$ and Crop FID from $33.22$ to $20.87$, although MIST achieves a slightly higher SSIM ($0.1945$ vs. $0.1892$). These results suggest that diffusion-based foundation models can improve perceptual realism in virtual staining, while still requiring target-domain adaptation to perform competitively

Across the reviewed methods, evaluation is dominated by pixel-based metrics such as MSE, PSNR, and SSIM, alongside perceptual and distributional feature-space metrics such as FID and LPIPS [6, 10, 16, 17]. These purely quantitative measures are often complemented by downstream task evaluation [6, 11] or qualitative assessment by expert pathologists [11, 13, 14]. For tasks where fine tissue structure must be preserved, pixel-based metrics alone are insufficient, a consideration that directly shapes the evaluation approach taken in this thesis.

# Chapter 4

# Dataset

The dataset consists of four matched WSI pairs of Picrosirius red (PSR)-stained tissue, with each tissue section acquired in both BF and cPOL. The BF image serves as the input modality and the corresponding cPOL image as the supervised target.

The images were acquired at Zeiss headquarters and provided by Tânia Costa at Karolinska Institutet, Huddinge, Sweden. The cPOL images exhibit strong collagen contrast, providing a high-quality structural target signal for model training.

The dataset consist of full tissue sections rather than tissue microarray cores, needle-biopsy cores, or pre-selected image crops. Images were acquired at $20\times$ magnification as 8-bit RGB whole-slide images with a pixel size of $0.173$ $\mu$m, each spanning tissue regions on the order of approximately $2$ cm per side and containing approximately 18–23 gigapixels. Direct processing at full resolution is computationally impractical; the WSIs are therefore divided into smaller paired patches before model training, as described in the following sections. All images were provided as multi-resolution pyramids with downsampling factors from $1$ to $256$, enabling access to both low-resolution slide overviews and full-resolution image data for patch extraction.

Before training, the data was partitioned at WSI level. Two WSIs were used for training, one for validation, and one was held out as the test set. WSI-level partitioning ensures that spatially correlated patches from the same tissue section do not appear in both training and evaluation sets, commonly referred to as spatial leakage.

## 4.1 Data Preprocessing

Full-resolution WSIs range from approximately $128{,}000 \times 140{,}000$ to $134{,}000 \times 170{,}000$ pixels. In an uncompressed 8-bit RGB representation, a single WSI would require approximately $54$-$68$ GB of memory, far exceeding what can be held in GPU memory during training or inference. Each WSI is therefore processed as a series of smaller tiles extracted at full resolution, i.e., pyramid level $0$.

WSIs are reconstructed by stitching individual field-of-view (FOV) images acquired as the slide is scanned. Inspection of the available image pairs reveal that this stitching process introduces spatially varying alignment errors between the BF and cPOL modalities, likely caused by differences in image content, as non-birefringent tissue structures visible in BF are not highlighted in cPOL, which may affect how the stitching algorithm registers neighbouring tiles.

As a result, alignment quality varies across the WSI, where some regions show near pixel-wise correspondence while others are shifted by several pixels in different

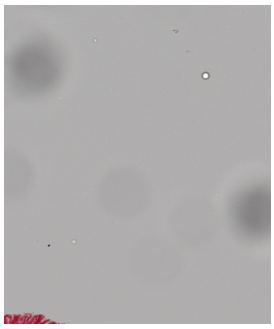
directions. A single global registration such as an affine transformation is therefore insufficient, while more flexible non-rigid methods risk introducing interpolation artifacts or unrealistic distortions.

To avoid these issues, tiling is performed on the raw unstitched FOV images rather than on the reconstructed WSIs. Since each BF and cPOL pair are acquired from the same tissue section, the FOV pairs are spatially consistent before stitching errors are introduced. The FOV metadata contains bounding-box information mapping each tile back to its location in the reconstructed WSI, making it possible to perform inference at FOV level while preserving spatial context within the full tissue section.

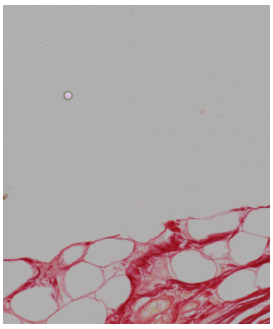
Each WSI is reconstructed from approximately $2500$–$3000$ FOV tiles, each with a resolution of $2056 \times 2464$ pixels. Many tiles consist mostly of background and are discarded before patch extraction to avoid redundancy in the training data. For each FOV tile-pair, a binary tissue mask is estimated from the BF representation in HSV color space.

The FOV tiles vary substantially in appearance, ranging from densely stained collagen-rich regions to pale non-collagen tissue and bright background, general-purpose foreground–background thresholding did not generalize reliably across this range. Tissue is therefore identified using fixed thresholds in HSV space. Each BF tile is scaled to $[0, 1]$ and converted from RGB to HSV. A pixel is classified as tissue if it is either sufficiently saturated, $S > \tau_s$, indicating stained tissue independently of brightness, or sufficiently dark, $V < \tau_v$, indicating non-background structure; near-black pixels with $V \leq 0.05$ are excluded to avoid black borders of the WSI. With $\tau_s = 0.02$ and $\tau_v = 0.97$, this retains both well stained collagen and faint tissue while rejecting the homogeneous bright background. The mask is then refined using connected-component analysis, removing foreground components smaller than $400$ pixels as a denoising step.

The tissue fraction of a FOV tile is computed as the mean of the binary tissue mask. Tiles with a tissue fraction below $10\%$ are discarded. Since background regions are highly homogeneous, a small number of low-tissue tiles are retained to ensure the model is exposed to realistic background during training. The threshold is set to avoid overrepresentation of uninformative regions. Figure 4.1 illustrates examples of tiles at different tissue fractions.



(a) Field-of-view tile with a computed tissue fraction of $1\%$.



(b) Field-of-view tile with a computed tissue fraction of $14\%$.

Figure 4.1: Examples of tiles with different tissue fractions computed during tile extraction. Tiles with insufficient tissue content were excluded to reduce the number of background-only samples.

For each accepted FOV pair, corresponding tiles are extracted from the BF and cPOL images at a size of $512 \times 512$ pixels. For training, a stride of 512 pixels is used, producing non-overlapping tile pairs; since the image dimensions are not exact multiples of the tile size, any tile that would extend beyond the image boundary is discarded, excluding a thin border region from the training set. For the held-out test partition, tiles are extracted at the same size with a nominal stride of 256 pixels, producing overlapping tiles where each pixel can contribute to multiple predictions. To preserve full coverage, the stride of the final step along each axis is reduced as needed so the last tile sits flush with the image edge, ensuring the entire image is tiled. The overlapping predictions are aggregated during reconstruction to reduce boundary artifacts between neighbouring tiles, enabling for evaluation on full FOV-level regions rather than on isolated patches. The two tiling strategies are illustrated in Figure 4.2, and the reconstruction procedure is described in Section 6.3.

After patch extraction and tissue filtering, this split produced 100,835 non-overlapping $512 \times 512$ training tiles and 65,299 validation tiles. The held-out test partition consists of 187,490 overlapping $512 \times 512$ tiles, which are reconstructed into 2604 microscopy FOV tiles; evaluation is performed at FOV, WSI, or ROI level. The large validation set relative to training reflects the natural variation in tissue content across WSIs. The dataset and its partition are summarised in Table 4.1.

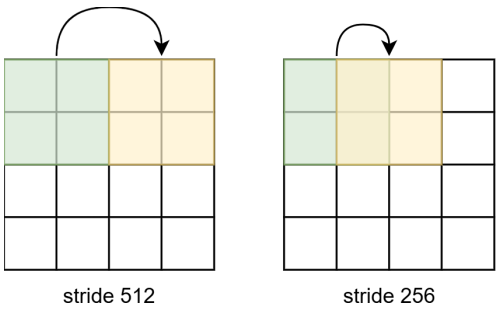



Figure 4.2: Patch extraction strategies for training (left) and evaluation (right). Training tiles ($512 \times 512$, stride 512) are non-overlapping, with each pixel included once. Test tiles use a stride of 256 pixels, producing overlapping patches where multiple predictions contribute to the same spatial locations, enabling for high quality FOV-level reconstruction.

Table 4.1: Summary of the dataset and its partition. Each WSI is a PSR-stained tissue section acquired in both BF (input) and cPOL (target), provided in Zeiss CZI pyramid format. Tile counts are reported after tissue filtering.

| **Split** | **WSIs** | **Patch size** | **Stride** | **Tiles** |
|---|---|---|---|---|
| Training | 2 | $512 \times 512$ | 512 (non-overlapping) | 100,835 |
| Validation | 1 | $512 \times 512$ | 512 (non-overlapping) | 65,299 |
| Test | 1 | $512 \times 512$ | 256 (overlapping) | 187,490: 2,604 (FOV-tiles) |
| **Total** | **4** | | | |

# Chapter 5

# Baseline & Evaluation Strategy

In paired image-to-image translation, pixel-wise similarity metrics such as MSE, PSNR, and SSIM provide standardised measures of low-level agreement and are useful during training where reconstruction losses constrain the output to remain close to the paired target. Their usefulness as final evaluation criteria depends on the modality and task objective

## 5.1 Baseline Pix2Pix

Since PSR-BF to cPOL translation has not, to our knowledge, been previously explored, there are no task-specific methods to compare against. Pix2Pix [18] is therefore used as a general paired image-to-image translation baseline, allowing the proposed models to be contextualised relative to an established method in image-translation literature.

Pix2Pix [18] is a conditional GAN developed for paired image-to-image translation, where the generator translates the input into a synthetic target-domain image and the discriminator learns to distinguish real from generated outputs conditioned on the corresponding input. This conditioning encourages the generated image to be both visually realistic and spatially consistent with the input structure.

The Pix2Pix objective combines an adversarial loss with a pixel-wise reconstruction loss, where the adversarial term encourages perceptual realism and the reconstruction term enforces correspondence with the paired ground truth. The generator objective is

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}}(G, D) + \lambda \mathcal{L}_1(G), \tag{5.1}$$

where $\mathcal{L}_{\text{adv}}$ denotes the conditional adversarial loss, $\mathcal{L}_1$ is the pixel-wise $L_1$ reconstruction loss between the generated image $\hat{y} = G(x)$ and the target image $y$, and $\lambda$ controls the relative weight of the reconstruction term. The $L_1$ loss is defined as

$$\mathcal{L}_1(G) = E_{x,y}\left[\|y - G(x)\|_1\right], \tag{5.2}$$

where the expectation over paired samples $(x, y)$ is approximated during training by averaging the loss over each mini-batch.

