Deep-Learning-Based Fast Optical Coherence Tomography (OCT) Image Denoising for Smart Laser Osteotomy

Yakub A. Bayhaqi, Arsham Hamidi, Ferda Canbaz, Alexander A. Navarini, Philippe C. Cattin, *Member, IEEE*, and Azhar Zam, *Member, IEEE*

***Abstract*— Laser osteotomy promises precise cutting and minor bone tissue damage. We proposed Optical Coherence Tomography (OCT) to monitor the ablation process toward our smart laser osteotomy approach. The OCT image is helpful to identify tissue type and provide feedback for the ablation laser to avoid critical tissues such as bone marrow and nerve. Furthermore, in the implementation, the tissue classifier's accuracy is dependent on the quality of the OCT image. Therefore, image denoising plays an important role in having an accurate feedback system. A common OCT image denoising technique is the frame-averaging method. Inherent to this method is the need for multiple images, i.e., the more images used, the better the resulting image quality. However, this approach comes at the price of increased acquisition time and sensitivity to motion artifacts. To overcome these limitations, we applied a deep-learning denoising method capable of imitating the frame-averaging method. The resulting image had a similar image quality to the frame-averaging and was better than the classical digital filtering methods. We also evaluated if this method affects the tissue classifier model's accuracy that will provide feedback to the ablation laser. We found that image denoising significantly increased the accuracy of the tissue classifier. Furthermore, we observed that the classifier trained using the deep learning denoised images achieved similar accuracy to the classifier trained using frame-averaged images. The results suggest the possibility of using the deep learning method as a pre-processing step for real-time tissue classification in smart laser osteotomy.**

***Index Terms*— Deep learning, image denoising, image processing, optical coherence tomography.**

Manuscript received 27 November 2021; revised 7 March 2022 and 4 April 2022; accepted 12 April 2022. Date of publication 20 April 2022; date of current version 30 September 2022. This work was supported by Werner Siemens Foundation (Switzerland) through the Minimally Invasive Robot-Assisted Computer-guided LaserosteotomE (MIRACLE Project). *(Corresponding authors: Yakub A. Bayhaqi; Azhar Zam.)*

Yakub A. Bayhaqi, Arsham Hamidi, and Ferda Canbaz are with the Biomedical Laser and Optics Group (BLOG), Department of Biomedical Engineering, University of Basel, 4123 Allschwil, Switzerland (e-mail: yakub.bayhaqi@unibas.ch; arsham.hamidi@unibas.ch; ferda.canbaz@unibas.ch).

Alexander A. Navarini is with the Digital Dermatology Group, Department of Biomedical Engineering, University of Basel, 4123 Allschwil, Switzerland, and also with the Department of Dermatology, University Hospital of Basel, 4031 Basel, Switzerland (e-mail: alexander.navarini@usb.ch).

Philippe C. Cattin is with the Center for Medical Image Analysis and Navigation (CIAN), Department of Biomedical Engineering, University of Basel, 4123 Allschwil, Switzerland (e-mail: philippe.cattin@unibas.ch).

Azhar Zam was with the Department of Biomedical Engineering, University of Basel, 4123 Allschwil, Switzerland. He is now with the Department of Biomedical Engineering, New York University, Brooklyn, NY 11201 USA, and also with the Division of Engineering, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates (e-mail: azhar.zam@nyu.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMI.2022.3168793, provided by the authors.

Digital Object Identifier 10.1109/TMI.2022.3168793

I. Introduction

LASER osteotomy offers many advantages over mechanical tools, such as a reduced risk of bacterial contamination (due to its contactless nature), less tissue loss, and high precision cutting [1]–[4]. In addition, the small focused cut of a laser osteotome enables surgeons to go beyond straight cuts and perform more complex cuts like circular, diamond, and dove-tail shapes [5]. Furthermore, the non-contact laser osteotome provides a possibility to introduce a feedback system to prevent cutting unwanted tissues or damaging critical tissues, such as bone marrow and nerve. Several such feedback systems have been developed, such as optical spectroscopy [6]–[8] and acoustic feedback induced by the laser ablation process [9]–[12]. However, these methods rely on signals emitted from the laser ablation process and permit some damage to the critical tissues. An alternative and ablation-free approach for monitoring the laser ablation process can be implemented by coupling the ablation laser with an Optical Coherence Tomography (OCT) imaging system.

OCT is an emerging technology that performs non-invasive cross-sectional tomography using light propagation properties in media and interference phenomena. This imaging technology is analogous to ultrasound imaging, except that it uses light instead of sound. The signal reconstruction is performed by measuring the magnitude and echo time delay of back-reflected or back-scattered light from internal micro-structures in the tissue. Thus, OCT is a viable alternative for real-time, high-resolution, and in-situ investigations of thin tissue structures [13], [14].

Critical tissues such as bone marrow and nerve must be avoided in laser osteotomy. Therefore, OCT could help to monitor tissue anatomy at the subsurface level during the laser ablation process. Fig. 1 presents a schematic diagram of our proposed system. This process consists of three main subprocesses. The first subprocess is the acquisition and denoising of the images. Next, the second subprocess is tracking the ablation crater. The tracked region of interest (image patch) from the OCT image will be used as an input

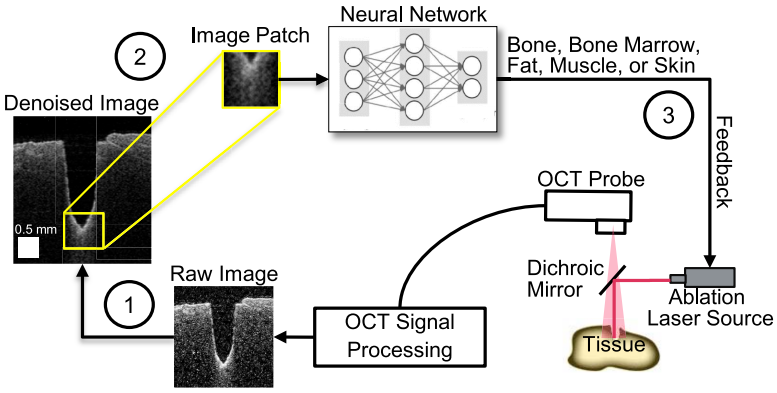

Fig. 1. Schematic of the proposed smart laser osteotomy system. Step 1: OCT images of tissue are acquired (light red) and denoised during laser ablation (dark red line). Step 2: A region of interest or image patch (yellow box) from the OCT image is used as input for the artificial neural network to classify the tissue types at the ablation position. Step 3: An artificial neural network provides feedback for the ablation laser to either stop or continue ablating based on the patch image.

for a classifier to classify tissue type at the ablation position. The third and final subprocess is the classification process. A deep-learning artificial neural network (ANN) is proposed to predict the current tissue type during the ablation and provides feedback to the ablating laser to stop or continue ablating.

In the proposed schematic, the tissue classifier's accuracy is dependent on the quality of the OCT image that is used for training. Although the ANN could identify tissue type from the raw image directly, we believe that the ANN could identify tissue type from the denoised image with better accuracy. Furthermore, besides using an ANN as a tissue classifier, our approach also proposed to use an ANN for the image denoising process. This paper demonstrated the benefit of using ANN as an image denoiser compared to the classical denoising methods. We also investigated the effect of denoising the OCT image to improve the tissue classifier's accuracy.

II. Speckle and Image Denoising in OCT

Like most narrow-band detection systems such as radar and ultrasound, speckle is the fundamental OCT image source. Speckle is the cause of the reconstructed image's grainy appearance. It depends on the size and temporal coherence of the light source and the tissue's structural characteristics. The phenomenon was found as the result of random interference between mutually coherent reflected waves from multiple back- and forward-scattering [15]. Consequently, speckle plays a dual role as a source of noise (speckle noise) and as a carrier of information about the tissue microstructure (signal-carrying speckle). Speckle is considered as noise when destructive interference happens and reduces the correspondence between the local density of scatterers and the intensity variations. If all the reflected waves from the tissue could be forced to interfere constructively, the noise would vanish, and the image contrast would be significantly improved. The OCT image denoising in this paper aimed toward this ideal of speckle-noise reduction.

Separating the signal carrier speckle from speckle noise (image despeckle) is an ongoing challenge in OCT. Several approaches have been suggested to address this problem. A common technique for reducing speckle noise in OCT is frame-averaging, where absolute magnitudes of repeated signals from the same location are averaged to form a new noise-reduced signal [16]. However, this method is susceptible to motion artifacts if compensation of movement is not resolved. Additionally, the resulting image quality depends on the number of repetitions. The more images used, the better the resulting image quality, but this inherently leads to increased acquisition time.

Other than that, several classical digital filters were also suggested to reduce speckle noise in OCT. Sparse and wavelet transform filtering approaches, for example, can be applied directly to a single frame image [17]–[20]. However, these methods are often computationally complex and remove small structural features from the image, resulting in lower image quality than the frame-averaging method.

In this work, besides using an ANN to classify the tissue type, we also used an ANN to reduce speckle noise in the OCT image. ANN is already known for its ability to classify, retrieve, detect, and segment images in image pattern recognition and computer vision. ANN has also been used to correct or denoise images, making them useful for most medical imaging applications and potentially leading to better diagnostic assessments. In ultrasound imaging, ANN has been shown to solve recovering ultrasound signals from under-sampled measurements by utilizing stacked autoencoders [21]. ANN has also been used to enhance low-dose Computed Tomography images, which may offer a solution for reducing X-ray radiation [22]. These endeavors' success arises from exploiting the spatial correlation at multiple resolutions, using a hierarchical network structure.

