A Comparative Analysis of Deep Learning Approaches for Virtual H&E Staining from Autofluorescence Imaging

Anuva Shah, Joshua Levy
Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center

Dartmouth Health

ABSTRACT

Hematoxylin and eosin (H&E) staining

- The most widely used tissue staining technique in pathology
- **Destructive**: once samples are stained, they cannot easily be used for further assays

Autofluorescence imaging

- **Non-destructive**: excites natural tissue fluorophores with specific light wavelengths, capturing structure without chemical stains

Autofluorescence highlights cellular structures in tissue, though it fails to provide the color contrast and clarity pathologists require for accurate diagnosis. To address this, **this study compared four deep learning methods for virtual H&E staining from autofluorescence**, identifying which method best preserved diagnostically relevant structures.

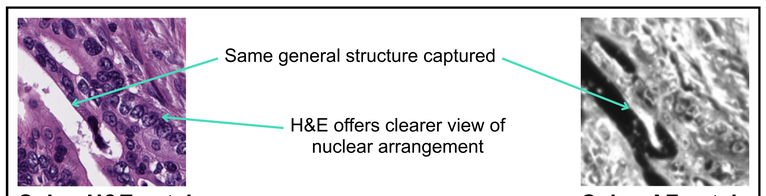

Figure 1: Comparison of paired H&E and AF patch structure.

INTRODUCTION

This study aimed to **develop and evaluate four different deep learning architectures for virtual staining**. A trustworthy digitized staining framework would have a number of **positive impacts** in clinical settings:

- **Tissue Preservation and Multi-Analysis**
  - Biopsies are vital and limited, especially with cancer patients
  - Physically staining samples with H&E damages tissue
  - Digitized stains preserve tissue for future diagnostics, allowing repeated non-destructive tissue analyses
  - A lab could perform AF imaging first, preserving the section for downstream assays that would otherwise compete with H&E for the same tissue
- **Stain Consistency and Reproducibility**
  - Real H&E stained samples vary based on lab, batch, protocol, and more; this is a known source of variability in diagnosis
  - A consistent model producing the same stain appearance each time eliminates this confound
- **Speed and Clinical Accessibility**
  - Current H&E staining requires reagents, staining infrastructure, and technician time; virtual staining shifts these requirements toward imaging infrastructure
  - H&E output would be generated in seconds by a model, reducing the amount of time patients wait for diagnostic reports
  - Relevant for low-resource settings where trained staff are not always immediately available

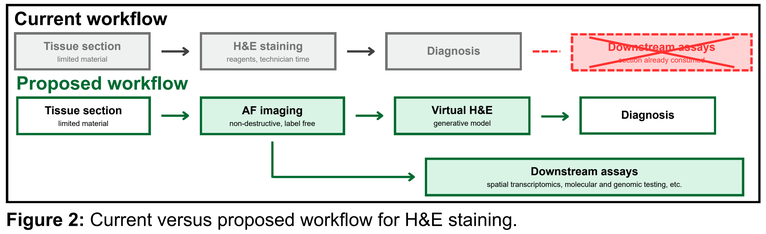

Figure 2: Current versus proposed workflow for H&E staining.

METHODOLOGY

Dataset & Preprocessing

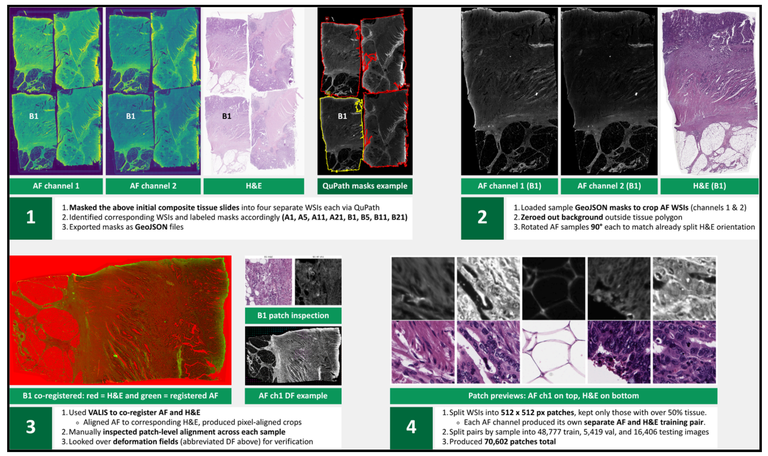

Figure 3: Dataset Preprocessing Workflow: (1) Masking composite tissue slides in QuPath, (2) cropping and rotating AF WSIs to match H&E, (3) co-registration with VALIS, and (4) tiling into 512 by 512 patches at over 50% tissue content.