The $\mathcal{L}_1$ term penalizes pixel-wise differences linearly rather than quadratically, making it less sensitive to large individual errors than MSE, discussed further in Section 5.2.1. The adversarial term complements this by encouraging sharper and more realistic outputs than a purely pixel-wise objective alone would produce.

## 5.2 Evaluation Metrics

The objective of this thesis is not only to reproduce broad image features, but to generate cPOL images preserving collagen structures relevant for downstream analysis. A prediction can obtain favourable MSE, PSNR, or SSIM values while still suppressing high-frequency collagen details important for fiber-level interpretation, smooth predictions reduce average pixel-wise error while losing the fine structural detail that matters most.

This limitation is consistent with previous work on perceptual image similarity. Zhang et al. [19] show that metrics such as PSNR and SSIM can disagree with human perceptual judgments, especially for structured image outputs where small pixel-wise differences do not correspond to perceptual or semantic similarity. Pixel-level metrics are therefore useful for measuring low-level agreement, but insufficient as standalone evaluation criteria when fine structure and interpretability are important. Figure 5.1 illustrates this directly: the smoother generated image obtains more favourable MSE, PSNR, and SSIM values while visibly suppressing fine fiber detail.

For this reason, MSE and PSNR are reported but treated as low-level fidelity metrics rather than definitive indicators of model quality. Final model assessment combines these with perceptual, distributional, and collagen-specific evaluation, as described in the following subsections.

Feature-space metrics such as LPIPS and FID are commonly reported in digital pathology image generation tasks [6, 10, 16, 17]. LPIPS measures perceptual similarity between generated and ground-truth images in a learned feature space, providing a measure less tied to pixel-wise agreement than MSE or PSNR. FID evaluates whether the distribution of generated cPOL images resembles that of real cPOL images at the dataset level. Together these metrics provide complementary information about visual realism and perceptual agreement, but neither directly measures whether collagen morphology is preserved

The most task-relevant evaluation is therefore collagen-specific downstream analysis. Fiber extraction using CTFIRE and color-based segmentation of birefringent collagen categories are used to assess whether generated images preserve structural features relevant for fiber analysis. Model performance is interpreted based on this combined evidence rather than any single metric.

The following subsections define the metrics and analyses discussed in the evaluation, grouped by what they measure: MSE, PSNR, and SSIM for low-level pixel-wise agreement; LPIPS and FID for perceptual and distributional similarity; and CTFIRE and color-based collagen segmentation for collagen-specific structural assessment.

### 5.2.1 Mean Squared Error

Mean squared error (MSE) is a pixel-wise metric that measures the average squared difference between a generated image and its corresponding ground truth. For an image of size $H \times W \times C$, MSE is defined as

$$\mathrm{MSE} = \frac{1}{HWC} \sum_{c=1}^{C} \sum_{i=1}^{H} \sum_{j=1}^{W} \left( I_{\mathrm{pred}}(i,j,c) - I_{\mathrm{gt}}(i,j,c) \right)^2 , \tag{5.3}$$

where $I_{\mathrm{pred}}$ and $I_{\mathrm{gt}}$ denote the predicted and ground-truth images respectively. $i$, $j$, and $c$ index the spatial dimensions and image channels.

In this thesis, images have a resolution of $512 \times 512$ with three channels ($C = 3$), MSE is computed across all pixel values and channels. MSE provides a direct measure

of average pixel-wise error. As a local metric, MSE is highly sensitive to small spatial misalignments and does not account for perceptual similarity, structural coherence, or preservation of fine collagen fiber morphology.

### 5.2.2 Peak Signal-to-Noise Ratio

Peak signal-to-noise ratio (PSNR) is a signal processing metric used to quantify reconstruction quality by comparing the magnitude of a signal to the error introduced during reconstruction. It measures the ratio between the maximum possible pixel value and the mean squared error.

PSNR is defined in terms of MSE as

$$\mathrm{PSNR} = 10 \log_{10} \left( \frac{MAX^2}{\mathrm{MSE}} \right), \tag{5.4}$$

where $MAX$ denotes the maximum possible pixel value of the image, for example $MAX = 255$ for 8-bit images or $MAX = 1$ for images normalized to the range $[0, 1]$.

PSNR is expressed in decibels (dB), where higher values indicate better reconstruction quality. Since PSNR is directly derived from MSE, it shares the same limitations, including sensitivity to small spatial misalignments and limited ability to capture perceptual or structural differences.

### 5.2.3 Structural Similarity Index

The structural similarity index (SSIM) [20] is a perceptual metric designed to measure similarity between two images based on luminance, contrast, and structural similarity. Given an image-patch pair, SSIM can be defined as

$$\mathrm{SSIM}(x, y) = l(x, y) \cdot c(x, y) \cdot s(x, y), \tag{5.5}$$

where the luminance, contrast, and structure comparisons are given by

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2}, \quad s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}. \tag{5.6}$$

Here, $\mu_x$ and $\mu_y$ denote mean intensities, $\sigma_x$ and $\sigma_y$ denote standard deviations, and $\sigma_{xy}$ denotes covariance between the two image patches. The constants $C_1$, $C_2$, and $C_3$ are included for numerical stability.

SSIM is commonly reported in image-to-image translation and virtual staining studies [6, 14, 21], and is used as part of the MSE-SSIM training objective (Section 6.4.1) to encourage similarity in local image statistics

SSIM is not used as a standalone evaluation criterion in this thesis. Zhang et al. [19] show that SSIM and related metrics can disagree with human perceptual judgments for structured image outputs, where small pixel-wise differences do not necessarily correspond to meaningful perceptual or semantic differences. During preliminary evaluation runs, SSIM was found to continuously favour smooth predictions that preserve broad luminance and contrast patterns while suppressing fine collagen detail, a significant limitation for cPOL images where high-frequency fiber structures are important. This limitation, shared with MSE and PSNR, is illustrated in Figure 5.1.

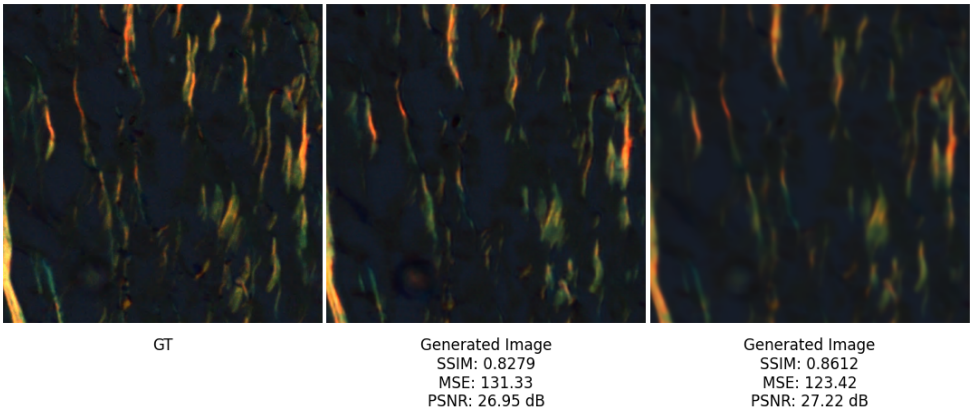



Figure 5.1: A randomly sampled $500 \times 500$ pixel crop illustrating the limitation of pixel-level metrics for evaluating generated cPOL images. From left to right: ground truth, sharp prediction, and smooth prediction. Despite visibly suppressing fine, high frequency detail, the smoother prediction obtains more favourable MSE, PSNR, and SSIM values, demonstrating that these metrics alone are insufficient for selecting the most structurally faithful output.

### 5.2.4 Fréchet Inception Distance

The Fréchet Inception Distance (FID) is a distributional metric that compares feature representations extracted from generated and real images using a pretrained Inception-v3 network. In this thesis, FID is computed using the Clean-FID implementation [22].

In this thesis, FID is computed at the FOV level by comparing the distribution of predicted cPOL FOVs with that of the ground-truth cPOL FOVs. Given feature sets $\mathbf{f}$ and $\hat{\mathbf{f}}$ extracted from real and generated images respectively, with means $\mu$, $\hat{\mu}$ and covariance matrices $\Sigma$, $\hat{\Sigma}$, FID is defined as

$$\mathrm{FID} = \|\mu - \hat{\mu}\|_2^2 + \mathrm{Tr}\left(\Sigma + \widehat{\Sigma} - 2(\Sigma\widehat{\Sigma})^{1/2}\right). \tag{5.7}$$

Here, $\mathrm{Tr}(\cdot)$ denotes the trace operator. In this thesis, $\mathbf{f}$ corresponds to the Inception features extracted from the ground-truth cPOL FOV tiles, while $\hat{\mathbf{f}}$ corresponds to the features extracted from the predicted cPOL FOV tiles.

Lower FID values indicate that the generated FOV tiles are closer to the ground-truth distribution in the Inception feature space, making it useful for assessing whether generated cPOL images reproduce the broader appearance, texture, and structural characteristics of real cPOL images across the held out FOV tiles.

This FOV-level evaluation is conceptually similar to crop-based FID [15], where FID is computed on local image regions rather than complete images, though here the FOVs are reconstructed predictions rather than randomly sampled crops.

### 5.2.5 Learned Perceptual Image Patch Similarity

Learned Perceptual Image Patch Similarity (LPIPS) [19] is a perceptual similarity metric that measures the distance between individual image pairs in a learned feature space extracted from a pretrained deep neural network. Unlike FID, that operates at the distribution level, LPIPS provides a per-image perceptual similarity score where lower values indicate greater perceptual similarity

Since LPIPS operates in a learned feature space rather than pixel space, it can capture perceptual differences not reflected by MSE, relevant here where preservation of visible fiber structure and texture matters. LPIPS is a general perceptual metric and does not explicitly measure collagen morphology or biological fidelity, and is therefore interpreted alongside FID, visual inspection, and collagen-specific downstream analysis.

### 5.2.6 CTFIRE-Based Evaluation

CTFIRE [23] is used for domain-specific evaluation of collagen fiber preservation. CTFIRE is a curvelet transform-based tool widely used for quantitative analysis of collagen fiber organization in microscopy images [23, 24], originally developed for SHG image data but applicable to other collagen imaging modalities.

CTFIRE enhances curvilinear structures and extracts individual fibers from tissue backgrounds, outputting quantitative descriptors for fiber orientation, alignment, length, and straightness. In this thesis, CTFIRE is applied to both generated and ground-truth cPOL images across 12 manually selected ROIs, and three summary measures are computed to compare the extracted orientation distributions.

First, the orientation error $\Delta_{\text{orient}}$ is the angular difference between the dominant fiber orientations in the generated and ground-truth ROI, computed as the axial angular difference modulo $180°$.

