Supplementary materials: Three-dimensional H&E histopathology powered by deep learning-assisted multimodal nonlinear microscopy

Selection of appropriate nonlinear microscopy channels for image-to-image translation

To optimize the network performance, it is important to determine which multimodal nonlinear microscopy channels are the most effective for generating realistic brightfield-equivalent images.

For the initial training step we chose a lightweight version of the U-Net considering individually the THG (Supplementary Figure 1a) and MPF (Supplementary Figure 1c) channels. THG produces strong signals at interfaces and regions with refractive index discontinuities, making it effective for highlighting nuclei and erythrocytes, whereas MPF imaging captures detailed representations of eosinophilic structures like ECM proteins and erythrocytes. The SHG modality was excluded from this step, as it predominantly highlights ordered collagenous structures with high specificity.

Initially, we spatially aligned the input grayscale THG and MPF images with the corresponding brightfield images (Supplementary Figure 1g) obtained from a conventional scanner. This alignment is crucial to enable the model to learn precise mappings between the nonlinear microscopy inputs and the corresponding target brightfield images.

For training and testing, the input and target images were subdivided into smaller patches of 128 by 128 pixels to alleviate the computational load. The model was trained in a feedforward manner using the mean squared error (MSE) loss function:

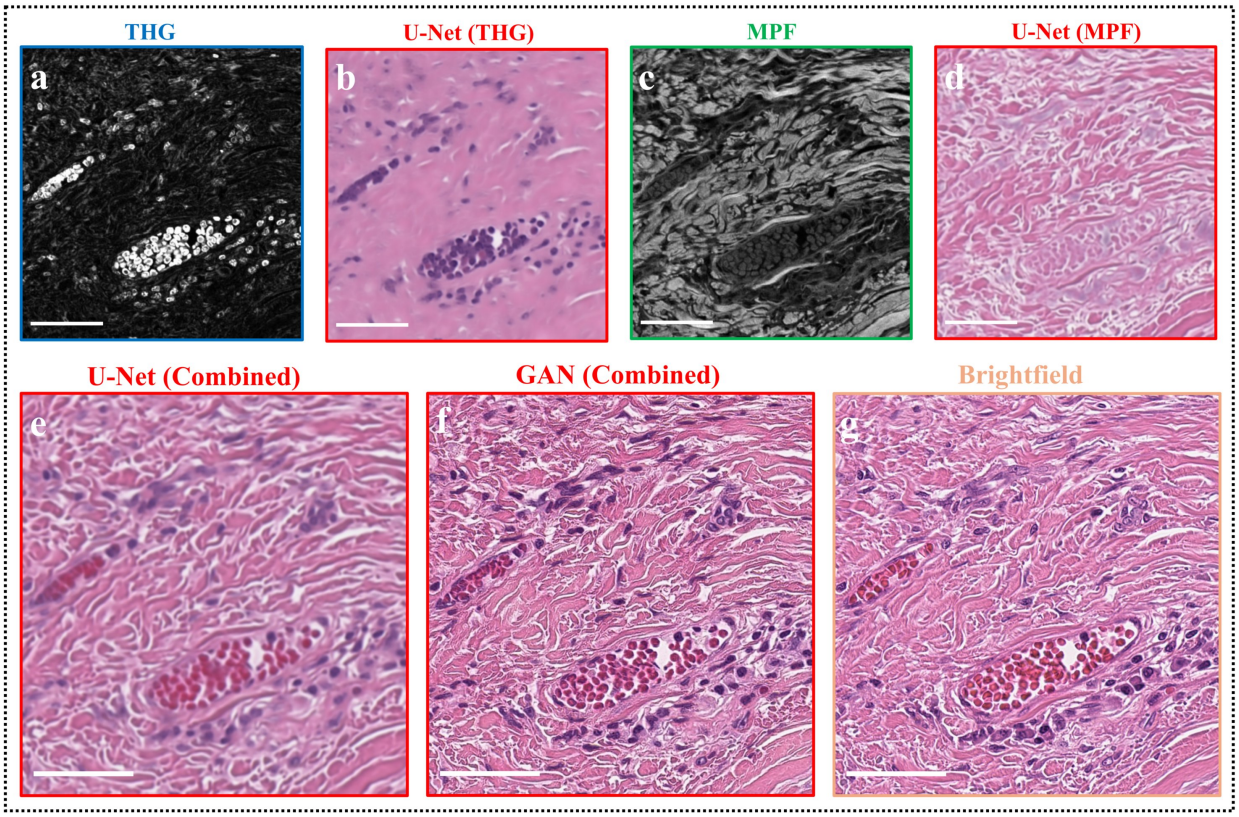
$$\mathscr{L}_{\mathrm{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(p_i - \hat{p}_i)^2, \tag{1}$$

where p_i is the target pixel value, $\hat{p}_i$ is the predicted pixel value, and n is the total number of pixels across all channels.

For the testing phase, in addition to subdivision, the input patches were sampled with a 75% overlap and full-scale generated images were obtained by combining the smaller generated patches using the Gaussian weighting scheme[1], described in the Methods section in more details.