OCT image denoising using ANNs has been proposed and implemented mostly based on the convolutional neural network (CNN) models [23]–[25]. The resulting images had similar image quality to those denoised with the frame-averaging method. The CNN could denoise retinal OCT images from a single image without blurring the retinal tissue structure's details, which reduces acquisition time. However, these references are only from the field of ophthalmology with their retinal OCT images. Here, the CNN model has to learn to denoise an image while retaining (memorize) the retinal tissue structure rather than just a general noise reduction. The CNN model might fail to denoise images in different OCT domains (e.g., intravascular, dental, or dermatology OCT images).

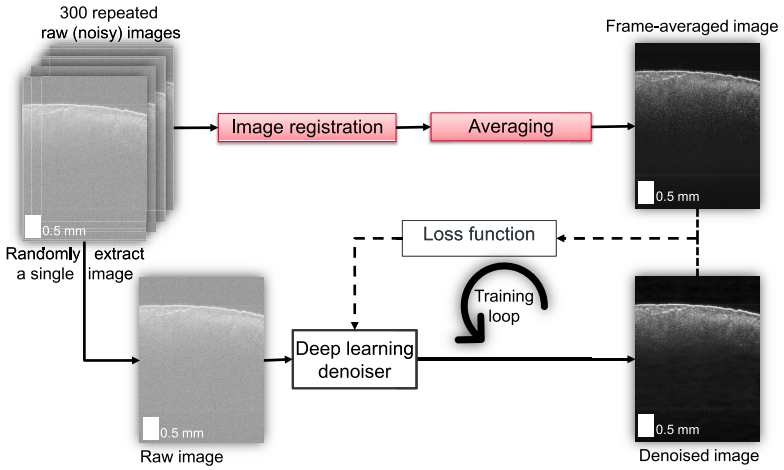

Fig. 2. Illustration of the training process using the deep learning method to denoise raw OCT images. The frame-averaged (reference) image is reconstructed by registering and averaging 300 repeated raw images of the same position on the sample. The deep learning denoiser is trained to modify the raw images to make them as similar as possible to the frame-averaged image, based on the loss function.

In our application, we trained and tested the performance of the CNN method using several tissues with various textures so that the CNN model learns more generalization of the noise. We extended the ability of the CNN to increase OCT image quality for normal tissues, which are encountered during laser osteotomy. Additionally, this paper gives an overview of the CNN model's performance to improve OCT image quality compared to frame-averaging and some classical digital filter methods. We also outlined the effectiveness of the image denoiser to increase the accuracy of a subsequent CNN tissue classifier.

III. OCT Image Denoising Methods

This work aimed to train a CNN that takes raw (noisy) images as input and generates images with the same quality level as the corresponding frame-averaged images. To achieve this, the CNN was trained to minimize the defined loss function, e.g., the mean squared error (MSE) between the corresponding raw and frame-averaged images. Furthermore, since noise in OCT also applies in the temporal domain, we also trained the CNN to generalize the temporal noise by using training images that were extracted in different (random) temporal locations over the repeated frames in the same spatial acquisition location. These training steps are shown in Fig. 2. At the end of the experiment, we investigated the benefit of denoising the image to improve the tissue classifier's accuracy.

A. Frame-Averaging Method

The frame-averaging method is still one of the most effective ways to reduce speckle noise in OCT imaging. In this paper, we used the frame-averaged image as the reference image or label. The reference image is generated by registering and averaging repeated scans of the same location. Averaging of N images improves the SNR by a factor $\sqrt{N}$ [26]. Hence, the higher the number of images, N, the lower the noise level. Nevertheless, since the frame-averaging method is susceptible to motion artifact, image registration is necessary before averaging the frames.

B. Deep-Learning Models

We compared two CNN models for denoising OCT images: a UNet autoencoder model [27] and a residual network (ResNet) model [28]. Additionally, we also compared two different loss functions for both CNN models. First, we trained both UNet and ResNet models with the MSE loss function only. Then, we investigated the model's performance by combining MSE, perceptual, and Wasserstein-adversarial loss. This combination was previously suggested in [22], [23], since the MSE loss function alone may skip some embedded details in the reference image.

1) UNet-Based Autoencoder: Starting with the UNet autoencoder, we adopted the structure of the CNN reported by Ronneberger *et al.* in 2015 [27]. We changed the size of the UNet input to the size of the acquired OCT images. The encoding path (left side) and decoding path (right side) were adapted accordingly, as shown in Fig. 3. Each side consists of five folded convolutional blocks. There are two convolutional layers (kernel size of 3×3) for each block; the number of filters gradually increases from 32, to 64, 128, 256, and 512, respectively, for the encoding path and vice versa for the decoding path. A 2×2 max-pooling layer (stride of two) downsamples the features after each convolutional block in the encoding path, except for the last (deepest) block. On the other hand, a 2×2 upsampling2D layer was applied after each convolutional block in the decoding path. Each convolutional block in the encoding path forwards a residual feature (copy) to the corresponding convolutional block in the decoding path. We equipped all the convolutional layers with a rectified linear unit (ReLU) as the activation function. We added a dropout layer (ratio of 0.1) after each convolutional block to prevent overfitting [29]. The final layer was a 1×1 convolutional layer to reconstruct the decoded image with a similar size as the input and activated with the sigmoid function.

2) Residual Network: The architecture of the residual network (ResNet) was suggested by He *et al.* [28]. The model consisted of a pre-residual layer, ten residual connecting blocks with identical structures, and a post-residual layer (shown in Fig. 4). The pre-and post-residual layers are 2D

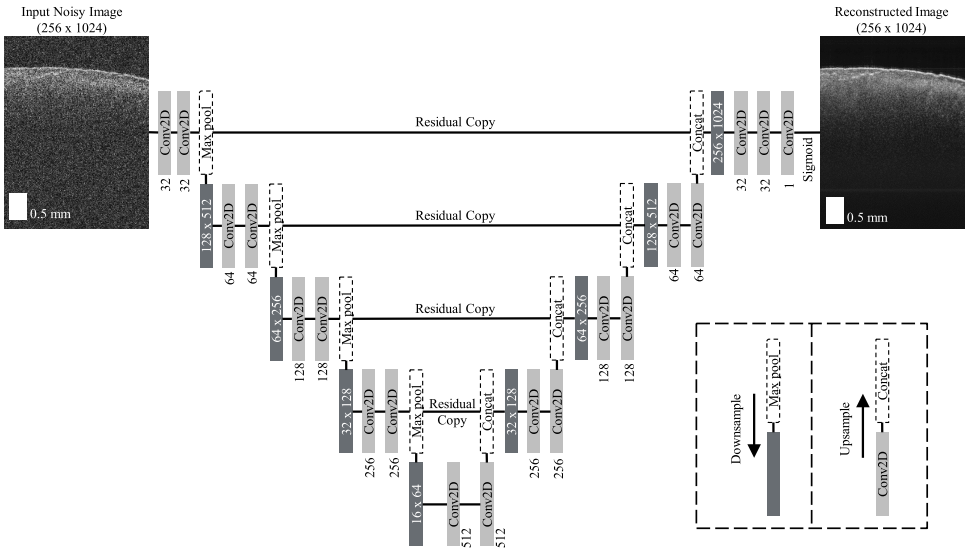

Fig. 3. The architecture of the UNet model. The main structure is separated between the downsampling encoder (left side) and the upsampling decoder (right side). Each side consists of four folded convolutional blocks. Each convolutional block in the upsampling concatenates a residual feature (copy) forwarded from the corresponding downsampling block. There are two convolutional layers for each block, the size of which depends on the depth of the factor of two, starting with 32 filters.

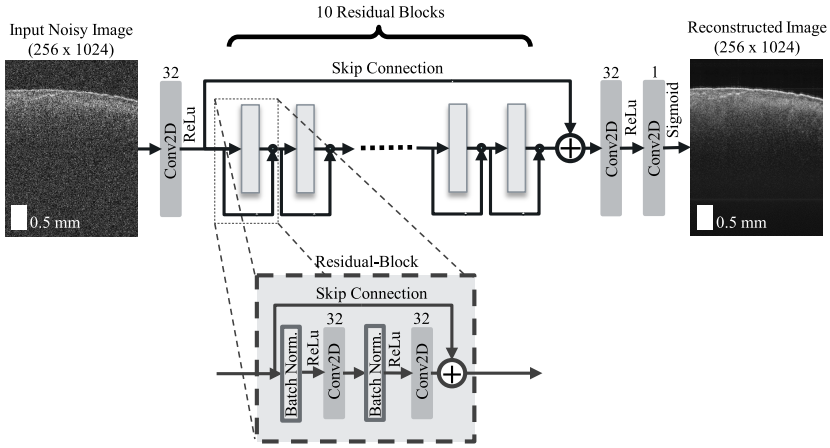

Fig. 4. The architecture of the ResNet model. It consists of 10 residual blocks placed between pre- and post- convolutional layers. Parallelly, a residual copy (skip) connects the pre- and post-layers. The residual block consists of two stacked convolutional layers. All the layers have 32 filters with a kernel size of 3 x 3 pixels.

convolutional layers. Each layer has 32 filters with 3 × 3 kernel size and is activated with the ReLU function. The residual blocks consisted of two stacked 2D convolutional layers of the same size as the pre-residual layer. Batch normalization is applied before each convolutional layer. A skip connection between blocks was introduced by He *et al.*. [28], which added a signal between the pre-and post-processing of a block (shown in Fig. 4). This identity mapping improves information flow through the network during feed-forward and back-propagation. Another skip connection was added between the signal before the residual blocks after processing through addition and followed by the post-residual layer. The final layer was the reconstruction layer, a 1 × 1 convolutional layer activated with a sigmoid function. The input size of the ResNet was set according to the size of the image in our datasets.