Second, the alignment error $E_{align}$ measures how well the predicted fiber directions match the ground-truth distribution by comparing alignment indices of the ROIs:

$$E_{\text{align}} = |A_{\text{GT}} - A_{\text{pred}}| \,. \tag{5.8}$$

The alignment index $A \in [0, 1]$ is computed per ROI distribution, simply put, a scalar describing how aligned the fibers are in one distribution.

$$A = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}\sin 2\theta_i\right)^2 + \left(\frac{1}{N}\sum_{i=1}^{N}\cos 2\theta_i\right)^2}, \tag{5.9}$$

where $N$ is the number of extracted fibers and $\theta_i$ is the orientation of the $i$-th fiber. Intuitively, if two distributions are identical, $E_{align}$ should remain 0.

Third, the histogram distance $D_{\text{hist}}$ compares the full orientation distributions by binning fiber orientations into normalised histograms over $[0°, 180°]$ with bin width $\Delta\theta = 5°$:

$$D_{\text{hist}} = \frac{1}{2}\sum_{b} |h_{\text{pred}}(b) - h_{\text{gt}}(b)|\,\Delta\theta. \tag{5.10}$$

Lower values for all three measures indicate closer agreement with the ground-truth cPOL image.

### 5.2.7 Color-Based Semantic Segmentation of Collagen

A complementary domain-specific evaluation is performed based on color-based segmentation of birefringent collagen. In cPOL imaging, collagen exhibits characteristic birefringent colors: green, yellow, and red-orange, associated with differences in fiber organization. Preserving this color composition in generated images is therefore an indicator of whether biologically meaningful collagen characteristics are accurately reproduced.

Segmentation is performed using a pixel classifier trained interactively in QuPath [25] by a domain expert with expertise in cancer biology and extracellular matrix research, a common workflow for collagen analysis in research settings. The classifier is an OpenCV artificial neural network (a shallow multilayer perceptron) operating on the RGB channels at full resolution (pixel size $0.173$ $\mu$m). For each pixel, a 12-dimensional feature vector is formed by applying four filters at a scale of $\sigma = 1.0$ to each of the three color channels: Gaussian smoothing, the maximum structure-tensor eigenvalue, structure-tensor coherence, and the maximum Hessian eigenvalue. This combines local color with measures of local structure, orientation, and anisotropy, allowing the classifier to separate the birefringent collagen colors both from one another and from non-collagen tissue. The network maps the 12 features to four classes: the three birefringent collagen categories (green, yellow, and red-orange) and an additional non-collagen/background class that is excluded from the composition analysis. The classifier assigns collagen-positive regions to the three color categories across the full held-out test WSI, processed in four parts due to memory constraints.

The final collagen composition is computed by aggregating classified areas across the four parts rather than averaging percentages. For a model $m$ and collagen category $c$, the composition percentage is defined as

$$P_{m,c} = \frac{\sum_{i=1}^{4} A_{m,i,c}}{\sum_{i=1}^{4} \sum_{c'} A_{m,i,c'}} \times 100, \tag{5.11}$$

where $A_{m,i,c}$ denotes the area assigned to category $c$ in image part $i$, and $c'$ indexes all collagen categories. Differences relative to the ground-truth cPOL image are computed in percentage points as

$$\Delta_{m,c} = P_{m,c} - P_{\mathrm{GT},c}, \tag{5.12}$$

where positive values indicate overestimation of a given category and negative values indicate underestimation.

# Chapter 6

# Methods

Figure 6.1 provides an overview of the proposed pipeline. (a) Paired PSR-BF and cPOL whole-slide images are first decomposed into FOV-tiles, from which $512 \times 512$ pixel patch pairs are extracted for training. (b) A generator network $G$ is trained to predict synthetic cPOL patches from BF input under two configurations: MSE-SSIM and MSE-GAN. (c) At inference, overlapping patch predictions are combined using Hann-window blending to reconstruct full FOV and WSI-level outputs. The following sections describe each stage in detail.

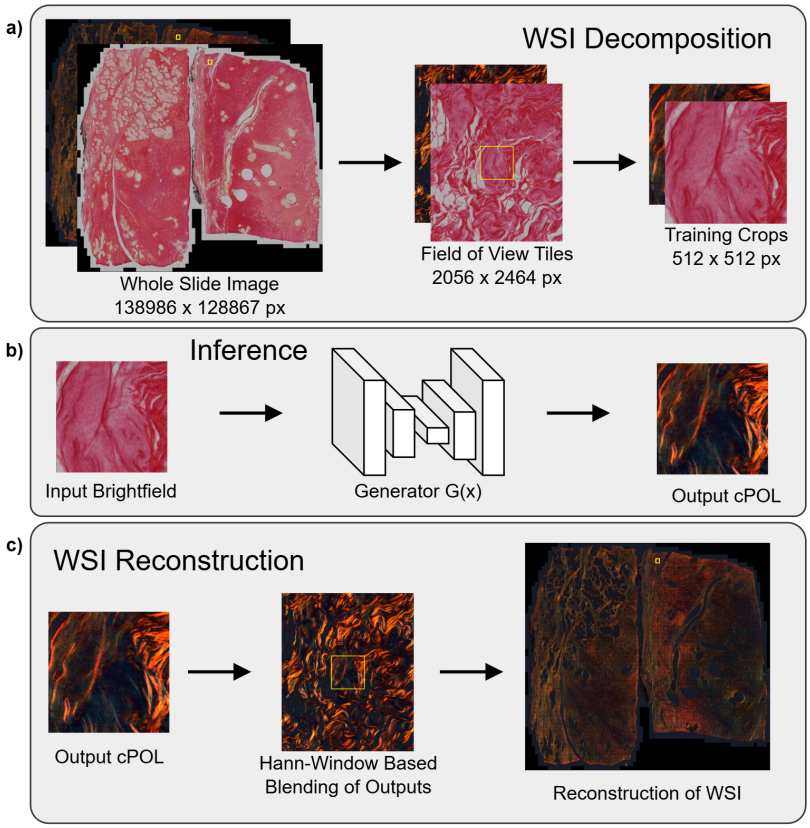



Figure 6.1: Overview of the proposed BF-to-cPOL translation pipeline. (a) Paired WSIs are decomposed into FOV tiles and $512 \times 512$ training patches. (b) A generator $G$ maps input BF patches to synthetic cPOL patches under MSE-SSIM and MSE-GAN training configurations. (c) Predicted patches are aggregated using Hann-window blending to reconstruct FOV and WSI-level outputs.

## 6.1 Stain Variability and Augmentation

Although stain normalization methods such as Macenko [26] and Reinhard [27] normalization are implemented and experimentally evaluated in this thesis, they are not included in the final preprocessing pipeline. These methods are designed to reduce color variability by mapping images toward a common stain or color distribution. Macenko normalization is developed for multi-stain images such as H&E, where the color distribution can be decomposed into two separable stain components using singular value decomposition in optical density space. PSR-stained images are dominated by a single red/pink collagen stain, i.e., PSR, meaning the second estimated stain direction does not correspond to a meaningful biological signal. The second stain direction may instead reflect illumination variation, noise, or local differences in non-collagen tissue components. Applying Macenko normalization in this setting therefore distorts the color distribution and reduces contrast in collagen-rich regions, producing washed-out images that suppress the collagen-related intensity needed to predict the cPOL signal.

In preliminary experiments, Reinhard-based tile-wise normalization introduced visible artifacts and local inconsistencies, this likely since normalization parameters are estimated independently per tile, causing the color transformation to vary between neighbouring regions of the same slide. Slide-level or lookup-table-based strategies present a different limitation: they assume that the color distribution of each tile is compatible with that estimated from the full WSI. This assumption does not hold for histological WSIs, where tiles can vary substantially in tissue composition, from dense collagen-rich, red regions to areas dominated by cells, adipose tissue, or background. Applying a fixed color transformation across this content risks suppressing tissue-specific intensity patterns that are informative for BF-to-cPOL translation.

Since the cPOL images were acquired under strict conditions using a consistent imaging protocol, their color and intensity characteristics are highly uniform across the dataset, making normalization of the target modality unnecessary. To expose the model to the mild stain and illumination variability present across PSR-BF acquisitions, light augmentation is applied to the input. The perturbation types and ranges were chosen to approximate the small variations in stain intensity and colour observed between slides, and were calibrated qualitatively so that augmented inputs remained visually consistent with plausible real acquisitions. The ranges were kept deliberately narrow: stronger perturbations produced inputs that no longer resembled genuine staining variation, and since the cPOL target is fixed, pushing the input toward such out-of-distribution appearances would only degrade the learned mapping rather than improve robustness.

Augmentation is applied to each training sample with probability $0.8$. When applied, the BF input is adjusted by a random brightness factor and a random contrast factor, each sampled uniformly from $[0.95, 1.05]$, after which each colour channel is independently scaled by a factor sampled uniformly from $[0.97, 1.03]$. Pixel values are finally clamped to $[0, 1]$.

## 6.2 Model Architectures

Most CNN-based approaches intended for biomedical image segmentation and image-to-image translation tasks are based on, or inspired by, the U-Net architecture proposed by Ronneberger et al. [8].

### 6.2.1 Keikhosravi et al.-Derived Residual U-Net

Keikhosravi et al. [6] demonstrated image translation from standard H&E-stained BF images to SHG-like collagen images. Since this modality pair is closely related to the BF-to-cPOL problem considered in this thesis, their architecture and training objective therefore formed the basis of the models implemented in this thesis.

The architecture follows the general U-Net encoder-decoder structure with skip connections, but replaces plain convolutional blocks with residual blocks, and uses strided convolutions for downsampling rather than separate max-pooling operations. Each encoder block contains a main branch consisting of a $3 \times 3$ convolution with stride 2 followed by a $1 \times 1$ convolution with stride 1, both followed by batch normalization and LeakyReLU activations. The main branch is then added to a projected shortcut. In the released implementation [28], used as the basis for this thesis, the first convolution in each encoder block uses a $4 \times 4$ kernel with stride 2 and padding 1 rather than the $3 \times 3$ kernel described in the paper.

In a plain convolutional block, the stacked layers are required to learn a full transformation $H(x)$ directly. With residual connections, the convolutional branch instead learns a residual correction. In the case where the shortcut branch is projected to match the output dimensions, the block output can be written as $H(x) = F(x) + P(x)$, where $P(x)$ denotes the projected shortcut branch and $F(x)$ represents the learned residual correction. This generally simplifies optimization, since convolutional layers can focus on learning the difference between the projected input representation and the desired output, rather than learning a full mapping. Residual connections also improve gradient flow, since the shortcut branch provides a more direct path through the network, reducing the risk that gradients vanish as they are propagated through many stacked convolutional layers.