The following four architectures were trained for 100 epochs on the same training and validation set:

- **U-Net Regression Baseline**
  - Encoder-decoder CNN
  - Direct MSE regression from AF to H&E
  - Simplest baseline with no adversarial loss
- **Pix2pix Conditional GAN**
  - U-Net generator and PatchGAN discriminator
  - Adversarial and weighted L1 loss with lambda = 100
  - Paired supervision anchored output directly to ground-truth H&E
- **CycleGAN**
  - Dual ResNet generators and PatchGAN discriminators
  - Adversarial and cycle-consistency loss
  - Nuclei-density loss
  - Added Exponential Moving Average (EMA)
- **Latent Diffusion Model**
  - Frozen VAE and conditional U-Net denoiser
  - AF-conditioned latent diffusion (DDPM training, DDIM sampling)

Then, models were evaluated on the testing set for both pixel-level accuracy and biological fidelity:

- **Pixel-level metrics**: Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR)
- **Biological fidelity metrics**: Nucleus detection and classification via HoVer-Net, embedding similarity via UNI2

RESULTS

SSIM and PSNR across models

Model	SSIM	PSNR (dB)
U-Net	0.35	15.17
pix2pix	0.26	13.25
CycleGAN	0.21	14.18
Latent Diffusion	0.18	13.45

HoVer-Net nucleus detection and typing across models

Model	Nuclei per patch	Count ratio	Class dist. L1
CycleGAN	20.8	0.972	0.119
Latent Diffusion	21.7	1.018	0.365
pix2pix	17.1	0.799	0.172
U-Net	6.4	0.301	0.311
Real	**21.4**	-	-

Figure 4: Pixel similarity and HoVer-Net nucleus content across 16,406 test patches. U-Net leads SSIM and PSNR but ranks last on nucleus count.

RESULTS

UNI2-h embedding similarity across models

Model	Cosine similarity	Fréchet distance
CycleGAN	0.504	89.5
pix2pix	0.355	161.5
Latent Diffusion	0.163	283.0
U-Net	0.074	283.8
Real	0.322	-

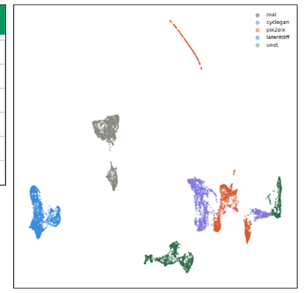

Figure 5: UNI2-h embedding similarity and UMAP projection, with 0.322 as the unrelated-patch cosine reference. CycleGAN (purple on map, green on table) sits nearest real; U-Net & LDM are separate entirely.

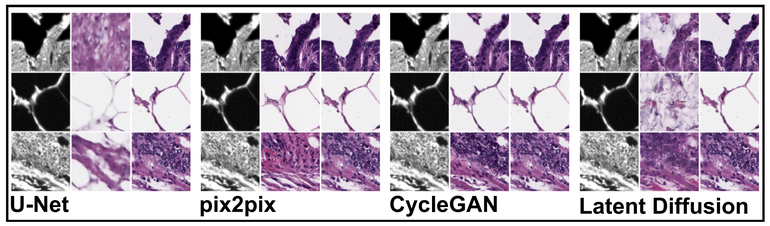

Figure 6: Generated H&E on three representative test patches: glandular epithelium (top), adipose (middle), and dense tumor (bottom). Each block shows AF input, virtually stained output, and real H&E. U-Net loses nuclear detail across all three, while CycleGAN preserves lumen structure and nuclear density closest to real.

CONCLUSION

CycleGAN proved the most viable architecture for virtual H&E staining, within clear limitations and with concrete paths forward.

- **Key Findings**
  - CycleGAN scored best on biological fidelity while U-Net scored best on pixel-level accuracy, supporting the secondary hypothesis that pixel-level and biological metrics rank models differently
  - The LDM, hypothesized to perform best, underperformed all three single-pass architectures under the training budget
  - CycleGAN is the most suitable for real-world use, though rare nuclei undergeneration and pixel-level accuracy need to be improved upon
- **Limitations**
  - Colon data came from two patients, collected by Dartmouth and Cedars-Sinai. This limits generalizability across organs and labs
  - CycleGAN was trained with a nuclei-density loss the other models did not receive, likely contributing to its HoVer-Net performance
  - The LDM operated inside a frozen VAE trained on natural images; encode-decode of real H&E alone reached only SSIM 0.71 (this caused a ceiling other models did not have)
- **Future Directions**
  - Joint multi-channel AF conditioning, as this study's splitting of channels discarded half the signal per pass
  - Intensity transformation (gamma or log) on AF input to expand contrast in lower-signal patches
  - Extend nuclei-density loss to the other architectures
- **Acknowledgements**: Dr. Joshua Levy, Dr. Khang Le, Emerging Diagnostics and Investigative Technologies Lab
- **References**: https://tinyurl.com/anuvaeditreferences

edit.