We first conducted experiments using each channel (either THG or MPF) individually to assess their effectiveness in the image-to-image translation task. Using the THG channel as input (Supplementary Figure 1a), the generated images (Supplementary Figure 1b) revealed detailed nuclear features, due to the strong signals from cell nuclei. However, the ECM lacked fine details, and the erythrocytes were indistinguishable from the nuclei — they were both mapped as blue-colored structures, resembling nuclei in H&E staining. This similarity occurred because both erythrocytes and nuclei produce intense THG signals, and the model could not differentiate between them based on the THG intensity and the contextual information provided by a single channel. On the other hand, utilizing only the MPF channel (Supplementary Figure 1c) resulted in images (Supplementary Figure 1d) with detailed ECM representation. However, for MPF the nuclei were absent in the reconstructed images since the nuclear regions emit little to no fluorescence signal due to hematoxylin-mediated fluorescence quenching[2]. Additionally, erythrocytes appeared similar in color to the ECM but exhibited a lighter shade, owing to weaker MPF signals originating from erythrocytes compared to the ECM.

Since a single channel is not sufficient to reproduce accurately H&E images, both THG and MPF channels were combined for image translation. The combined two-channel input utilizes the complementary strengths of both modalities and generates images that closely resemble standard H&E-stained brightfield images (Supplementary Figure 1e). The ECM exhibits fine details, nuclei are clearly visible and correctly colored, and erythrocytes are assigned the appropriate rich red color.



Supplementary Figure 1. H&E brightfield image generation from the THG **(a)** and MPF **(c)** contrasts employing different neural networks. U-Net was used to generate THG **(b)** and MPF **(d)** images. Among the U-Net architectures, the performance was maximized by using both THG and MPF inputs **(e)**, but the generated images were blurry, as expected. The highest overall quality was achieved by training the U-Net model in an adversarial manner using the GAN framework **(f)**, which resulted in generated images most similar to the target brightfield images **(g)**. Scale bar: 50 µm.

U-Net performance optimization

We proceeded with the following network optimization steps to enhance the U-Net image-to-image translation performance. Since the literature suggests that alternative loss functions based on the structural similarity index measure (SSIM) and the mean absolute error, also known as the Manhattan distance (L1), can yield better reconstruction results[3], we compared the performance of the lightweight U-Net, trained using $\mathscr{L}_{\text{MSE}}$ and a new version of the same network trained using a combined loss function that involved a weighted sum of $\mathscr{L}_{\text{SSIM}}$ and $\mathscr{L}_{\text{L1}}$ losses. In the main text, this loss is referred to as content loss $\mathscr{L}_{\text{cont}}$.

To quantitatively assess the performance of the trained networks, besides MSE, we used several other evaluation metrics:

- **mean absolute error (L1):**

$$\text{L1} = \frac{1}{n}\sum_{i=1}^{n}|p_i - \hat{p}_i|. \tag{2}$$

- **peak signal-to-noise ratio (PSNR):**

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{\text{MSE}}\right), \tag{3}$$

where MAX_I is the maximum possible pixel value of the images, which in our case was equal to 1, since the networks were designed to operate on normalized images.

- **structural similarity index measure (SSIM):**

$$\mathrm{SSIM}(A,B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{AB} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_A^2 + \sigma_B^2 + C_2)}, \tag{4}$$

where μ_A and μ_B are the mean pixel values of images A and B, σ_A^2 and σ_B^2 are the variances, σ_{AB} is the covariance, and C_1 and C_2 are stabilization constants[4].

This definition holds for grayscale images. For colored images SSIM is calculated for each channel separately and averaged:

$$\mathrm{SSIM}_{\mathrm{RGB}} = \frac{\mathrm{SSIM}_{\mathrm{R}} + \mathrm{SSIM}_{\mathrm{G}} + \mathrm{SSIM}_{\mathrm{B}}}{3}, \tag{5}$$

where $\mathrm{SSIM}_{\mathrm{R}}$, $\mathrm{SSIM}_{\mathrm{G}}$ and $\mathrm{SSIM}_{\mathrm{B}}$ are structural similarity index measures for red, green and blue channels, respectively.

- **gradient-SSIM (GSSIM):**

Similar to SSIM but applied to image gradients to assess structural similarity in edge information[5]:

$$\mathrm{GSSIM}(A,B) = \frac{(2\mu_A\mu_B + C_1)(2\sigma_{G_AG_B} + C_2)}{(\mu_A^2 + \mu_B^2 + C_1)(\sigma_{G_A}^2 + \sigma_{G_B}^2 + C_2)}, \tag{6}$$

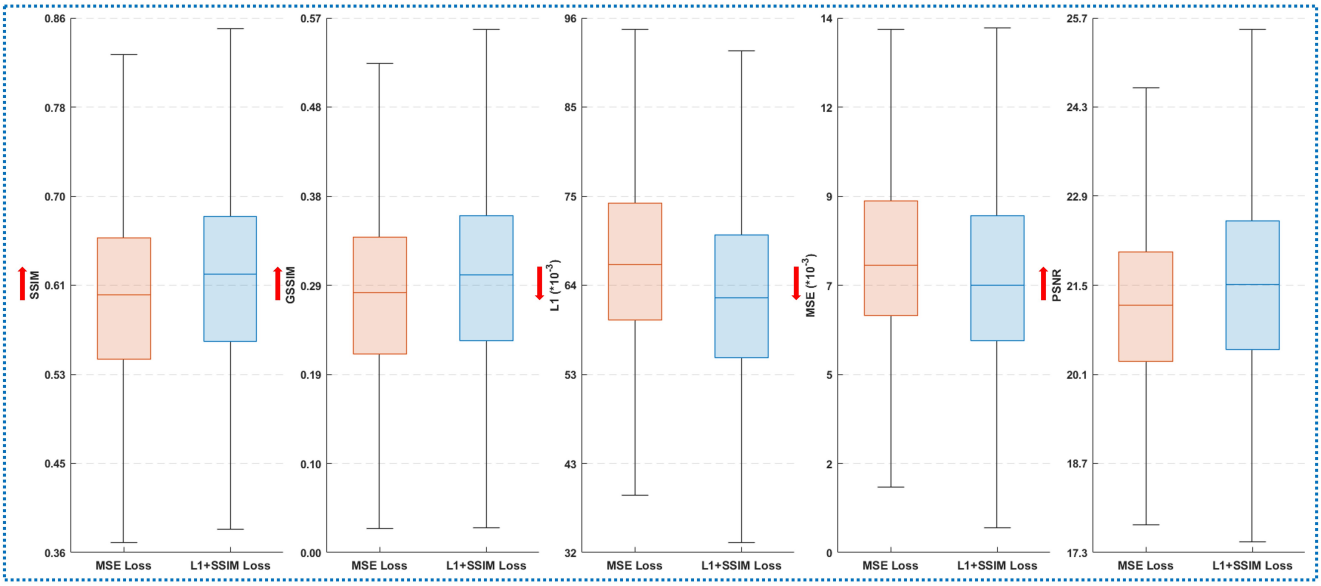
where G_A and G_B are the gradient magnitudes of images A and B.