3) Mean Squared Error Loss: The most intuitive way of measuring the similarity between two images is by using the MSE. MSE measures the quadratic mean of the overall pixel difference between the corresponding reference image and the predicted image. This measurement is defined as:

$$MSE = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(\hat{p}_{i,j} - p_{i,j})^2}{mn} \quad (1)$$

The MSE is the squared mean deviation of the pixel value (p) in the frame-averaged image and the pixel value ($\hat{p}$) in the denoised image at the i, j-th position with the same width m and height n. The value shows the general similarity per pixel between these images. The aim of training the UNet and ResNet is to minimize MSE to zero. However, studies show that the MSE loss may result in over-smoothed image [22].

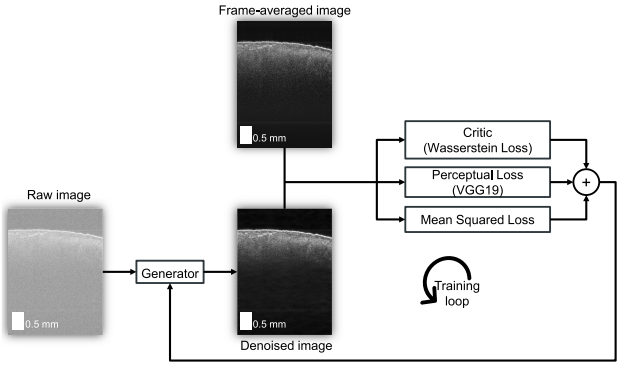

Fig. 5. Illustration of the adversarial and perceptual learning process. The training involves measuring Wasserstein, perceptual, and mean squared error loss of the denoised image produced by a generator. The critic is trained in a binary manner that tries to distinguish frame-averaged and denoised image [29]. Meanwhile, the perceptual loss is the measure of the relative perspective difference between the frame-averaged and denoised images, which is extracted from the high-level feature representation of a pre-trained VGG19 image-Net [30].

4) Adversarial and Perceptual Learning: First, we trained the models using the MSE loss function alone. We defined these models as UNet-MSE and ResNet-MSE, respectively. In this specific situation, the models were trained to generate pixel values similar to those of the frame-averaged image, based on the MSE measurements. Furthermore, Yang *et al.* [22] proposed to train the CNN models by introducing the Wasserstein loss (critics) [29] and perceptual loss [30]. The solution raised to tackle MSE loss problems which are associated with over-smoothed edges and loss of details. A MSE-based CNN overlooks subtle tissue texture in the image, which is critical for human perception. These additional losses have been demonstrated to improve the CNN for denoising images with better image quality and statistical properties than the MSE-based CNN. Furthermore, Halupka *et al.* [23] used the same method to denoise OCT retinal images. Our application of these losses in the CNN learning process is illustrated in Fig. 5. Therefore, we trained additional UNet and ResNet models with the Wasserstein and perceptual loss. We defined them as UNet-WGAN and ResNet-WGAN, respectively.

A generative adversarial network (GAN) consists of a discriminator (D) and a generator (G) network. During training, the discriminator learns to distinguish between the frame-averaged image and the image denoised with the generator. Simultaneously, the generator will try to generate a high-quality denoised image from a raw image that would be indistinguishable by the discriminator. The discriminator network architecture is illustrated in Fig. 6. On the other side, the generator network is the investigated UNet or ResNet.

We applied the improved version of the original GAN, which used the Wasserstein distance [29] as the discriminator loss function to criticize or score the performance of the generator. A Wasserstein GAN (WGAN) would have more stability during the training process compared to the original GAN. Gulrajani *et al.* also suggested that using the gradient penalty term to enforce the Lipschitz constraint would even

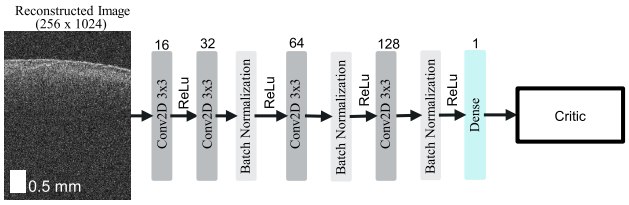

Fig. 6. The discriminator network for measuring the Wasserstein distance (critic) of the reconstructed image.

improve more training stability [31]. Both the discriminator and generator are trained with the min-max objective defined as:

$$\min_G \max_D L_{WGAN}(D,G) = \mathop{\mathbb{E}}_{\tilde{I}_{raw}}[D(\tilde{I}_{raw})] - \mathop{\mathbb{E}}_{I_{ref}}[D(I_{ref})] + \lambda \mathop{\mathbb{E}}_{\hat{I}_{raw}}[(\|\nabla_{\hat{I}_{raw}} D(\hat{I}_{raw})\|_2 - 1)^2] \tag{2}$$

The Wasserstein distance is calculated between the denoised image ($\tilde{I}_{raw}$) and the frame-averaged image (I_{ref}). The denoised image is the reconstructed image by the generator from the raw image ($\tilde{I}_{raw} = G(I_{raw})$). The final term is the gradient penalty which enforces the Lipschitz constraint to have gradient norm ($\|\nabla_{\hat{I}_{raw}} D(\hat{I}_{raw})\|_2$) at most 1, where,

$$\hat{I}_{raw} = pI_{ref} + (1-p)\tilde{I}_{raw}. \tag{3}$$

p is a uniform random number between 0 and 1 ($p \sim U[0,1]$). The gradient penalty is weighted with a coefficient λ.

The perceptual loss is calculated from the high-level features extracted from a pre-trained VGG19 network [30]. The perceptual loss measures the similarity (MSE) of feature representations between the frame-averaged and the denoised images. This loss offers a more robust training approach because the feature extracted from the VGG19 network represents an external or alternate perspective, such as the content or style of the image. The perceptual loss function obliges the generator to denoise raw images with similar feature representations rather than requiring the pixels to match exactly the pixel of the frame-averaged image. The feature representation loss is the mean squared error (Euclidean distance) between the extracted features:

$$L_{VGG/i,j} = \frac{1}{W_{i,j} H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} (\varphi_{i,j}(I_{ref})_{x,y} - \varphi_{i,j}(\hat{I}_{raw})_{x,y})^2 \tag{4}$$

where, $\varphi_{i,j}$ indicates the ReLU activated feature map obtained by the j-th convolutional layer before the j-th pooling layer within the VGG19 network. $W_{i,j}$ and $H_{i,j}$ describe the dimensions of the respective feature maps of the corresponding layer. In the implementation, we used the fifth convolutional and pooling layers to measure the perceptual loss. Furthermore, since the pre-trained VGG19 network worked with 3-channels image, we converted the 1-channel OCT image to 3-channels image by repeating the first channel to the second and third channel. The adversarial network model minimized the combination of MSE, perceptual, and Wasserstein loss, and

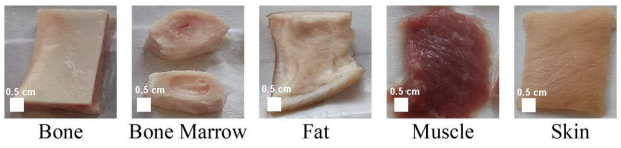

Fig. 7. Examples of the tissue types used in the experiments.

established a ratio between them [22], [23]. Thus, the min-max objective of the WGAN training is given by:

$$Loss = \min_{G} \max_{D} \omega_1 L_{WGAN}(D, G) + \omega_2 L_{VGG(G)} + L_{MSE} \tag{5}$$

Here, ω_1 and ω_2 are the weights balancing the different contributions of the loss function. This loss will force the generator to match exactly the corresponding pixel value and force the output to be perceptually similar to the reference image.

IV. Experiment

A. Image Acquisition and Processing

A custom-made swept-source OCT system was used for this experiment. The system was equipped with an Axsun swept laser source with a center wavelength (λ_0) of 1060 nm and sweep rate of 100 kHz. The sensitivity (SNR_{max}) of our OCT system was 96.46 dB. The OCT B-scan image sizes are 1024 pixels (3.6 mm) deep and 256 pixels (4.1. mm) wide. The output image format is 16-bit grayscale with TIFF-formatted image.

The OCT image datasets for the denoising experiment were acquired from three different pigs, which were used for training, validation, and testing datasets, respectively. The tissue sample types being investigated include the femur bone, bone marrow, fat, muscle, and skin tissues (Fig. 7). We randomly selected the scan locations over the surface area of each tissue. Thus, the images obtained vary in shape and surface location. The training image datasets were acquired from the first pig at 140 scan locations for each tissue type, yielding a total of 700 scans. Meanwhile, the validation image datasets were acquired from the second pig at only 28 scan locations for each tissue type. Therefore, the validation datasets contain only 140 scans. Furthermore, the testing image datasets were acquired from the last pig with similar number of scans as the training image datasets (700 scans).

We acquired 300 repeated B-scan frames for each scan location, the number of B-scans was selected to provide the highest signal strength (as a default setting by the OCT scanner software). These images at each scan location were then registered to remove motion artifacts. Rigid image registration was used because motion artifacts in our OCT images mainly originate from the object's translational motion. We used the fast normalized cross-correlation similarity measure to detect shifts between two images [32], [33]. This method was demonstrated for its application for fast-image-template matching. This registration method was implemented using the *normxcorr2* function in MATLAB® to calculate the correlation coefficient matrix of two images. We used the first frame for each scan as the static (fixed) image and the other frame as the moving image. The predicted translation is given by the location of the maximum correlation coefficient [34]. The motion-corrected frames were then averaged (frame-averaged) and labeled as the "ground truth" image.

Moreover, we randomly extracted 10 raw images from the 300 repeated images for each scan location and paired them with the same ground-truth frame-averaged images. We called them the raw images and used them for training, as explained in Section III and Fig. 2.

