

Restoration of metabolic functional metrics from label-free, two-photon human tissue images using multiscale deep-learning-based denoising algorithms

Nilay Vora^{Ⓜ, a, †} Christopher M. Polleys^{Ⓜ, a, †} Filippos Sakellariou,^b
Georgios Georgalis^{Ⓜ, c} Hong-Thao Thieu,^{d, ‡} Elizabeth M. Genega,^{e, §}
Narges Jahanseir,^e Abani Patra^{Ⓜ, c, f} Eric Miller^{Ⓜ, g, h} and Irene Georgakoudi^{Ⓜ, a, *}

^aTufts University, Department of Biomedical Engineering, Medford, Massachusetts, United States

^bAnatolia College, Thessaloniki, Greece

^cTufts University, Data Intensive Studies Center, Medford, Massachusetts, United States

^dTufts University School of Medicine, Tufts Medical Center, Department of Obstetrics and Gynecology, Boston, Massachusetts, United States

^eTufts University School of Medicine, Tufts Medical Center, Department of Pathology and Laboratory Medicine, Boston, Massachusetts, United States

^fTufts University, Department of Mathematics, Medford, Massachusetts, United States

^gTufts University, Department of Electrical and Computer Engineering, Medford, Massachusetts, United States

^hTufts University, Tufts Institute for Artificial Intelligence, Medford, Massachusetts, United States

ABSTRACT. Significance: Label-free, two-photon excited fluorescence (TPEF) imaging captures morphological and functional metabolic tissue changes and enables enhanced understanding of numerous diseases. However, noise and other artifacts present in these images severely complicate the extraction of biologically useful information.

Aim: We aim to employ deep neural architectures in the synthesis of a multiscale denoising algorithm optimized for restoring metrics of metabolic activity from low-signal-to-noise ratio (SNR), TPEF images.

Approach: TPEF images of reduced nicotinamide adenine dinucleotide (phosphate) (NAD(P)H) and flavoproteins (FAD) from freshly excised human cervical tissues are used to assess the impact of various denoising models, preprocessing methods, and data on metrics of image quality and the recovery of six metrics of metabolic function from the images relative to ground truth images.

Results: Optimized recovery of the redox ratio and mitochondrial organization is achieved using a novel algorithm based on deep denoising in the wavelet transform domain. This algorithm also leads to significant improvements in peak-SNR (PSNR) and structural similarity index measure (SSIM) for all images. Interestingly, other models yield even higher PSNR and SSIM improvements, but they are not optimal for recovery of metabolic function metrics.

Conclusions: Denoising algorithms can recover diagnostically useful information from low SNR label-free TPEF images and will be useful for the clinical translation of such imaging.

*Address all correspondence to Irene Georgakoudi, irene.georgakoudi@tufts.edu

†These two authors contributed equally to this work.

‡Current affiliation: Department of Obstetrics and Gynecology, Newton-Wellesley Hospital, Newton, MA 02462, USA.

§Current affiliation: Department of Pathology and Laboratory Medicine, Emory University Hospital, Atlanta, GA 30322, USA.

© The Authors. Published by SPIE under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI. [DOI: [10.1117/1.JBO.28.12.126006](https://doi.org/10.1117/1.JBO.28.12.126006)]

Keywords: multiscale denosing; metabolic imaging; label-free; deep learning; two-photon excited fluorescence; biomedical optics

Paper 230199GR received Jul. 12, 2023; revised Oct. 23, 2023; accepted Nov. 28, 2023; published Dec. 22, 2023.

1 Introduction

Metabolism refers to the set of chemical reactions that occur within a cell to produce energy and to build the necessary macromolecules to sustain life.¹ The energetic and macromolecular demands of a cell often change with aging and the onset of several diseases, including cancer, diabetes, neurodegenerative disorders, and cardiovascular diseases.² Therefore, it is clear that understanding the nature of such metabolic changes at the cellular level to characterize heterogeneity and dynamic interactions among different cell populations is critical for the development of improved diagnostic and treatment methods.³ However, established methods to assess metabolic function in the clinic and the laboratory either lack resolution⁴ or are destructive.⁵

One approach that is capable of probing the tissue metabolic state with high three-dimensional resolution in a non-destructive manner is two-photon excited fluorescence (TPEF) microscopy.⁶ TPEF is a non-linear imaging technique that benefits from intrinsic optical sectioning and the ability to penetrate hundreds of micrometers into bulk tissue.⁷ TPEF is also uniquely suited to capture images from endogenous fluorophores, such as NAD(P)H and FAD.⁸ NADH and FAD are coenzymes that facilitate energy generation and biomolecular synthesis via a number of pathways.⁹ Several of these pathways, including the tricarboxylic acid cycle, glutaminolysis, fatty acid oxidation, and oxidative phosphorylation, occur in the mitochondria.¹⁰ NADPH plays an important role in anti-oxidant pathways and has similar fluorescence characteristics to those of NADH.¹¹ Thus, the term NAD(P)H is used throughout this paper to refer to the fluorescence of both NADH and NADPH. A large fraction of the flavin-associated cellular fluorescence is attributed to FAD bound to lipoamide dehydrogenase (LipDH), even though contributions from free FAD and FAD bound to complex II (electron transfer flavoprotein) may also be significant. Here, we use the term FAD to refer to all flavin-associated fluorescence detected from cells.

Despite the lack of specificity in the origins of the fluorescence signals, the ratio of FAD/NAD(P)H or its normalized definition of FAD/(NAD(P)H+FAD) has been shown to correlate to the oxido-reductive state of the cells in many studies.^{12–15} Mitochondria are also characterized by the ability to fuse and fission to enhance energy production and delivery in response to stress or to facilitate the removal of damaged mitochondria.¹⁶ Such differences in mitochondrial organization have also been quantified based on analysis of NAD(P)H TPEF images.^{17,18} NAD(P)H fluoresces more efficiently when bound to enzymes typically in the mitochondria; therefore variations in NAD(P)H TPEF intensity fluctuations can be exploited for label-free quantitative assessments of mitochondrial organization (clustering) in cells, tissues, and living humans.^{17,19} Changes in mitochondrial organization have in turn been attributed to metabolic function changes.^{20–22} The heterogeneity of parameters such as the redox ratio and mitochondrial clustering within a tissue have also been identified as important indicators of metabolic state.^{23–25} A number of studies have already highlighted the diagnostic potential of such assessments in living humans, and there is growing interest in performing such measurements at the bedside or via endoscopes to expand the range of diagnostic applications to several organs beyond the skin.^{26–29} Fast image acquisition in these settings is critical; however, endoscope designs typically include relatively low numerical aperture (0.5 to 0.7) objectives and are not as efficient in the generation and collection of TPEF.³⁰ As a result, low resolution, noise, and other degradations may mask the diagnostically useful functional features. Thus, approaches to enhance label-free, TPEF images could play a transformative role in the successful translation of this technique to improve tissue metabolic function assessments in the context of diagnosis or treatment.

Traditionally, both standard image processing methods as well as inverse techniques have been used to enhance the interpretability of TPEF data.³¹ These methods are most appropriate

when the forward signal model is known (preferably linear) and when stochasticity is either additive noise with a known distribution or the data are Poisson with the mean given by the forward signal model.³² Neither is the case for TPEF sensing in which the interaction of light with tissue leads to a highly complex forward model and the data are a mix of Poisson statistics and additive Gaussian noise.³³ Motivated by these challenges as well as the recent success of machine learning methods for addressing a range of image analysis and interpretation problems, we consider the use of deep-learning methods for enhancing TPEF images to improve the extraction of metabolically relevant information.

Deep-learning-based methods have already been shown to enhance the quality and resolution of a wide range of images, including label-free two-photon images.^{34–38} Convolutional neural network-based content-aware image restoration (CARE), residual channel attention networks (RCAN), and super-resolution generative adversarial networks (SRGAN) have been developed for this purpose.^{34–36} Although these models have been applied to fluorescence microscopy data, their use has been limited to exogenously labeled samples, which have enhanced contrast compared to label-free images. However, recently, Shen et al. demonstrated the application of a generative adversarial networks (GAN) for the restoration of label-free multimodal nonlinear images.³⁷ We note that, in these and related studies, standard metrics, such as peak SNR (PSNR) and structural similarity index measure (SSIM), are used widely as indicators of the quality of image restoration, even though they may not always match the human visual system's assessment of image quality (measured by mean opinion score).^{34,35,37,38}

Here, we report on the ability of deep-learning based denoising approaches to restore functional metabolic metrics extracted from label-free TPEF images. Specifically, we consider the recovery of average and depth dependent variations in the redox ratio (FAD/(NAD(P)H + FAD)) and mitochondrial clustering extracted from analysis of TPEF images acquired from freshly excised human cervical epithelia, including healthy and precancerous lesions. Accurate extraction of the depth-dependent variations in these metrics have been identified as important diagnostic biomarkers of metabolic function, especially in stratified squamous epithelia, such as the skin and the cervix, where the proliferative and differentiation state of cells at distinct depths is expected to vary significantly in healthy tissues.^{22,25,29,39–41} In addition, we assess whether PSNR and SSIM improvements are correlated with the restoration of the functional metabolic metrics. We consider CARE (a U-Net), GANs (SRGAN), and RCAN networks and assess nine loss functions, including mean average error (MAE), mean square error (MSE), SSIM, frequency focal loss (FFL), coefficient of variation (R2), redox ratio loss, and three combinations of these loss functions (see Discussion S1 and Table S2 in the [Supplementary Material](#)). We also examine whether training on FAD or NAD(P)H images impacts the successful restoration of metabolic function metrics from the corresponding denoised images.

We find that a novel combination of a one level wavelet transformation and CARE models trained to denoise each of the four wavelet domain sub-bands yields denoised images that enable optimized recovery of all metabolic function metrics. Interestingly, we observe that the architecture most successful in recovering metabolic metrics is not optimal in terms of more standard metrics used to measure performance, such as PSNR and SSIM. Thus, our results indicate that deep-learning based denoising algorithms may require distinct multiscale training and testing approaches for the recovery of functional metrics needed for improved diagnosis, understanding the drivers of disease, and the development of novel therapeutics, instead of traditional morphological image quality metrics.

2 Materials and Methods

2.1 Sample Acquisition

All activities pertaining to cervical tissue biopsy handling were done in accordance with approved Tufts Health Sciences IRB protocol #10283. Patients over the age of 18 with a recent low-grade squamous intraepithelial lesion (LSIL) or high-grade squamous intraepithelial lesion (HSIL) pap smear diagnosis undergoing a colposcopy or loop electrosurgical excision procedure were recruited to the study. Informed consent was acquired from all study subjects before participation. During the routine procedure, a second biopsy from a colposcopically abnormal region of the cervix was taken and placed in a custom-built tissue carrier containing keratinocyte

serum-free media (Lonza). Biopsies were transported via personal vehicle to the Tufts Advanced Microscopy Imaging Center for imaging. All imaging was conducted within 4 h post-biopsy. Immediately after imaging, biopsies were fixed in 10% neutral buffered formalin. Biopsies were returned within 5 business days to the Tufts Medical Center Department of Pathology for standard histopathological diagnosis.