As an example, for image A gradient magnitude is defined as:

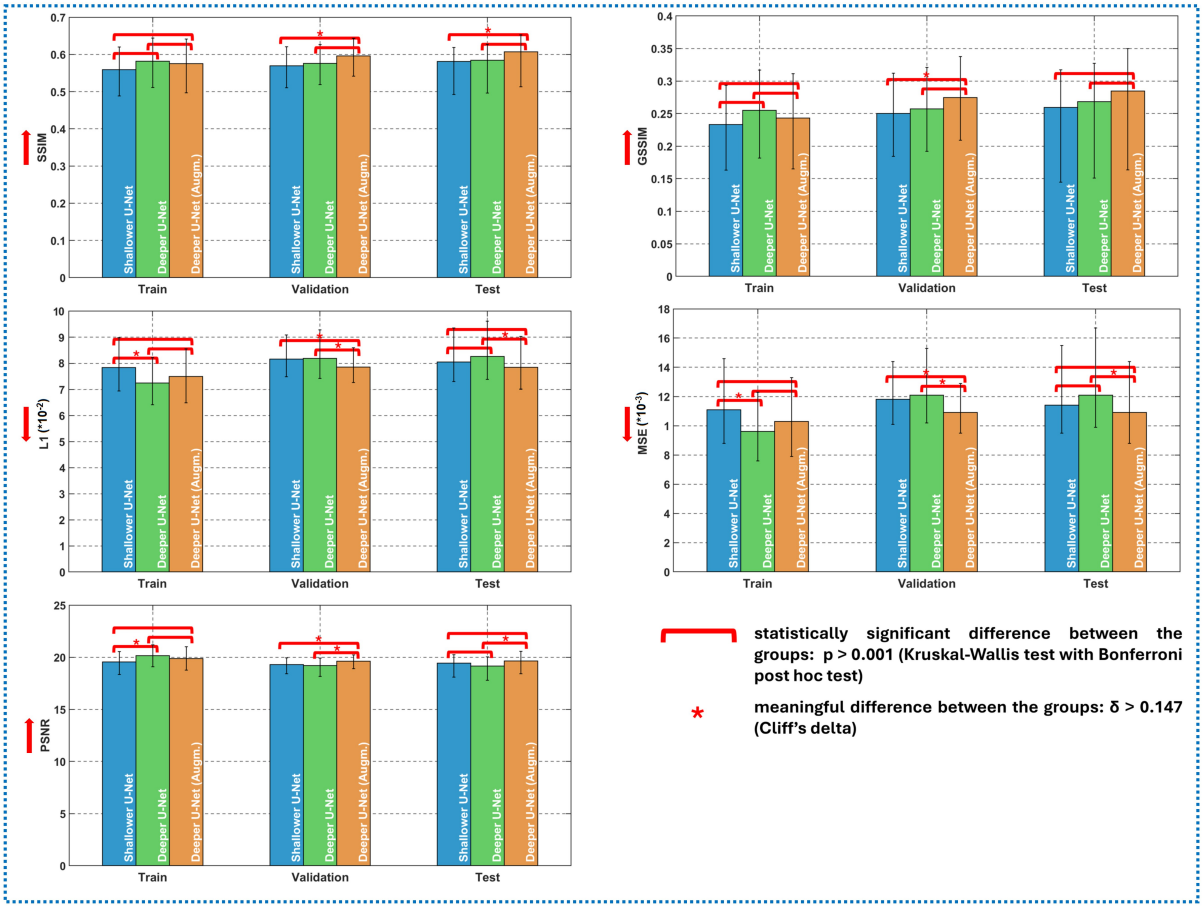
$$G_A = |G_A^x| + |G_A^y|, \tag{7}$$

where $|G_A^x|$ and $|G_A^y|$ are image gradients in x and y directions obtained by using horizontal and vertical edge Sobel operators. Following the SSIM protocol, GSSIM is computed per RGB channel independently and the final score is obtained by averaging the results across channels.

The network trained with the content loss function demonstrated improved performance across all five metrics (Supplementary Figure 2), indicating that the content loss function enables the network to learn a more detailed and accurate representation of the data.



Supplementary Figure 2. Loss function influence on the network performance. The red arrows next to each metric indicate whether a higher (up-arrow) or lower (down-arrow) value corresponds to better generated image quality.



