MACHINE INTELLIGENCE-ENABLED MULTIMODAL BIOMEDICAL IMAGING

By

Aniwat Juhong

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Electrical and Computer Engineering – Doctor of Philosophy

2025

ABSTRACT

Due to the rapid development of computational technologies, deep-learning-based approaches have emerged as practical and promising remedies for a wide range of biomedical applications. This dissertation demonstrates the utilization of deep learning approaches across multiple modalities in the field of biomedical applications: histopathology image analysis, multispectral optoacoustic tomography (MSOT), computed tomography (CT), magnetic particle imaging (MPI), and Raman spectroscopy. The first deep learning application is convolutional neural networks (CNNs) for resolution enhancement and nuclei segmentation of hematoxylin and eosin (H&E) images. This deep learning-based approach could facilitate cancer diagnosis using H&E images acquired by a low resource setting. The second application is based on hybrid recurrent and convolutional neural networks to generate sequential cross-sectional MSOT images in order to reduce the acquisition time. Essentially, the proposed deep learning model can generate the missing sequential MSOT images in the data acquired by a large step size setting, resulting in a comparable resolution to the data acquired by a small step size setting. The third application is an efficient end-to-end deep learning model based on U-Net architecture and a multi-head attention mechanism for MPI-CT image segmentation. This proposed model can directly segment the MPI signal from the co-registered MPI-CT image with promising performance. Lastly, it is a custom-made Raman spectrometer together with computer vision-based positional tracking and monocular depth estimation using deep learning for the visualization of 2D and 3D surface-enhanced Raman Scattering (SERS) nanoparticles (NPs) imaging, respectively. The combination of Raman spectroscopy, image processing, deep learning, and SERS molecular imaging shows the robust and feasible potential for clinical applications.

ACKNOWLEDGMENTS

I would like to express my gratitude for the support and collaboration provided by numerous people who helped me to refine my Doctor of Philosophy and complete this dissertation. First and foremost, I wish to wholeheartedly thank Prof. Zhen Qiu for his unwavering faith in my ability to carry out this project, his financial support, and his spirit of adventure. He is an exceptional mentor throughout my Ph.D. journey. This dissertation is a result of his vision to develop deep learning for numerous useful biomedical applications. I also appreciate invaluable suggestions from my committee members: Prof. Nelson Sepúlveda, Prof. Wen Li, Prof. Ming Han, and Prof. Xuefei Huang to complete this dissertation. Without their evaluation time, feedback, and guidance, I would not improve my research skills and broaden my horizons. Secondly, I am profoundly grateful to Prof. Christopher H. Contag and Prof. Wibool Piyawattanametha for their support and providing me with an opportunity to work at the Institute for Quantitative Health Science and Engineering (IQ), Michigan State University. Moreover, I would like to thank my colleagues, including Dr. Bo Li, Dr. Cheng-You Yoa, Dr. Chia-Wei Yang, Dr. Kunli Liu, Dr. Brett Volmert, Yifan Liu, and A.K.M. Atique Ullah for their collaboration. Last but not least, this acknowledgement would not be complete without mentioning my parents and family. Their substantial encouragement, affection, and support led me to rise above all the difficulties during my Ph.D. study.

TABLE OF CONTENTS

CHAPTER 1: Introduction

1.1 Deep learning overview

Deep learning is one of machine learning approaches that utilize multiple layers of data representation to effectively capture the unique features of the input data at different stages, demonstrating exceptional performance in a wide range of applications such as image classification, image segmentation, natural language processing, data generative, etc. As a result, deep learning has been rapidly developed in recent years, encompassing methodological constructions and actual implementation. Indeed, deep learning employs computational models consisting of numerous layers of processing to acquire and represent data with higher level of abstraction, and it can implicitly capture complex patterns in extensive datasets. The growing amount of data that can be gathered through biomedical and clinical data needs the advancement of deep learning techniques to handle, such as Convolution Neural networks (CNNs), Recurrent neural networks (RNNs), Attention mechanisms, and Transformer based neural networks to process and evaluate the data. Some examples of biomedical devices that commonly apply deep learning include Computed Tomography (CT), Magnetic Resonance imaging (MRI), Magnetic Particle imaging (MPI), Ultrasound, photoacoustic tomography, optical microscopy and tomography and so on. Specifically, this dissertation demonstrates deep learning for biophotonics and molecular imaging applications, which are multidisciplinary life sciences, combining the principles of optics, photonics and biology to investigate biological systems at tissue, cellular, and molecular levels. The field of biophotonics is one of the essential parts for the development of unprecedented diagnostic and therapeutic approaches in the biomedical field; therefore, it has been significantly improved over decades, particularly the use of deep learning techniques to empower biophotonics research by enabling advanced image analysis, improved image and signal

processing, and the ability to comprehensively analyze biophotonics data.

1.2 Organization of the dissertation

This dissertation is divided into four chapters for four different modalities and applications. Additionally, there is a fifth chapter addressing future research. Chapter 2 demonstrates approaches based on deep learning for super-resolution and segmentation for histology images. The two proposed deep learning models in this chapters were jointly trained together to reach the join optimization to perform both resolution enhancement and segmentation for breast cancer H&E images. In chapter 3, a deep learning application for generating the sequential multispectral sequential Multispectral Optoacoustic Tomography (MSOT) Imaging is presented. The aim of this work is to reduce the acquisition time without any hardware modifications. In this work, the mice injected with ICG-conjugated superparamagnetic iron oxide nanoworms particles (NWs-ICG) were scanned under the MSOT system providing three imaging modalities: photoacoustic, ultrasound, and NWs-ICG acoustic images. The proposed deep learning can reduce the acquisition time of volumetric imaging for these three modalities. Chapter 4 shows the MPI signal segmentation of MPI-CT images, which is significantly important for MPI quantification. This work proposed a novel architecture based on U-Net architecture and attention mechanisms that can surpass other state-of-the-art models. Lastly, chapter 5 shows the application of depth map estimation based on deep learning in tandem with surface enhanced Raman scattering (SERS) for the image-guidance surgery application. With depth information, the SERS is more practical for a real clinical application. The final chapter concludes the dissertation, on-going work related to biomedical applications as well as possible future work.

CHAPTER 2: Super-resolution and Segmentation Deep Learning for Breast Cancer Histopathology Image Analysis

Traditionally, a high-performance microscope with a large numerical aperture is required to acquire high-resolution images. However, images' size is typically tremendous. Therefore, they are not conveniently managed and transferred across a computer network or stored in a limited computer storage system. As a result, image compression is commonly used to reduce image size resulting in poor image resolution. Here, we demonstrate custom convolution neural networks (CNNs) for both super-resolution image enhancement from low-resolution images and characterization of both cells and nuclei from hematoxylin and eosin (H&E) stained breast cancer histopathological images by using a combination of generator and discriminator networks so-called super-resolution generative adversarial network-based on aggregated residual transformation (SRGAN-ResNeXt) to facilitate cancer diagnosis in low resource settings. The results provide high enhancement in image quality where the peak signal-to-noise ratio and structural similarity of our network results are over 30 dB and 0.93, respectively. The derived performance is superior to the results obtained from both the bicubic interpolation and the well-known SRGAN deep-learning methods. In addition, another custom CNN is used to perform image segmentation from the generated high-resolution breast cancer images derived with our model with an average Intersection over Union of 0.869 and an average Dice Similarity Coefficient of 0.893 for the H&E image segmentation results. Finally, we propose the jointly trained SRGAN-ResNeXt and Inception U-net Models, which applied the weights from the individually trained SRGAN-

ResNeXt and Inception U-net Models as the pre-trained weights for transfer learning. The jointly trained model's results are progressively improved and promising. We anticipate these custom CNNs can help resolve the inaccessibility of advanced microscopes or whole slide imaging (WSI) systems to acquire high-resolution images from low-performance microscopes located in remote-constraint settings.

2.1 Introduction

Pathology diagnosis is routine work usually performed by a skilled pathologist or cytologist. The diagnosis process begins with staining (typically hematoxylin and eosin or H&E) of a specimen on a glass slide and observing it under a high-resolution (HR) microscope. Typically, the diagnosis process for each biopsy slide could take up to 15-20 mins per slide which is very time-consuming. Pathologists must visually scan over a vast field of view to find any abnormalities on each slide. Therefore, whole slide imaging (WSI) has been introduced to solve this main problem [1]. The WSI refers to scanning a complete microscope slide and creating a single high-resolution digital file. This is commonly achieved by capturing many small HR image tiles or strips and then montaging them to create a full image of a histological section. The WSI equipped with pathological image diagnosis software is changing the workflow of many laboratories. Specimens on glass slides can now be transformed into HR digital files that can be efficiently stored, accessed, and analyzed. The latter is due to the advancement of computer vision and convolution neural networks (CNNs) algorithms in digital pathological image analysis [2, 3].

However, in resource-constraint settings, accessibility of both HR microscope and WSI is a crucial obstacle to delivering quality health care, frequently resulting in undertreatment and overtreatment of infectious diseases based on clinical assessment alone [4]. Laboratory infrastructure is typically clustered in urban settings and is relatively inaccessible in regions where

significant portions of the affected population reside [5]. Many of the neglected diseases in particular, are more prevalent in rural areas, far from these diagnostic centers [6]. Therefore, novel, simple, and inexpensive approaches to perform digital pathological diagnoses are needed in both clinical and public health environments. Potential solutions are to provide a software-based solution to help transform low-resolution (LR) to either HR or super-resolution (SR) images.