In summary, the training image datasets consisted of 7000 raw images with 700 corresponding ground truth images (7000 image pairs). The validation image datasets consisted of 1400 raw images with 140 corresponding ground-truth frame-averaged images (1400 image pairs). Furthermore, the testing image datasets consisted of only a single image randomly (instead of 10) extracted from each of the 300 repeated frame images. Therefore, only 700 image pairs were used for the testing.

B. Performance Comparison Methods

We compared the image quality and measured the similarities between the images that were denoised using the defined CNN models (UNet-MSE, Resnet-MSE, UNet-WGAN, and Resnet-WGAN) and the reference (frame-averaged) images. The performance of the CNN models were also compared with three classical digital filters—the median filter, block-matching 3D (BM3D) [17], and double-density complex wavelet transform (DD-CDWT) [19]. The image quality evaluation of the denoised images were done quantitatively and qualitatively. Processing time comparisons were also investigated to show the possibility of using the CNN for real-time image denoising.

Additionally, we also investigate the role of training dataset size to the performance of the CNN models. We trained each of CNN model with three variations of dataset size (2000, 5000, and 7000 image pairs). This investigation is intended to demonstrate the ability of the CNN models to generalize noise in the OCT images. Here, the WGAN based loss is expected to improve the noise generalization better than the MSE only loss. In summary, 12 CNN models were defined and compared along with the BM3D and DD-CDWT denoising methods. Definition of the evaluated denoising methods is explained in Table I.

C. CNN Model Training Details

We train all the CNN models in Keras-GPU environment with TensorFlow backend [35]. The training took place on an NVIDIA DGX A100 workstation equipped with NVIDIA A100 GPUs, which enabled us to perform parallel computations to speed up the training process. We trained all of our models using 1000 epochs. The training was done in mini-batches, with a batch size of 8. We selected the adaptive learning rate optimization algorithm (Adam) as the training optimizer [36], with the step size $\alpha = 1 \text{ x } 10^{-5}$ and decay parameters $\beta_1 = 0.9$ and $\beta_2 = 0.9$. The loss-weighting parameters, ω_1 and ω_2, for the WGAN-based models were similarly set to $1 \text{ x } 10^{-3}$. The gradient penalty (λ) was set

TABLE I
DEFINITION OF THE IMAGE AND DENOISING METHODS THAT WERE INVESTIGATED

Method	Definition
Raw Image	The raw image taken from single scan
Frame-Averaged	The frame-averaged image from 300 scans on the same location
Median Filter	The median filtered image of the raw image
BM3D	Block-matching 3D filter
DD-CDWT	Double-density complex wavelet transform
UNet-MSE 1	UNet model with MSE loss function and trained with 2000 images.
UNet-MSE 2	UNet model with MSE loss function and trained with 5000 images.
UNet-MSE 3	UNet model with MSE loss function and trained with 7000 images.
UNet-WGAN 1	UNet model with WGAN loss function and trained with 2000 images.
UNet-WGAN 2	UNet model with WGAN loss function and trained with 5000 images.
UNet-WGAN 3	UNet model with WGAN loss function and trained with 7000 images.
ResNet-MSE 1	ResNet model with MSE loss function and trained with 2000 images.
ResNet-MSE 2	ResNet model with MSE loss function and trained with 5000 images.
ResNet-MSE 3	ResNet model with MSE loss function and trained with 7000 images.
ResNet-WGAN 1	ResNet model with WGAN loss function and trained with 2000 images.
ResNet-WGAN 2	ResNet model with WGAN loss function and trained with 5000 images.
ResNet-WGAN 3	ResNet model with WGAN loss function and trained with 7000 images.

to 10, as suggested by the original paper [31]. Furthermore, online data augmentation was performed by small random geometrical (translation) shifts and flipping each image horizontally. We used the reflection mode to fill the points outside the boundaries of the image after translation. The data augmentation generators for all classifiers were set to have similar random seeds for fair training.

D. Quantitative Image Quality and Similarity Evaluation

1) Image Quality Metrics: We used the signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) to measure the image quality. These metrics show the noise magnitude of an image. The SNR is defined as the logarithmic ratio of the mean pixel values to the standard deviation of pixel values over the image foreground as follows:

$$SNR = 20\log\frac{\mu_{fg}}{\sigma_{fg}} \tag{6}$$

The mean μ_{fg} and the standard deviation σ_{fg} were measured over a defined foreground area of the image, consisting of the tissue structure. We applied the Canny edge detection algorithm to define the foreground and background areas [37], [38]. Defining the area enabled us to measure the contrast to noise ratio between tissue textural features and general noise, defined as:

$$CNR = 10\log\frac{\mu_{fg} - \mu_{bg}}{\sqrt{\sigma_{fg}^2 + \sigma_{bg}^2}} \tag{7}$$

The tissue textural feature is defined as the deviation of the mean value of the foreground μ_{fg} and background μ_{bg}. The general noise is defined as the square root of the total foreground μ_{fg} and background noise μ_{bg}.

2) Similarity Metrics: In addition to the image quality, we also measured the relative similarity of the images to evaluate the denoising methods in terms of noise suppression performance. The similarity is the relative measurement between the frame-averaged and the denoised image. In this work, we defined similarity based on three metrics. The first metric is the peak signal-to-noise ratio (PSNR). Unlike the previous SNR, PSNR is defined as the logarithmic ratio, which is a relative measurement with respect to the reference image:

$$PSNR = 10\log\frac{MAX_I^2}{MSE} \tag{8}$$

where MAX_I is the peak intensity or maximum pixel value that exists in the frame-averaged image and MSE is the mean squared error between the frame-averaged and the denoised image.

The second metric that we used to measure the similarity was the structural similarity index (SSIM). SSIM is the measure of the perceived visual difference between two images, which was difficult to estimate with PSNR alone. The metric describes similarity based on three main properties: luminance, contrast, and structure [39]. The simplified version of the SSIM is:

$$SSIM(I, \hat{I}) = \frac{(2\mu_I\mu_{\hat{I}} + C_1)(2\sigma_{I\hat{I}} + C_2)}{(\mu_I^2 + \mu_{\hat{I}}^2 + C_1)(\sigma_I^2 + \sigma_{\hat{I}}^2 + C_2)} \tag{9}$$

where μ_I, $\mu_{\hat{I}}$, σ_I, $\sigma_{\hat{I}}$, and $\sigma_{I\hat{I}}$ are the local means, standard deviations, and cross-covariance for images I and $\hat{I}$. The constants C_1 and C_2 are the regularization coefficients, used to avoid instability in image regions where the local mean or standard deviation is close to zero. In this work, we set the C_1 and C_2 parameters to $(0.01 \text{ x } L)^2$ and $(0.03 \text{ x } L)^2$, respectively, where L is the maximum possible pixel intensity range (65535) of our particular OCT image.

The final metric was the edge preservation index (EPI), proposed by Sattar *et al.* [40] to measure edge preservation between the denoised image and the corresponding frame-averaged image. The EPI is defined as follows:

$$EPI = \frac{\Gamma(\Delta s - \overline{\Delta s}, \widehat{\Delta s} - \overline{\widehat{\Delta s}})}{\sqrt{\Gamma(\Delta s - \overline{\Delta s}, \Delta s - \overline{\Delta s}).\Gamma(\widehat{\Delta s} - \overline{\widehat{\Delta s}}, \widehat{\Delta s} - \overline{\widehat{\Delta s}})}} \tag{10}$$

where Δs and $\widehat{\Delta s}$ are the Laplacian filtered version of the frame-averaged and denoised images, respectively. The gamma function $\Gamma(x, y)$ is the pixel-wise summation function and defined as follows:

$$\Gamma(x, y) = \sum_{i=1}^{w}\sum_{j=1}^{h} x(i, j).y(i, j) \tag{11}$$

with w and h are the image width and height, respectively.

In addition to measure the image quality and similarity between the frame-averaged and denoised images, we also

measured the processing performance of each method. This was done by averaging the processing time of each image over the testing dataset.

E. Qualitative Image Evaluation

The evaluation of the denoising performance was also conducted with qualitative experiments. Experts mean opinions were collected to quantify the image quality subjectively. The experts were selected who mostly work with medical image processing. This selection was chosen because it is difficult to find an expert who is working specifically on OCT images for normal tissues. Therefore, we defined three non-diagnostic evaluation points to rank the denoised images.

1) Sharpness: This first point evaluates if the tissue surface (border) is clearly visible or blurred. The highest score of 6 indicates that the image has clearly visible (sharp). On the other hand, the lowest score of 1 indicates difficulties in distinguishing tissue surfaces due to blurred images.

2) Contrast Details: This point evaluates the ability to discriminate structures below the tissue surface. The highest score of 6 indicates a clearly visible structural pattern. The lowest score of 1 indicates no structural pattern was noticeable.

3) Noise Level: This last point evaluates the noise level, such as shot noise, salt and paper noise, and Gaussian noise, on both the background and foreground area of the image. The highest score of 6 indicates a noise-free image. The lowest score of 1 indicates that the noise level is too high and hides the tissue structure.

Each expert was asked to evaluate a set of denoised images for each tissue type. A set of denoised images consisted of 17 images as described in section III.B (Table I).

F. Accuracy Comparison of Tissue Classifiers

Tissue classification is the primary aim of our proposed smart laser osteotomy scheme. Apart from showing the performance of the CNN for image denoising, we also explored the effect of image denoising on the accuracy of the tissue classifier. The CNN denoisers are expected to surpass the frame-averaging method with faster processing time and similar image quality, without significant change in its effectiveness of increasing the tissue classifier's accuracy.