Supplementary Figure 3. Generated image quality metrics across different networks: shallower U-Net vs. deeper U-Net vs. deeper U-Net with augmentation. The bar height corresponds to the median value of each metric with error bars spanning from Q1 (1st quartile) to Q3 (3rd quartile). The red arrows next to each metric indicate whether a higher (up-arrow) or lower (down-arrow) value corresponds to better reconstructed image quality.

The next step involved comparing the performance of different network architectures and training strategies to further improve the quality of the generated images. We assessed three configurations: the initial lightweight U-Net, a deeper version of the U-Net, and a deeper U-Net paired with image augmentation techniques. In this phase, we partitioned our dataset into training, validation, and testing subsets, comprising 80%, 10%, and 10% of the total data, respectively. Early stopping was employed to prevent overfitting.

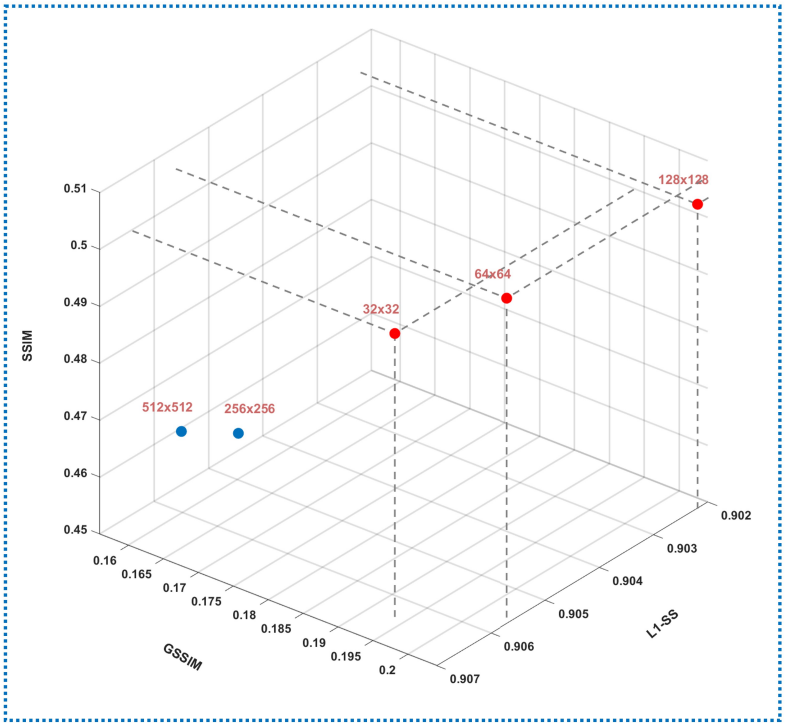
The performance of all three models was evaluated on the training, validation, and testing datasets using the same five metrics (Supplementary Figure 3). Statistical analysis using the Kruskal-Wallis test with Bonferroni post hoc correction revealed a statistically significant improvement across all metrics when using the deeper network architecture, with an even larger improvement observed when image augmentation was applied ($p < 0.001$ in most cases). However, Cliff's delta effect size analysis indicated that network depth alone did not always result in a meaningful difference, whereas the addition of image augmentation often led to a more significant improvement. Based on these findings, we decided to proceed with the deeper U-Net architecture paired with image augmentation for our final model.

The final optimization step focused on determining the optimal patch size for image subdivision. We evaluated the network performance with patch sizes ranging from 32×32 pixels to 512×512 pixels, maintaining a consistent patch overlap of 75%. We observed that the networks trained on larger patches reached the convergence elbow sooner, but the quality of the generated images favored smaller patches.

To identify the optimal patch size, we concentrated on the median values of three key metrics: SSIM, GSSIM, and L1 similarity score (L1-SS), which corresponds to 1-L1, making it a maximization-oriented metric. We plotted these metrics in a three-dimensional space and conducted a Pareto efficiency analysis to determine the Pareto-optimal solutions (Supplementary Figure 4). The analysis revealed that the patch sizes of 32×32, 64×64, and 128×128 pixels were Pareto-optimal in terms of reconstruction quality. Since neural networks operating on larger patch sizes also exhibited consistently higher convergence

speed, from Pareto-optimal solutions we selected the 128×128 pixel patch size for our final generator model.

Subsequently, the optimized U-Net generator was integrated into the GAN framework, which serves as the backbone of the N-H&E image-to-image translation method.



Supplementary Figure 4. Patch size influence on the network performance. Red dots correspond to the patch sizes forming the Pareto front.

Validation of the Gaussian weighting scheme for patch blending

The Gaussian weighting scheme approach for blending the overlapping patches generated by the N-H&E method was validated by comparing patch-wise quality metrics with those of the fully blended images, using the corresponding ground-truth images as a reference. All comparison metrics are summarized in Supplementary Table 1 and suggest no notable deterioration of the blended image quality due to the blending procedure.

Supplementary Table 1. Patch-wise and whole blended image quality metrics. Q1 corresponds to the 1st quartile, Q3 - 3rd quartile, IQR - interquartile range. The red arrows next to each metric indicate whether a higher (up-arrow) or lower (down-arrow) value corresponds to better quality of the generated images.