Patients over the age of 18 undergoing hysterectomies for benign gynecological disease were also recruited to the study as healthy controls. The only difference between healthy and precancerous biopsy acquisition was in the actual biopsy excision. Healthy biopsies were sampled from the resected cervix by a pathologist after macroscopic inspection to rule out abnormalities.

2.2 Deep Learning Dataset Details

A total of 151 regions of interest (ROIs) (image stacks) were collected from 54 patients. The training and validation sets were comprised of 100 ROIs featuring 5 to 50 optical sections (OSs) per ROI. 75% of the ROIs were randomly selected for training, and the remaining 25% were set aside as the validation set (1657 training OSs and 554 validation OSs). To prevent data leakage, training and validation OSs were separated on an ROI basis. The test set featured 51 ROIs (with 10 to 50 OSs per ROI) and was excluded from all training (1018 OSs). For k -fold validation, the 100 training and validation ROIs were shuffled and split again using the same 75:25 ratio for up to five times to ensure robustness of denoising on a constant test set (see Fig. S5 in the [Supplementary Material](#)). The dataset features images from tissues with three diagnoses: benign, LSIL, and HSIL. The test set was composed of 25 benign ROIs (49.02%), 14 LSIL ROIs (27.45%), and 12 HSIL ROIs (23.53%). The training and validation sets were composed of 55 benign ROIs (54.45%), 25 LSIL ROIs (24.75%), and 21 HSIL ROIs (20.79%). Based on training/validation splitting seed, these values could range from 52% to 57.3% benign, 25.3 to 26.7% LSIL, and 18.7% to 22.7% HSIL in the training set and 48% to 64% benign, 20% to 24% LSIL, and 16% to 28% HSIL in the validation set. An alternative training scheme was initially attempted. In this scheme, only benign ROIs were used in training with 112 ROIs of mixed diagnosis being used in the test set and 39 benign ROIs being used for training. The training set was later modified as it became evident that training on a mixture of diagnoses resulted in superior restoration of downstream metrics (Fig. 5).

2.3 Optical Instrumentation and Image Acquisition

Images were collected using a commercially available Leica SP8 inverted microscope system equipped with an Insight fs laser. Tissue biopsies were placed epithelial side down onto a glass bottom dish and light was delivered using an epi-illumination scheme. 10 to 60 mW of power were delivered at the tissue surface with power being varied linearly through the depth of the tissue during acquisition. The rate of power increase was determined based on achieving minimal pixel saturation in the most superficial and most basal OSs. The 60 mW threshold was motivated by the finding that this power equates to a 0.6 minimal erythma dose. The threshold of sunburn development is a 1.0 minimal erythma dose.⁴² The maintenance of safe excitation power levels supports the translational relevance of the present dataset. Images were acquired using bidirectional scanning with a 600 Hz line scan frequency and a 400 ns pixel dwell time. Tissue biopsies were excited with 755 and 860 nm light. Two hybrid photodetectors were set up to collect the two-photon autofluorescence signal from NAD(P)H (460 ± 25 nm) and FAD (525 ± 25 nm). Hybrid photodetectors were used in place of photomultiplier tubes due to their increased sensitivity to low intensity fluorescence signal compared with photomultiplier tubes.^{43,44} Two photomultiplier tubes were set up to collect the second harmonic generation signal from collagen fibers (430 ± 12 nm) and the red autofluorescence signal (624 ± 20 nm). Light was delivered and collected using a 40X/1.1 NA water-immersion objective lens (290×290 - μm^2 field-of-view). Images were collected through the full thickness of the epithelium using a depth-sampling rate of 4 μm . Six individual frames were collected at each depth. On average, 3 to 5 ROIs were sampled from each biopsy.

2.4 Morphological and Functional Metrics

Images were calibrated and processed as described in detail previously to extract images that represented NAD(P)H and FAD TPEF intensity fluctuations.^{23–25,45} At each optical depth,

NAD(P)H and FAD images were used to define a corresponding redox ratio for each pixel of the field, given as

$$\text{Optical Redox Ratio (RR)} = \frac{\text{FAD}}{(\text{FAD} + \text{NAD(P)H})}. \quad (1)$$

From the RR distributions for each OS, we calculated the mean RR and the interquartile range (IQR) as metrics of the overall oxidation-reduction tissue state and the corresponding heterogeneity, respectively. The mean and sample variance (variability) of the mean OS RR and the OS RR IQR for all images in an epithelial stack were calculated to assess the depth-dependence of these metrics.

NAD(P)H images were analyzed as described previously^{17,18,21,22} using a Fourier based approach to extract a value for the parameter β as a metric of the level of mitochondrial fragmentation and networking, which also depends highly on the metabolic activity of the tissue. Briefly, an inverse power law was fit to the power spectral density (PSD) of the two-dimensional Fourier transform of the cytoplasmic NAD(P)H intensity fluctuation images, given as

$$R(k) = Ak^{-\beta}, \quad (2)$$

where R is the fit to the PSD, k is the magnitude of the spatial frequency, β is the power law exponent, and A is a constant. The mean and sample variance of β were assessed as a function of depth for each image stack.

2.5 Deep Learning Model Description

The basic structure of the CARE network has been described extensively [Fig. S3(a) in the [Supplementary Material](#)].³⁴ The network was implemented through Keras and TensorFlow.^{46,47} A copy of the CSBDeep repository (available in a Github repository: <https://github.com/CSBDeep/CSBDeep>) was locally imported into an anaconda environment.⁴⁸ The network was configured to take a $256 \times 256 \times 1$ input image and generate a $256 \times 256 \times 1$ denoised image. A 40-gigabyte Nvidia Tesla A100 GPU card was used for all training and evaluation. Typically, a $1 \times 512 \times 512 \times z$ -depth image stack was split into $4 \times 256 \times 256 \times z$ -depth image patches before training using a 2×2 grid as previously described.⁴⁹ A starting learning rate of 1×10^{-5} was used with an Adam optimizer.⁵⁰ Training was allowed to continue for 300 epochs with a scheduler reducing the learning rate when the network performance stagnated for more than 20 epochs. Early stopping was not implemented to allow the model to improve with lower learning rates. Model improvement was generally observed to stagnate after 75 to 100 epochs with the best model weights being saved. The loss functions were varied to find the optimal function to improve the downstream analysis performance. The loss functions used include SSIM loss, R2 loss, FFL, MAE (L1) loss, MSE (L2) loss, Redox ratio loss, and combined losses, such as a combined SSIM + L2, SSIM + FFL, SSIM + R2 loss.⁵¹ Combined loss functions were weighted using an α term to control the contribution of SSIM loss and L2, FFL, R2 loss. An $\alpha = 0.84$ was used in this study, in line with other studies, to balance the contribution of both loss functions.⁵² Six down-sampling and up-sampling layers were generated with the first layer expanding the single-channel images to 32 channels. Residual connections were used to preserve encoded information from each down sampled layer and pass it forward to the decoder layers (see Fig. S3 in the [Supplementary Material](#)).

For the wavelet U-Net (WU-net) architecture, four CARE networks, one per sub-band, were built as described above. A discrete wavelet transform (DWT) was used to decompose a $1 \times 256 \times 256$ optical section patch (OSP) into $1 \times 128 \times 128 \times 4$ frequency band images. The four frequency bands would then be individually input to each CARE network for denoising. After denoising, an inverse DWT (iDWT) was used to reconstruct the $1 \times 256 \times 256$ OSP (for greater detail see Fig. S4 in the [Supplementary Material](#)).

The training time typically varied from 1 to 2 h, with an evaluation time of ~ 24 s per image stack. For all trained CARE networks, 3D SSIM, PSNR, mean β , β variability, mean RR, RR variability, RR IQR, and RR IQR variability were analyzed. All final metrics were assessed using a single frame input, denoised, and ground truth (six frame averaged) images with built-in and custom MATLAB (MathWorks; Natick, Massachusetts) functions.

2.6 Statistics

For Figs. 3–5(c) and Table 7, the Fisher r -to- z transformation was used to convert Pearson's correlation coefficients (r) to z_r values.⁵³ This transform was calculated as

$$z_r = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right). \quad (3)$$

The z_r value, unlike r , belongs to a normal distribution, allowing for the calculation of a Z -statistic to determine confidence intervals. The test Z -statistic for comparison of z_r values to determine significance was calculated as

$$Z_{\text{test}} = \frac{z_{r1} - z_{r2}}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}}, \quad (4)$$

where n_1 and n_2 are the sample size of r_1 and r_2 , respectively.⁵⁴ The Z_{test} value was then compared to the critical Z -values to determine the significance and p -values using a two-tailed distribution.

3 Results

3.1 Identification of the Optimal Deep-Learning Model Architecture for Denoising Label-Free, Optical TPEF Images to Enable Recovery of Metabolic Function Metrics

Human cervical tissue biopsies were collected from 54 patients and imaged immediately upon excision, as described in Sec. 2.3 (Fig. 1). Several ROIs were imaged from each biopsy. Multiple OSs were imaged from each ROI at distinct depths. At each OS, we acquired TPEF images at a combination of two excitation wavelengths (755 and 860 nm) and three or four emission bands. Images collected at 755 nm excitation and 435 to 485 nm emission were attributed primarily to NAD(P)H, whereas images at 860 nm excitation and 500 to 550 nm emission were considered to contain signal primarily from FAD and FAD bound to lipoamide dehydrogenase. Six frames were acquired at each wavelength setting. To reduce the contribution of noise, these six frames were averaged together. The decision to use six frames was made based on data collected from freshly excised rat oral squamous epithelial tissue. Analysis of this dataset highlighted that six-frame images conveyed quite accurately the depth dependence of the metabolic function metrics when compared with the analysis of 32 frame-averages (see Fig. S6 in the [Supplementary Material](#)). Metrics extracted from these averaged images were previously observed to enable highly sensitive and specific detection of cervical pre-cancer.²⁵ The averaged image was therefore considered the ground truth used for training and testing the denoising success of single frames. Single frames, the corresponding denoised images, and ground truth images were analyzed using established procedures to extract the RR and mitochondrial clustering (β) (Fig. 1). All models (Fig. 1) were trained and evaluated with identically generated image stacks. Various combinations of model architectures, loss functions, data transformations, and training data combinations, as outlined in Table 1, were evaluated on 3229 total OSs (1657 training OSs + 554 validation OSs + 1018 testing OSs) representing healthy/benign cervical tissues as well as precancerous (low-grade and high grade) squamous intraepithelial lesions (LSIL and HSIL, respectively).