Due to the rapid development of computational technologies, deep-learning-based diagnosis has become a sought-after technique for digital pathology image analysis implementation [2, 3]. Depending on the analysis, the technique can be divided into supervised and unsupervised learning. Supervised learning aims to define a function that can map input images to their outputs or labels (normal cells, abnormal cells, cancer cells, and other parameters) such as classification or segmentation problems. On the other hand, the purpose of unsupervised learning is to define another function that can extract the latent features and structures from unlabeled data such as clustering problems, dimensional reduction, and super-high-resolution problems. Several studies use CNNs for nuclei segmentation [7-11]. Those methods can surpass the traditional methods such as Otsu segmentation [12], Watershed method [13], and K-mean clustering [14] since the traditional methods are sensitive to parameter setting and could be effective for specific data types. CNNs based approaches have become practical tools for nuclei and cell segmentation tasks as they can achieve a resounding success. HoverNet [15] is one of the effective CNNs for nuclei segmentation. The model predicts horizontal and vertical distance between a nucleus centroid to its corresponding foreground pixels. Masker-controlled watershed is then applied as the post-processing method to obtain nucleus instances. However, the HoverNet results can be sensitive to the noise in the distance maps because of the marker-controlled watershed. StarDIST [16] is another CNNs for nuclei segmentation that predicts centroid probability maps to localize the

nuclei. The predicted centroids are applied to generate polygons to determine the boundary and the number of the cells. The downside of the StarDIST is that polygons are only predicted using the centroid pixels' features. These results in a lack of contextual information for large nucleus instances and could affect prediction accuracy. CPP-Net[17] extends the StarDIST by integrating the rich contextual information from a sampled point set for each centroid pixel and applying the Shape-Award Perceptual loss that constrains CPP-Net's predictions regarding the nucleus shape.

U-net architecture is a renowned convolution neural network architecture for image segmentation. It is widely used for biomedical image segmentation [18]. Its structure is simple convolution blocks, and the skip connections are added from decoder to encoder. The U-net architecture allows for simultaneously using global location and context and it works with very few samples to improve the model performance. In addition, it is an end-to-end process for the entire image in the forward pass and directly generates the segmentation image. Its structure is also simple to be modified or assembled with other models. Potentially, the performance of the U-net can be improved by using other effective convolution architectures to replace the simple convolution blocks. In recent years, CNNs have also been applied for super high-resolution biomedical images with a wide range of imaging modalities [19-25] such as fluorescence imaging, light-sheet imaging, and color imaging of pathological slides. However, those works employed the same concept of SRGAN [26] that the generator is built using the ResNet architecture or residual structure[27]. Indeed, several architectures can surpass the residual structure. Exploring one of them and applying it to the generative adversarial network (GAN) will be more worthwhile. For instance, the DenseNet [28] network is applied as the backbone for SGAN namely ESRGAN [29] showing the impressive result and surpassing the original SRGAN model. According to the Top-1 and Top-5 accuracy vs. computational complexity testing reported on Benchmark Analysis of

Representative Deep Learning Neural Networks Architectures [30], the ResNeXt CNNs architectures can outperform state-of-the-art (SOTA) architectures such as ResNet, DenseNet, Inception, etc., even the complexity of ResNeXt is somewhat less than others. Recently, deep learning techniques based on transformer architectures [31] have emerged as an alternative to the CNNs architectures since they can provide better results on large datasets. However, the transformer architectures are more complicated and require a high computation cost. If the model is excessively complicated, it will be challenging to build the jointly trained models to simultaneously update the weights of the joint models due to the restriction of computing resources (time, memory, speed, etc.).

To overcome limitations in digital pathological diagnosis, we describe a novel method for transforming LR digital pathological images derived from low-cost microscopes to super-resolution (SR) images (equivalent to a 40x magnification) with a super-resolution generative adversarial convolution neural network technique based on ResNeXt architecture [32] (SRGAN-ResNeXt) [22]. Most SRGAN deep learning works for biomedical image enhancement used a single residual network (ResNet) in each layer to capture and extract image features, while our deep learning used the ResNext architecture instead. Typically, the ResNet architecture can exceptionally perform on very deep convolution layers since the skip connection in the ResNet adds the input information to the output of the convolution layers. Therefore, the output of ResNet contains the representative features from the convolution operation and the critical information from the original input. Moreover, the skip connection allows the gradient to effortlessly backpropagate and update the weight to minimize the loss value. However, the single residual block might be insufficient to capture all significant features. Therefore, to increase the model capability, we apply residual blocks in parallel (stacking the same topology blocks) for each layer

(ResNeXt architecture). Utilizing the ResNeXt architecture not only improves the feature capturing but also reduces the complexity of the model in preference to make it deeper since hyper-parameters (width, filter sizes, etc.) are shared. This approach can provide considerable resolution enhancement for poor-quality images. Training the SRGAN-ResNeXt Model requires a dataset consisting of high-resolution images (ground truth) and corresponding low-resolution images. We used a commercial microscope (Nikon Eclipse Ci) to prepare a dataset for training this model. Peak Signal to Noise Ratio (PSNR) and Structural image similarity method (SSIM) was used to evaluate the generated images from our model, which are 32.92 dB and 0.93, respectively. These are promising results as they are higher than the original SRGAN Model's evaluation results that were trained on the same data set (H&E images). Furthermore, we applied the Inception U-net Model [33], the improved U-net Model by using Inception architecture as a backbone in the U-net network for H&E image segmentation. To train the Inception U-net Model, a large number of H&E images are required to be accurately masked on nuclei areas which are very time-consuming. Thus, we used a dataset from a cancer imaging archive [34] to train our Inception U-net Model. Our inception U-net Model's Union (IoU) and Dice Similarity Coefficient (DSC) are 0.869 and 0.893, respectively. Since the SRGAN-ResNeXt and Inception U-net model were separately trained, the performance of both models could be improved by jointly training them together as the segmentation loss and the generator loss could be effectively back propagated to update the weights for the generator model and Inception U-net model with a joint optimization.

Figure 1 shows the overall workflow of the models. First, the breast tumor H&E slides were prepared on biopsy slides (Figure 1(a)-(b)) to be imaged with a 40x magnification (Figure 1(c)), then acquired the images' quality was downgraded by downsampling and adding blurring noise. Therefore, the model has both corresponding ground truth (high-resolution images) and low-

resolution images for training the SRGAN-ResNeXt (Figure 1(d)-(f)). Eventually, the well-trained generator model from the SRGAN-ResNeXt (Figure 1(h)) was applied to the unseen low-resolution image (Figure 1(g)) to enhance its quality by generating the high-resolution image (Figure 1(i)). Furthermore, the generated high-resolution image was characterized as its resolution was substantially improved and contained considerable details that were impossible to perform before applying the model. In other words, our approach can tackle those low-resolution images by applying the Inception U-net Model (Figure 1(j)) to the generated high-resolution images (the output of the generator model from SRGAN-ResNeXt). As a result, the newly generated image can be segmented and quantified to characterize the nuclei's density, size, and morphology.

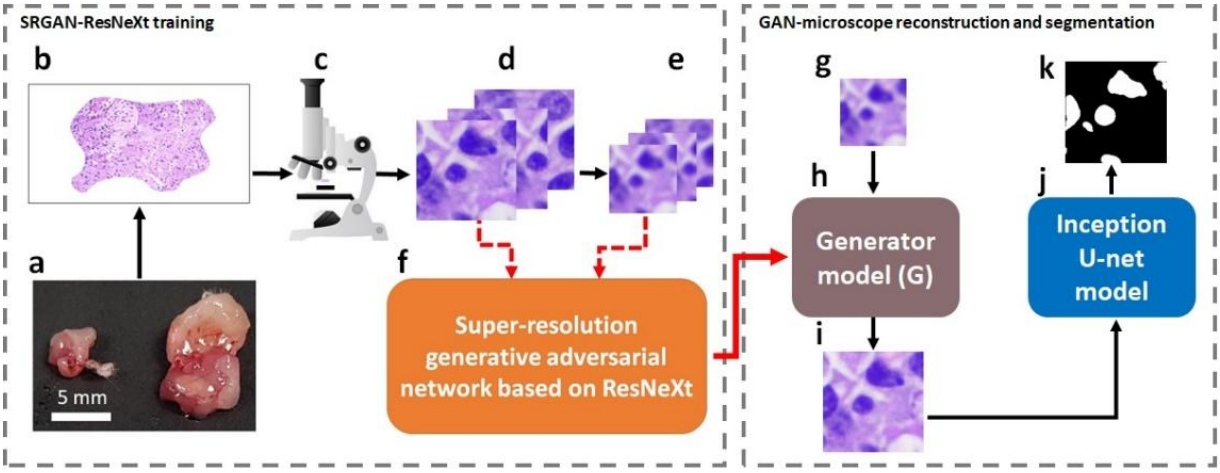


Figure 1. The workflow of super-resolution and segmentation deep learning. (a) Fresh breast tumor tissues. (b) The corresponding H&E stained tissue slides. (c) A commercial microscope (Nikon Eclipse Ci) for capturing the H&E stained tissue slide images. (d) High-resolution images acquired by the microscope. (e) Simulated low-resolution images. (f) The training SRGAN- ResNeXt network. (g) The unseen low-resolution image. (h) The generator model from SRGAN-ResNeXt. (i) The generated high-resolution image. (j) The Inception U-net Model for segmentation. (k) The segmented H&E image.

2.2 Methods

2.2.1 Proposed SRGAN-ResNeXt architecture

Here, we propose SRGAN-ResNeXt architecture built from scratch to synthesize super-resolution images from low-resolution images. The concept of the SRGAN-ResNeXt is similar to the

traditional GAN that consists of generator and discriminator models. The generator and discriminator models of our SRGAN-ResNeXt are depicted in Figure 2(a) and Figure 2(b), respectively. The generator model takes a low-resolution image as the input and generates a high-resolution image after passing through the convolution, ResNeXt, and upsampling layers. The discriminator model is utilized to distinguish the generated image from the ground-truth image by taking them as the input and providing probability as the output. The ultimate goal of SRGAN-ResNeXt is to train the generator model to synthesize the image that can fool the discriminator completely. To achieve this, we need to design the generator model properly, use a large number of images as the dataset to train the models, and fine-tune the hyperparameters thoroughly. To train SRGAN-ResNeXt, we first trained the discriminator model by freezing the generator model. Next step, we used an adversarial network to train the generator model. The adversarial network (Figure 2(c)) is the combined models, which are the generator model, discriminator model, and VGG19-the latter works as the feature extractor [35].