Patch-wise comparison (median [Q1, Q3, IQR])				
SSIM ↑	**GSSIM** ↑	**L1** ↓	**MSE** ↓	**PSNR*** ↑
0.60 [0.52, 0.65, 0.13]	0.27 [0.20, 0.34, 0.15]	0.077 [0.067, 0.087, 0.021]	0.010 [0.008, 0.013, 0.005]	19.9 [18.8, 21.0, 2.2]
Whole image comparison (median [Q1, Q3, IQR])				
SSIM ↑	**GSSIM** ↑	**L1** ↓	**MSE** ↓	**PSNR*** ↑
0.61 [0.55, 0.65, 0.10]	0.28 [0.23, 0.33, 0.11]	0.075 [0.068, 0.08, 0.016]	0.010 [0.009, 0.013, 0.004]	20.0 [19.0, 20.6, 1.6]

* PSNR *values are expressed in decibels* (dB).

GAN-generated image quality assessment

The unified GAN-based model generated high-quality images with minimal quality degradation compared to tissue-specific expert models, as shown in the Supplementary Table 2.

Supplementary Table 2. Model performance metrics across different tissue types. The table summarizes the median, first quartile (Q1), third quartile (Q3), and interquartile range (IQR) of the model evaluation metrics given in the square brackets, with the best values highlighted in bold. Among the expert models, the model tailored for colon tissue demonstrated superior performance across all evaluation metrics, while the unified model achieved the highest SSIM in colon tissue images, the highest GSSIM was observed in skin tissue images, whereas the remaining metrics favored prostate tissue. The metric values showed minimal decrease when using the unified model, indicating good generalizability. The red arrows next to each metric indicate whether a higher (up-arrow) or lower (down-arrow) value corresponds to better quality of the generated images.

Expert model performance metrics (median [Q1, Q3, IQR])

	SSIM ↑	GSSIM ↑	L1 ↓	MSE ↓	PSNR* ↑
Skin	0.61 [0.53, 0.66, 0.13]	0.32 [0.24, 0.38, 0.14]	0.078 [0.068, 0.091, 0.023]	0.012 [0.009, 0.015, 0.006]	19.4 [18.2, 20.6, 2.4]
Colon	**0.66 [0.60, 0.72, 0.12]**	**0.33 [0.26, 0.39, 0.14]**	**0.069 [0.058, 0.078, 0.021]**	**0.008 [0.006, 0.011, 0.005]**	**20.9 [19.8, 22.3, 2.5]**
Cervix	0.51 [0.42, 0.58, 0.16]	0.19 [0.12, 0.27, 0.15]	0.089 [0.077, 0.100, 0.024]	0.013 [0.010, 0.017, 0.007]	18.9 [17.8, 20.0, 2.2]
Prostate	0.55 [0.48, 0.61, 0.13]	0.21 [0.14, 0.28, 0.14]	0.071 [0.062, 0.081, 0.019]	0.009 [0.007, 0.011, 0.004]	20.5 [19.4, 21.5, 2.1]

Unified model performance metrics (median [Q1, Q3, IQR])

	SSIM ↑	GSSIM ↑	L1 ↓	MSE ↓	PSNR* ↑
Skin	0.61 [0.54, 0.66, 0.13]	**0.31 [0.24, 0.38, 0.14]**	0.075 [0.066, 0.087, 0.022]	0.010 [0.008, 0.014, 0.006]	19.8 [18.5, 20.9, 2.4]
Colon	**0.63 [0.58, 0.67, 0.09]**	0.29 [0.23, 0.35, 0.12]	0.075 [0.065, 0.083, 0.018]	0.010 [0.007, 0.012, 0.004]	20.2 [19.4, 21.3, 1.9]
Cervix	0.50 [0.40, 0.57, 0.17]	0.18 [0.10, 0.26, 0.15]	0.091 [0.080, 0.102, 0.022]	0.014 [0.011, 0.017, 0.006]	18.7 [17.7, 19.6, 2.0]
Prostate	0.55 [0.48, 0.61, 0.13]	0.20 [0.13, 0.27, 0.14]	**0.072 [0.064, 0.082, 0.018]**	**0.009 [0.007, 0.012, 0.004]**	**20.3 [19.3, 21.4, 2.1]**

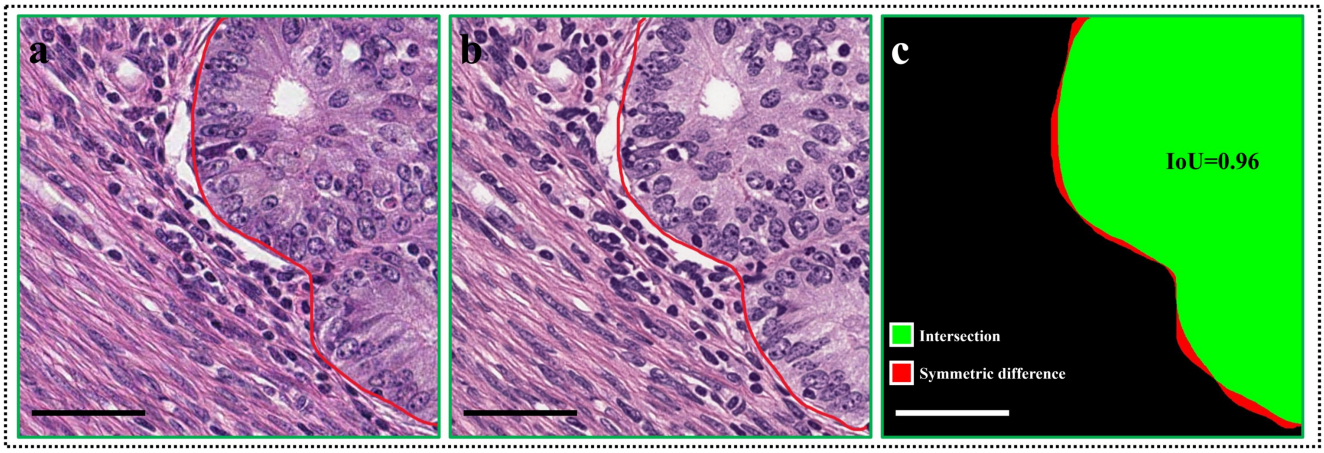
* PSNR *values are expressed in decibels (*dB*).*