The decoder path is also modified compared with the original U-Net. The decoder upsamples feature maps by interpolation before passing them through a residual-style convolutional block. Keikhosravi et al. describe this upsampling as bicubic interpolation; however, the released implementation uses bilinear interpolation. Since this difference is unlikely to meaningfully affect translation quality, the released implementation was retained without modification, as to stay consistent with the codebase.

The corresponding encoder features are not concatenated directly, but are first processed through separate skip-connection branches consisting of convolution, batch normalization, and LeakyReLU operations before being concatenated with the decoder features. The final upsampling step is replaced by a pixel shuffle layer, which rearranges feature map channels into spatial dimensions to reconstruct the full-resolution output.

Two adaptations were made to the released implementation. All convolutional blocks in the encoder, decoder, and skip connections were modified to use reflection padding instead of zero padding, avoiding tile-edge artifacts that arise from artificial border values in fine-detail images such as cPOL. The output layer was also adapted for three-channel RGB output rather than the original single-channel SHG output. The final convolution produces $4 \cdot C_{\text{out}}$ channels, which are rearranged by the pixel shuffle layer into a full-resolution output, with $C_{\text{out}} = 3$, this constructs a three-channel image.

A schematic overview of the architecture is shown in Figure 6.2, visualizing the encoder blocks, processed skip connections, decoder blocks, and final pixel shuffle reconstruction. The same generator architecture was used for the supervised MSE–

SSIM model and the adversarial model. The distinction between these configurations lies in the training objective rather than in the generator topology.

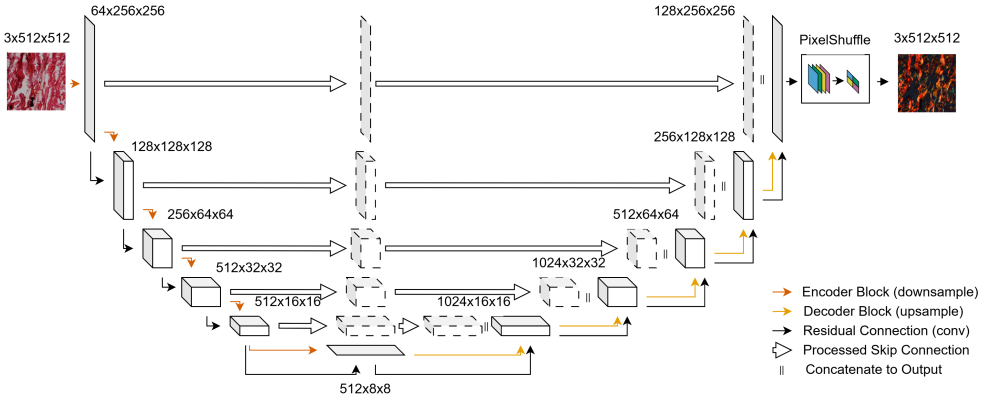



Figure 6.2: Schematic overview of the implemented residual U-Net architecture. The network consists of encoder blocks, decoder blocks, processed skip connections, and a final pixel shuffle layer that reconstructs the full-resolution three-channel output image.

### 6.2.2 Discriminator Architecture

The MSE-GAN model, presented in Section 6.4.2, uses an image-level discriminator that maps each input to a single scalar realism score for the whole image.

The network begins with a $3 \times 3$ convolution, mapping the three-channel input to 64 feature channels, followed by a LeakyReLU activation with negative slope $0.1$. This is followed by five discriminator blocks. Each block consists of two $3 \times 3$ convolutional layers with LeakyReLU activations, where the first convolution in each block uses stride 1 and the second uses stride 2, downsampling the feature map by a factor of two. The number of feature channels increases across blocks: 64, 128, 256, 512, 1024, and 2048.

After the convolutional feature extractor, adaptive average pooling reduces the feature map to a $2048 \times 1 \times 1$ global encoding of the input. The resulting vector is then passed through two fully connected layers: the first maps from 2048 to 2048 features with a LeakyReLU activation, while the second maps the representation to a single output value. A sigmoid activation is applied to produce a scalar score in the range $[0, 1]$, representing the discriminators estimate of whether the input is real or generated. A schematic overview of the discriminator architecture is presented in Figure 6.3.

## 6.3 Field-of-View and WSI Reconstruction

During test set evaluation, each FOV is divided into $512 \times 512$ patches with a stride of 256. The trained generator is applied independently to each patch. Since neighbouring patches overlap, multiple predictions contribute to the same pixel locations in the reconstructed FOV. These overlapping predictions are combined using a Hann-window weighting scheme, following Pielawski and Wählby [29], who show that window-based blending reduces edge effects in patch-based CNN prediction

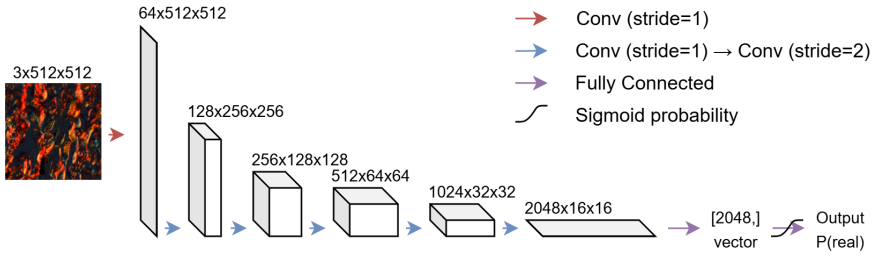



Figure 6.3: Schematic overview of the discriminator architecture, employed as training objective in the MSE-GAN implementation.

Figure 6.4 illustrates the reconstruction for a simplified overlapping prediction example. Each tile prediction is multiplied by the Hann window before being accumulated, the final reconstruction is obtained by dividing the weighted sum by the accumulated weights at each pixel. Compared with direct addition, this effectively eliminates sharp transitions between neighbouring patches by assigning lower weight to patch borders and higher weight to patch centres.

For a patch size of $512 \times 512$, a one-dimensional Hann window $w$ is first computed and then extended to two dimensions using an outer product:

$$W(i,j) = \max\big(w(i)\,w(j),\ \varepsilon\big), \qquad \varepsilon = 10^{-6}, \tag{6.1}$$

where $W$ is the two-dimensional weighting window. The window is subsequently normalized so that its maximum value is one. The floor $\varepsilon = 10^{-6}$ keeps all weights strictly positive, preventing division by zero in the per-pixel normalization during reconstruction.

For each predicted tile $\hat{y}_k$, located at position $(p_x, p_y)$ in the FOV canvas, the reconstruction canvas is updated by adding the weighted prediction:

$$C(i,j) \leftarrow C(i,j) + W_k(i,j)\hat{y}_k(i,j), \tag{6.2}$$

and a separate weight accumulation map is updated as

$$A(i,j) \leftarrow A(i,j) + W_k(i,j). \tag{6.3}$$

After all overlapping predictions for an FOV have been accumulated, the final reconstructed image is obtained by normalizing the weighted prediction sum by the accumulated weights:

$$\hat{Y}(i,j) = \frac{C(i,j)}{A(i,j)}. \tag{6.4}$$

Reconstruction is performed at the FOV level rather than directly on the full WSI, enabling both quantitative and qualitative evaluation at FOV level as well as in the context of the full tissue section. Each FOV contains bounding-box metadata indicating its position in the reconstructed WSI, preserving the spatial relationship between predicted FOVs and the original whole-slide coordinate system.

To reconstruct a full WSI from predicted FOVs, each FOV is placed onto a canvas at the position indicated by its bounding-box metadata. Since the full-resolution canvas would require tens of gigabytes of memory, it is never fully loaded into RAM. Instead, predictions are written directly to disk in a memory-mapped format, allowing individual FOV predictions to be inserted sequentially without holding the entire canvas in memory.

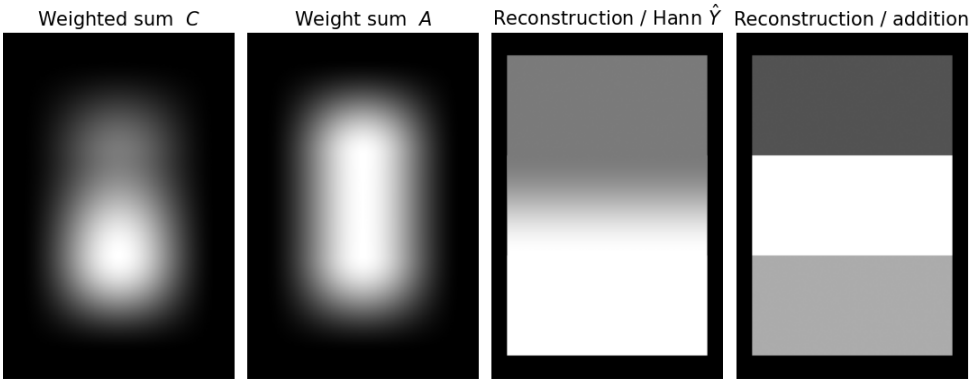



Figure 6.4: Illustration of Hann-window reconstruction for overlapping tile predictions, shown for a simplified example of three overlapping tiles. From left to right: the weighted-sum canvas $C = \sum_k W_k \hat{y}_k$, which accumulates each prediction $\hat{y}_k$ multiplied by its Hann window $W_k$; the weight-sum canvas $A = \sum_k W_k$, which accumulates the corresponding window weights; the final Hann reconstruction $\hat{Y} = C/A$; and, for comparison, the result of directly adding the overlapping predictions without window weighting. The Hann reconstruction $\hat{Y}$ varies smoothly across overlaps, whereas direct addition produces visible step artifacts at the tile boundaries.

## 6.4 Model Training

Let $x$ denote an input PSR-BF patch and $y$, the corresponding cPOL target. The objective is to optimize a generator network $G$ such that the predicted image $\hat{y} = G(x)$ approximates the paired target $y$.

All models were trained on $512 \times 512$ patch pairs extracted from two WSIs, with one WSI held out for validation. Training used mini-batch stochastic gradient descent with a batch size of 12, limited by available GPU memory given the $512 \times 512$ patch size and intermediate feature representations held during backpropagation. Augmentation was applied to the training set as described in Section 6.1; the validation set was kept deterministic. Validation was performed every 2000 global training steps and checkpoints were saved throughout.

The model configurations differ in their training objectives. Pixel-wise reconstruction losses provide direct supervision against the paired target, but do not necessarily capture structural or perceptual similarity. Additional loss terms are therefore introduced to encourage outputs that better match the structural appearance of real cPOL images. All training is conducted on an NVIDIA L40 GPU on the Uppmax high-performance computing cluster.