2.2.2 Generator model

The generator network is a deep convolution network containing the pre-residual layer, 16 parallel-residual layers (ResNeXt), a post-residual layer, two upsampling layers, and the final convolution layer as shown in Figure 2(a). To assemble the generator model, the pre-residual block is the first block, which contains a single 2D convolution layer and ReLU is used as the activation function. The second block is 16 parallel-residual layers (ResNeXt architecture). Each layer after convolution layers is followed by a batch normalization with 0.8 of momentum value and the activation function is also ReLU. For the ResNeXt block, the size of transformation sets or branch numbers is defined as cardinality. Increasing the number of cardinalities can improve and better the performance of the convolution neural network. However, the excessive number of

cardinalities could lead to expensive computation. Thus, we use eight cardinalities for our generator model [Figure 2(a)], which is the optimal number of our task. The next block is the post-residual block, the simple convolution layer, and batch normalization (momentum =0.8). After that, the fourth block is the upsampling block, which has two sub-pixel convolution layers [36], upsampling the scale by four times. Lastly, the last convolution layer uses the Tanh activation function to form the generated image with R, G, and B color channels. To train the generator model, we need to use the joint model, which is the adversarial network [Figure 2(c)]. The discriminator and VGG19 models are untrainable during training the generator model.

2.2.3 Discriminator model

The discriminator network [37] is a relatively simple convolution network, comprising eight convolutional layers and two fully connected layers, designed to evaluate the similarity between the ground truth and generated images. After each convolution block, a batch normalization layer is used, followed by an activation function named the Leaky ReLU function (α=0.2). The number of 3x3 filter kernels increases by a factor of 2 from 64 (the first layer) to 512 (the eighth layer) kernels similar to the VGG network. The last two layers are dense layers working as a classification block, predicting the probability of an image being either real or fake. We have to freeze the generator model or make it untrainable to train the discriminator model. The learning progress of the discriminator model is remarkably faster than the generator model. Therefore, during the training generator model, it must be slowed down learning progress which will be further discussed in the next section below.

2.2.4 Loss functions

The perceptual loss function (I^{SR}) is highly significant to the performance of the generator model in the SRGAN-ResNeXt network. It is the weighted sum of a content loss (VGG19 loss, $\mathrm{I}_{\mathrm{X}}^{\mathrm{SR}}$) and adversarial loss (Discriminator loss, $\mathrm{I}_{\mathrm{Gen}}^{\mathrm{SR}}$) as shown in Equation (1) as

$$I^{SR} = I_X^{SR} + C_w I_{Gen}^{SR}. \quad (1)$$

The generator exploits this loss function to optimize and update its trainable parameters. To achieve the well-trained generator model, the weight, C_w, was assigned to the loss value from the discriminator model to slow down the learning progress since the discriminator model can be trained faster than the generator model. If the discriminator model can excessively perform well to distinguish between the generated image and the ground truth image, we would not be able to come up with the exceptional generator model since the generated image cannot fool the discriminator model. In the original SRGAN training, C_w is a constant for the whole learning process. However, this weight started from 0.5 and increased to 0.05 for every 10,000 epochs in our model. Since the generator model will gradually improve its performance and capability, we have to balance the performance of both the generator and discriminator models. The total number of epochs for training our model was 50,000. Therefore, C_w was varied from 0.5 to 0.7.

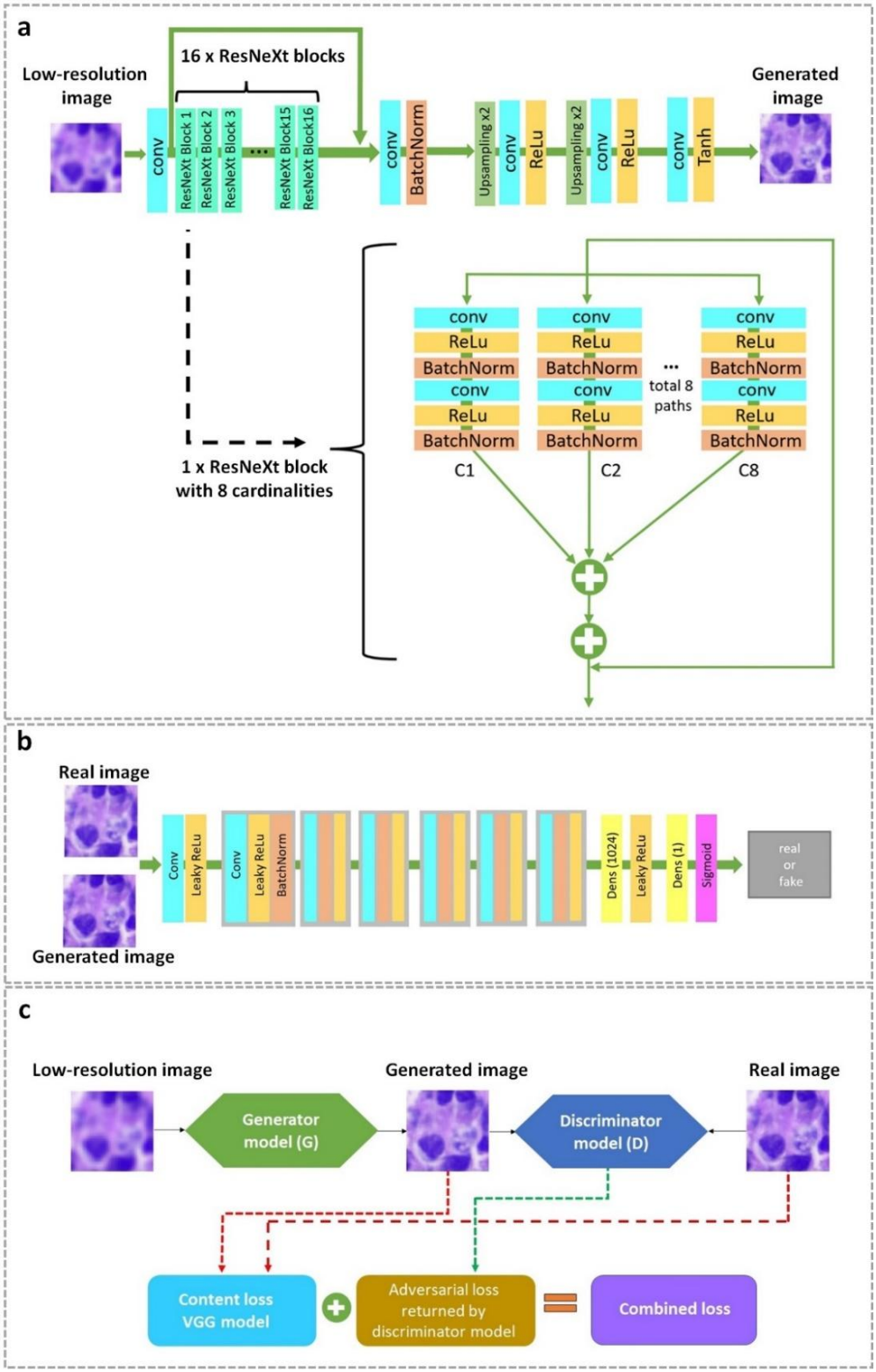


Figure 2. Super-resolution generative adversarial network-based on SRGAN-ResNeXt. (a) The architecture of the generator. (b) The architecture of the discriminator. (c) The combined models so-called adversarial model for training Generator model.

Albeit using the pixel-wise mean square error (MSE) to distinguish between the ground truth and the reconstructed image is undemanding to optimize, it returns a poor-quality image in terms of human perception. The output of MSE is the average features' difference of two data. Therefore, it cannot extract high-dimensional features. However, the content loss or VGG loss (I_X^{SR}) , is defined as the Euclidean distance between the feature map of the generated image $G_{\theta G}(I^{HR})$ and the ground truth, I^{HR}, can help solve this problem. The I_X^{SR} loss is based on ReLU activation layers of the pre-train 19-layer VGG network and it can be calculated following Equation (2) as shown as

$$I_{VGG}^{SR} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}(\emptyset_{i,j}(I^{HR})_{x,y} - \emptyset_{i,j}(G_{\theta G}(I^{LR})_{x,y}), \quad (2)$$

where $W_{i,j}\ and\ H_{i,j}$ describe the dimensions of the respective feature maps within the VGG network. The features map ($\emptyset_{i,j}$), can be obtained by the j-th convolution before the i^{th} maxpooling layer within the VGG19 network. Apart from using a feature map from VGG loss, the adversarial loss (I_{Gen}^{SR}) is also employed to differentiate the similarity of the two images. It is defined as the probabilities varying from 0 to 1, which is the result of the discriminator model ($D_{\theta_D}(G_{\theta_G}(I^{LR}))$) as shown in Equation (3) below as

$$I_{Gen}^{SR} = \sum_{n=1}^{N} -logD_{\theta_D}(G_{\theta_G}(I^{LR})). \quad (3)$$

The perceptual loss effectively leverages the combination of these two loss functions to train the generator model that can generate high-detailed images.