Expert pathologist evaluation of the N-H&E-generated image quality

In addition to quantitative assessment of N-H&E image-to-image translation, the generated images of skin, colon, cervical, and prostate tissues were qualitatively evaluated by three expert pathologists. The evaluation focused on image quality, structural accuracy, and the feasibility of making accurate clinical assessments based on the generated images.

The first part of the study tested the comparative diagnostic evaluation by assessing a pathologist's ability to identify tumor margins in the generated images. One of the three pathologists was provided with pairs of original and generated brightfield images of skin and cervical tissue. The dataset included 32 skin tissue images, with visible tumor margins present in 10 of them, and 26 cervical tissue images, with visible margins in 9. Initially, the pathologist marked the tumor margins in the generated images without having seen the original brightfield images. After being provided with the corresponding original brightfield images, the pathologist repeated the task. Binary masks were derived from the pathologist's annotations to separate tumor and non-tumor areas in the generated and original brightfield images. The masks were compared pairwise using the intersection over union (IoU) metric. The IoU ranged from 0.96 to 0.99, reflecting an almost perfect agreement and indicating that the quality of the generated images did not hinder the pathologist's ability to accurately delineate the tumor margins. Any inaccuracies were attributable to slight variability in the manual delineation rather than issues with image quality (Supplementary Figure 5).

In the second part of the evaluation, all three pathologists were asked to rate a total of 96 pairs of original and generated brightfield images of skin, colon, cervical, and prostate tissues on a scale from 1 (poor) to 10 (excellent) based on image quality, realism, and structural accuracy, which was defined by correct representation of anatomical features and tissue-specific landmarks. The generated images consistently received scores of 9 or 10 across all tissue types, with pathologists noting that key structural features, such as collagen fibers, nuclei, and erythrocytes, were accurately rendered and clearly visible in the generated images. In some instances, particularly for colon and cervical tissue, the pathologists preferred the generated images, citing enhanced details and improved contrast with minor criticism raised regarding the resolution and hue accuracy in certain generated images (the evaluation summary is provided in Supplementary Table 3). It is important to stress that since hue variability is common even in original brightfield images due to slight inconsistencies in staining protocols or imaging parameters, the model, trained on a diverse dataset, learns to standardize the stain rather than mimic it.



Supplementary Figure 5. Tumor margin delineation comparison in original **(a)** and N-H&E generated **(b)** brightfield images. The red lines in **(a, b)** correspond to the tumor margin, as specified by an expert pathologist. Intersection (green) and symmetric difference (red) areas are represented in **(c)**. Scale bar: 50 µm.

Supplementary Table 3. Expert pathologist evaluation scores (1-10) for each tissue type, where 10 indicates an excellent reconstructed image quality. The comments column includes brief justifications for each score. The first data row of the table details the number of ground-truth and generated image pairs of each tissue type used for evaluation by each pathologist.

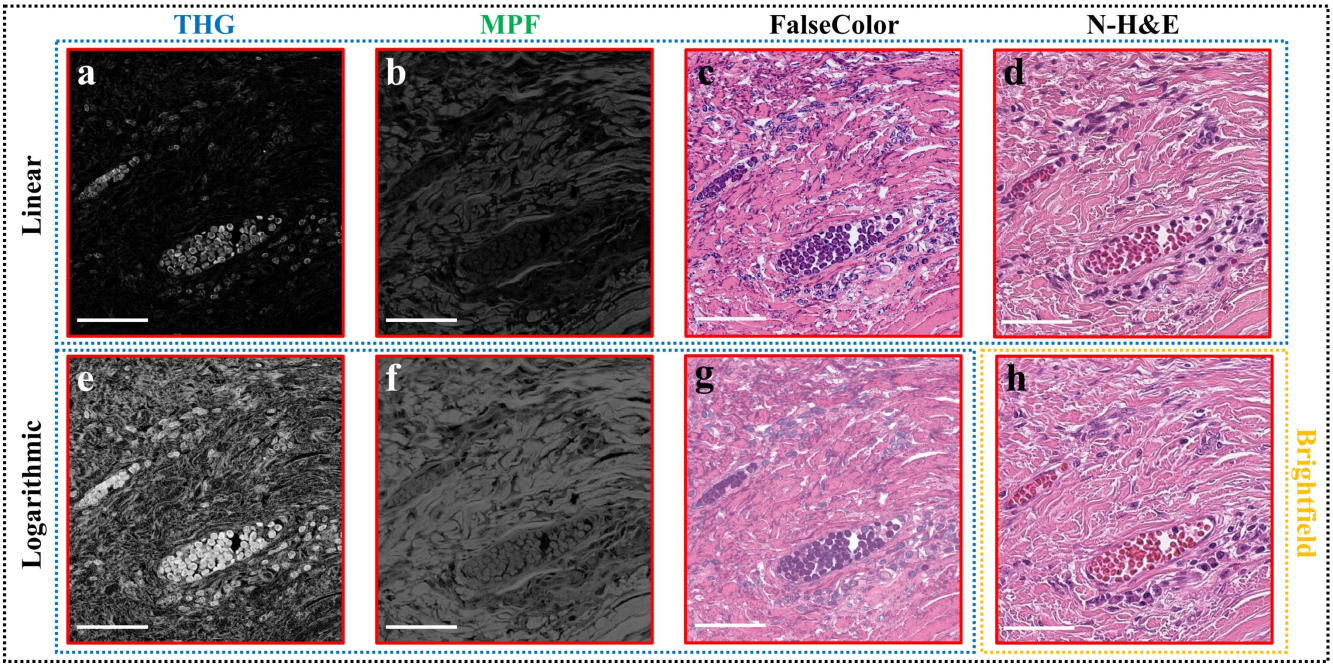
Qualitative expert pathologist evaluation (score 1-10)					
	Skin	**Colon**	**Cervix**	**Prostate**	**Comments**
# of pairs	32	22	26	16	
Pathologist 1	10*	10	10	10	**No noticeable differences between the original and generated images; in some cases, image resolution could be improved.**
Pathologist 2	9-10	10	10	9-10	**All structural elements of the tissues are rendered correctly; some generated images exhibit better contrast and provide more details than the original.**
Pathologist 3	9	9	9	9	**Minimal hue differences as compared to the original images.**