We acquired additional image datasets to evaluate changes in the tissue classifier's accuracy due to image denoising. Tissue images of a pig were used to train and validate the classifier. The number of tissues and scan locations were comparable to the datasets used to train the CNN image denoisers (five tissue types and 140 scan locations per tissue). Both raw and frame-averaged image pairs were extracted for each scan location. In total, 700 image pairs (scan locations) were used to train and validate the classifier (with a fraction of 0.7 and 0.3, respectively).

For each scan location, we also denoised the raw images using each denoising method investigated in this study, separately. However, unlike the denoising performance experiments, we only tested the deep learning models which were trained with 7000 training datasets. In summary, each scan

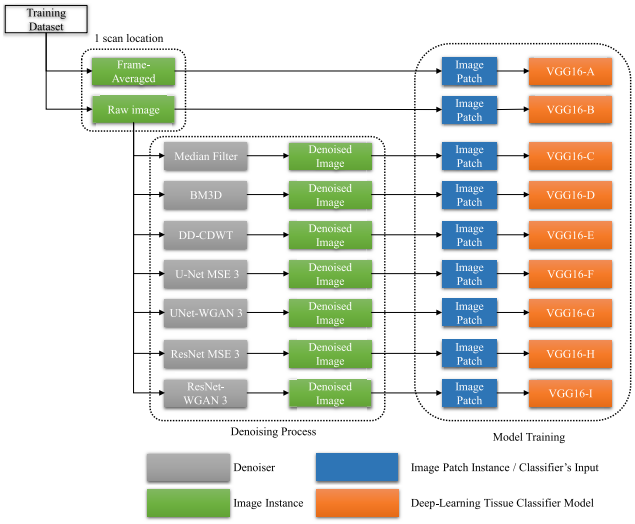

Fig. 8. Training Process of the tissue classifier. A pair of raw and frame-averaged images were acquired for each scan location. Then, the raw image was denoised using the denoisers that were investigated in this paper (median-filtered, BM3D,DD-CDWT, UNet-MSE 3, UNet-WGAN 3, ResNet-MSE 3, and ResNet-WGAN 3). A total of nine image instances (green) were acquired for each scan location. A tissue classifier (orange) was trained to differentiate tissue type based on an image patch (blue) input that is extracted from each image instance (green). The input of the classifier was an image patch (blue) with sized of 128 $\times$ 128 pixels grayscale image.

location or image pair consists of nine images (raw, frame-averaged, median-filtered, BM3D, DD-CDWT, UNet-MSE 3, UNet-WGAN 3, ResNet-MSE 3, and ResNet-WGAN 3). Additionally, another tissue image from a different pig was used to test the classifier. Similarly, with the classifiers' training datasets, additional 700 image pairs (five tissue types and 140 scan locations per tissue) were acquired and used for testing.

Furthermore, different from the deep-learning models for image denoising, which uses the full (1024 x 256 pixels) OCT image as the input, the input for the classifier was an image patch. The image patch was a 128 x 128 pixels grayscale image, selected at the center of the tissue surface border on the full OCT image (Fig. 1, step 2). This location represents the location of the laser ablation which is designed to be coaxially aligned with the center of the OCT image. We defined the ablation spot as always in the lateral center of the image. Vertical Canny edge detection method was used to trace the tissue surface in axial direction because of its simplicity and low sensitivity to noise [37], [38]. To have a similar patch location across the classifiers' dataset, we extracted an image patch at the same location for all nine images in each scan location.

Therefore, we trained all of the nine classifiers separately and compared their accuracies in classifying tissue type. The process of denoising and model training is illustrated in Fig. 8. We trained the first classifier to classify tissue type using the frame-averaged image patches. Then, we trained the second classifier using the raw image patches to classify tissue type. We used the second classifier as a reference classifier to compare the performance of the other classifiers. Next, the

third classifier was trained to classify tissue type using the median filtered image patches. The fourth and fifth classifiers were trained separately to classify tissue type based on image patches that denoised using the BM3D and DD-CDWT, respectively. Furthermore, we also trained classifiers using the image patches that were denoised with the CNN models. The sixth and seventh classifiers were trained separately to classify tissue based on image patches that were denoised with the UNet-MSE 3 and UNet-WGAN 3, respectively. Finally, we trained the eighth and ninth classifiers separately based on image patches that were denoised with the ResNet-MSE 3 and ResNet-WGAN 3, respectively.

We used the VGG16-Net [41] model as the base model for all classifiers which are often used for image-based object recognition. We customized the size of the input layers to fit the image patch size. The output layer was a soft-max activated layer, consisting of five neurons for classifying the tissue type (bone, bone marrow, fat, muscle, and skin). All hidden layers were equipped with the Rectified Linear Unit (ReLU) activation function. All the VGG16-Net models were trained and tested using the same workstation to train the CNN denoising models. We initialized all the weight in the models with a similar seed value to ensure fair training. We trained the models with 1000 epochs. The classifiers training was also done in mini-batches, with a batch size of 16 images, to fit our GPU's memory capacity. We defined the cross-categorical entropy as the training loss function and Adam (learning rate = 1.0×10^{-4}) as the training optimizer. We implemented weight decay (L_2 penalty multiplier set to 5.0×10^{-4}) regularizers for all convolutional layers and dropout regularizers for the last two fully connected layers (dropout ratio set to 0.2) as described in the reference [41].

V. Results and Discussion

After training all of the defined CNN denoiser models, we then applied them to denoise a set of testing images. Comparisons of the denoised images for the bone tissue are shown in Fig. 9. More image comparisons for the other tissue types are shown in the supplementary materials (Fig. 4-8). In this section, we discussed the benefit of using the CNN models for improving the raw image quality. We also compared the results with the frame-averaging, median filter, BM3D, and DD-CDWT. We also measured the similarity between the frame-averaged images with the images that were denoised with the CNN models, median filter, BM3D, DD-CDWT, and the raw images. Furthermore, we measured and compared the processing time for these denoisers to denoise the raw images. Finally, we trained a tissue classifier for each denoiser which used the corresponding denoised image as the training dataset. We compared the changes in the classifier's accuracy relative to a classifier that was trained with the frame-averaged images.

A. Quantitative Evaluation Results

The quantitative measurements were done to compare the SNR and CNR of the denoised images. Table II shows the averaged SNR and CNR of the raw images and the denoised images using the denoisers investigated in this paper. As the reference method, the frame-averaging effectively improves the image quality of the raw images. However, the frame-averaged images have less SNR compared to almost all other denoised images. The frame-averaging method calculates the mean individual pixel intensity over the temporal domain (frames) and maintains the speckle information in the image. Meanwhile, the other denoising methods effectively reduce the noise (including speckle) in the spatial domain. Here, the SNR is calculated based on the mean and noise in the spatial domain. Nevertheless, the frame-averaged images have higher CNR compared to almost all of the denoised images.

TABLE II
The Mean SNR and CNR of the Raw Images, Frame-Averaged Images, and the Images That Were Denoised With the CNN Models, Median Filter, BM3D, and DD-CDWT. Best Value Is Denoted With Bold Text

Method	Quality Metrics	
	Mean SNR (dB)	Mean CNR (dB)
Raw	52.830	0.855
Frame-Averaged	55.858	3.022
Median Filter	58.964	2.598
BM3D	56.338	2.228
DD-CDWT	57.244	2.105
UNet-MSE 1	**61.759**	2.656
UNet-MSE 2	60.027	2.858
UNet-MSE 3	57.923	2.889
UNet-WGAN 1	55.087	1.582
UNet-WGAN 2	53.677	3.419
UNet-WGAN 3	57.281	2.922
ResNet-MSE 1	56.019	**3.566**
ResNet-MSE 2	57.127	2.972
ResNet-MSE 3	57.038	2.899
ResNet-WGAN 1	60.629	2.570
ResNet-WGAN 2	57.299	2.741
ResNet-WGAN 3	57.830	1.999

Most of the deep learning methods also improved the SNR of the raw images better than frame-averaging method, except the UNet-WGAN 1 and UNet-WGAN 2. Moreover, the UNet-WGAN 2 and ResNet-MSE 1 improved the CNR of the raw images better than frame-averaging method. The CNN models learned to denoise images through convolutional spatial filters that minimize the loss function. Similar to the median filter, the convolutional filters also consider the neighboring pixels (kernel) for each individual pixel in the image. Additionally, the deep learning regularizers give better generalization to reduce the noise without over-smoothing the speckle pattern and edge details.

Investigation on the generalization capacity of the CNN models has also been conducted by variation of training data sizes. Among the deep learning models with MSE loss, the UNet-MSE models have lower SNR when trained with a higher number of training sizes. Meanwhile, although there is a slight decrease of CNR between the ResNet-MSE 2 and ResNet-MSE 3, the ResNet-MSE models tend to have higher SNR with higher number of training sizes. We also observed that the UNet-MSE models tend to have higher CNR with higher number of training sizes. In contrast, lower CNR is noticeable with higher number of training sizes for the

TABLE III
THE SIMILARITY METRICS OF THE FRAME-AVERAGED IMAGES WITH THE IMAGES THAT WERE DENOISED WITH THE CNN MODELS, MEDIAN FILTER, BM3D, AND DD-CDWT IMAGES. BEST VALUE IS DENOTED WITH BOLD TEXT

Method	Similarity Metrics		
	Mean EPI	Mean PSNR (dB)	Mean SSIM
Median Filter	0.267	34.950	0.979
BM3D	0.466	34.347	0.973
DD-CDWT	0.473	34.404	0.971
UNet-MSE 1	0.529	36.419	0.986
UNet-MSE 2	0.512	37.136	0.986
UNet-MSE 3	0.577	38.906	**0.992**
UNet-WGAN 1	0.542	35.051	0.988
UNet-WGAN 2	0.572	32.195	0.986
UNet-WGAN 3	**0.677**	39.984	**0.992**
ResNet-MSE 1	0.665	39.324	**0.992**
ResNet-MSE 2	0.671	**40.094**	**0.992**
ResNet-MSE 3	0.667	40.033	**0.992**
ResNet-WGAN 1	0.558	33.888	0.986
ResNet-WGAN 2	0.615	32.743	0.989
ResNet-WGAN 3	0.629	30.466	0.989

ResNet-MSE models. On the other hand, we observed a non-linear relation between variation of training data size with both SNR and CNR for all deep learning models with WGAN loss.