2.2.5 Dataset for training SRGAN-ResNeXt Model

To obtain breast cancer H&E images, the female MUC1 double-transgenic mice with breast tumors [38] were euthanized and their tumors were sent out to the histopathology lab (MSU-IHPL Research facility) to prepare the H&E stained breast tumor slides. All procedures performed on

animals were approved by the University's Institutional Animal Care & Use Committee (AUF 06/18-082-00) and were within the guideline of human care of laboratory animals. Four tumor mice were euthanized, and a tumor of each mouse was surgically removed to prepare four different tumor H&E slides. The H&E slides were then imaged by the commercial microscope (Nikon Eclipse Ci) with 40x magnifications to prepare the dataset for training SRGAN-ResNeXt. The size of each whole slide image is greater than 80,000 x 80,000 pixels and the image patches with a size of 256 x 256 pixels were extracted from each whole slide image with a 50 % overlapping area. The data augmentation was applied to these extracted image patches. The total number of image patches including the augmented images is over 13,000 images, which were used for training only. To prepare the low-resolution images, we downed sampling 4 times from the original high-resolution image patch and added blurring noise using the normalized boxed filter with kernel shown in Equation (4) below. We increased the kernel size until we could not discriminate the nuclei boundary and the simulated low-resolution images are even worse than some native low-resolution images.

$$K = \frac{1}{ksize.width * ksize.height}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}, \tag{4}$$

Where K is the blurring normalized boxed filter, ksize.width is the kernel width, and ksize.height is the kernel height. Figure 3(a) shows the cropping area from the large FOV H&E images. Figure 3(b) are the small patches that were cropped from the large FOV image.

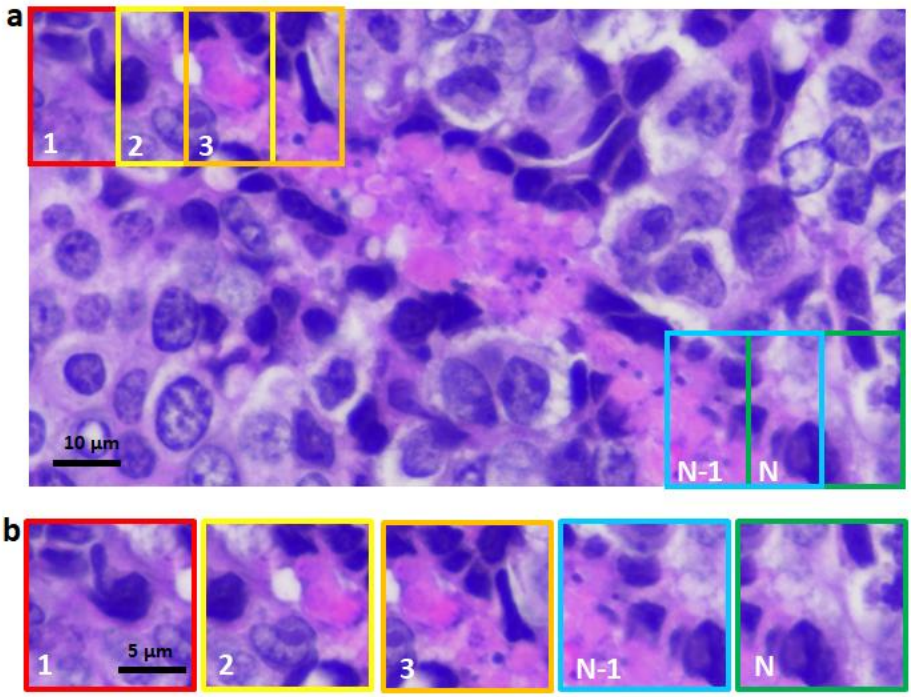


Figure 3. Data set preparation for training SRGAN-ResNeXt, cropped image with 50% overlapping area. (a) Large field of view H&E image, (b) The small patches of the large image (a) with 50% overlapping area.

2.2.6 The Inception U-net architecture

The conventional CNNs for image segmentation tasks have two main components: an encoder and a decoder. Similarly, the U-net architecture has these two parts, but the skip connection is the crucial mechanism that allows U-net to surpass the conventional method and perform better. This concept is akin to the residual block that the input (encoder part) will concatenate to the output (decoder part) at the same dimension. However, each layer of the original U-net architecture is a simple convolution block, which might be insufficient to extract some crucial information. For this reason, the Inception architecture [39] was applied to improve the capability of the U-net Model. Inception architecture uses a wide range of kernel sizes for the same input to simultaneously extract global and local features. A larger kernel size is suitable for the information distributed globally, whereas a smaller kernel size is appropriate for the information distributed locally. Consequently, the Inception CNN architecture can be satisfactorily performed to extract the feature from the data.

Here, we applied four different kernel sizes of the Inception blocks in our U-net Model as shown in Figure 4 below by replacing each convolution block in the original U-net architecture with the Inception blocks.

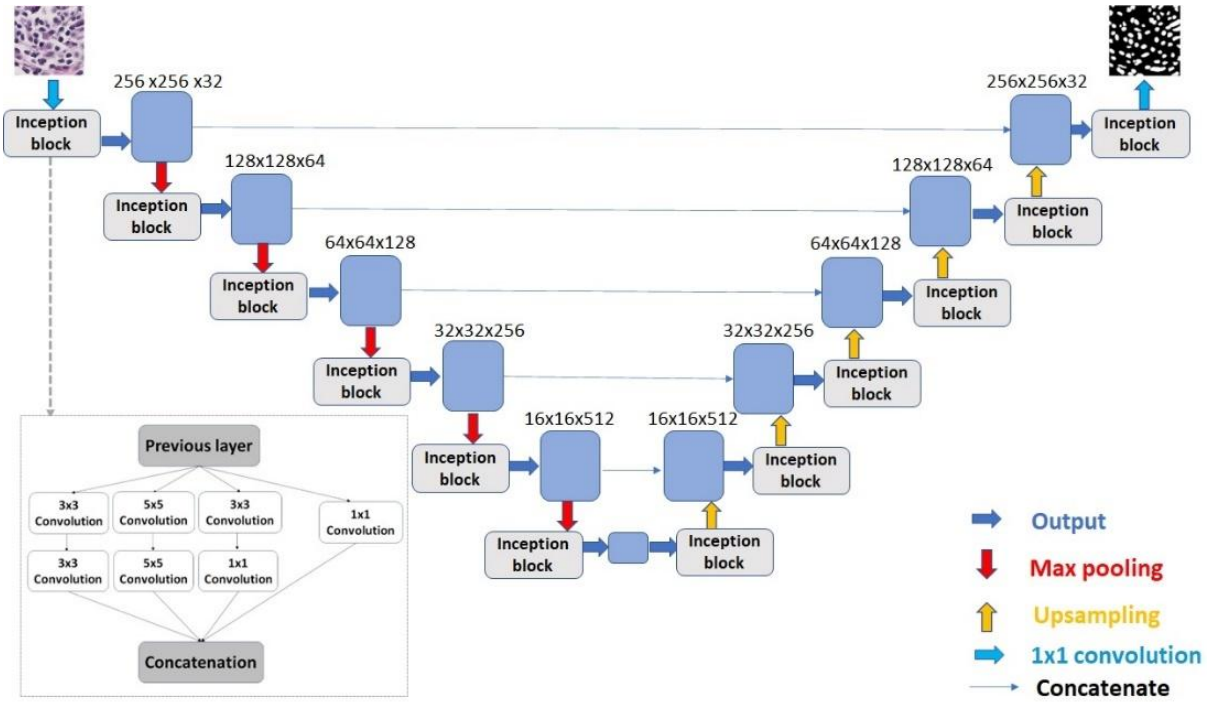


Figure 4. Inception U-net architecture for H&E image segmentation. Every single blue box corresponds to a multi-channel feature map. The value over the boxes represents the number of channels.

Figure 4 illustrates the Inception U-net architecture. The first part is the encoder (the left side of Figure 4) where the Inception convolution blocks are utilized instead of the simple convolution blocks. All Inception blocks in this part consist of different sizes (3x3, 5x5, and 1x1) parallel filters (Inception structure) followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with the stride of 2 steps for downsampling, respectively and this is the repeated process. The number of feature channels is double at each downsampling step. The second part is the decoder (the right side of Figure 4). It consists of a feature map upsampling followed by a 2x2 up-convolution (halving the number of feature channels), a corresponding concatenation from the decoder part, and Inception blocks. The ReLU activation is used for each block. The H&E images

and their corresponding segmentation masks are implemented to train this model as input and output, respectively. The loss function for U-net is a mean squared error (MSE) function as shown in Equation (5) shown below as

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \quad (5)$$

where the MSE is the average of the squared differences between ground truth (y_i) and predicted value from our model ($\hat{y}_i$) and N is the number of samples.

2.2.7 Data set for training the segmentation models

Since image segmentation is a supervised task, the outputs or targets need to be labeled, which is expensive and time-consuming. Fortunately, several datasets provide the H&E images and their corresponding nuclei masks. Here, we used the dataset from the cancer imaging archive[34]. This dataset provides nucleus segmentation for the whole cancer slide over 1,000 images in the cancer genome atlas (TCGA) repository. These images are from 10 different cancer types such as bladder urothelial carcinoma (BLCA), invasive breast carcinoma (BRCA), cervical squamous cell carcinoma, and endocervical adenocarcinoma (CESC).