PSNR and SSIM improvements are standard metrics of image visual quality and have been used in other studies focused on denoising biomedical images as indicators of model success.^{34,35,37,38} We aim to assess whether images restored by models that yield optimized PSNR and SSIM values result in accurate recovery of metabolic metrics (Fig. 2). For evaluation of the model architecture, loss function, and signal type, only results from models trained on NAD(P)H data from tissues of known benign status were included.

Leading denoising model architectures were selected for evaluation based on a comprehensive literature search. CARE, RCAN, and SRGAN (Fig. 1) models were trained as described in Sec. 2.5 and Supplementary Methods: Deep Learning Performance Benchmark in the [Supplementary Material](#). A representative OS from an LSIL biopsy is shown in Fig. 3(a). The results shown were generated by models trained using an SSIM + MSE (or L2) loss function. A summary of all parameters used to generate the figures and tables is listed in Table 2.

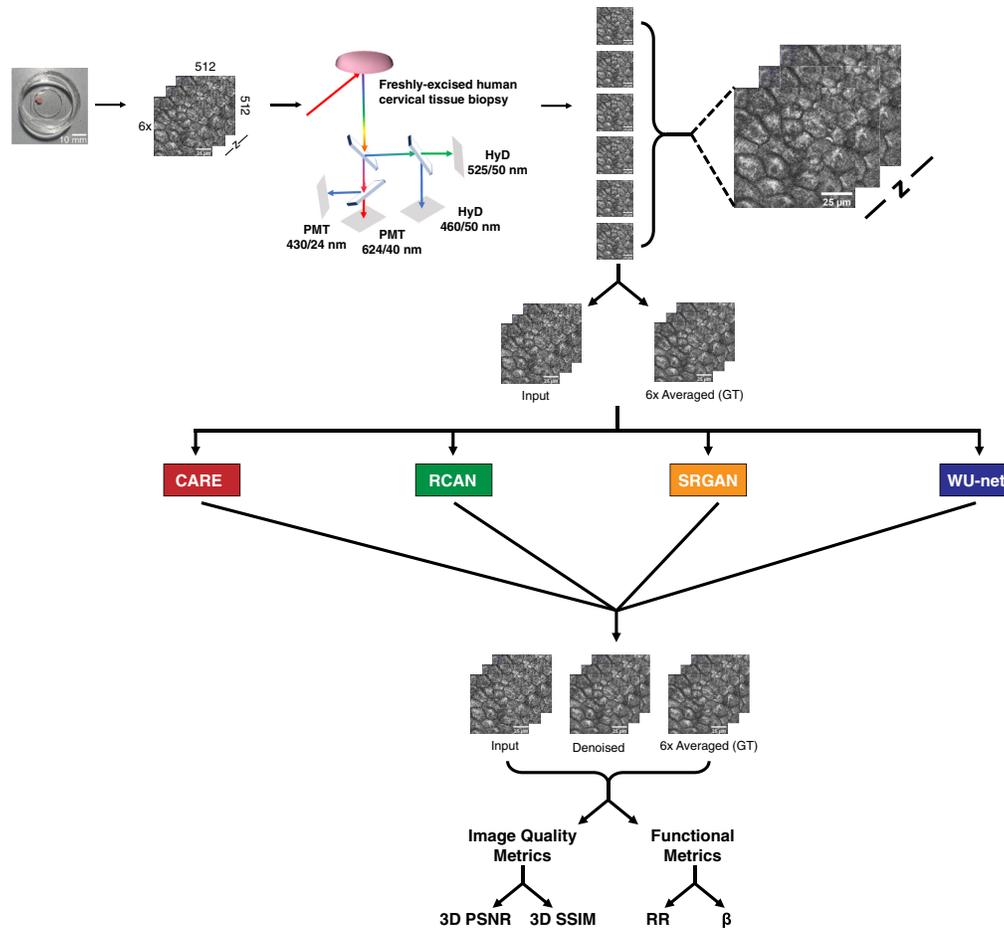


Fig. 1 Summary of deep learning pipeline. Human cervical tissue biopsies are collected and subsequently imaged within 4 h post-excision. Collected biopsies are plated on glass bottom dishes and imaged using a Leica SP8 commercial microscope. At a minimum, three ROIs are imaged per sample. At each ROI, multiple OSs are imaged at distinct depths through the epithelium. Depth-resolved, two-photon OSs are collected using two excitation wavelengths and several bandpass-filtered detectors. Six images are captured for each excitation/emission wavelength and every OS at a given depth, z . These six images are averaged together to generate the ground truth image set. A random image from the six per depth z is selected as the input (RAW) image. The paired image stacks are provided to the neural network for training and denoising. Four-leading denoising networks are used in this study to denoise input images: a previously described CARE model, an RCAN model, an SRGAN model, and a WU-Net.^{34–36,38} Denoised images and input images are compared against 6× averaged images to determine 3D PSNR and SSIM along with metabolic metrics. Scale bar = 25 μm .

Table 1 Summary of all parameters explored during training and optimization of the final model (highlighted in bold). Results shown below are focused on the optimized model, but all combinations were trained and evaluated.

| | | | | | | |
|-------------------------------------|--------------------------|-----------------------------------|---------|-----------|------------|------------------|
| Model architecture | CARE | RCAN | SRGAN | — | — | — |
| Loss functions | MAE L1 | MSE L2 | RR loss | SSIM + L2 | SSIM + FFL | SSIM + R2 |
| Signal pre-processing method | Wavelet transform | None | — | — | — | — |
| Training data format | Healthy data only | Healthy and diseased (mix) | — | — | — | — |
| Training data type | NAD(P)H data | FAD data | — | — | — | — |

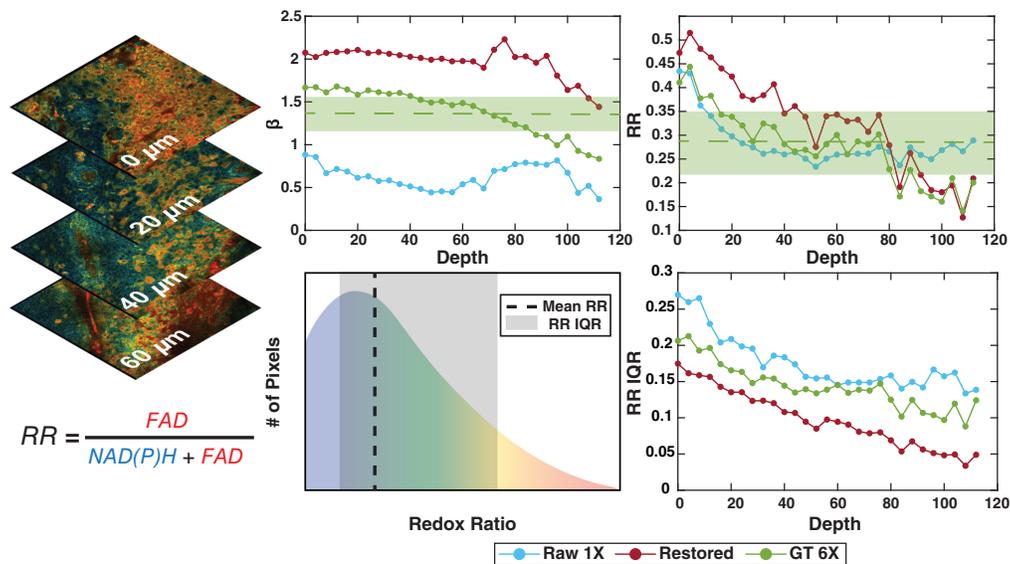


Fig. 2 β and RR metrics are extracted from each OS (see Sec. 2.4 for greater detail). Depth-dependent trends across the multiple cell layers of the cervical squamous epithelium are assessed for input images (RAW 1X), denoised images (restored) and six-frame averaged, ground truth images (GT 6X). Measurements of mean values and corresponding variability across all depths are shown as a dashed line and shaded region in the mitochondrial clustering, β , and RR (FAD/(NAD(P)H+FAD)) panels for the GT 6X image. The distribution of RR values for each OS is used to extract the IQR, representing the range of values within the 25% and 75% of the RR distributions and providing an assessment of intra-field RR heterogeneity. IQR variability is a metric of inter-field (depth-dependent) RR heterogeneity.

Prior to denoising, standard image quality metrics were calculated for input (RAW 1X) images by comparing the RAW 1X images to ground truth (GT 6X) images. PSNR and SSIM values were calculated using the GT 6X image as a reference and RAW 1X or denoised images as the distorted image.⁵⁵ Across all images, FAD image PSNR was greater than NAD(P)H image PSNR (Table 3), even though FAD images featured lower cytoplasmic signal compared with NAD(P)H images [Fig. 3(a)]. During PSNR calculation, the reduced signal intensity led to smaller differences between RAW 1X and GT 6X images and yielded a greater observed PSNR value. This observation was also consistent with the results from other studies.³⁷ SSIM values were consistent between NAD(P)H and FAD images (Table 3). Corruption of the GT 6X images for both channels by noise was expected to have similar effects on structural similarity and calculated SSIM values.

We used 777 and 109 RAW 1X NAD(P)H OSs for training and validation of the models, respectively. Each 512×512 OS was patched into four- 256×256 image patches (OSP) prior to training and validation (3108 and 436 OSPs, respectively). All three models were trained before being evaluated on an independent set of 2343 OSs (9372 OSPs). Metrics of image quality and metabolic function were calculated as described in Supplementary Methods: Deep Learning Metrics in the [Supplementary Material](#) and Sec. 2.4.