** Values correspond to evaluation scores assigned by expert pathologists.*

Comparison of H&E image reconstruction techniques

Images generated by the N-H&E method were compared with H&E images reconstructed from THG and MPF inputs using the FalseColor technique[6]. The FalseColor technique is commonly used to reconstruct H&E images from linear (first order) excitation fluorescence images. FalseColor reconstruction was carried out using THG and MPF channels. The image generation was evaluated for both linear (Supplementary Figure 6a,b) and logarithmic (Supplementary Figure 6e,f) intensity scales. Due to the high dynamic range of the THG and MPF signals, a manual threshold-based tissue-background segmentation followed by a flat-field correction was applied to ensure visibility of both high- and low-intensity tissue features. The reconstructed images from both linear (Supplementary Figure 6c) and logarithmic (Supplementary Figure 6g) scales were not visually accurate, which may be attributable to the limitations of the Beer-Lambert absorption model on which the FalseColor method is based. Although suitable for linear optical processes, this model does not account for nonlinear optical phenomena such as multiphoton absorption, higher harmonic generation, and fluorescence quenching, leading to suboptimal reconstruction performance.

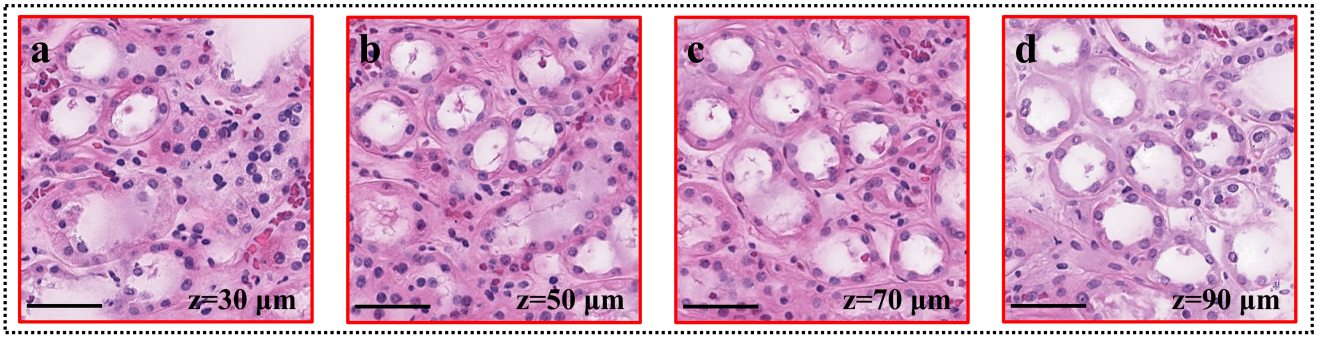
In contrast, GAN-based N-H&E method produced reconstructions (Supplementary Figure 6d) that closely matched the histological ground truth H&E image (Supplementary Figure 6h) with correctly rendered key histological features including cell nuclei, erythrocytes, and ECM structures.



Supplementary Figure 6. Comparison of H&E stain reconstruction techniques. FalseColor method[6] was used to digitally stain the THG and MPF grayscale images in linear scale **(a,b)** with flat-field pre-processing correction and the same unmodified images in logarithmic scale **(e,f)**. Visual assessment revealed that, in both cases, FalseColor method failed to accurately render the H&E stained images **(c,g)**, while GAN-based N-H&E technique produced the images **(d)** closest to their ground truth target images **(h)**. Scale bar: 50 µm.