2.2.8 Jointly trained SRGAN-ResNeXt and Inception U-net Models

The SRGAN-ResNeXt and Inception U-net Models were jointly trained by using the separately trained weights of the SRGAN-ResNeXt Model and the Inception U-net Model as the pre-trained weights for transfer learning. Figure 5(a) shows the joint models for training the generator model. The conception of the jointly trained generator (JTG) Model is akin to the adversarial model shown in Figure 2(c). Still, the JTG Model employs not only the content loss (returned by the VGG19 Model) and the adversarial loss (returned by the discriminator model) but also the segmentation loss of the generated high-resolution image and ground truth high-resolution image (returned by

the jointly trained Inception U-net). The combined loss of the JTG Model is shown in Equation (6) as

$$I^{JG} = I_X^{SR} + C_w I_{Gen}^{SR} + C_{w2} I_{GenS}^{SRS}, \quad (6)$$

Where I^{JG} is the combined loss of the jointly trained generator model, $\mathrm{I}_{\mathrm{X}}^{\mathrm{SR}}$ is the content loss (VGG19 loss), $\mathrm{I}_{\mathrm{X}}^{\mathrm{SR}}$ is the adversarial loss (Discriminator loss), I_{GenS}^{SRS} is the segmentation loss (Jointly trained Inception U-net loss), and C_w & C_{w2} are hyperparameters. The VGG19 Model, the discriminator model, and the jointly trained Inception U-net Model are fixed as untrainable during training the JTG Model.

The jointly trained Inception U-net (JTIU) Model was trained using the generated high-resolution image (returned by the JTG Model) and the ground truth of the high-resolution image as the model's inputs. The outputs of both inputs have the same ground truth to calculate the loss value. Therefore, the JTIU can learn how to generate the same quality segmentation image from both generated high-resolution images and native high-resolution images. During training the JTIU Model, the JTG Model was fixed as well.

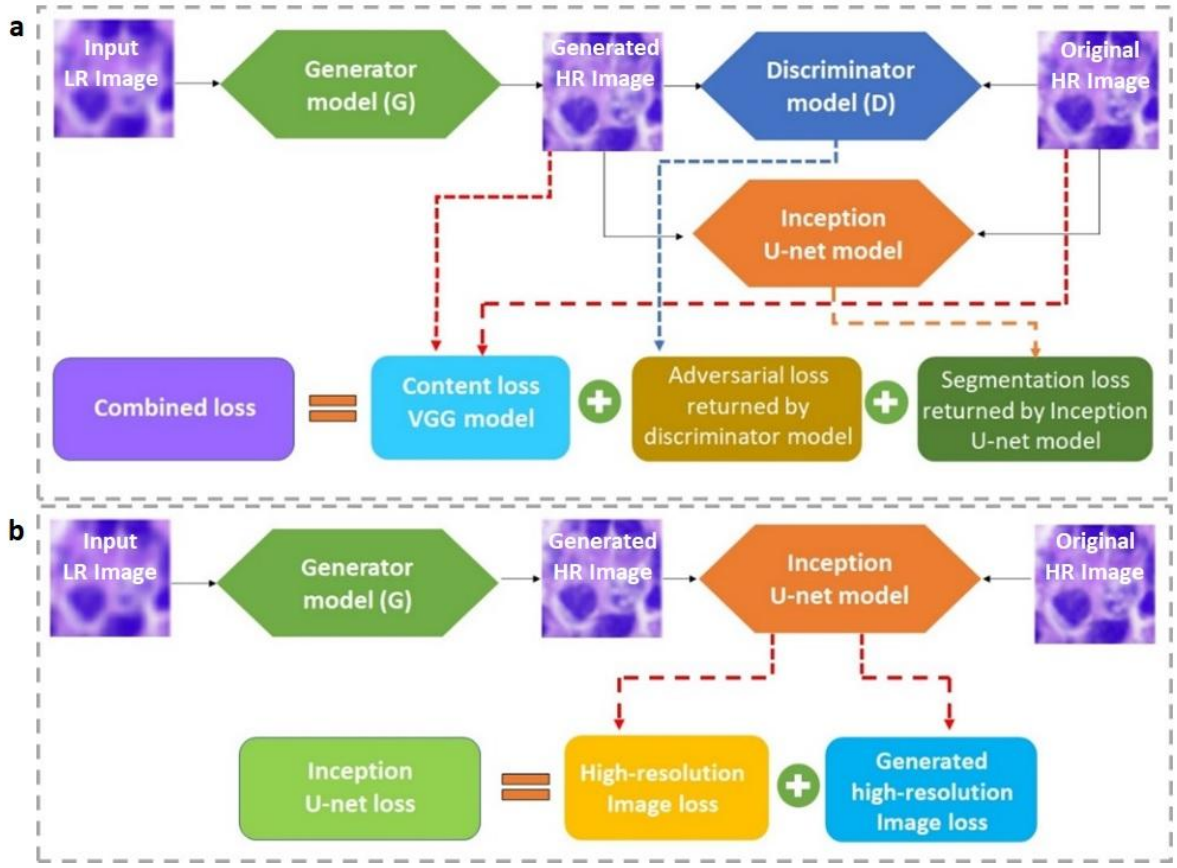


Figure 5. Jointly trained SRGAN-ResNeXt Model and Inception U-net Model. (a) The assembled models for the jointly trained generator (JTG) Model. (b) The assembled models for the jointly trained Inception U-net (JTIU) Model.

2.2.9 Data set for the jointly trained Models

Two other tumor mice were sacrificed, and a tumor of each mouse was prepared for H&E slides. Therefore, we have two tumor H&E slides from different mice for training the jointly trained models. The 220 image patches with a size of 256 x 256 pixels were randomly extracted from these H&E slides (110 patches per slide). 210 and 10 patches were used for training and testing, respectively. Each image patch was manually labeled for the ground truth of segmentation. Thus, this dataset contains low-resolution, high-resolution and segmentation images.

2.2.10 Training implementations

The separately trained SRGAN-ResNeXt and Inception U-net models were trained on Google Colaboratory-Pro (or Google Colab-pro) and implemented on the computer with a 9th Gen Intel Core i7-9750H CPU, 16 GB RAM, and an NVIDIA RTX 2060 graphic card. Since the jointly trained models require more resources for training due to the combination of several models, they were trained on Google Colaboratory-Pro+ (Google Colab Pro+), which provides Faster GPUs and significantly more memory than Google Colab-pro.

2.3 Results and discussion

2.3.1 Super high-resolution image reconstruction and segmentation.

The goal of SRGAN-ResNeXt is to have a well-trained generator model to reconstruct high-resolution images. We could not feed the large image into the generator model due to the computation restriction during implementation. Therefore, the large images were divided into serval small images. Furthermore, the overlapping area between these divided images was required to stitch them back to obtain the same field of view (FOV) as the original large image. Figure 6 shows the results of applying both the SRGAN-ResNeXt and the Inception U-net Models to a breast tumor H&E image. Figure 6(a1), 6(b1), and 6(c1) are the small patches of the whole slide image from different areas. All these small images were downscaled and added blurring noise as shown in Figure 6(a2), 6(b2), and 6(c2). The SRGAN-ResNeXt Model was employed to enhance these low-resolution images by synthesizing high-resolution images (Figure 6(a3), 6(b3), and 6(c3)). The Inception U-net was then applied to these generated high-resolution images for segmentation (Figure 6(a4), 6(b4), and 6(c4)).

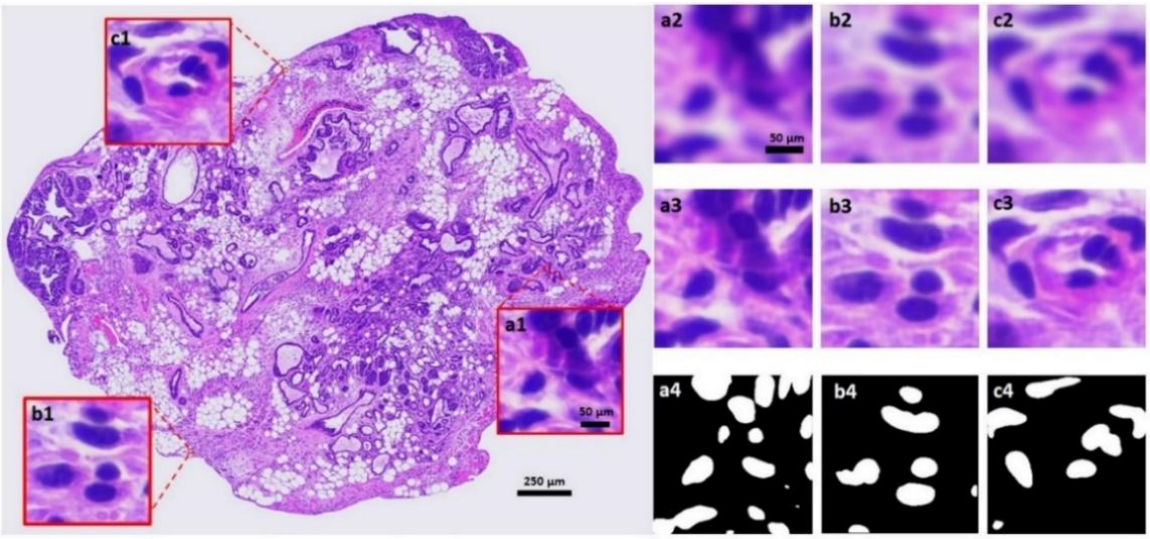


Figure 6. The whole slide image (WSI) of a breast tumor H&E slide and the result of our deep learning model. (a1, b1, and c1) The high-resolution images of the WSI from different areas. (a2, b2, and c2) The low-resolution images. (a3, b3, and c3) The reconstructed high-resolution images using our deep learning model (SRGAN-ResNeXt). (a4, b4, and c4) The corresponding nuclei segmentation to (a3, b3, and c3) using the Inception U-net Model.

Figure 7(a1) and 7(b1) show the low-resolution image and the enhanced-resolution image generated by the SRGAN-ResNeXt model, respectively. They were fed into the Inception U-net Model for nuclei segmentation. Figure 7(a2) shows the segmentation result of the low-resolution image and Figure 7(b2) shows the segmentation result of the enhanced image. It is relatively demanding to perform the image segmentation for the low-resolution image without enhancing its resolution first. The CNNs cannot extract meaningful features from the blurry pixels resulting in unsatisfactory segmentation performance. The mean square error (MSE) of blurry images and generated high-resolution images are 21.24 and 2.75, respectively. The MSE of the blurry image is significantly higher than the generated high-resolution image. To circumvent this issue, we propose to apply the SRGAN-ResNeXt Model to improve the poor-quality image before characterizing or performing segmentation to obtain better results. Figure 7(c1) and 7(c2) show the ground truth for high-resolution image and segmentation image, respectively.