CARE-generated image stacks demonstrated higher PSNR for FAD images and higher SSIM for both NAD(P)H and FAD images compared with restored-image stacks generated by RCAN and SRGAN. Across all test set images, standard metrics of image quality (Table 3) and visual inspection [Fig. 3(b)] suggested the RCAN- and CARE-denoised images had similar image quality. Across the entire test set, we observed that SRGAN failed to restore cellular features in the RAW 1X that are found within the GT 6X images [Fig. 3(b)] and underperformed even relative to RAW 1X images in standard image quality metrics (Table 3). Perceptual loss was believed to impact content restoration in the SRGAN architecture.³⁵ Inputs for perceptual loss calculations have been shown to impact significantly the SRGAN performance and were likely the cause of SRGAN's poor recovery of the image quality.³⁵

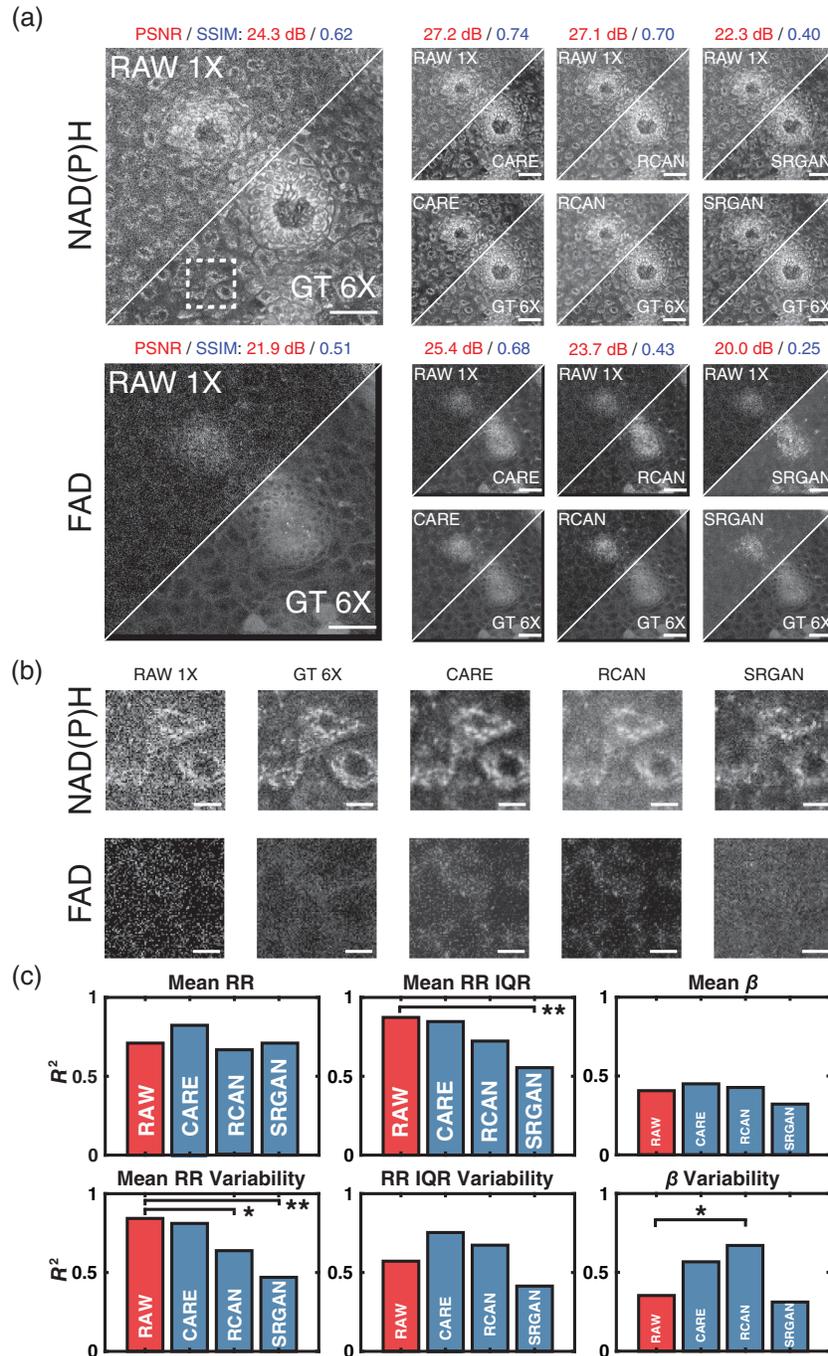


Fig. 3 (a) A $290 \times 290 \mu\text{m}^2$ field of view from an LSIL cervical tissue biopsy. NAD(P)H and FAD images for the same region are shown along with the corresponding denoised image from each of the three trained models (CARE, RCAN, and SRGAN). Scale bar = $50 \mu\text{m}$. (b) A $44.2 \times 44.2 \mu\text{m}^2$ field of view (white square in a) of three cells. NAD(P)H and FAD images are shown for all models and the input and ground truth images. Scale bar = $10 \mu\text{m}$. (c) Bar plots of the coefficient of determination of all downstream metrics for images denoised by all models and RAW 1X versus the GT 6X image. Fisher r to z transformation was used to measure significance. * $p < 0.05$ and ** $p < 0.01$.

To assess the restoration of metabolic activity, depth-dependent optical RR and mitochondrial clustering (β) values were calculated for the restored images, input (RAW 1X) images, and ground truth (GT 6X) images (Fig. 2). Perturbations in metric values are reflective of changes in the metabolic functional state of the different cell layers in tissue.^{21,22,24,25,39,40} Pearson correlation coefficient values were calculated between the metabolic function metrics from the GT 6X and

Table 2 Summary of parameters used to generate Figs. 3–6 and Tables 3–6. Parameters are bolded when all combinations from Table 1 are used.

| | Model architecture | Loss function | Preprocessing method | Training data format | Training data type |
|----------------|--------------------|-------------------|----------------------|----------------------|--------------------|
| Fig. 3/Table 3 | All | SSIM + L2 loss | None | Healthy only | NAD(P)H |
| Fig. 4/Table 4 | CARE | SSIM + R2 | All | Healthy only | NAD(P)H |
| Fig. 5/Table 5 | CARE | SSIM + R2 | Wavelet transform | All | All |
| Fig. 6/Table 6 | CARE | SSIM + R2 MSE MAE | All | All | All |

Table 3 Summary of standard metrics of image quality for RAW 1X images and denoised images generated from various model architectures. Values are reported for mean performance (\pm standard deviation) across all test set ROIs.

| Model name | NAD(P)H images | | FAD images | |
|------------|----------------------------------|-----------------------------------|----------------------------------|-----------------------------------|
| | PSNR (dB) \uparrow | SSIM \uparrow | PSNR (dB) \uparrow | SSIM \uparrow |
| RAW 1X | 19.2 \pm 2.8 | 0.48 \pm 0.09 | 23.1 \pm 5.5 | 0.49 \pm 0.13 |
| CARE | 22.7 \pm 2.9 | 0.63 \pm 0.08 | 26.8 \pm 3.1 | 0.60 \pm 0.07 |
| RCAN | 23.1 \pm 1.7 | 0.62 \pm 0.08 | 24.3 \pm 2.0 | 0.51 \pm 0.12 |
| SRGAN | 19.6 \pm 1.1 | 0.31 \pm 0.08 | 20.2 \pm 1.5 | 0.25 \pm 0.07 |

Note: bold values indicate the greatest metric performance.

either the RAW 1X or restored images. Statistical significance was derived from Fisher- r -to- z transformation for all metrics of interest. Interestingly, analysis of the RAW 1X images led to very high correlations with metrics of RR intra- and inter-field variability compared with GT 6X images. We hypothesized that similar sources of noise in both FAD and NAD(P)H images led to this outcome because RR metrics were calculated using a ratio of FAD and NAD(P)H intensity measurements. It was for this reason that, in this initial comparison, we trained models on NAD(P)H images and applied the weights to NAD(P)H and FAD images. RCAN-generated images demonstrated a statistically significant recovery of β variability ($\sigma^2(\beta)$) [Fig. 3(c)]. However, recovery of mean RR variability by this model was poor [Fig. 3(c)]. CARE-denoised images, overall, featured higher (albeit not statistically significant) correlations with RR metrics compared with all other models [Fig. 3(c)]. Thus, the U-Net architecture of CARE was utilized for all further optimization steps.

3.2 Multiscale Image Transformation Enhances Quantification of Mitochondrial Clustering

Although denoising improved the restoration of $\sigma^2(\beta)$, the mean β ($\bar{\beta}$) values of the denoised images were not well correlated with the values from the GT 6X images. We considered DWT to enhance high spatial frequency restoration necessary for β metric calculations. A single level DWT transformed each image into four sub-band images: a coarser scale approximation and three detail images, one horizontal, one vertical, and one diagonal.⁵⁶ To generate the three sub-band images, a basis function, called a mother wavelet, was convolved along both dimensions of the original image, and an associated scaling function was used to generate the coarser approximation. During standard wavelet-based denoising, thresholds are used to remove noise from wavelet-transformed detail images, before implementing an inverse-transform to recover the restored image.^{57,58} The DWT has been shown to be advantageous compared with traditional low-pass filtering as the pixel-by-pixel convolution with the mother wavelet preserves correlations of high frequency features. In this study, we used deep learning models trained on each of the transformed images to adaptively learn the best threshold for denoising of low frequencies