Although all of the denoiser performed well in improving the image's SNR and CNR, the image quality measurements were insufficient in quantifying the extent to which denoiser images preserved better structural details, especially for comparing between the deep learning models with WGAN loss. Additional comparison metrics have been done to show the averaged similarity between the frame-averaged images with the images that were denoised with the CNN models, median filter, BM3D, DD-CDWT, and the raw images, respectively. The measurements considered the image's edge (EPI), structural (SSIM), and intensity (PSNR) preservation (or similarity) concerning the frame-averaged images. These metrics give clearer information about the preservation of the tissue structural information, which was previously inferred from the frame-averaged images. The measurement results are shown in Table III.

The results show that the median filter improved the SNR better than the frame-averaging method. However, although the median filter reduced the speckle noise, the median filtered images have the lowest EPI compared to the other denoiser, which indicates the loss of sharp edge details. Meanwhile, the deep learning methods have higher EPI and SSIM than the median filter, indicating better preservation of the sharp edge details. Most of the deep learning denoised images have higher PSNR than the median filtered, BM3D, and DD-CDWT denoised images, except those denoised by the UNet-WGAN 2, ResNet-WGAN 1, ResNet-WGAN 2, and ResNet-WGAN 3. The similarity measurement also shows that the BM3D and DD-CDWT denoised images have higher EPI, which better preserved the edge details than the median filter. These methods still kept some residual noise on the resulting image, leading to lower PSNR and SSIM than the median filter. The SSIM of the BM3D and DD-CDWT denoised images were lower than the deep learning denoising methods.

The image comparisons in Fig. 9 show that the deep learning models with the MSE loss respond to uncertainty with smoothing (blurring) [30], [42]. Although this problem could be solved by increasing the number of training data sizes. The WGAN based model performed better generalization by keeping the speckle noise as one feature to distinguish between the reference and generated images. Therefore, the WGAN-based models kept a small amount of artificial speckle noise in the generated images.

The results also show that the image quality and similarity were slightly higher for the ResNet-based models than UNet-based models. However, since the image quality and similarity differences between the deep learning models are relatively small, it is difficult to conclude the best model. We believe that these image quality and similarity metrics are insufficient to measure the difference between the deep learning models. Further qualitative measurements are needed to visually inspect the tissue's complex anatomical structures preserved by the proposed deep learning methods.

B. Qualitative Evaluation Results

Evaluation of the denoised image quality based on the quantitative metrics alone may be insufficient to show the visual improvement of the denoised images. Additional qualitative evaluations to show the performance of the CNN models were also conducted. Visual inspection of the denoised images was conducted by six experts. The experts were shown 17 image versions of the same image as described in Table I. The image file names were number coded to hide the corresponding denoising method, which was used to denoise the image before being shown to experts. The experts were asked to evaluate the sharpness, contrast, and noise level of the images as explained in section III.E. The experts survey results for the image sharpness, contrast, and noise level are shown in Figs 10, 11, and 12, respectively.

The survey shows that the experts voted the frame-averaged images to have the highest sharpness. In contrast, the median filter method has been voted to show a low preservation of sharpness. The UNet-MSE 1 and 2 are also voted almost similarly with the median filter. Figs. 9f and 9g confirmed that the images denoised with the UNet-MSE 1 and 2 are over-smoothed. On the other hand, UNet-WGAN 1 and ResNet-WGAN 1 were partly voted to preserve the image sharpness. Although Figs. 9i and 9o show sharp images, loss of tissue-specific textural features are noticeable.

In the second part of the survey, we asked the experts to compare the perceived contrast of the images. The experts voted that the UNet-MSE 1 and 2 have low contrast, even lower than the median filter. Similar to the sharpness evaluation, the UNet-WGAN 1 and ResNet-WGAN 1 were also partly voted to preserve the image contrast. Almost similar performance was observed between the frame-averaging methods and the ResNet-WGAN 2 models in the eyes of the experts.

The noise level evaluation showed that the UNet-WGAN 3 model reduced the noise as well as the frame-averaging

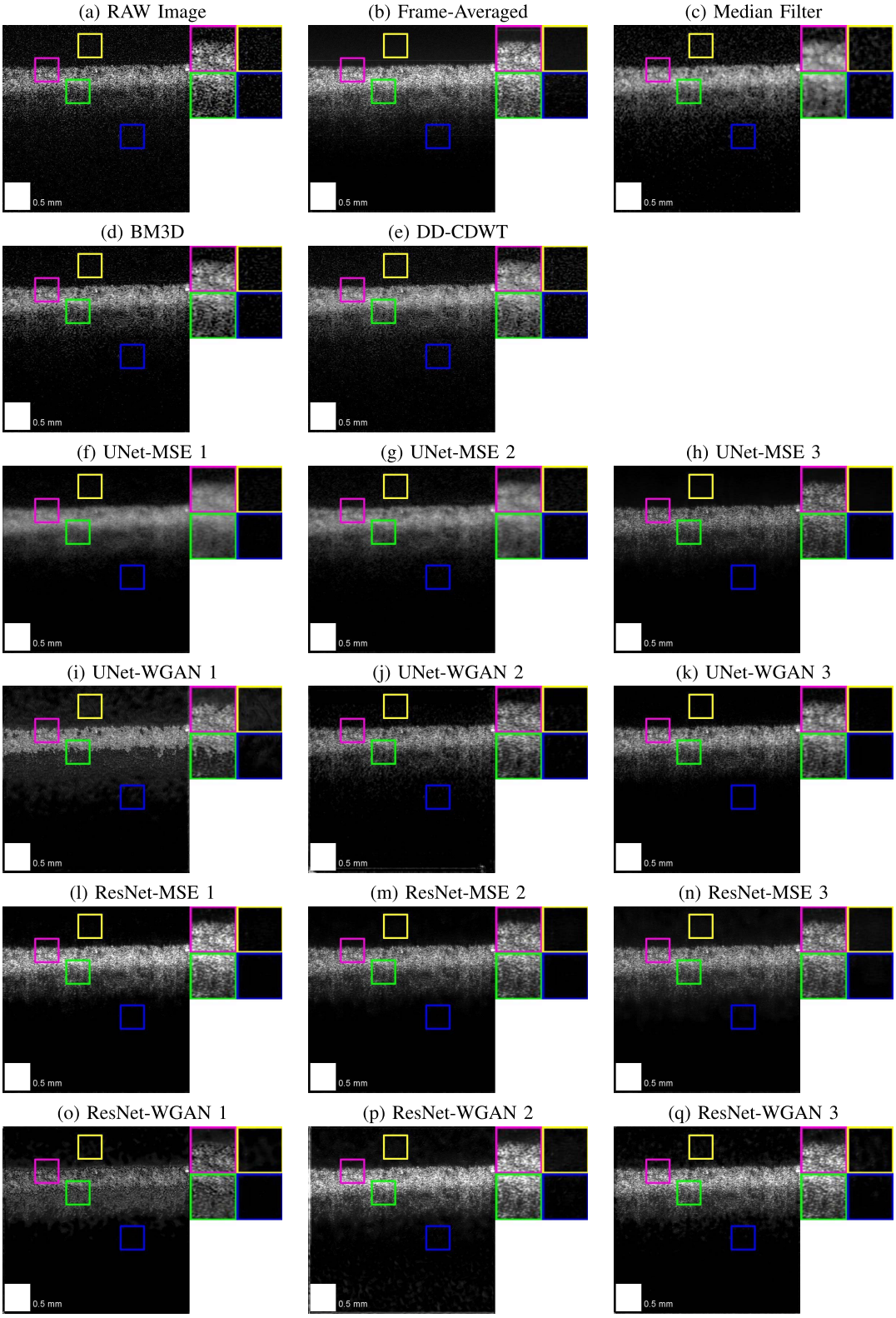

Fig. 9. Image comparison for bone tissue. From top left to bottom right, starting with the raw-image (a), frame-averaged image (b), and the raw images that were denoised using the median filter (c), BM3D (d), DD-CDWT (e), UNet-MSE 1 (f), UNet-MSE 2 (g), UNet-MSE 3 (h), UNet-WGAN 1 (i), UNet-WGAN 2 (j), UNet-WGAN 3 (k), ResNet-MSE 1 (l), ResNet-MSE 2 (m), ResNet-MSE 3 (n), ResNet-WGAN 1 (o), ResNet-WGAN 2 (p), and ResNet-WGAN 3 (q). The colored boxes show the zoomed version of regions inside the images. Yellow box shows the background region above the tissue surface. Magenta box shows the surface region of the tissue. Green box shows the region inside the tissue with high signal. Blue box shows the background region with low to no signal inside the tissue.