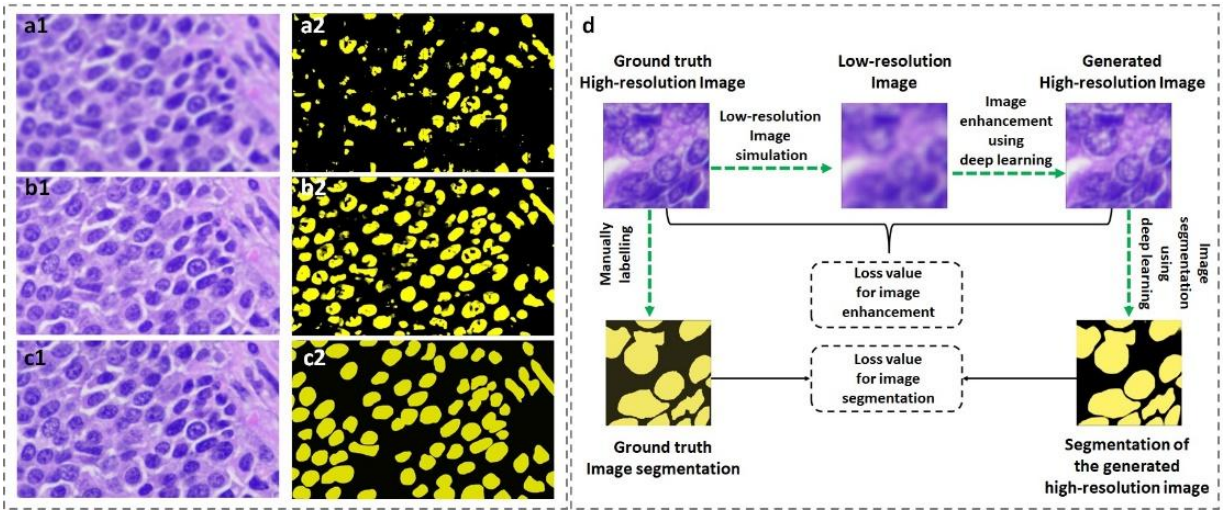


Figure 7. The H&E image segmentation of the low-resolution image and the enhanced-resolution image. (a1-a2) The low-resolution image and its segmentation image (output of the Inception U-net). (b1-b2) The enhanced-resolution image (output of the SRGAN-ResNeXt) and its segmentation image (output of the Inception U-net). (c1-c2) The ground truth of the high-resolution image and the segmentation image. (d) Ground truth preparation for both of the high-resolution image and the segmented image.

2.3.2. Performance of the SRGAN-ResNeXt Model

Peak signal to noise ratio (PSNR) is one of the ubiquitous methods used to quantify the quality of the generated image compared to the original image (ground truth) [31]. It is a ratio between the maximum possible power of a signal and the power of distorting noise, affecting its representation quality. The higher the PSNR, the better the quality of the generated image. To compute the PSNR, we have to calculate the mean squire error (MSE) first and use the Equation (7) below to define PSNR as

$$PSNR = 20log_{10}(\frac{MAX_f}{\sqrt{MSE}}) . \quad (7)$$

The MSE is defined as the following

$$MSE = \frac{1}{mn}\sum_0^{m-1}\sum_0^{n-1}\|f(i,j) - g(i,j)\|^2, \quad (8)$$

Where f is the matrix data of the ground truth,

g is the matrix data of the generated image,

m is the number of rows of pixels of the images,

i represents the index of that row,

n is the number of columns of pixels of the image,

j represents the index of that column, and

MAX_f is the maximum signal value that exists in our ground truth.

Structural similarity index measure (SSIM) is a perception-based model. It considers image distortion in terms of perceived change structural information (loss of correlation, luminance distortion, and contrast distortion) [40].

$$SSIM\ (x, y) = \frac{(2\mu_x\mu_y+c_1)(2\sigma+c_2)}{(\mu_x^2+\mu_y^2+c_1)(\sigma_x^2+\sigma_y^2+c_2)}, \quad (9)$$

Where

μ_x denotes the average of x,

μ_y denotes the average of y,

σ_x^2 denotes the variance of x,

σ_y^2denotes the variance of y,

σ denotes the covariance of x and y,

and c_1 and c_2 are two variables to stabilize the division with a weak denominator.

Here, we calculated the PSNR [dB] and SSIM index between the generated images reconstructed by our model and high-resolution images (ground truth) by using data from two different H&E breast cancer slides, which are not used to train the model (unseen data). For each slide, we used the random 54 small low-resolution images with a size of 64x64 pixels to reconstruct high-resolution images with a size of 256x 256 pixels compared to the ground. The results of PSNR/SSIM are shown in Table 1 below. In order to compare the performance of the generator models with different backbone architectures (ResNet (original SRGAN), Transformer, DenseNet, and ResNeXt), we trained them with the same dataset we acquired from the breast cancer H&E

slides. The proposed model can provide better results, which the average PSNR/SSIM of the data from both H&E slides is over 30 dB/0.92, whereas the average result from the traditional method (Bicubic interpolation), the typical SRGAN, SRGAN-DenseNet, and SRGAN-Transformer are 24.10 dB/0.848, 27.51 dB/0.915, 27.55 dB/0.93, and 18.50 dB/0.69, respectively.

Table 1. PSNR/SSIM compares results between the generated high-resolution images and the ground truth (realistic high-resolution images) from the testing dataset.

PSNR/SSIM	Breast cancer1 40x	Breast cancer2 40x	Average
Bicubic interpolation	24.13 dB/ 0.84	24.07 dB/0.86	24.1 dB/0.85
SRGAN Model	27.84 dB/0.91	27.18 dB/0.92	27.51 dB/0.915
SRGAN-DenseNet	27.96 dB/0.93	27.15 dB/0.93	27.55 dB/0.93
SRGAN-Transformer	18.68 dB / 0.69	18.33 dB /0.68	18.50 dB/ 0.69
Our model (SRGAN-ResNeXt)	**32.34** dB/ **0.93**	**31.92** dB/**0.93**	**32.13 dB/0.93**
Ground truth (high-resolution image)	∞/1	∞/1	∞/1

Figure 8 compares the reconstruction results of the typical SRGAN, SRGAN-Transformer, SRGAN-DenseNet, and our SRGAN-ResNeXt. Figure 8(a) and 8(b) illustrate the original high-resolution (ground truth) breast tumor H&E image and bicubic interpolation of a low-resolution image, respectively. Figure 8(c), 8(d), 8(e), and 8(f) show the generated high-resolution H&E images reconstructed by the traditional SRGAN, the SRGAN-Transformer, the SRGAN-DenseNet, and our SRGAN-ResNeXt, respectively. The contrast of some areas of SRGAN-DenseNet results looks slightly better than SRGAN, and SRGAN-ResNeXt results. However, some small details of the SRGAN-DenseNet results are missing as shown in Figure 8(g) pointed out by the red arrows. For the SRGAN-Transformer, it cannot surpass the SRGAN based on CNNs architectures by training with our limited custom dataset and computational resource. The model based on the Transformer architecture can potentially overcome the CNNs models if the dataset is

sufficiently large and the computational resources have high performance enough to increase the model complexity (increasing the number of attention heads, Transformer encoders, multilayer perceptron, etc.)