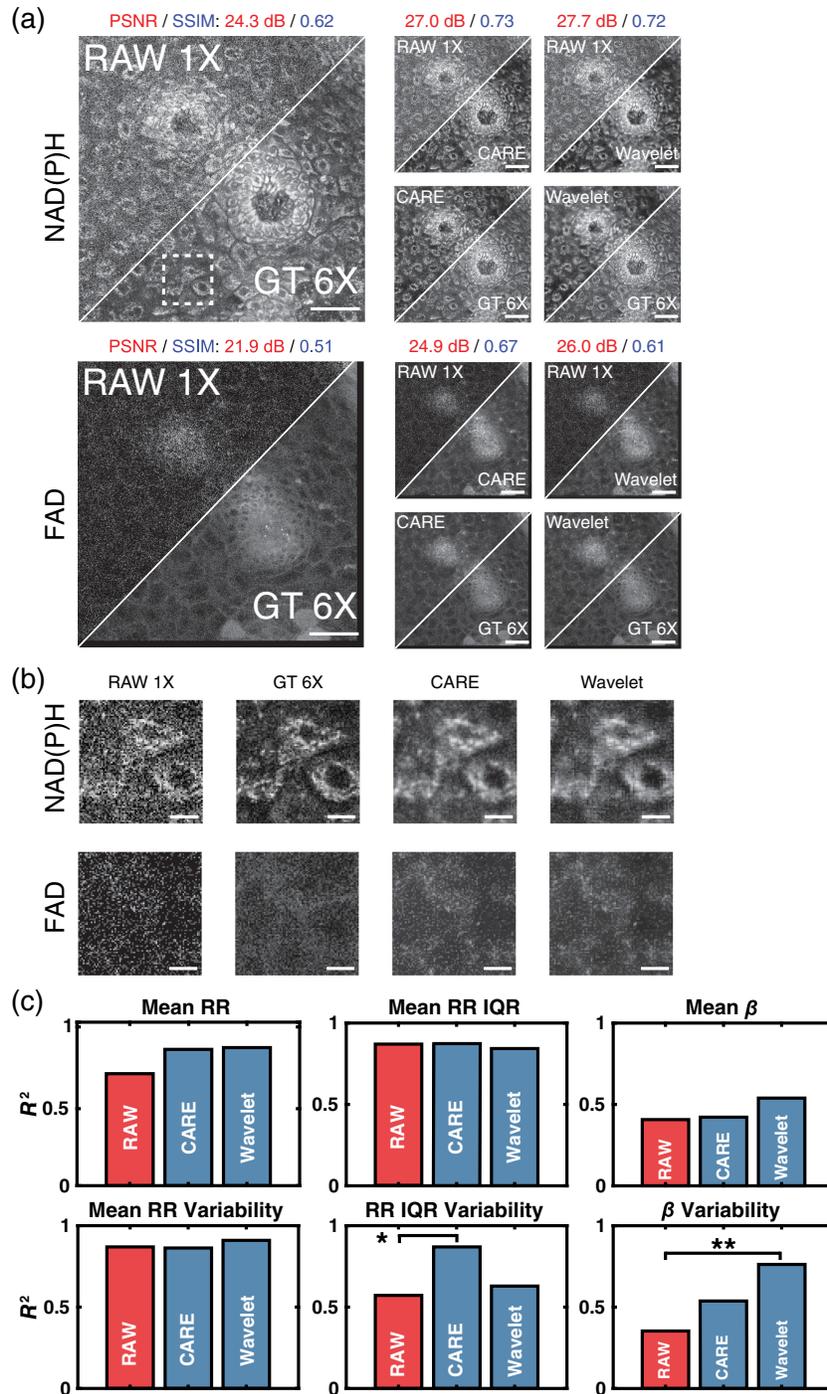


Fig. 4 (a) A $290 \times 290 \mu\text{m}^2$ field of view from an LSIL cervical tissue biopsy. NAD(P)H and FAD images for the same region are shown along with the corresponding denoised image from each signal pre-processing method utilized (single frame image and wavelet transformation). Scale bar = $50 \mu\text{m}$. (b) A $44.2 \times 44.2 \mu\text{m}^2$ field of view (white square in a) of three cells. NAD(P)H and FAD images are shown for all models based on the corresponding signal pre-processing method used during training and the input and ground truth images. Scale bar = $10 \mu\text{m}$. (c) Bar plots of the coefficient of determination of all downstream metrics for images denoised by all models trained based on the corresponding signal pre-processing method used during training and RAW 1X versus the GT 6X image. Fisher r to z transformation was used to measure significance. $*p < 0.05$ and $**p < 0.01$.

Table 4 Summary of standard metrics of image quality for RAW 1X images and denoised images generated after signal pre-processing. Values are reported for mean performance (\pm standard deviation) across all test set ROIs.

| Model name | NAD(P)H images | | FAD images | |
|------------|----------------------------------|-----------------------------------|----------------------------------|-----------------------------------|
| | PSNR (dB) \uparrow | SSIM \uparrow | PSNR (dB) \uparrow | SSIM \uparrow |
| RAW 1X | 19.2 \pm 2.8 | 0.48 \pm 0.09 | 23.1 \pm 5.5 | 0.49 \pm 0.13 |
| CARE | 22.8 \pm 3.0 | 0.63 \pm 0.08 | 27.2 \pm 3.7 | 0.62 \pm 0.07 |
| Wavelet | 23.6 \pm 2.3 | 0.63 \pm 0.08 | 26.1 \pm 2.3 | 0.57 \pm 0.09 |

Note: bold values indicate the greatest metric performance.

(approximation) and high frequencies (details) rather than relying on arbitrary thresholding for denoising (see Discussion S2 in the [Supplementary Material](#)).⁵⁹ As with any DWT-denoising model, the selection of the correct mother wavelet played a significant role in model performance. For all models, mother wavelets from the biorthogonal, coiflets, and Daubechies families were evaluated. These mother wavelet families were selected due to their frequent use in denoising tasks.⁶⁰ Multiple models were trained and evaluated, with biorthogonal 1.1 yielding the highest recovery of metabolic metrics (data not shown). As such, biorthogonal 1.1 was used for all subsequent model optimization.

The application of DWT before training four CARE models and iDWT after evaluation yielded images with improved FAD and NAD(P)H PSNR with slight decreases in SSIM [Fig. 4(a)]. Across the entire test set, NAD(P)H PSNR improved using WU-Net, whereas FAD PSNR and SSIM both decreased compared with CARE (Table 4). All loss functions were evaluated for WU-Net with SSIM + R2 loss (results shown in Fig. 4) and SSIM + FFL loss (see Table S4 in the [Supplementary Material](#)) yielding the best overall performance. WU-Net denoised NAD(P)H images extracted similar cellular structures as the CARE derived images but featured a lower background signal and small fluctuations in cytoplasmic signal, leading to the observed higher PSNR values [Fig. 4(b)]. WU-Net led to statistically significant improvements in the correlation of extracted σ^2 (β) with GT 6X images relative to the analysis of the RAW 1X images. Extracted $\bar{\beta}$ values were also better correlated to GT 6X images, albeit improvements were not significant.

Comparing WU-Net with an identical CARE model, we observed that WU-net achieved an improved performance on β metrics while maintaining the recovery of RR metrics [Fig. 4(c)]. The overall improved β restoration suggested that WU-Net was better able to capture true signals from noise in the high spatial frequencies found in NAD(P)H images. WU-Net further preserved the relationship between NAD(P)H and FAD channel images, enabling high correlations for RR metrics. Due to the observed performance of WU-Net for β metric recovery, we explored further optimization of WU-Net, which could be achieved by varying the training datasets.

3.3 Selection of Training Data

Initial model development focused on a limited training set of cervical tissues of known benign status (healthy). Benign tissue samples comprise cell layers with consistent changes in differentiation as a function of depth among image stacks. Training on such images was expected to enable the model to learn characteristics of noise without having to account for feature heterogeneity found in pre-cancerous cervical tissue samples. We further sought to assess whether training on a data set that was expanded to include image stacks from tissues with both benign and pre-cancerous lesions (Mix) impacted performance. In this new training set, 1657 and 554 RAW 1X NAD(P)H OSs (6628 and 2216 OSPs) were used for training and validation of the models, respectively. An independent test set of 1018 OSs (4072 OSPs) was used to evaluate the model performance after training.

An additional consideration that we explored was the impact of the source of image contrast, i.e., NAD(P)H or FAD, used for training. NAD(P)H images featured greater structural information compared with FAD images, and they were utilized in our analysis for the extraction of

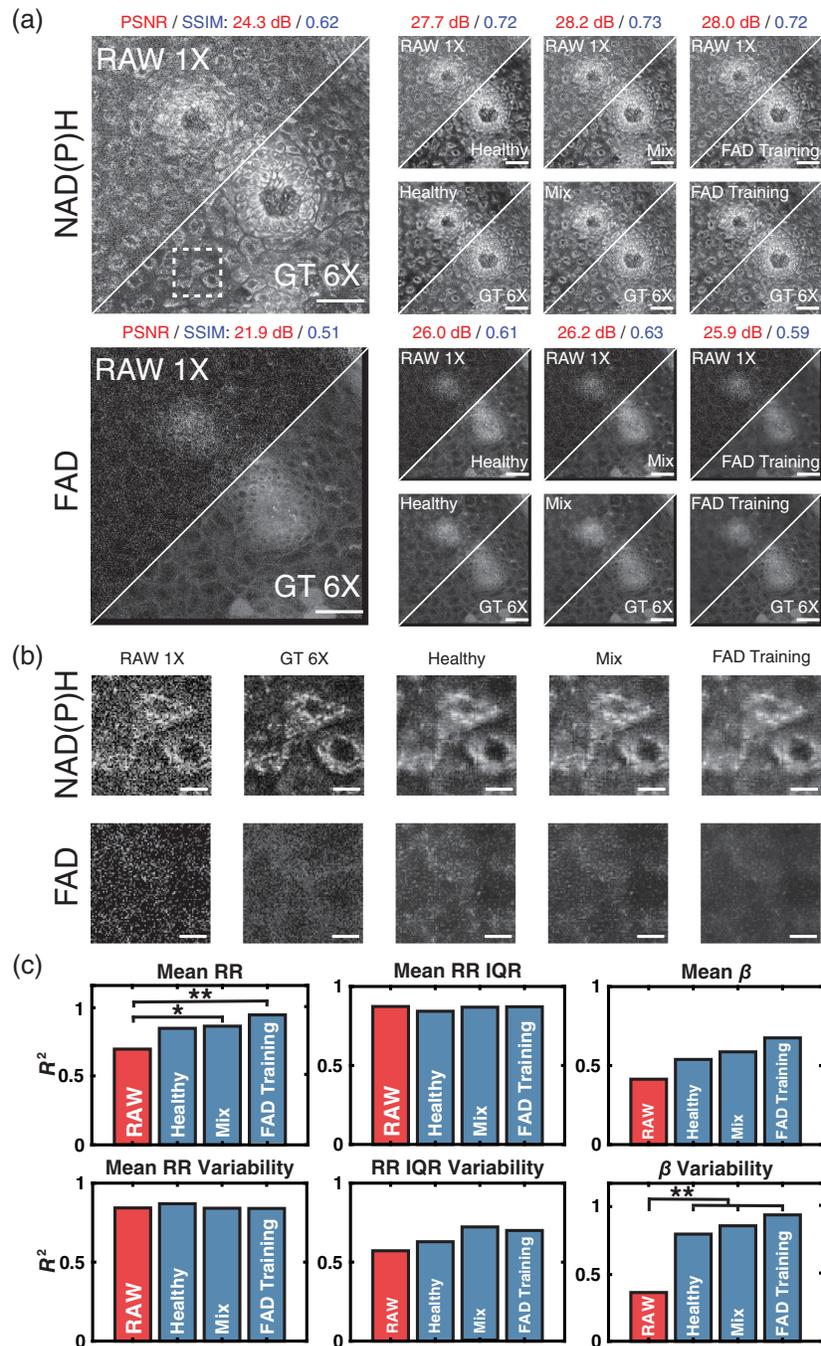


Fig. 5 (a) A $290 \times 290 \mu\text{m}^2$ field of view from an LSIL cervical tissue biopsy. NAD(P)H and FAD images for the same region are shown along with the corresponding denoised image from each data type used as training data for the WU-Net model (NAD(P)H healthy only-, NAD(P)H mixed diagnosis-, and FAD mixed diagnosis-wavelet transformed images). All models were equally constructed with only the data type and diagnosis type varied. Scale bar = $50 \mu\text{m}$. (b) A $44.2 \times 44.2 \mu\text{m}^2$ field of view (white square in a) of three cells. NAD(P)H and FAD images are shown for all data types used during training and the input and ground truth images. Scale bar = $10 \mu\text{m}$. (c) Bar plots of the coefficient of determination of all downstream metrics for images denoised by models trained using varying data types and diagnosis types and RAW 1X versus the GT 6X image. Fisher r to z transformation was used to measure significance. * $p < 0.05$ and ** $p < 0.01$.