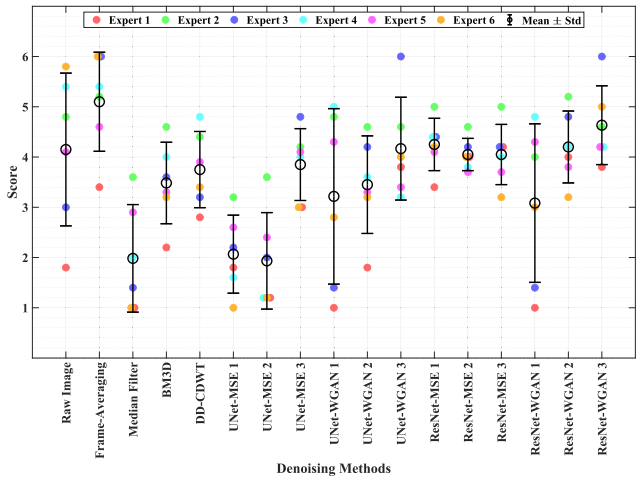

Fig. 10. Error bar plot represents the survey results for the sharpness of the raw and denoised images. The mean scores are indicated in black circles, with the whiskers indicating one standard deviation of the scores (equally up and down). Individual scores from each expert are shown in dots with the same color. Scores with similar values are shown side by side horizontally to avoid overlapping.

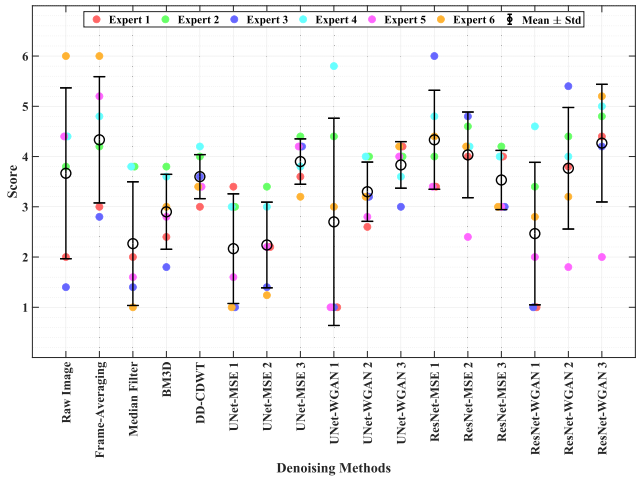

Fig. 11. Error bar plot represents the survey results for the contrast of the raw and denoised images. The mean scores are indicated in black circles, with the whiskers indicating one standard deviation of the scores (equally up and down). Individual scores from each expert are shown in dots with the same color. Scores with similar values are shown side by side horizontally to avoid overlapping.

method. The UNet-WGAN 1 only received two votes, which indicates that the denoised images have relatively higher noise level, confirmed in Fig. 9i. The images showed that the denoiser completely hid the textural features in the tissue. Other than that, the median filtered and BM3D has almost similar noise level as the raw images.

In summary, it is difficult to conclude which denoising methods performed best due to the high variation of scores among the experts. It is because the experts work with different medical imaging modalities, including OCT for ophthalmology and ultrasound imaging.

C. Improvement in Classifier Accuracy

We collected 700 additional OCT image pairs to demonstrate the effect of image denoising on a tissue classifier's

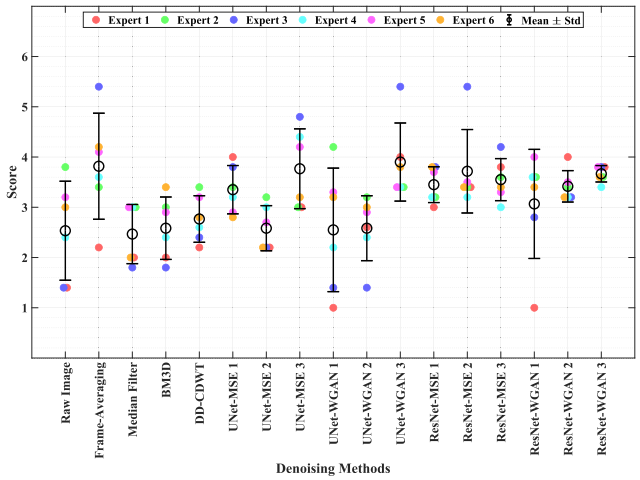

Fig. 12. Error bar plot represents the survey results for the noise level of the raw and denoised images. The mean scores are indicated in black circles, with the whiskers indicating one standard deviation of the scores (equally up and down). Individual scores from each expert are shown in dots with the same color. Scores with similar values are shown side by side horizontally to avoid overlapping.

accuracy. Each image pair consists of raw and frame-averaged images. We denoised the raw images for each image pair or scan location using each denoisers investigated in this study separately. However, we only investigated the deep learning models which were trained with 7000 images. Images patches were extracted from the denoised images and used as input for the tissue classifier. We trained a tissue classifier for each denoising method and used the corresponding denoised image patches as the training dataset. In total, nine VGG16-Net tissue classifiers were trained separately. After training the models, we tested them with the test dataset consisting of 700 denoised OCT image patches. A detailed explanation regarding the training process is explained in section IV-F.

The results in Table IV show that all of the denoiser increased the accuracy of the tissue classifier except the median filtering method. The tissue classifier trained with raw images has an accuracy of 86.43 %. As a reference, the tissue classifier trained with frame-averaged images has an accuracy of 91.29 %, the highest accuracy compared to the other tissue classifiers. Although the frame-averaged image still contained an element of the speckle noise, they also contained more information about the tissue structure, gathered from several image frames, compared to single raw denoised images. Therefore, the tissue classifier trained with frame-averaged images could learn more features than that trained with single raw denoised images. However, the tissue classifiers trained using deep learning denoised images have the closest accuracies to the tissue classifier trained using the frame-averaged images. In comparison, the ResNet models improved the tissue classifier's accuracy more than the UNet models. This is because the ResNet models reduced the speckle noise better than the UNet models. Therefore, the tissue classifier could focus more on the signal-carrying speckle to differentiate tissue type.

Here, it is also shown that the MSE-based deep learning models have higher accuracy compared to the WGAN-based models. This discrepancy appears because the WGAN

TABLE IV
THE AVERAGE ACCURACY OF THE CLASSIFIER TRAINED USING THE RAW IMAGES, FRAME-AVERAGED IMAGES, AND THE IMAGES THAT IS DENOISED USING BM3D, DD-CDWT, UNET-MSE 3, UNET-WGAN 3, RESNET-MSE 3, AND RESNET-WGAN 3. BEST VALUE IS DENOTED WITH BOLD TEXT

Denoising Method	Classifier Average Accuracy (%)
Raw	86.43
Frame-Averaging	**91.29**
Median Filter	85.57
BM3D	88.86
DD-CDWT	88.71
UNet-MSE 3	90.83
UNet-WGAN 3	89.73
ResNet-MSE 3	91.03
ResNet-WGAN 3	89.86

generated more artificial noise, lowering the image quality and similarity, as discussed in the previous section.

The BM3D and DD-CDWT methods also increased the tissue classifier's accuracy. However, the increases are lower than the frame-averaging and deep learning because of the residual noise in the denoised images. On the other hand, the median filter even reduced the tissue classifier's accuracy. These results indicate that the deep learning-based methods could replace the classical digital filtering methods without significantly altering the structural information in the image for tissue classification. Nevertheless, further investigation is needed to optimize the tissue classifier's performance by training and testing with more data samples.

D. Processing Time

One aim of this work was to find a fast algorithm for denoising the OCT images. We evaluated the performance of the trained models on our OCT computer with the following specifications: 2.8 GHz Intel Core i7 processor, 16 GB 1867 MHz DDR3 memory, and equipped with a GPU NVIDIA GTX 1050Ti. We compared the average processing time of the proposed denoising methods. However, the processing performance comparison of the CNN models was done between UNet and Resnet only, as the generator in both MSE and WGAN based loss model. The results are given in Table V. Our OCT system required 9 msec to acquire a single B-Scan (raw) image. Therefore, the frame-averaging was the longest denoising method which required 300 raw images. The median filter was the fastest algorithm as it only calculates a small-sized median kernel over the input image. However, this achievement must be considered alongside the median filter's image similarity result, which erased information over the image. In the second position, the deep learning methods denoised the raw image faster than the BM3D and DD-CDWT methods. The deep learning based models were faster since they could be run in parallel on the GPU, whereas the classical digital filtering methods were run on the CPU. Furthermore, among the deep learning methods, the ResNet models denoised a single raw image in only 0.078 s, which is faster than the UNet-based models that denoised a single raw image in 0.084 s.

TABLE V
THE AVERAGE TIME REQUIRED TO DENOISE THE OCT IMAGES USING THE FRAME-AVERAGING, BM3D, DD-CDWT, UNET-MSE, UNET-WGAN, RESNET-MSE, AND RESNET-WGAN. NOTE THAT DEEP LEARNING ALGORITHMS WERE RUN USING GPU. THE MEASUREMENT ALSO INCLUDES ACQUISITION TIME OF 300 RAW IMAGES FOR THE FRAME-AVERAGING METHOD. BEST VALUE IS DENOTED WITH BOLD TEXT

Denoising Method	Average Time (s)
Single Raw	**0.009**
Frame-Averaging	2.731
Median Filter	0.010
BM3D	2.495
DD-CDWT	0.202
UNet	0.084
ResNet	0.078

Our findings confirm the possibility of achieving real-time image denoising for our smart laser osteotomy approach. All of the deep learning models denoise the raw image below 90 msec. Additionally, the tissue classification using VGG16-Net needs an average time of 0.034 s. This leads to a total processing time of 0.112 s for the acquisition, denoising with ResNet, and classification steps, which is slightly slower than the optimum repetition rate of the ablation laser. Our previous experiment showed that the optimum pulse repetition rate of our laser ablation was 10 Hz (100 msec per pulse) [43].

VI. CONCLUSION AND FUTURE WORKS

This work demonstrates the ability of deep learning methods to mimic the frame-averaging method for denoising OCT images of five normal tissue types. The deep learning methods produced better image quality and similarity than the classical digital filtering methods. Even though the median filter could increase the SNR and CNR better than some of the deep learning methods, it failed to maintain the structural information of the image. Furthermore, the processing speed of the deep learning-based method was also faster than the BM3D, DD-CDWT, and frame-averaging methods. The quantitative and qualitative experiment results suggest that the deep learning methods are a feasible alternative to the frame-averaging method for real-time OCT image denoising, a necessary subprocess for our smart laser osteotomy approach.