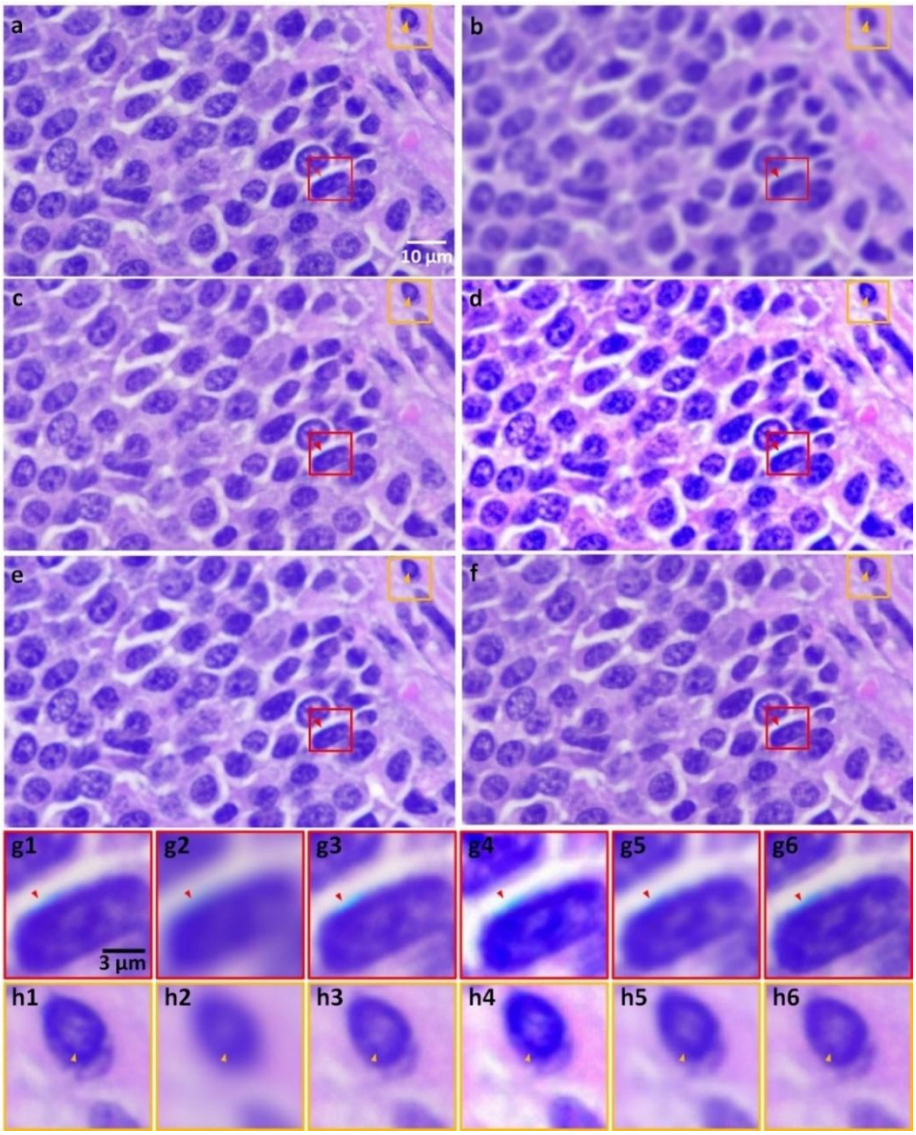


Figure 8. Comparison of the results for our deep-learning model based on ResNeXt against bicubic interpolation of the low-resolution image, SRGAN, SRGAN-Transformer, and SRGAN-DenseNet. (a) The original ground truth image. (b) Bicubic interpolation of the low-resolution image. (c) The SRGAN result. (d) The SRGAN-Transformer result. (e) the SRGAN-DenseNet result. (f) Our model result. (g1-g6) Enlarged image in the red boxes from (a-f), respectively. (h1-h6) Enlarged images in the yellow boxes from (a-f), respectively.

2.3.3 Performance of the Inception U-net architecture

Intersection over Union (IoU) as known as the Jaccard index is the benchmark used to evaluate the similarity between a predicted segmentation area and its labeled area (ground truth) [41]. The concept of IoU is to measure of pixels common between the target and predictions mask (intersection) divided by the total number of pixels present across both the prediction mask and ground truth (union) as shown in the equation below

$$IoU = \frac{target \cap prediction}{target \cup prediction}. \quad (10)$$

The IoU ranges from 0 -1 (0-100%) with 0 indicating that there is no overlapping area, whereas 1 indicates an impeccably overlapping area.

Dice similarity coefficient (DSC) is another well-known parameter used to evaluate the similarity between the predicted area (our output) and ground truth [32]. The DSC can be calculated following the equation below

$$DSC = \frac{2\,|X \cap Y|}{|X|+|Y|}. \quad (11)$$

It is remarkably similar to the IoU. They are positively correlated. The unseen H&E cancer images from the cancer imaging archive [34] were used to evaluate the performance of our Inception U-net and the typical U-net Models. Table 2 shows their performance that the IoU and DSC from the Inception U-net Model are higher than the ones from the U-net Model. According to this result, Inception U-net Model can surpass the original U-net Model by using the Inception architecture as a core structure instead of a simple convolution block.

Although the Inception U-net can slightly surpass the original U-net, these improvements will have a tremendous impact on the histopathology analyses because the histopathology image analysis needs to perform on the vast area of H&E images (whole slide image), the small accurate and inaccurate segmented nuclei of each small patch will be accumulated and lead to the correct and

incorrect diagnosis results. For example, one of the criteria to determine tumor stages is the density of inflammatory cells. The segmentation area can be used to determine it. Suppose there is a small error in the segmentation of inflammatory cells in every small H&E image patch. In that case, the total number of inflammatory cells on the whole slide image might be less accurate than the actual one, so a pathologist could wrongly diagnose the tumor stage.

Table 2. The comparison of tumor cell nuclei segmentation performances using U-net and Inception U-net architectures.

	U-net	Inception U-net
IoU/Jaccard index	0.720	0.869
DSC/F1score	0.875	0.893

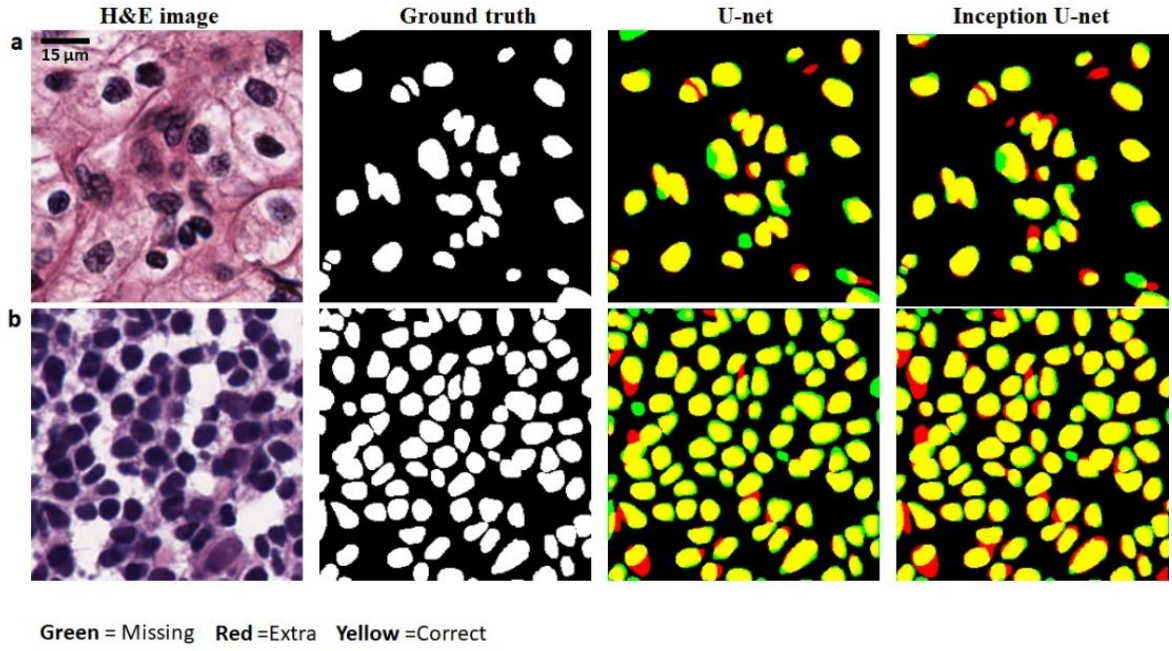


Figure 9. Comparison results between the traditional U-net and Inception U-net by using H&E images and ground truth from the dataset [34]. (a) A low density of nuclei H&E image. (b) A high density of nuclei H&E image. The results from both models have been colored code such that green denotes false negative, yellow denotes true positive, and red denotes false positive pixels.

2.3.4 Performance of the jointly trained SRGAN-ResNeXt and Inception U-net Models

After jointly training SRGAN-ResNeXt and Inception U-net Models on another unseen dataset, the performance of the ResNeXt generator was slightly improved due to the limited number of

data (220 patches). Still, the performance of the Inception U-net was considerably enhanced as shown in Figure 10, Table. 3, and Table. 4 below.

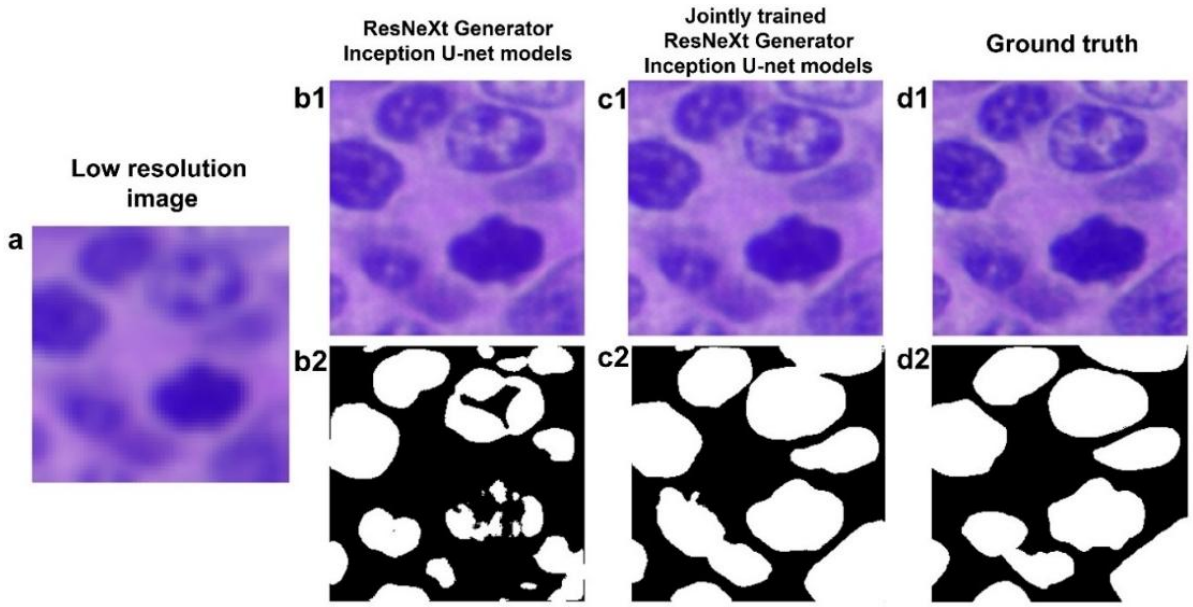


Figure 10. The improvement of the SRGAN-ResNeXt and Inception U-net after training them jointly. (a) Low-resolution image input. (b1-b2) The ResNeXt generator and Inception U-net models' results. (c1-c2) The jointly trained models' results. (d1-d2) High-resolution and segmentation ground truth images.