### 6.4.1 MSE-SSIM Training

The MSE-SSIM model uses the residual U-Net generator described in Section 6.2.1, trained with a combined pixel-wise and structural similarity objective following Keikhosravi et al. [6]. The training configuration was adapted from their original implementation: MSE is employed, rather than $\mathcal{L}_1$ and the loss weighting is kept fixed rather than ramped up during training, both of which were found to produce more stable results in our setting. The training objective is defined as

$$\rho \mathcal{L}_{\text{MSE}} + (1 - \rho)\mathcal{L}_{\text{SSIM}}, \tag{6.5}$$

where $\rho$ controls the relative weighting between pixel-wise reconstruction and structural similarity. Following Keikhosravi et al. [6], $\rho = 0.7$ is used, giving higher weight to the MSE term while still including SSIM as a structural similarity component. In this implementation, the weighting is kept fixed throughout training. The SSIM loss is implemented as

$$\mathcal{L}_{\text{SSIM}} = 1 - \text{SSIM}(G(x), y). \tag{6.6}$$

The model is optimized using Adam with a learning rate of $2 \cdot 10^{-4}$. Full validation is conducted every 2000 global training steps, and the best checkpoint for this configuration is selected according to the combined validation loss.

MSE is used as the reconstruction term since the task requires accurate recovery of continuous image intensities from a spatially aligned target. Compared to $\mathcal{L}_1$, MSE penalizes large pixel-wise deviations more strongly, encouraging the network to correct prominent intensity errors early in training. This is well suited to the paired supervised setting, where the objective is to approximate a specific aligned target rather than generate one of many possible outputs. In preliminary experiments, MSE produced more accurate reconstructions than $\mathcal{L}_1$ and was therefore retained.

The SSIM term is included to encourage similarity in local luminance, contrast, and structural statistics, following the original training objective of Keikhosravi et al. [6]. Its suitability for cPOL image translation is discussed in Section 5.2.3. Training and validation curves are shown in Figure 6.5.


Figure 6.5: Training and validation combined MSE-SSIM loss. The curves follow a similar trajectory to the MSE alone, stabilising around step 12,000 with minor fluctuations.

### 6.4.2 MSE-GAN Training

The MSE-GAN configuration uses the same residual U-Net generator (Section 6.2.1) but adds adversarial supervision through an unconditioned discriminator and removes the SSIM objective. As established in Chapter 5, pixel-wise objectives favour over-smoothed outputs, the adversarial term counteracts this. The discriminator receives either a real cPOL image $y$ or a generated image $G(x)$ and predicts which distribution it belongs to. Being unconditioned, it judges only whether the output resembles real cPOL images, not its structural consistency with the input.

The two losses play distinct roles: the MSE term enforces pixel-wise correspondence with the paired target, while the adversarial loss acts as a distribution-level regularizer that pushes outputs toward a realistic cPOL appearance without overriding the reconstruction. The unconditioned design is a deliberately conservative choice, keeping generated images grounded in the paired data, avoiding hallucinated collagen structures that would corrupt downstream analysis, at the cost of a weaker adversarial signal than a conditional discriminator.

The reconstruction term is defined as the MSE between the generated image $\hat{y} = G(x)$ and the paired target $y$. The discriminator is trained using a least-squares GAN objective:

$$\mathcal{L}_D = E_y\left[(1 - D(y))^2\right] + E_x\left[D(G(x))^2\right], \tag{6.7}$$

where $E_y$ and $E_x$ denote expectations over the target and input distributions respectively, approximated in practice by averaging over each mini-batch. This objective drives $D(y)$ towards 1 for real images and $D(G(x))$ towards 0 for generated ones.

The adversarial generator loss is defined using the discriminator score for the generated image:

$$\mathcal{L}_{\text{adv}}(G, D) = E_x\left[(1 - D(G(x)))^2\right]. \tag{6.8}$$

To reduce isolated pixel-level noise in the generated images, total variation regularization is included. This term penalizes large pixel-to-pixel intensity changes in the generated image:

$$\mathcal{L}_{\text{TV}}(G) = E_x\left[(G(x)_{i+1,j} - G(x)_{i,j})^2 + (G(x)_{i,j+1} - G(x)_{i,j})^2\right]. \tag{6.9}$$

where $G(x)_{i,j}$ denotes the generated pixel value at spatial location $(i, j)$, and the two terms sum the squared differences between vertically and horizontally adjacent pixels.

The final generator objective combines the three terms:

$$\mathcal{L}_G = \mathcal{L}_{\text{MSE}} + \lambda_{\text{TV}}\mathcal{L}_{\text{TV}} + \alpha_{\text{adv}}\mathcal{L}_{\text{adv}}, \tag{6.10}$$

where $\mathcal{L}_{\text{MSE}}$ is the pixel-wise reconstruction loss between $\hat{y} = G(x)$ and $y$, $\lambda_{\text{TV}}$ controls the strength of the total variation regularization, and $\alpha_{\text{adv}}$ controls the contribution of the adversarial loss. In the final implementation, $\lambda_{\text{TV}} = 10^{-5}$ and $\alpha_{\text{adv}} = 0.2$.

For each training batch, the discriminator is updated once, followed by two generator updates with the discriminator frozen. This asymmetric schedule was found to stabilise training in our setting, mitigating early discriminator dominance. Both networks are optimised with Adam at a learning rate of $1 \cdot 10^{-5}$. Validation is performed every 2000 steps, and checkpoints are selected by validation MSE rather than the adversarial loss, since MSE directly measures pixel-wise correspondence with the

paired target, whereas the adversarial loss reflects only the discriminator's current state rather than agreement with the target.

The training and validation MSE curves (Figure 6.6) show efficient learning over the first 18,000 steps before plateauing. The discriminator loss (Figure 6.7) remains stable at approximately $0.5$ throughout, indicating that the discriminator cannot reliably distinguish real from generated cPOL images, the expected operating point in adversarial training. The best checkpoint is selected at step 36,000.

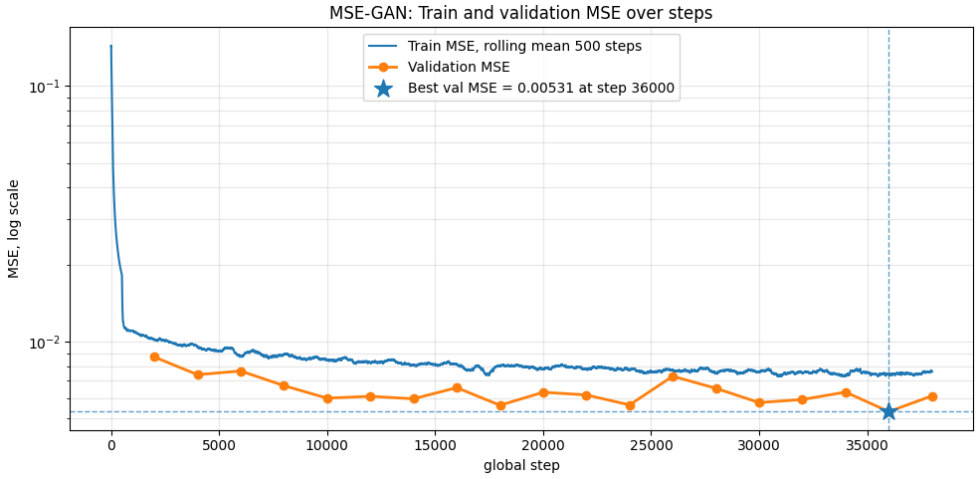



Figure 6.6: Training and validation MSE for the MSE-GAN configuration, declining rapidly before plateauing around step 18,000. Best checkpoint at step 36,000.

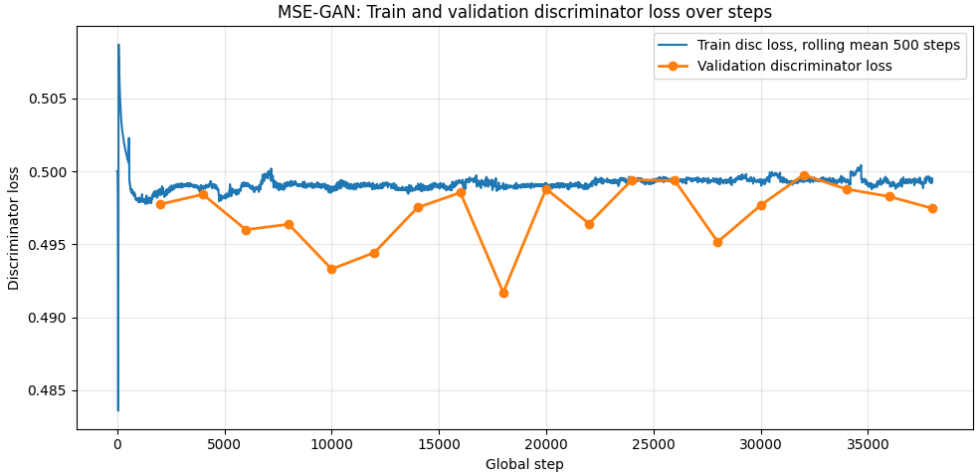



Figure 6.7: Discriminator loss for the MSE-GAN configuration, stable at approximately $0.5$ throughout training.

# Chapter 7

# Results

All results are computed on previously unseen held-out test data at full spatial resolution. Evaluation is conducted either at microscopy FOV level, on manually selected ROIs of $5000 \times 5000$ pixels, or at whole-slide level, as described in Chapter 5.

## 7.1 Quantitative Image Similarity Results

Quantitative image similarity metrics were computed on 2604 paired FOV tiles extracted from a held-out test WSI, not seen during training. Each FOV measured $2056 \times 2464$ pixels. With a pixel size of $0.173\mu$m, this corresponds to a physical area of approximately $426 \times 356\mu$m, or $0.152\text{mm}^2$. All tiles contained at least $10\%$ tissue.

Table 7.1: Quantitative image similarity results on 2604 paired FOV tiles from the held-out test WSI. Bold indicates best performance per metric.

| **Model** | **MSE** ↓ | **PSNR** ↑ | **FID** ↓ | **LPIPS** ↓ | | |
|---|---|---|---|---|---|---|
| | | | | Mean | Min | Max |
| Pix2Pix | 0.0030 | 28.9 | 23.36 | 0.182 | 0.076 | 0.547 |
| MSE-SSIM | **0.0020** | **31.2** | 28.26 | 0.256 | 0.161 | 0.618 |
| MSE-GAN | 0.0025 | 29.5 | **12.42** | **0.166** | **0.065** | **0.483** |

The quantitative results show that different metrics favour different model configurations. The MSE-SSIM model obtains the lowest MSE and highest PSNR, indicating the closest average pixel-wise agreement with the ground-truth cPOL images. The MSE-GAN obtains the lowest FID and LPIPS by a substantial margin, suggesting stronger distributional and perceptual similarity to real cPOL images.