Moreover, we also showed that denoising the OCT image increased the tissue classifier's accuracy. The frame-averaging method improved the tissue classifier's accuracy better than the other denoising methods. Furthermore, the tissue classifier has better accuracy when trained using the images that are denoised by deep learning methods than the classical digital filter methods. It proves that the deep learning methods could mimic the frame-averaging method better than the classical digital filter methods.

In the future, we will integrate the deep learning denoising method into our OCT device for monitoring laser ablation in real-time. We are aware that integrating the tissue classifier and image denoising problem will further increase the processing time. The average total time for acquisition, denoising, and

classification steps is slightly slower than the optimum repetition rate of the ablation laser. Further study will involve optimization of the tissue classifier for better accuracy and faster prediction time. One of the optimizations includes using a faster CPU and GPU. Other than that, we also plan to use the classification accuracy directly as one of the loss functions for deep learning denoising models. Thus, reducing the processing time for both denoising images and predicting tissue type simultaneously seem feasible.

REFERENCES

[1] K.-W. Baek *et al.*, "A comparative investigation of bone surface after cutting with mechanical tools and Er: YAG laser," *Lasers Surg. Med.*, vol. 47, no. 5, pp. 426–432, Jul. 2015.

[2] A. Trompeter, J. Dabis, O. Templeton-Ward, A. E. Lacey, and B. Narayan, "The history, evolution and basic science of osteotomy techniques," *Strategies Trauma Limb Reconstruction*, vol. 12, no. 3, pp. 169–180, Nov. 2017.

[3] S. Kondo *et al.*, "Thermological study of drilling bone tissue with a high-speed drill," *Neurosurgery*, vol. 46, no. 5, pp. 1162–1168, 2000.

[4] S. Stübinger, "Advances in bone surgery: The Er: YAG laser in oral surgery and implant dentistry," *Clin., Cosmetic Investigational Dentistry*, vol. 2, pp. 47–62, Jun. 2010.

[5] K.-W. Baek *et al.*, "Clinical applicability of robot-guided contact-free laser osteotomy in cranio-maxillo-facial surgery: *In-vitro* simulation and *in-vivo* surgery in minipig mandibles," *Brit. J. Oral Maxillofacial Surg.*, vol. 53, no. 10, pp. 976–981, Dec. 2015.

[6] H. Abbasi, G. Rauter, R. Guzman, P. C. Cattin, and A. Zam, "Laser-induced breakdown spectroscopy as a potential tool for autocarbonization detection in laserosteotomy," *J. Biomed. Opt.*, vol. 23, no. 7, p. 1, Mar. 2018.

[7] R. Kanawade *et al.*, "Qualitative tissue differentiation by analysing the intensity ratios of atomic emission lines using laser induced breakdown spectroscopy (LIBS): Prospects for a feedback mechanism for surgical laser systems," *J. Biophotonics*, vol. 8, nos. 1–2, pp. 153–161, Jan. 2015.

[8] F. Mehari *et al.*, "Investigation of laser induced breakdown spectroscopy (LIBS) for the differentiation of nerve and gland tissue—A possible application for a laser surgery feedback control mechanism," *Plasma Sci. Technol.*, vol. 18, no. 6, pp. 654–660, Jun. 2016.

[9] H. N. Kenhagho, G. Rauter, R. Guzman, P. C. Cattin, and A. Zam, "Optoacoustic tissue differentiation using a Mach–Zehnder interferometer," *IEEE Trans. Ultrason., Ferroelectr., Freq. Control*, vol. 66, no. 9, pp. 1435–1443, Sep. 2019.

[10] N. Kenhagho *et al.*, "Characterization of ablated bone and muscle for long-pulsed laser ablation in dry and wet conditions," *Materials*, vol. 12, no. 8, p. 1338, Apr. 2019.

[11] E. Bay, A. Douplik, and D. Razansky, "Optoacoustic monitoring of cutting efficiency and thermal damage during laser ablation," *Lasers Med. Sci.*, vol. 29, no. 3, pp. 1029–1035, May 2014.

[12] V. Periyasamy, C. Özsoy, M. Reiss, X. L. Deán-Ben, and D. Razansky, "*In vivo* optoacoustic monitoring of percutaneous laser ablation of tumors in a murine breast cancer model," *Opt. Lett.*, vol. 45, no. 7, pp. 2006–2009, 2020.

[13] M. E. Brezinski *et al.*, "Optical coherence tomography for optical biopsy," *Circulation*, vol. 93, no. 6, pp. 1206–1213, 1996.

[14] A. Hamidi, Y. A. Bayhaqi, F. Canbaz, A. A. Navarini, P. C. Cattin, and A. Zam, "Long-range optical coherence tomography with extended depth-of-focus: Avisual feedback system for smart laser osteotomy," *Biomed. Opt. Exp.*, vol. 12, no. 4, pp. 2118–2133, Apr. 2021.

[15] J. Schmitt, S. Xiang, and K. Yung, "Speckle in optical coherence tomography," *J. Biomed. Opt.*, vol. 4, no. 1, pp. 95–105, 1999.

[16] W. Wu, O. Tan, R. R. Pappuru, H. Duan, and D. Huang, "Assessment of frame-averaging algorithms in oct image analysis," *Ophthalmic Surg. Lasers Imag. Retina*, vol. 44, no. 2, pp. 75–168, 2013.

[17] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, "Image denoising by sparse 3-D transform-domain collaborative filtering," *IEEE Trans. Image Process.*, vol. 16, no. 8, pp. 2080–2095, Aug. 2007.

[18] S. Huang, C. Tang, M. Xu, Y. Qiu, and Z. Lei, "BM3D-based total variation algorithm for speckle removal with structure-preserving in OCT images," *Appl. Opt.*, vol. 58, no. 23, pp. 6233–6243, Aug. 2019.

[19] I. W. Selesnick, "The double-density dual-tree DWT," *IEEE Trans. Signal Process.*, vol. 52, no. 5, pp. 1304–1314, May 2004.

[20] H. Liu, S. Lin, C. Ye, D. Yu, J. Qin, and L. An, "Using a dual-tree complex wavelet transform for denoising an optical coherence tomography angiography blood vessel image," *OSA Continuum*, vol. 3, no. 9, pp. 2630–2645, Sep. 2020.

[21] D. Perdios, A. Besson, M. Arditi, and J.-P. Thiran, "A deep learning approach to ultrasound image recovery," in *Proc. IEEE Int. Ultrason. Symp. (IUS)*, Sep. 2017, pp. 1–4.

[22] Q. Yang *et al.*, "Low-dose CT image denoising using a generative adversarial network with Wasserstein distance and perceptual loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1348–1357, Jun. 2018.

[23] K. J. Halupka *et al.*, "Retinal optical coherence tomography image enhancement via deep learning," *Biomed. Opt. Exp.*, vol. 9, no. 12, pp. 6205–6221, Dec. 2018.

[24] Y. Huang, N. Zhang, and Q. Hao, "Real-time noise reduction based on ground truth free deep learning for optical coherence tomography," *Biomed. Opt. Exp.*, vol. 12, no. 4, pp. 2027–2040, Apr. 2021.

[25] Z. Mao *et al.*, "Deep learning based noise reduction method for automatic 3D segmentation of the anterior of lamina cribrosa in optical coherence tomography volumetric scans," *Biomed. Opt. Exp.*, vol. 10, no. 11, pp. 5832–5851, 2019.

[26] B. Baumann *et al.*, "Signal averaging improves signal-to-noise in oct images: But which approach works best, and when?" *Biomed. Opt. Exp.*, vol. 10, no. 11, pp. 5755–5775, 2019.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Cham: Switzerland: Springer, 2016, pp. 630–645.

[29] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 214–223.

[30] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 105–114.

[31] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of Wasserstein gans," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5769–5779.

[32] J. Sarvaiya, S. Patnaik, and S. Bombaywala, "Image registration by template matching using normalized cross-correlation," in *Proc. Int. Conf. Adv. Comput., Control, Telecommun. Technol. (ACT)*, 2009, pp. 819–822.

[33] S. Makita, M. Miura, S. Azuma, T. Mino, T. Yamaguchi, and Y. Yasuno, "Accurately motion-corrected Lissajous oct with multi-type image registration," *Biomed. Opt. Exp.*, vol. 12, no. 1, pp. 637–653, 2021.

[34] MATLAB. *Registering an Image Using Normalized Cross-Correlation*. [Online]. Available: https://www.mathworks.com/help/images/registering-an-image-using-normalized-cross-correlation.html

[35] F. Chollet *et al.* (2015). *Keras*. [Online]. Available: https://keras.io

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–15.

[37] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[38] S. Luo, J. Yang, Q. Gao, S. Zhou, and C. A. Zhan, "The edge detectors suitable for retinal oct image segmentation," *J. Healthc. Eng.*, vol. 2017, p. 3978410, 2017.

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[40] F. Sattar, L. Floreby, G. Salomonsson, and B. Lövström, "Image enhancement based on a nonlinear multiscale method," *IEEE Trans. Image Process.*, vol. 6, no. 6, pp. 888–895, Jun. 1997.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015, pp. 1–14.

[42] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," 2016, *arXiv:1603.08155*.

[43] L. M. B. Bernal *et al.*, "Optimizing controlled laser cutting of hard tissue (bone)," *Automatisierungstechnik*, vol. 66, no. 12, pp. 1072–1082, Dec. 2018.