Table 5 Summary of standard metrics of image quality for RAW 1X images and denoised images generated after training models on various data types. Values are reported for mean performance (\pm standard deviation) across all test set ROIs.

| Model name | NAD(P)H images | | FAD images | |
|---------------|----------------------------------|-----------------------------------|----------------------------------|-----------------------------------|
| | PSNR (dB) \uparrow | SSIM \uparrow | PSNR (dB) \uparrow | SSIM \uparrow |
| RAW 1X | 19.2 \pm 2.8 | 0.48 \pm 0.09 | 23.1 \pm 5.5 | 0.49 \pm 0.13 |
| Healthy | 23.6 \pm 2.3 | 0.63 \pm 0.08 | 26.1 \pm 2.3 | 0.57 \pm 0.09 |
| Mixed NAD(P)H | 23.4 \pm 2.5 | 0.63 \pm 0.08 | 26.3 \pm 3.1 | 0.57 \pm 0.08 |
| Mixed FAD | 23.5 \pm 2.6 | 0.62 \pm 0.09 | 24.8 \pm 3.7 | 0.52 \pm 0.08 |

Note: bold values indicate the greatest metric performance.

mitochondrial clustering-focused metabolic function metrics (Figs. 3–5). Thus, training was focused on NAD(P)H images, and optimized model weights from NAD(P)H image training were used to denoise FAD images for extraction of RR-based metrics. However, because similar noise characteristics were assumed to be present in both RAW 1X NAD(P)H and FAD images, we sought to confirm that training on NAD(P)H images was optimal. Thus, we used FAD images to train WU-Net models using the same hyperparameters and settings as the ones used when NAD(P)H images were used. Post-training, NAD(P)H images were denoised using the weights of the FAD image trained model to extract RR and mitochondrial clustering-based metrics.

The use of training sets with mixed diagnosis images resulted in minimal differences in the denoised images when compared with training just on healthy sample images [Fig. 5(a)]. PSNR and SSIM values for images were observed to be nearly identical because of these insignificant differences (Table 5). Both models led to denoised images with consistent cell boundary and intracellular structures given the same RAW 1X images [Fig. 5(b)] and had similar levels of restoration of downstream metrics, with the mixed diagnosis dataset leading to slightly improved correlations in most cases [Fig. 5(c)]. The increase in correlation could be attributed to the large training set available for a mixture of diagnoses compared with only training on healthy data.

An identical model was trained using the FAD image data from the mixed diagnosis dataset. The denoised images from the FAD-trained model looked like those from the corresponding NAD(P)H-trained model [Figs. 5(a) and 5(b)]; however, the standard metrics of image quality were slightly lower. Images denoised by the FAD-trained model demonstrated a higher background signal compared with images denoised by NAD(P)H-trained models [Fig. 5(b)]. However, despite FAD images lacking much of the structural and morphological information of their NAD(P)H counterparts, their use in training led to further improvements in β metric recovery and mean RR restoration from the RAW 1X images [Fig. 5(c)]. We hypothesize that high frequency information in the FAD images originated primarily from noise in comparison with NAD(P)H images. As a result of the high frequency information containing primarily noise, the model improved in its learning of noise characteristics in the images, enabling improved denoising and recovery of metrics of metabolic activity [Fig. 5(c)].

3.4 Summary of Final Model Performance

Across all models, image quality improved after denoising based on PSNR and SSIM (Table 6). Based on standard image quality metrics of all models discussed in this study, it could be assumed that models trained using NAD(P)H images and the CARE architecture with standard loss functions of MAE and MSE would perform best at the restoration of downstream metrics [Fig. 6(a)]. CARE models trained with MAE and MSE loss functions both demonstrated statistically significant improvement in denoised FAD and NAD(P)H image PSNR and SSIM ($p < 0.001$). Comparatively, wavelet-transformed-FAD images denoised using WU-Net with SSIM + R2 loss had poorer standard metric performance (Table 6). Images restored with this model did not achieve statistically significant improvement of FAD image PSNR and SSIM [Fig. 6(a)]. As PSNR and SSIM are commonly used as indicators of model performance, it was

Table 6 Summary of standard metrics of image quality (PSNR and SSIM) for RAW 1X images, standard implementation of CARE, and the best performing model from this study. Values are reported for mean performance (\pm standard deviation).

| Model name | NAD(P)H images | | FAD images | |
|--|----------------------------------|-----------------------------------|----------------------------------|-----------------------------------|
| | PSNR (dB) \uparrow | SSIM \uparrow | PSNR (dB) \uparrow | SSIM \uparrow |
| RAW 1X | 19.2 \pm 2.8 | 0.48 \pm 0.09 | 23.1 \pm 5.5 | 0.49 \pm 0.13 |
| Healthy NAD(P)H CARE MAE | 23.6 \pm 2.3 | 0.63 \pm 0.08 | 26.9 \pm 2.7 | 0.59 \pm 0.08 |
| Healthy NAD(P)H CARE MSE | 23.7 \pm 2.6 | 0.64 \pm 0.08 | 26.8 \pm 2.7 | 0.59 \pm 0.08 |
| Mixed FAD CARE SSIM + R2 wavelet transform | 23.5 \pm 2.6 | 0.62 \pm 0.09 | 24.8 \pm 3.7 | 0.52 \pm 0.08 |

Note: bold values indicate the greatest metric performance.

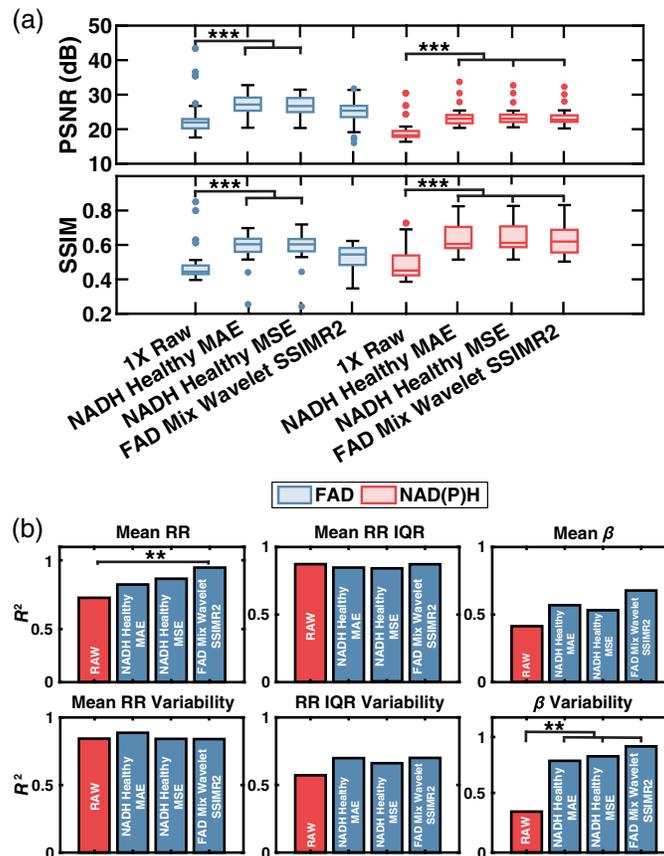


Fig. 6 (a) Box and whisker plots of PSNR and SSIM of 40 test set ROIs. Denoised images demonstrated an improvement in standard metrics of image quality. (b) Bar plots of the coefficient of determination of all downstream metrics for images denoised by models trained using various data types, loss functions, and diagnosis types versus the GT 6X image. A one-way ANOVA with Tukey Kramer post-hoc test was used to measure significance of PSNR and SSIM. Fisher r to z transformation was used to measure significance of improvement in metabolic metric correlations $^*p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$.

expected that improvements in these metrics would correspond to better recovery of downstream metabolic metrics. However, the WU-net model trained on mixed diagnosis, FAD images led to denoised images with extracted metabolic metrics that were consistently correlated with the metrics extracted from GT 6X images [Fig. 6(b)]. The final correlations of the models shown in Fig. 6 are reported in Table 7.

Table 7 Correlation values of models in Fig. 6. Fisher r to z transformation was used to measure significance.

| Downstream metrics | | | | | | |
|--|--------------------|---------------------------------|------------------------|-------------------------------------|-------------------------|---------------------------------------|
| Final model | Mean RR \uparrow | σ^2 (mean RR) \uparrow | Mean RR IQR \uparrow | σ^2 (mean RR IQR) \uparrow | Mean β \uparrow | σ^2 (mean β) \uparrow |
| RAW 1X | 0.71 | 0.84 | 0.87 | 0.57 | 0.40 | 0.33 |
| Healthy NAD(P)H CARE MAE | 0.82 | 0.89 | 0.85 | 0.70 | 0.57 | 0.78* |
| Healthy NAD(P)H CARE MSE | 0.87 | 0.84 | 0.84 | 0.66 | 0.53 | 0.81* |
| Mixed FAD CARE SSIM + R2 wavelet transform | 0.96* | 0.84 | 0.87 | 0.70 | 0.68 | 0.90* |

Note: bold values indicate the greatest metric performance.

* $p < 0.01$.

4 Discussion

Tissue morphological and functional metrics extracted from label-free, two-photon microscopy (2PM) images could provide significant clinical utility for disease diagnosis.²⁵ Neural networks will likely play a critical role in enabling the accurate extraction of such metrics from images that are likely to be acquired in an *in vivo* imaging setting. Previous studies by multiple groups have demonstrated that deep learning-based denoising models can be used to improve the PSNR and SSIM of fluorescence images acquired using 2PM.^{34,36,37} Here, we demonstrated that PSNR and SSIM, although relevant in the assessment of image quality, were not representative of functional metric recovery needed for clinical utility (Fig. 6).

Different algorithms have been reported for denoising of fluorescence images; however, only Shen et al. reported a network used for denoising of label-free autofluorescence images.³⁷ In this study, a modified enhanced SRGAN model was used to denoise *ex-vivo*, multi-modal label-free images of human ovarian cancer tissue sections.³⁷ The trained GAN demonstrated a 4.5 dB improvement in PSNR and a 79% improvement in SSIM after denoising.³⁷ In comparison, we demonstrated 4.3 and 2.7 dB improvements in PSNR and 30% and 6% improvements in SSIM for NAD(P)H and FAD images, respectively [Fig. 6(a)]. Although improvement in image PSNR and SSIM were lower, RAW 1X and denoised images in this study have higher PSNR and SSIM for all images, suggesting that the differences in enhancement are due to limits in image improvement and not a lack of network performance (Table 6).