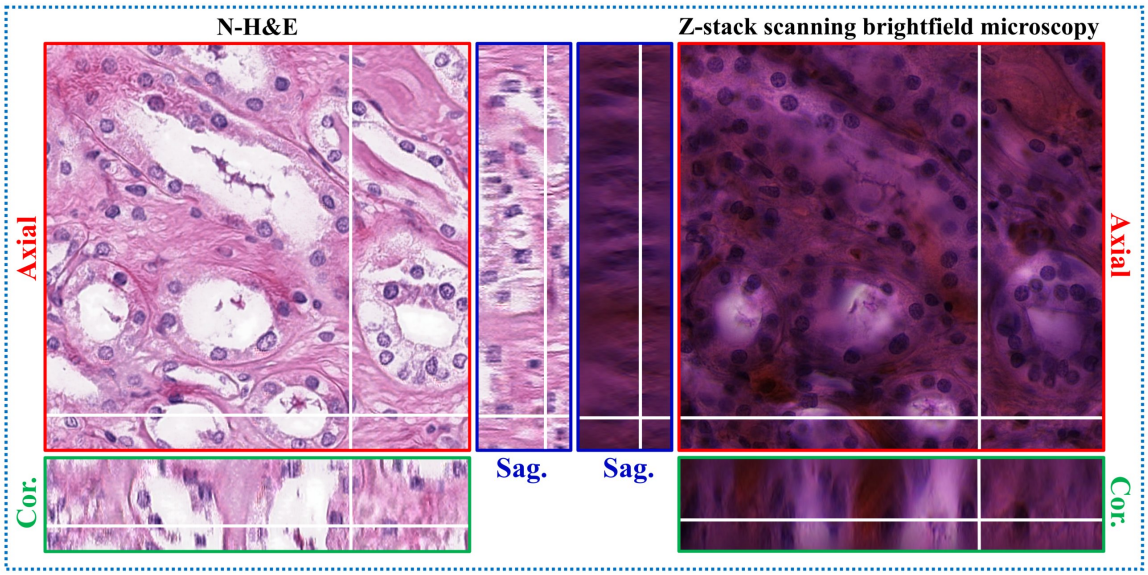
Quality assessment of deep tissue 3D N-H&E imaging and benchmarking against Z-stack scanning brightfield microscopy

The quality of N-H&E image reconstruction from THG and MPF optical sections at various depths was evaluated using a 100 µm-thick renal tissue sample. Supplementary Figure 7 demonstrates high-quality H&E image reconstruction even at a depth of 90 µm, indicating that reliable H&E images can be generated from deep tissue sections. The thickness of the imaged cryosection was constrained by the maximum preparation thickness attainable by the slicer.



Supplementary Figure 7. The 2D N-H&E image reconstruction at different depth of a 100 µm-thick renal tissue from THG and MPF microscopy optical sections. The reconstructed images are shown for the depth of 30 micrometers **(a)**, 50 micrometers **(b)**, 70 micrometers **(c)** and 90 micrometers **(d)**. All relevant structures, including extracellular matrix, cell nuclei, erythrocytes, and renal tubules, are clearly visualized and accurately rendered in color throughout the entire tissue depth, as evaluated by expert pathologists based on visual inspection. Scale bar: 50 µm.

The 3D imaging performance of the N-H&E method was benchmarked against Z-stack scanning brightfield microscopy based on the Pannoramic Scan II (3DHISTECH) (Supplementary Figure 8). While brightfield microscopy can directly acquire volumetric H&E images of thin sections and does not require specific fluorescent staining, its performance in thick tissue samples is limited. This is due to significant absorption and scattering of visible light in biological tissues, resulting in optical sections that are often dark and noisy because of out-of-focus contributions to the signal (Supplementary Figure 8b). Furthermore, sagittal and coronal views reveal poor axial resolution, with cell nuclei appearing elongated and blurred. In contrast, the N-H&E technique delivers high-quality optical sectioning with minimal out-of-focus noise and significantly improved axial resolution, producing sharp and well-defined nuclear structures at all depths (Supplementary Figure 8a).



Supplementary Figure 8. Comparison of 3D H&E image volumes produced by the N-H&E technique **(a)** and the Z-stack scanning brightfield microscopy (Pannoramic Scan II, 3DHISTECH) **(b)** throughout the whole 50 µm tissue thickness. The superior image quality of the N-H&E technique is particularly evident in the coronal and sagittal views **(a)**, where multimodal nonlinear microscopy achieved superior axial resolution, enabling clear identification of tissue structures, such as individual cell nuclei, within the whole tissue volume.

References

1. Jiang, R., Duke, B., Flament, F. & Aarabi, P. Synthesizing ultraviolet skin images via gan with gaussian weighted patch blending. In *2022 IEEE International Symposium on Multimedia (ISM)*, 157–158, DOI: https://doi.org/10.1109/ISM55400.2022.00033 (IEEE, Italy, 2022).

2. Tuer, A. E. *et al.* Nonlinear multicontrast microscopy of hematoxylin-and-eosin-stained histological sections. *J. Biomed. Opt.* **15**, 2855–2867, DOI: https://doi.org/10.1117/1.3382908 (2010).

3. Zhao, H., Gallo, O., Frosio, I. & Kautz, J. Loss functions for image restoration with neural networks. *IEEE Transactions on Comput. Imaging* **3**, 47–57, DOI: https://doi.org/10.1109/TCI.2016.2644865 (2016).

4. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Process.* **13**, 600–612, DOI: https://doi.org/10.1109/TIP.2003.819861 (2004).

5. Chen, G. H., Yang, C. L. & Xie, S. L. Gradient-based structural similarity for image quality assessment. In *2006 International Conference on Image Processing*, 2929–2932, DOI: https://doi.org/10.1109/ICIP.2006.313132 (IEEE, Atlanta, GA, USA, 2006).

6. Serafin, R., Xie, W., Glaser, A. K. & Liu, J. T. Falsecolor-uppercasePython: a rapid intensity-leveling and digital-staining package for fluorescence-based slide-free digital pathology. *PLOS One* **15**, 1–17, DOI: https://doi.org/10.1371/journal.pone.0233198 (2020).