Notably, the model with the best pixel-wise metrics performs worst on perceptual and distributional measures, reinforcing the case for complementary evaluation.

## 7.2 Collagen Feature Evaluation

To assess how well collagen fibers are preserved in generated cPOL images, the held-out WSI is reconstructed for each model and compared with the corresponding ground-truth using CTFIRE-based collagen feature analysis [23]

Twelve ROIs were manually selected by a domain expert with expertise in cancer biology and extracellular matrix research to provide representative regions for

collagen fiber analysis. Each ROI measures $5000 \times 5000$ pixels, corresponding to a physical side length of approximately $865\mu$m and an area of approximately $0.75\text{mm}^2$ at a pixel size of $0.173\mu$m.

The selected ROIs are presented in Figure 7.1, where each region is marked with a yellow bounding box. CTFIRE was applied to both generated and ground-truth cPOL images for each ROI, and the three measures were computed for each model. Table 7.2 summarises the results across all 12 ROIs.

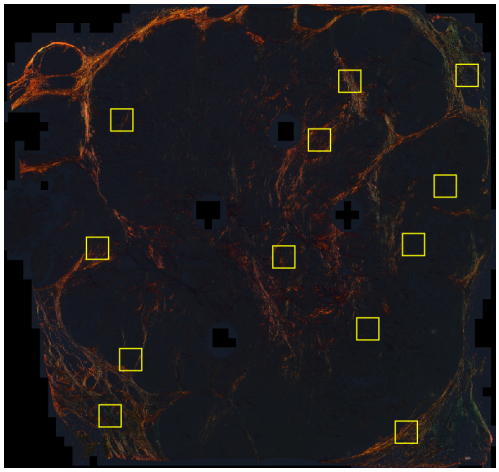


Figure 7.1: Held-out whole-slide image with the 12 selected regions of interest marked in yellow. These ROIs are used for collagen fiber analysis of ground-truth and generated cPOL images.

Table 7.2: CTFIRE-based collagen orientation agreement across 12 ROIs, reported as mean $\pm$ standard deviation. Lower values indicate closer agreement with the ground-truth cPOL image. Best values are shown in bold.

| **Model** | **Orientation error (°)** $\downarrow$ | **Alignment error** $\downarrow$ | **Histogram distance** $\downarrow$ |
|---|---|---|---|
| Pix2Pix | $\mathbf{1.93 \pm 1.42}$ | $0.041 \pm 0.029$ | $0.063 \pm 0.039$ |
| MSE–SSIM | $9.56 \pm 5.58$ | $0.105 \pm 0.069$ | $0.078 \pm 0.045$ |
| MSE-GAN | $2.23 \pm 1.23$ | $\mathbf{0.028 \pm 0.020}$ | $\mathbf{0.044 \pm 0.016}$ |

The MSE-GAN obtains the lowest alignment error and histogram distance, indicating the closest agreement with ground-truth fiber organisation and orientation distributions across the 12 ROIs. Pix2Pix achieves a slightly lower mean orientation error, though the difference is small and the MSE-GAN outperforms Pix2Pix on the remaining two measures. The MSE-SSIM model shows the largest deviation across all three measures. Together with the quantitative image similarity results, this indicates that the MSE-GAN provides the strongest overall preservation of collagen fiber structure among the evaluated models.

## 7.3 Color-Based Collagen Category Evaluation

Color-based collagen composition was evaluated on the full held-out test WSI using the method described in Section 5.2.7. The results are summarised in Table 7.3.

Table 7.3: Color-based collagen composition for the ground-truth cPOL image and generated model outputs. Composition is reported as the percentage of classified collagen area assigned to each color category. Deltas are computed relative to the ground truth in percentage points (pp). Best values shown in bold.

| | **Collagen composition** (%) | | | **Δ vs. Ground Truth** (pp) | | |
|---|---|---|---|---|---|---|
| **Model** | Green | Yellow | Red-orange | Green | Yellow | Red-orange |
| GT | 38.76 | 7.44 | 53.81 | – | – | – |
| Pix2Pix | 36.16 | 9.21 | 54.64 | -2.60 | +1.77 | +0.83 |
| MSE-SSIM | 15.80 | 3.35 | 80.85 | -22.96 | -4.09 | +27.04 |
| MSE-GAN | 40.55 | 6.41 | 53.03 | **+1.80** | **-1.02** | **-0.77** |

The MSE-GAN output most closely matches the ground-truth collagen composition across all three color categories, differing by only $+1.80$ percentage points for green, $-1.02$ percentage points for yellow, and $-0.77$ percentage points for red-orange collagen. Pix2Pix also shows relatively close agreement, with larger deviations for green and yellow.

The MSE-SSIM model exhibits a very different pattern, with substantially lower green and yellow proportions and a much higher red-orange proportion than the ground truth, red-orange collagen is overestimated by $+27.04$ percentage points. This indicates that the predictions are strongly skewed towards red birefringent collagen. The adversarial models produce a considerably more realistic color-category distribution, with the MSE-GAN most closely matching the ground truth.

## 7.4 Qualitative Assessment of Generated cPOL Images

Differences between the model outputs emerge only at finer pyramidal levels, where high-frequency detail must be preserved. At the coarsest scale (Figure 7.2), all three models match the ground truth with no noticeable deviation.

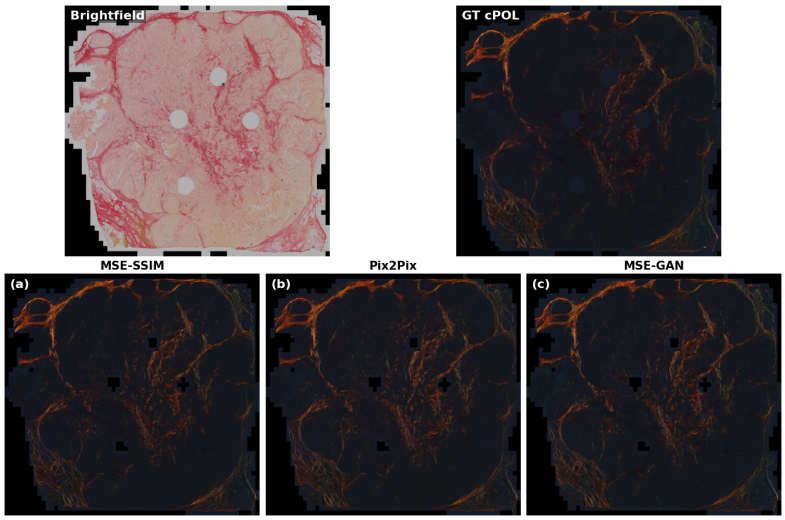



Figure 7.2: Top row: brightfield input (left) and ground-truth cPOL (right) for the held-out test WSI. Bottom row: outputs from (a) MSE-SSIM, (b) Pix2Pix, and (c) MSE-GAN. Shown at the lowest-resolution pyramidal level.

Rendered at full resolution, a randomly sampled FOV tile, along with a zoomed-in $700 \times 700$ pixel region of the tile, is presented for each implementation in Figures 7.3, 7.4, and 7.5. For each figure, the top row presents the ground truth FOV, and the bottom row the prediction for the corresponding model.

The MSE-SSIM implementation, Figure 7.3, produces an output visually similar to the ground truth at the FOV level, although upon zooming in, the prediction is clearly blurred. This blurring introduces a red-dominant appearance in regions of mixed fiber orientation, as the finer yellow partitions of the collagen signal are dampened. Implementations incorporating adversarial losses retain sharper birefringence gradients. The Pix2Pix output, Figure 7.4, accurately captures fiber edges but exhibits a checkerboard artefact pattern. The MSE-GAN model, Figure 7.5, most closely matches the ground truth overall, though green birefringent regions appear slightly elevated in intensity compared to the reference image.

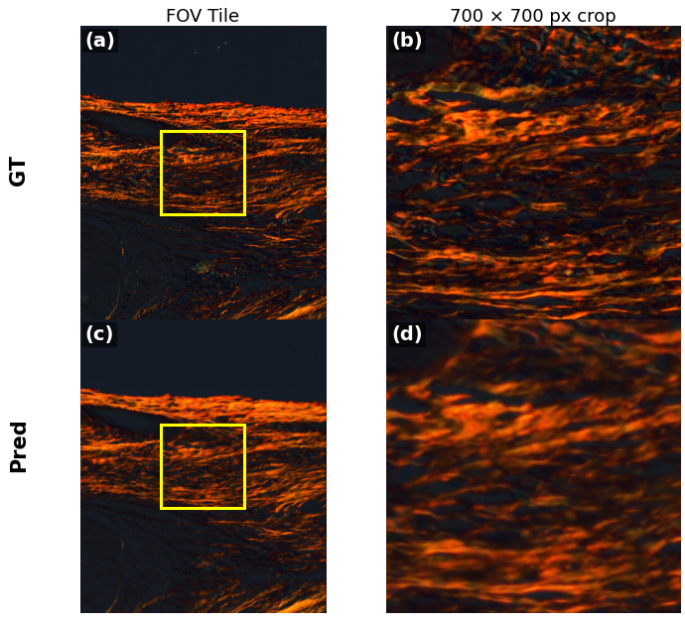



Figure 7.3: Ground truth FOV tile (top row) and MSE-SSIM output (bottom row). Left column: randomly sampled tile with the zoomed region marked by a yellow bounding box. Right column: $700 \times 700$ pixel crop of the indicated region. The zoomed in view enables for visualization of the missing detail in predicted output.

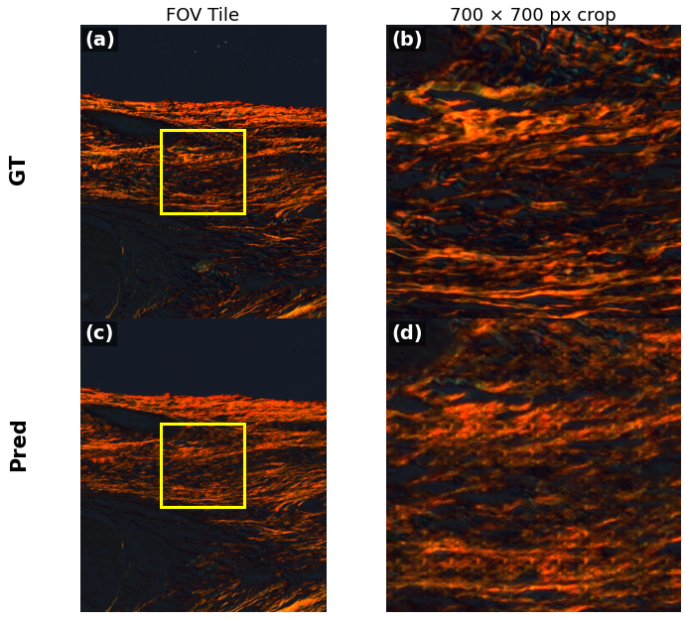



Figure 7.4: Ground truth FOV tile (top row) and Pix2Pix output (bottom row). Left column: randomly sampled tile with the zoomed region marked by a yellow bounding box. Right column: $700 \times 700$ pixel crop of the indicated region. A clear checkerboard pattern is visible in the zoomed in view.