We further observed that GAN models did not perform well on our dataset. GANs aim to emulate characteristics of high SNR images in low SNR images through an adversarial training process.³⁵ To improve the image quality, GANs learn the manifold of high SNR data, which is assumed to be composed of images that have similar image quality metrics.⁶¹ Thus, it is important for image quality to be consistent across all high SNR images. High-SNR images from a single depth in a thin OS, such as those used to train the enhanced SRGAN model in Shen et al., have similar image quality for all ground truth images, leading to improved GAN performance.³⁷ In our study, bulk tissues were imaged at multiple depths, leading to inconsistent image quality in our ground truth images as SNR is known to change as a function of depth. As such, we hypothesize that the GAN model implemented in this study failed to learn the manifold of high SNR images and improve our images, whereas the enhanced SRGAN model implemented by Shen et al. succeeded.

Although multiple studies demonstrate models capable of improving PSNR and SSIM, the assessment of morphofunctional metrics of metabolic activity after denoising has not been examined previously.^{34–38} Here, we calculate restored image PSNR and SSIM along with metabolic metric recovery and observe that higher PSNR and SSIM values did not ensure the greatest restoration of RR and β metrics (Fig. 6). Although PSNR and SSIM values between models are observed to be within <5% of each other (Tables 3–5), many studies indicate a maximum improvement of PSNR and SSIM values as indicators of model performance.^{34–38} In this study, we observe that models with optimal PSNR and SSIM values did not yield the greatest recovery

of metabolic metrics. Together, PSNR and SSIM are not well suited for the assessment of model performance on label-free 2PM images, necessitating further validation using metrics of metabolic activity.

The application of denoising algorithms on label-free 2PM datasets to date has been limited by the lack of available large clinical datasets.^{37,62} Deep learning models have shown promise with small datasets (Shen et al. used only 24 paired images) in image restoration; however, larger datasets are needed for consistent reconstruction of high-SNR images.^{37,61} Here, we present a denoising network trained on a larger training set of 1657 OSs (6628 OSPs) and evaluate on an independent test set of 1018 OSs (4072 OSPs).

Using CARE, we observed improvements in image quality based on standard metrics (Table 3). However, the pre-packaged, standard models showed poor recovery of β metrics. Custom-loss functions improved metabolic metric recovery by penalizing models for both failing to generate similar images and reducing pixel correlation (see Tables S2 and S3 in the [Supplementary Material](#)). More interestingly, we observed that using DWT to separate the frequency information in an image before training independent models (WU-Net) produced images that had high metabolic metric correlations with GT 6X metrics [Fig. 4(c)]. By training on independent frequency-band images, the models were forced to learn the noise characteristics of different frequency bands without convolving the bands.⁵⁹

A key advantage of WU-Net, in comparison with identically trained (non-wavelet) U-Nets, was the denoising of higher frequencies at which noise was expected to be dominant. Denoising of high frequency noise led to enhanced recovery of β metrics as WU-Net was more consistent in reducing noise in these frequencies (see Discussion S3 and Fig. S2 in the [Supplementary Material](#)). WU-Net led to a statistically significant decrease in high frequencies compared with a comparable CARE model (see Fig. S2 in the [Supplementary Material](#)). Further, the incorporation of SSIM + R2 as a loss function promoted the restoration of similar frequencies from the GT 6X image in the denoised image while minimizing the loss of correlation between pixels. Further, we observed that models trained on FAD images outperform their NAD(P)H counterparts [Fig. 5(c)]. To explain this phenomenon, we examined the correlation of optical RR metrics between RAW 1X images and GT 6X images. RR metrics from RAW 1X images correlated well with RR metrics from GT 6X images, suggesting that the noise characteristics in FAD and NAD(P)H images are similar. However, as the FAD images contain less signal compared with their NAD(P)H paired images, high-frequency contributions are mostly noise in the RAW 1X FAD images. Thus, training on FAD images likely improved the model's learning of noise characteristics. This led to an improvement in downstream metric recovery and the translation of model weights to NAD(P)H images. Additional models were trained on both FAD and NAD(P)H images to see if the performance could be improved. Despite the increase in training data, the recovery of downstream metrics did not improve, and in most cases, the performance was lower (data not shown). It is hypothesized that differences in signal contribution in FAD and NAD(P)H images are responsible for the reduction in performance that we observe; however, further hyperparameter optimization is needed to confirm this hypothesis.

WU-Net with a custom loss (SSIM + R2) function and training on FAD data demonstrates improved restoration of most metrics of metabolic activity from label-free, 2PM images (Table 7); however, further improvements in the restoration of $\bar{\beta}$ are desired. One potential method of improving $\bar{\beta}$ restoration is a loss function that minimizes the differences in the PSD maps of paired images that are used for β calculation. A challenge of such a method is the computational time required for generating these maps.^{22,24,25} Future studies may examine simpler predictors of mitochondrial clustering using a modified GAN network, in which the discriminator network estimates β from the input images and optimizes the generator to achieve accurate β metric recovery. Further, the optimization of the RR loss function is still needed. In this study, the loss function was set to calculate RR based on unmasked images due to the need for a differentiable loss function for back-propagation of model weights. The use of true RR images for loss calculation would require the development of a non-differentiable optimization method, which was beyond the scope of this study. Future work will explore the development of such a loss function for improved performance.

In this study, we specifically focused on the restoration of morphological and functional metrics from label-free, 2PM images of human cervical tissue, relying on a single denoising

algorithm. Future studies will examine the application of the trained denoising model and model architecture on datasets acquired from different microscope systems, objective lenses, and tissue types. Validation of the model on these datasets will support the broad use of WU-Net for denoising label-free 2PM images. The successful implementation of pre-trained models on other datasets will reduce the need for large clinical datasets.^{37,62} As the model advances, improvements in ground truth data collection are needed. Ground truth data used in this study contain noise and therefore are not truly representative of mitochondrial signal. The averaging of six-frames was the limit to what we could represent with our collected data, but future studies will aim to assess what further enhancements can be achieved as it is important that we acquire images as fast as possible in the clinic. Alternative techniques for image acquisition such as a slower line scan speed could be utilized to improve the ground truth image quality.

In summary, we demonstrated that maximizing standard metrics of image quality (PSNR and SSIM) did not necessarily lead to improved recovery of functional tissue metrics, especially ones associated with mitochondrial organization (Table 7). Using WU-Net with a custom loss function, we demonstrated improved recovery of functional metrics of metabolic activity, even though PSNR and SSIM metrics were not optimal. Results from this study support the application of deep learning algorithms for the restoration of RR and β metrics from low-SNR 2PM images. As more data become available from varying microscope systems, objective lenses, and tissue types, a more robust algorithm could be generated for rapid image collection and classification, eventually improving patient health during *in vivo* image collection.

Disclosures

The authors declare no conflicts of interests.

Code and Data Availability

The raw datasets used for model generation in the current study along with the trained model weights are available from the corresponding author on reasonable request. Codes for network training and prediction are publicly available at <https://github.com/Tufts-University/Denoising2PMImages>.

Supplementary Material

See [Supplementary Material](#) for supporting content.

Acknowledgments

We would like to thank the National Institute of Biomedical Imaging and Bioengineering (R01 EB030061), the National Institutes of Health, Office of the Director (S10 OD021624), and the National Cancer Institute for funding this work (R03 CA235053). The authors acknowledge the Tufts University High Performance Compute Cluster,⁶⁷ which was utilized for the research reported in this paper. We would also like to thank Jasmine Kwan for her help in preparing figures for this paper.

References

1. W. G. Kaelin and C. B. Thompson, "Clues from cell metabolism," *Nature* **465**, 562–564 (2010).
2. R. J. DeBerardinis and C. B. Thompson, "Cellular metabolism and disease: what do metabolic outliers teach us?" *Cell* **148**, 1132–1144 (2012).
3. J. Kim and R. J. DeBerardinis, "Mechanisms and implications of metabolic heterogeneity in cancer," *Cell Metab.* **30**, 434–446 (2019).
4. V. D. Gialleonardo, D. M. Wilson, and K. R. Keshari, "The potential of metabolic imaging," *Semin. Nucl. Med.* **46**, 28–39 (2016).
5. X. Lu et al., "LC-MS-based metabonomics analysis," *J. Chromatogr. B* **866**, 64–76 (2008).
6. P. T. C. So et al., "Two-photon excitation fluorescence microscopy," *Annu. Rev. Biomed. Eng.* **2**, 399–429 (2000).
7. F. Helmchen and W. Denk, "Deep tissue two-photon microscopy," *Nat. Methods* **2**, 932–940 (2005).
8. I. Georgakoudi and K. P. Quinn, "Optical imaging using endogenous contrast to assess metabolic state," *Annu. Rev. Biomed. Eng.* **14**, 351–367 (2012).
9. A. A. Heikal, "Intracellular coenzymes as natural biomarkers for metabolic activities and mitochondrial anomalies," *Biomark. Med.* **4**, 241–263 (2010).

10. J. B. Spinelli and M. C. Haigis, "The multifaceted contributions of mitochondria to cellular metabolism," *Nat. Cell Biol.* **20**, 745–754 (2018).
11. T. S. Blacker, M. R. Duchon, and A. J. Bain, "NAD(P)H binding configurations revealed by time-resolved fluorescence and two-photon absorption," *Biophys. J.* **122**, 1240–1253 (2023).
12. B. Chance et al., "Oxidation-reduction ratio studies of mitochondria in freeze-trapped samples. NADH and flavoprotein fluorescence signals," *J. Biol. Chem.* **254**, 4764–4771 (1979).
13. A. Mayevsky and B. Chance, "Oxidation-reduction states of NADH *in vivo*: from animals to clinical use," *Mitochondrion* **7**, 330–339 (2007).
14. O. I. Kolenc and K. P. Quinn, "Evaluating cell metabolism through autofluorescence imaging of NAD(P)H and FAD," *Antioxidants Redox Signal.* **30**, 875–889 (2019).
15. I. Georgakoudi and K. P. Quinn, "Label-free optical metabolic imaging in cells and tissues," *Annu. Rev. Biomed. Eng.* **25**, 413–443 (2023).
16. L. Tilokani et al., "Mitochondrial dynamics: overview of molecular mechanisms," *Essays Biochem.* **62**, 341–360 (2018).
17. J. M. Levitt et al., "Diagnostic cellular organization features extracted from autofluorescence images," *Opt. Lett.* **32**, 3305 (2007).
18. J. Xylas et al., "Improved Fourier-based characterization of intracellular fractal features," *Opt. Express* **20**, 23442 (2012).
19. K. Blinova et al., "Mitochondrial NADH fluorescence is enhanced by complex I binding," *Biochemistry* **47**, 9636–9645 (2008).
20. T. Wai and T. Langer, "Mitochondrial dynamics and metabolic regulation," *Trends Endocrinol. Metab.* **27**, 105–117 (2016).
21. J. Xylas et al., "Noninvasive assessment of mitochondrial organization in three-dimensional tissues reveals changes associated with cancer development," *Int. J. Cancer* **136**, 322–332 (2015).
22. D. Pouli et al., "Imaging mitochondrial dynamics in human skin reveals depth-dependent hypoxia and malignant potential for diagnosis," *Sci. Transl. Med.* **8**, 367ra169 (2016).
23. K. P. Quinn et al., "Quantitative metabolic imaging using endogenous fluorescence to detect stem cell differentiation," *Sci. Rep.* **3**, 3432 (2013).
24. A. Varone et al., "Endogenous two-photon fluorescence imaging elucidates metabolic changes related to enhanced glycolysis and glutamine consumption in precancerous epithelial tissues," *Cancer Res.* **74**, 3067–3075 (2014).
25. D. Pouli et al., "Label-free, high-resolution optical metabolic imaging of human cervical precancers reveals potential for intraepithelial neoplasia diagnosis," *Cell Rep. Med.* **1**, 100017 (2020).
26. M. Balu et al., "In vivo multiphoton microscopy of basal cell carcinoma," *JAMA Dermatol.* **151**, 1068 (2015).
27. S. You et al., "Label-free deep profiling of the tumor microenvironment," *Cancer Res.* **81**, 2534–2544 (2021).
28. E. Pshenay-Severin et al., "Multimodal nonlinear endomicroscopic imaging probe using a double-core double-clad fiber and focus-combining micro-optical concept," *Light Sci. Appl.* **10**, 207 (2021).
29. J. Shiu et al., "Multimodal analyses of vitiligo skin identifies tissue characteristics of stable disease," *JCI Insight* **7**, e154585 (2022).
30. V. Kučikas et al., "Two-photon endoscopy: state of the art and perspectives," *Mol. Imaging Biol.* **25**, 3–17 (2021).
31. L. Fan et al., "Brief review of image denoising techniques," *Visual Comput. Ind. Biomed. Art* **2**, 7 (2019).
32. S. V. Mohd and S. N. George, "Biomedical signal processing and control a review on medical image denoising algorithms," *Biomed. Signal Process. Control* **61**, 102036 (2020).
33. K. Monakhova et al., "Dancing under the stars: video denoising in starlight," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. and Pattern Recognit.*, June, pp. 16220–16230 (2022).
34. M. Weigert et al., "Content-aware image restoration: pushing the limits of fluorescence microscopy," *Nat. Methods* **15**, 1090–1097 (2018).
35. C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conf. Comput. Vis. and Pattern Recognit. (CVPR)*, January, pp. 105–114 (2016).
36. J. Chen et al., "Three-dimensional residual channel attention networks denoise and sharpen fluorescence microscopy image volumes," *Nat. Methods* **18**, 678–687 (2021).
37. B. Shen et al., "Deep learning autofluorescence-harmonic microscopy," *Light Sci. Appl.* **11**, 76 (2022).
38. Y. Zhang et al., "Image super-resolution using very deep residual channel attention networks," *Lect. Notes Comput. Sci.* **11211**, 294–310 (2018).
39. A. Ferenczy, "Ultrastructure of the uterine cervix," in *The Physiology and Biochemistry of the Uterus in Pregnancy and Labor*, G. Huszar, Ed., pp. 239–260, CRC Press (1986).
40. I. Pavlova et al., "Microanatomical and biochemical origins of normal and precancerous cervical autofluorescence using laser-scanning fluorescence confocal microscopy," *Photochem. Photobiol.* **77**, 550–555 (2007).
41. J. Shiu et al., "Non-invasive imaging techniques for monitoring cellular response to treatment in stable vitiligo," bioRxiv (2023).

42. F. Fischer et al., "Assessing the risk of skin damage due to femtosecond laser irradiation," *J. Biophotonics* **1**, 470–477 (2008).
43. A. Fukasawa et al., "High speed HPD for photon counting," *IEEE Trans. Nucl. Sci.* **55**, 758–762 (2008).
44. W. Becker et al., "FLIM and FCS detection in laser-scanning microscopes: increased efficiency by GAASP hybrid detectors," *Microsc. Res. Tech.* **74**, 804–811 (2011).
45. K. P. Quinn et al., "Characterization of metabolic changes associated with the functional development of 3D engineered tissues by non-invasive, dynamic measurement of individual cell redox ratios," *Biomaterials* **33**, 5341–5348 (2012).
46. M. Abadi, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," <https://www.tensorflow.org/> (2015).
47. F. Chollet and Others, "Keras," <https://github.com/fchollet/keras> (2015).
48. "Anaconda software distribution," in *Anaconda Documentation*, Version 2-2.4.0, Anaconda Inc. (2020).
49. N. Vora et al., "Denoising of depth-resolved, label-free, two-photon images using deep-learning-based algorithms," *Proc. SPIE* **12391**, 1239107 (2023).
50. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *3rd Int. Conf. Learn. Represent., ICLR 2015 – Conf. Track Proc.*, pp. 1–15 (2014).
51. L. Jiang et al., "Focal frequency loss for image reconstruction and synthesis," in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 13899–13909 (2020).
52. H. Zhao et al., "Loss functions for image restoration with neural networks," *IEEE Trans. Comput. Imaging* **3**(1), 47–57 (2017).
53. R. A. Fisher, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika* **10**(4), 507 (1915).
54. R. A. Fisher, "On the "probable error" of a coefficient of correlation as deduce from a small sample," *Metron* **1**, 205–235 (1921).
55. A. Hore and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *20th Int. Conf. Pattern Recognit.*, IEEE, pp. 2366–2369 (2010).
56. B. Ergen, "Signal and image denoising using wavelet transform," in *Advances in Wavelet Theory and Their Applications in Engineering, Physics and Technology*, D. Baleanu, Ed., InTech (2012).
57. L. Hongqiao and W. Shengqian, "A new image denoising method using wavelet transform," in *Int. Forum on Inf. Technol. and Appl.*, IEEE, Vol. 1, pp. 111–114 (2009).
58. F. Luisier, T. Blu, and M. Unser, "A new SURE approach to image denoising: interscale orthonormal wavelet thresholding," *IEEE Trans. Image Process.* **16**, 593–606 (2007).
59. C. Aytekin et al., "A sub-band approach to deep denoising wavelet networks and a frequency-adaptive loss for perceptual quality," in *IEEE 23rd Int. Workshop on Multimedia Signal Process. (MMSP)*, pp. 1–6 (2021).
60. B. Han, "Wavelet filter banks," in *Framelets and Wavelets: Algorithms, Analysis, and Applications*, pp. 67–151, Springer International Publishing, Cham, Switzerland (2017).
61. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press (2016).
62. S. K. Melanthota et al., "Deep learning-based image processing in optical microscopy," *Biophys. Rev.* **14**, 463–481 (2022).
63. Z. Wang et al., "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.* **13**, 600–612 (2004).
64. O. Ronneberger, P. Fischer, and T. Brox, "U-Net: convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, N. Navab et al., Eds., pp. 234–241, Springer International Publishing, Cham, Switzerland (2015).
65. J. Gui et al., "A review on generative adversarial networks: algorithms, theory, and applications," *IEEE Trans. Knowl. Data Eng.* **14**, 3313–3332 (2021).
66. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd Int. Conf. Learn. Represent., ICLR 2015 – Conf. Track Proc.*, pp. 1–14 (2014).
67. Tufts University, "High Performance Cluster," 2023, <https://it.tufts.edu/high-performance-computing> (accessed 01 June 2022).

Nilay Vora received his PhD in biomedical engineering in 2023 from Tufts University, where he was a member of Optical Diagnostics for Diseased and Engineered Tissues (ODDET) Lab under the advisement of Professor Irene Georgakoudi. His work is focused on the development of machine learning (ML) based pipelines for analyzing low signal-to-noise ratio biomedical image datasets. His research interest lies in the translational of imaging systems from the benchtop to the bedside, using ML as a tool to improve patient outcomes.

Christopher M. Polleys is a 6th year PhD student in the ODDET Lab at Tufts University under the advisement of Professor Irene Georgakoudi. His work is focused on developing a novel set of label-two-photon biomarkers for the real-time detection of cervical precancer. His research

interests lie in the use of non-invasive, label-free imaging methods to understand the underlying biochemistry and biomechanics associated with disease progression in living systems.

Filippos Sakellariou is an undergraduate student at Vassar College in Poughkeepsie, New York. He has been involved in programming and developing apps. Among other things, he has written two proposals for the improvement of the Swift programming language.

Georgios Georgalis research includes different aspects of data science, with a focus on ML, uncertainty quantification, and image processing. Current projects include the development of deep learning and surrogate models for large scale hybrid rocket motor simulations, timeseries analysis for optical datasets, and image processing for laparoscopic images. He completed his doctoral studies at Purdue University in 2021 and has work experience in AI applications for nuclear reactors, small satellites, and rocket propulsion systems.

Hong-Thao Thieu is a full-time obstetrician gynecologist in Newton, Massachusetts, who practices the entire spectrum of clinical care, with specific interests in pediatric and adolescent gynecology and complex family planning. She is a graduate of Vanderbilt University School of Medicine and Tufts OBGYN Residency Training Program.

Elizabeth Genega: Biography is not available.

Narges Jahanseir did her anatomic pathology residency at Mayo Clinic in Rochester, Minnesota, following by a fellowship in breast and gynecologic pathology at Brown University, in Providence, Rhode Island. She is currently an assistant professor of pathology at Tufts University and the director of breast and gyn pathology at Tufts Medical Center.

Abani Patra received his PhD in computational mathematics in 1995 and has subsequently held faculty positions at SUNY at Buffalo and Tufts University and directed research and education centers. His primary research interests are in mathematical and computational modeling and uncertainty quantification of complex physical systems.

Eric Miller received his SB, SM, and PhD degrees in EECS from MIT. He is a professor of ECE at Tufts. He holds secondary appointments in mathematics, computer science, and biomedical engineering and is a senior scientist at the Je USDA Human Nutrition Research. His research focuses on physics and statistically based signal and image processing. He is a member of Tau Beta Pi, Phi Beta Kappa, and Eta Kappa Nu and is an IEEE fellow.

Irene Georgakoudi is a professor of biomedical engineering at Tufts University. She received her BA degree in physics from Dartmouth College, her PhD in biophysics from the University of Rochester and performed postdoctoral work at the MIT and the Wellman Center for Photomedicine. Her lab is interested in the development of label-free methods to assess quantitatively different aspects of normal and diseased development of human tissues, especially in the context of cell metabolism and matrix organization.