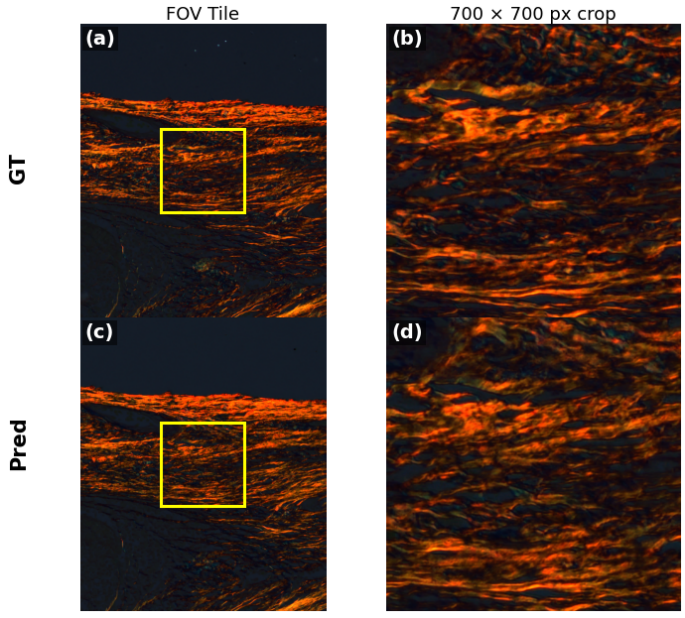



Figure 7.5: Ground truth FOV tile (top row) and MSE-GAN output (bottom row). Left column: randomly sampled tile with the zoomed region marked by a yellow bounding box. Right column: $700 \times 700$ pixel crop of the indicated region. The images match closely, although green birefringent collagen appears more dominant in the predicted image.

In a more comparative view, Figure 7.6 presents five randomly sampled $500 \times 500$ crops, one per column, with each row corresponding to a model or ground truth

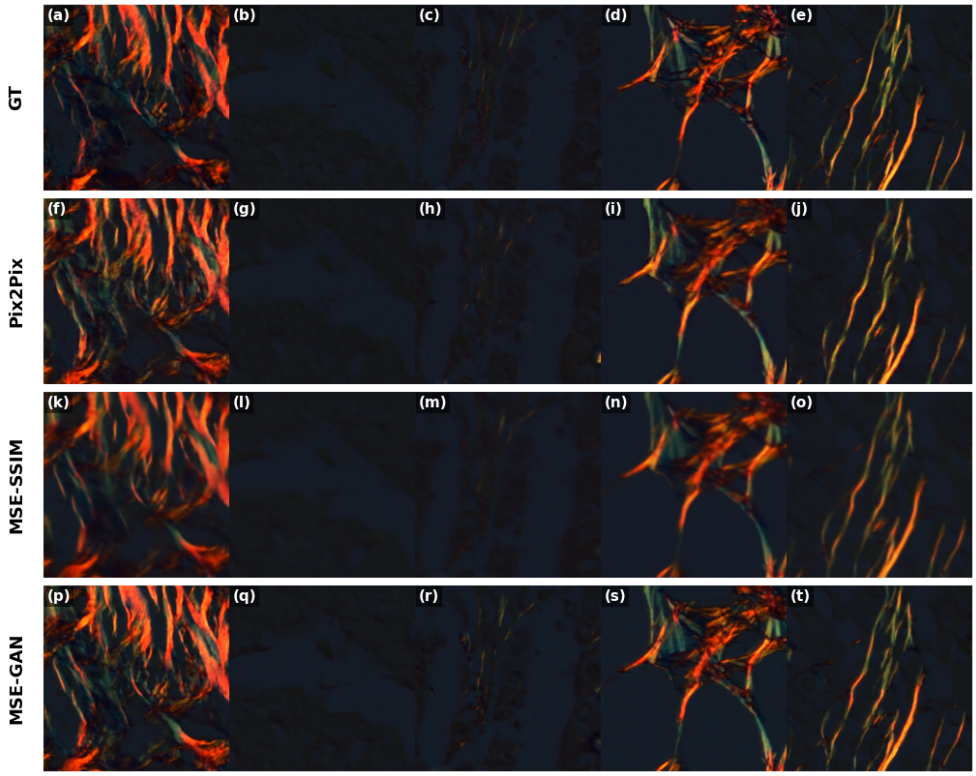



Figure 7.6: Five $500 \times 500$ pixel crops from five randomly sampled FOV tiles. Each column corresponds to a single crop location. Each row corresponds to a model or reference. (a-e) Ground truth cPOL, (f-j) Pix2Pix, (k-o) MSE-SSIM, (p-t) MSE-GAN.

reference. Pix2Pix and MSE-SSIM appears less sharp, matching the ground truth less closely than MSE-GAN. This is most pronounced for MSE-SSIM implementation, which exhibits a consistent blurring artefact across all crops.

# Chapter 8

# Discussion

Evaluating generative models is inherently more challenging than evaluating discriminative ones. A precision or F1 score computed against discrete labels directly reflects task performance, but in image-to-image translation, different measures operate at different levels of abstraction and can disagree substantially. The results presented in this thesis make this concrete, as the three models rank differently depending on which metric is used, and the model with the best pixel-level scores produces the worst biological outputs.

The MSE-SSIM model's strong pixel-wise performance is a direct consequence of its training objective. By minimizing average squared error, the model is rewarded for predicting smooth, spatially averaged outputs that reduce per-pixel deviation at the cost of suppressing local structure. In practice, this collapses the output toward red-dominant birefringence: fine yellow and green signals encoding distinct collagen populations are suppressed, while the broad red signal is preserved and amplified. The color-category evaluation makes this concrete, with red-orange collagen overestimated by $+27.04$ percentage points.

The MSE-GAN avoids this by complementing pixel-wise supervision with an adversarial regularization, outputs not resembling cPOL images are penalized at distributional level. This pushes the generator away from the blurring local minimum and toward outputs that preserve the sharp birefringence gradients characteristic of real cPOL. The result is a model that performs worst on MSE and PSNR, yet most faithfully reproduces the collagen structure that matters for downstream analysis.

The domain-specific evaluation confirms that the BF input contains sufficient structural information for the learned mapping to recover broad fiber organisation: all three models preserve primary fiber orientation to a reasonable degree. The models do diverge in how faithfully they reproduce finer aspects of fiber alignment and birefringent color composition, where the adversarial objective provides a clear advantage. That highly aligned collagen regions are associated with stromal remodelling linked to tumor progression [4–6] essentially makes accurate reproduction of alignment a prerequisite for the generated images to be useful in practice.

The inversion between pixel-level and perceptual evaluation observed here reflects a fundamental property of MSE as a loss function. A blurred prediction that spreads intensity across neighbouring pixels incurs lower squared error per pixel than a sharp prediction that preserves structure but displaces it by even a single pixel. This well-known local minimum results in that any model trained primarily with MSE will tend toward over-smoothed outputs, and any evaluation that relies primarily on MSE or its derivatives (PSNR, SSIM) will systematically favour such

models. The practical implication is that pixel-based metrics are not only incomplete for this task, they are actively misleading.

## 8.1 Limitations

The dataset consists of four WSI pairs acquired under controlled conditions at a single imaging site, limiting the diversity of tissue appearance, staining variation, and collagen organisation presented during training. While each WSI produces a large number of patches, these cannot be treated as independent samples, as they originate from neighbouring regions of the same tissue section. The generalisability of the learned mapping to tissue from different patients, staining protocols, or imaging systems therefore remains untested.

A deeper limitation inherent to the generative modelling approach itself: The models have learned a mapping from observed inputs to known outputs, and when presented with tissue patterns not well represented in the training data, they can only extrapolate from learned relationships. Generated images should therefore be interpreted as conditioned predictions rather than direct measurements of birefringence, a distinction that is particularly important in medical and research contexts where outputs may be used to inform biological interpretation.

Finally, the evaluation does not include expert pathological review, which could provide insight into whether generated images preserve biologically meaningful morphology beyond what the quantitative measures capture.

# Chapter 9

# Conclusions

This thesis demonstrated that CNN-based image-to-image translation can synthesize picrosirius red (PSR) stained circularly polarized light (cPOL) images from routinely available PSR-stained brightfield (BF) WSIs, and that the generated images can support downstream collagen fiber analysis with results comparable to those obtained from real cPOL acquisitions.

Of the models evaluated, the MSE-GAN performed substantially better across nearly all dimensions of evaluation. It achieved the lowest FID and LPIPS scores by a wide margin, best reproduced fiber alignment and orientation distributions across 12 expert-selected ROIs, and matched the ground-truth birefringent color composition to within two percentage points across all three collagen categories. The MSE-SSIM model, despite achieving the best pixel-wise scores, failed the collagen-specific evaluation by overestimating red-orange collagen by over 27 percentage points and producing outputs that suppress the fine birefringence structure necessary for fiber analysis.

A central finding is that pixel-wise metrics are not only insufficient but actively misleading for evaluating image-to-image translation in the current domain. The evaluation framework introduced here, combining perceptual and distributional metrics with CTFIRE-based fiber analysis and color-based collagen segmentation, provides a meaningful basis for model selection and transfers naturally to related cross-modality translation tasks in computational pathology. Code for the models and evaluation pipeline is publicly available on GitHub.

## 9.1 Future Work

The most obvious extension is a larger and more diverse training dataset spanning multiple patients, tissue types, staining conditions, and imaging systems. The current four-WSI dataset limits generalisability, and broader validation across acquisition sites and patient cohorts would be required before the approach could be applied in a retrospective research or clinical setting.

A conditional discriminator, receiving both the generated image and the corresponding BF input, would provide a stronger adversarial signal than the unconditioned design used here, explicitly penalizing structural inconsistency between the input and the generated output, though at the cost of a less interpretable training objective. A comprehensive ablation study, evaluating the effects of the augmentation strategy and the chosen hyperparameters, would further strengthen the findings.

Expert pathological review of generated image quality would strengthen the

evaluation beyond what quantitative measures alone can capture, and is a natural next step toward validating the method for exploratory use in collagen fiber analysis.

# Code Availability

Code for preprocessing, model inference, and evaluation is available at https://github.com/joelsundin/bf2cpol/. The repository includes the residual U-Net and discriminator implementations, model checkpoints, the field-of-view tiling and Hann-window reconstruction pipeline, and the collagen-specific evaluation code. The repository is under active development.

# Use of Generative AI

Generative AI tools (ChatGPT) were used during the preparation of this thesis. The tools were used primarily in place of a conventional search engine. Providing descriptions of technical concepts, library documentation, and background material. The generative tools were also used for assistance with language, and LaTeX formatting.

All research questions, experimental design, implementation, analysis, and conclusions presented in this work are the author's own. Generative AI was not used to generate results or data, or to formulate the scientific content of the thesis. Any text suggested by such tools was reviewed, edited, and verified by the author, who takes full responsibility for the final content of this work.

# Bibliography


[1] Emily A. Ricke, Karin Williams, Yi-Fen Lee, Suzana Couto, Yuzhuo Wang, Simon W. Hayward, Gerald R. Cunha, and William A. Ricke. Androgen hormone action in prostatic carcinogenesis: stromal androgen receptors mediate prostate cancer progression, malignant transformation and metastasis. *Carcinogenesis*, 33(7):1391–1398, July 2012. ISSN 1460-2180. doi: 10.1093/carcin/bgs153.

[2] Gerald R. Cunha, Will Ricke, Axel Thomson, Paul C. Marker, Gail Risbridger, Simon W. Hayward, Y. Z. Wang, Annemarie A. Donjacour, and Takeshi Kurita. Hormonal, cellular, and molecular regulation of normal and neoplastic prostatic development. *The Journal of Steroid Biochemistry and Molecular Biology*, 92(4): 221–236, November 2004. ISSN 0960-0760. doi: 10.1016/j.jsbmb.2004.10.017.

[3] Zizhao Mai, Yunfan Lin, Pei Lin, Xinyuan Zhao, and Li Cui. Modulating extracellular matrix stiffness: a strategic approach to boost cancer immunotherapy. *Cell Death & Disease*, 15(5):307, May 2024. ISSN 2041-4889. doi: 10.1038/s41419-024-06697-4. URL https://www.nature.com/articles/s41419-024-06697-4.

[4] Mikala Egeblad, Morten G Rasch, and Valerie M Weaver. Dynamic interplay between the collagen scaffold and tumor evolution. *Current Opinion in Cell Biology*, 22(5):697–706, October 2010. ISSN 0955-0674. doi: 10.1016/j.ceb.2010.08.015. URL https://www.sciencedirect.com/science/article/pii/S0955067410001389.

[5] Mitsuo Yamauchi, Thomas H. Barker, Don L. Gibbons, and Jonathan M. Kurie. The fibrotic tumor stroma. *The Journal of Clinical Investigation*, 128(1):16–25, January 2018. ISSN 0021-9738. doi: 10.1172/JCI93554. URL https://www.jci.org/articles/view/93554.

[6] Adib Keikhosravi, Bin Li, Yuming Liu, Matthew W. Conklin, Agnes G. Loeffler, and Kevin W. Eliceiri. Non-disruptive collagen characterization in clinical histopathology using cross-modality image synthesis. *Communications Biology*, 3(1):414, July 2020. ISSN 2399-3642. doi: 10.1038/s42003-020-01151-5. URL https://www.nature.com/articles/s42003-020-01151-5.

[7] Laure Rittié. Method for Picrosirius Red-Polarization Detection of Collagen Fibers in Tissue Sections. In Laure Rittié, editor, *Fibrosis*, volume 1627, pages 395–407. Springer New York, New York, NY, 2017. ISBN 978-1-4939-7112-1 978-1-4939-7113-8. doi: 10.1007/978-1-4939-7113-8_26. URL http://link.springer.com/10.1007/978-1-4939-7113-8_26. Series Title: Methods in Molecular Biology.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. URL http://arxiv.org/abs/1505.04597. arXiv:1505.04597 [cs].

[9] Yair Rivenson, Hongda Wang, Zhensong Wei, Kevin de Haan, Yibo Zhang, Yichen Wu, Harun Günaydın, Jonathan E. Zuckerman, Thomas Chong, Anthony E. Sisk, Lindsey M. Westbrook, W. Dean Wallace, and Aydogan Ozcan. Virtual histological staining of unlabelled tissue-autofluorescence images via deep learning. *Nature Biomedical Engineering*, 3(6):466–477, June 2019. ISSN 2157-846X. doi: 10.1038/s41551-019-0362-y. URL https://www.nature.com/articles/s41551-019-0362-y.

[10] Shuting Liu, Baochang Zhang, Yiqing Liu, Anjia Han, Huijuan Shi, Tian Guan, and Yonghong He. Unpaired Stain Transfer Using Pathology-Consistent Constrained Generative Adversarial Networks. *IEEE Transactions on Medical Imaging*, 40(8):1977–1989, August 2021. ISSN 1558-254X. doi: 10.1109/TMI.2021.3069874. URL https://ieeexplore.ieee.org/document/9389763/.

[11] Shikha Dubey, Yosep Chong, Beatrice Knudsen, and Shireen Y. Elhabian. VIMs: Virtual Immunohistochemistry Multiplex staining via Text-to-Stain Diffusion Trained on Uniplex Stains, July 2024. URL http://arxiv.org/abs/2407.19113. arXiv:2407.19113 [eess].

[12] CurveAlign – Laboratory for Optical and Computational Instrumentation – UW–Madison. URL https://loci.wisc.edu/curvealign/.

[13] Kevin de Haan, Yijie Zhang, Jonathan E. Zuckerman, Tairan Liu, Anthony E. Sisk, Miguel F. P. Diaz, Kuang-Yu Jen, Alexander Nobori, Sofia Liou, Sarah Zhang, Rana Riahi, Yair Rivenson, W. Dean Wallace, and Aydogan Ozcan. Deep learning-based transformation of H&E stained tissues into special stains. *Nature Communications*, 12(1):4884, August 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25221-2. URL https://www.nature.com/articles/s41467-021-25221-2.

[14] Yijie Zhang, Luzhe Huang, Nir Pillar, Yuzhu Li, Hanlong Chen, and Aydogan Ozcan. Pixel super-resolved virtual staining of label-free tissue using diffusion models. *Nature Communications*, 16(1):5016, May 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-60387-z. URL https://www.nature.com/articles/s41467-025-60387-z.

[15] Srikar Yellapragada, Alexandros Graikos, Zilinghan Li, Kostas Triaridis, Varun Belagali, Tarak Nath Nandi, Karen Bai, Beatrice S. Knudsen, Tahsin Kurc, Rajarsi R. Gupta, Prateek Prasanna, Ravi K. Madduri, Joel Saltz, and Dimitris Samaras. PixCell: A generative foundation model for digital histopathology images, December 2025. URL http://arxiv.org/abs/2506.05127. arXiv:2506.05127 [eess].

[16] Fangda Li, Zhiqiang Hu, Wen Chen, and Avinash Kak. Adaptive Supervised PatchNCE Loss for Learning H&E-to-IHC Stain Translation with Inconsistent Groundtruth Image Pairs, March 2023. URL http://arxiv.org/abs/2303.06193. arXiv:2303.06193 [cs].

[17] Bytez.com, Kexin Sun, Zhineng Chen, Gongwei Wang, Jun Liu, Xiongjun Ye, and Yu-Gang Jiang. Bi-Directional Feature Fusion Generative Adversarial Net..., June 2023. URL https://bytez.com/docs/cvpr/22488/paper.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks, November 2018. URL http://arxiv.org/abs/1611.07004. arXiv:1611.07004 [cs.CV].

[19] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00068. URL https://ieeexplore.ieee.org/document/8578166/.

[20] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1941-0042. doi: 10.1109/TIP.2003.819861. URL https://ieeexplore.ieee.org/document/1284395/.

[21] Meharban M S, Sabu M K, and T Santhanakrishnan. T1W MRI to T2W MRI Image Synthesis Using SSIM-CycleGAN. In *2024 11th International Conference on Advances in Computing and Communications (ICACC)*, pages 1–6, November 2024. doi: 10.1109/ICACC63692.2024.10845374. URL https://ieeexplore.ieee.org/document/10845374.

[22] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11400–11410, New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01112. URL https://ieeexplore.ieee.org/document/9880182/.

[23] Jeremy S. Bredfeldt, Yuming Liu, Carolyn A. Pehlke, Matthew W. Conklin, Joseph M. Szulczewski, David R. Inman, Patricia J. Keely, Robert D. Nowak, Thomas R. Mackie, and Kevin W. Eliceiri. Computational segmentation of collagen fibers from second-harmonic generation images of breast cancer. *Journal of Biomedical Optics*, 19(1):016007, January 2014. ISSN 1083-3668. doi: 10.1117/1.JBO.19.1.016007. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3886580/.

[24] Sanja Z. Despotović, Đorđe N. Milićević, Aleksandar J. Krmpot, Aleksandra M. Pavlović, Vladimir D. Živanović, Zoran Krivokapić, Vladimir B. Pavlović, Steva Lević, Gorana Nikolić, and Mihailo D. Rabasović. Altered organization of collagen fibers in the uninvolved human colon mucosa 10 cm and 20 cm away from the malignant tumor. *Scientific Reports*, 10(1):6359, April 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-63368-y. URL https://www.nature.com/articles/s41598-020-63368-y.

[25] Peter Bankhead, Maurice B. Loughrey, José A. Fernández, Yvonne Dombrowski, Darragh G. McArt, Philip D. Dunne, Stephen McQuaid, Ronan T. Gray, Liam J. Murray, Helen G. Coleman, Jacqueline A. James, Manuel Salto-Tellez, and Peter W. Hamilton. QuPath: Open source software for digital pathology image analysis. *Scientific Reports*, 7:16878, December 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-17204-5. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5715110/.

[26] Marc Macenko, Marc Niethammer, J. S. Marron, David Borland, John T. Woosley, Xiaojun Guan, Charles Schmitt, and Nancy E. Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, Boston, MA, USA, June 2009. IEEE. ISBN 978-1-4244-3931-7. doi: 10.1109/ISBI.2009.5193250. URL http://ieeexplore.ieee.org/document/5193250/.

[27] E Reinhard, M Ashikhmin, B Gooch, and P Shirley. Color transfer between images. *IEEE Computer Graphics and Applications*, 21 (5):34–41, September 2001. ISSN 1558-1756. doi: 10.1109/38.946629.

[28] uw-loci. GitHub - uw-loci/he_shg_synth_workflow at v1.0.0. URL https://github.com/uw-loci/he_shg_synth_workflow/tree/v1.0.0.

[29] Nicolas Pielawski and Carolina Wählby. Introducing Hann windows for reducing edge-effects in patch-based image segmentation. *PLOS ONE*, 15(3): e0229839, March 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0229839. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0229839.