Table 3 and Table 4 show the performance improvement of the jointly trained SRGAN-ResNeXt and Inception U-net Models, respectively. Since the jointly trained models require to apply the dataset that contains not only low-resolution and high-resolution images but also the corresponding segmentation masks, preparing large data is expensive. Although the joint models were trained on the small dataset (220 patches from two different tumor mice), the results look promising. The performance of the jointly trained models can be potentially improved by training them on the larger dataset.

Table 3. PSNR/SSIM compares results between the high-resolution generated and the ground truth (realistic high-resolution image) dataset of the SRGAN-ResNeXt model and the jointly trained **SRGAN-ResNeXt.**

PSNR/SSIM	SRGAN-ResNeXt	Jointly trained SRGAN-ResNeXt
PSNR/SIIM	31.56 dB/ 0.91	**31.63 dB/0.92**

Table 4. The comparison of tumor cell nuclei segmentation performances using U-net and Inception **U-net architectures.**

	Inception U-net	Jointly trained Inception U-net
IoU/Jaccard index	0.50	**0.84**
DSC/F1score	0.75	**0.91**

2.4 Conclusion

In this work, we demonstrated a practical approach to enhancing low-resolution H&E stained images by using the state-of-the-art SRGAN-ResNeXt network. The model can deeply learn how to map the low-resolution images to their corresponding high-resolution images. Even though cell images contain sophisticated patterns and structures, the SRGAN-ResNeXt Model can still provide high-quality reconstruction results. Moreover, it can outperform the original SRGAN Model. Therefore, we take these advantages to characterize and quantify the nuclei from the generated high-resolution images. The nuclei from those generated images were segmented using another neural network: the Inception U-net architecture. Since we have generated both high-resolution H&E images and their nuclei segmentation, we can derive both nuclei area, pixel intensity, and other essential parameters to assist pathologists' diagnosis. If the resolution of H&E images is poor and unfavorable, the characterization could be inaccurate leading to misdiagnosis. Moreover, the individually well-trained weights of SRGAN-ResNeXt and Inception U-net Models can be applied as the pre-trained weights (transfer learning) for the jointly trained SRGAN-ResNeXt and Inception U-net Models. The performance of the jointly trained models is noticeably improved and promising. We anticipate this work can be applied in broad applications such as retrieving image quality from a compressed archiving image for transferring among data networks and enhancing image quality from a low-cost microscope. For the latter, these custom CNNs can help solve the inaccessibility of advanced microscopes to acquire high-resolution images from low-performance microscopes located in most remote clinical settings in developing nations. In future

work, we intend to apply the proposed CNNs to decrease image acquisition time for a WSI H&E scanner which typically uses a high NA objective lens in combination with a slow scan to acquire a high-resolution image.

CHAPTER 3: Recurrent and Convolution Neural Networks for Sequential Multispectral Optoacoustic Tomography (MSOT) Imaging

Volumetric optoacoustic imaging is a beneficial technique for diagnosing and analyzing biological samples since it provides meticulous details in anatomy and physiology. However, acquiring high through-plane resolution volumetric images is time-consuming, requiring a precise motorized stage to move samples under the optoacoustic system along the z-axis. Here, we propose deep learning based on hybrid recurrent and convolution neural networks to generate sequential cross-sectional optoacoustic images. A multispectral optoacoustic tomography (MSOT) system was utilized to acquire the dataset from breast tumors for training our deep learning model. This system can simultaneously acquire the sequential images (cross-sectional images) of MSOT and ultrasound. Furthermore, it provides a spectral unmixing algorithm applied to the MSOT images for extracting the sequential images of a specific exogenous contrast agent. This study used ICG-conjugated superparamagnetic iron oxide nanoworms particles (NWs-ICG) as the contrast agent. Our deep learning model applies to all three modalities (multispectral optoacoustic imaging at a specific wavelength, ultrasound, and NWs-ICG optoacoustic imaging). The generated 2D sequential images were compared to the ground truth 2D sequential images acquired using a small step size. The results of these three modalities can achieve excellent image quality where the average of peak-signal-to-noise ratio and summation absolute errors between the ground truths and the generated images is over 75 dB and less than 2,000. Instead of acquiring seven images

with a step size of 0.1 mm, we can receive two images with a step size of 0.6 mm as input images for the proposed deep learning model. The deep learning model can generate or interpolate other five images with the step size of 0.1 mm between these two input images meaning we can save acquisition time by approximately 71%.

3.1 Introduction

Multispectral Optoacoustic Tomography (MSOT) is an *in vivo* optical imaging modality for molecular, anatomical, and functional imaging Fields [43, 44]. The principle of MSOT is based on the optoacoustic effect, i.e., a molecule is excited by an ultra-short laser pulse, which can penetrate through tissue several centimeters [45, 46], resulting in thermoelastic expansion surrounding the molecule that generates a photoacoustic wave [47]. The ultrasound traducer is then used to detect this wave as an ultrasound signal. The difference of absorption contrast of tissue in single wavelength images is employed to reconstruct anatomical images. Using multiple wavelengths to excite the tissue, we can obtain multispectral images from intrinsic and extrinsic signals. A laser between 680 nm and 980 nm is the predominant source for intrinsic signals such as deoxygenated hemoglobin, oxygenated hemoglobin, melanin, myoglobin, bilirubin, fat, etc. Extrinsic signals do not usually occur in cells, tissue, or animals. Agents that can absorb in the near-infrared (NIR) range such as indocyanine green, fluorescence proteins, nanoparticles, etc., can increase the optoacoustic signal (extrinsic signal). Thus, they can be distinguished from intrinsic tissue background signals by using effective spectral unmixing algorithms such as linear regression, guided independent comment (ICA), and principal component analysis (PCA) [48, 49]. MSOT is widely used for several studies such as cancer research [50-54], drug development [55, 56], and nanoparticle [57-60]. However, using multiwavelength excitation to scan the sample is time-consuming, especially cross-sectional scanning for 3D image reconstruction. Imaging needs to

sweep all the wavelengths with every single scanning position. For *in vivo* experiments, this might lead to image degradation from motion artifacts and potential lethality from prolonged anesthesia.

In recent years, deep learning-based approaches have played a vital role in optoacoustic imaging, and they have been widely used in several applications such as image classification, segmentation [61-65], quantitative photoacoustic imaging [66-70], image enhancement [71-75], etc. One main advantage of deep learning for those applications is that it depends less on hardware modifications. In addition, most of those deep learning techniques were designed to use a single 2D image as their input and apply convolution architectures for feature extraction. For instance, deep learning for automatic segmentation of optoacoustic ultrasound (OPUS) images [76] used the U-net architecture [18] to perform the image segmentation. U-net is a well-known convolution neural network (CNN) architecture for image segmentation, particularly biomedical images [77-80].

Nevertheless, there are no techniques based on deep learning to reduce the acquisition time of cross-sectional scanning for 3D photoacoustic imaging. Herein, we propose the hybrid architecture of convolution neural network (CNN) and recurrent neural network (RNN) for generating sequential optoacoustic, unmixed optoacoustic of a specific contrast agent, and ultrasound images to extend the stack of cross-sectional images and reduce acquisition time by approximately 71%. This hybrid architecture is called Inception Generator Long Sort-Term Memory (I-Gen-LSTM). The Inception Generator is a CNN model designed based on the Inception U-net architecture. Inception is a convolution layer [81] that convolves the input in parallel with different kernel sizes extracting more features than a simple convolution layer. RNN is a robust and effective approach for sequential problems. It is a feed-forward neural network with internal memory and performs the same function for every data input. In addition, the output of

the current input depends upon the previous output. However, the original RNN has drawbacks regarding exploding and vanishing gradients from backpropagation to update weights, particularly long sequential inputs. Long Short-Term Memory (LSTM) networks [82] are improved RNN networks capable of learning long-term dependencies by adding a forget gate, input gate, and output gate. Therefore, we leverage Inception Generator and LSTM networks to generate sequential images. Our results demonstrate that the I-Gen-LSTM model is a versatile method that can generate not only sequential optoacoustic images but also sequential unmixed optoacoustic and ultrasound images.

3.2 Methods

3.2.1 Data acquisition

A commercial multispectral optoacoustic tomography (MSOT) system (inVision 512-echo, iThera Medical GmbH, Munich, Germany) was used to acquire the data for training the I-Gen-LSTM model. The MSOT system has a 270-degree ultrasound transducer tomographic array, which can acquire signals from multiple angles around an object. This tomographic array enables the system for imaging complex shapes since it can capture 2-dimensional signals in the imaging plane. Figure 11(a) shows the detection and illumination geometry in the imaging chamber of the MSOT system. In addition, this system provides a tunable laser with a range of 660-1,300 nm, which is particularly suitable for most biological samples. The excitation pulse laser is used to illuminate the sample. The sample absorbs this pulse and converts it to heat, which results in a transient thermoelastic expansion that generates an acoustic wave. The ultrasound transducer is then used to detect this acoustic wave, and the back-projection algorithm [83] is applied to the detected optoacoustic wave to reconstruct the images. For the dataset preparation, transgenic mice [84] with breast tumors were intravenously injected with indocyanine green (ICG)-conjugated superparamagnetic iron

oxide nanoworms (NWs-ICG) [85], which accumulate in tumors longer than pure ICG through the enhanced permeability and retention (EPR) effect [86]. Twenty-four hours after injection, the mice were euthanized and the tumors were removed and dissected for this study. All procedures performed on animals were approved by the University's Institutional Animal Care & Use Committee and were within the guidelines of humane care of laboratory animals. To acquire images of the tumors, 4 mg of agarose powder was dissolved in 40 mL of warm deionized water. The breast tumor was put in this dissolved agarose solution, allowing approximately 15 minutes for the solution to solidify. The hardened agarose with the tumor inside shown in Figure 11(b), was grasped by the holder and then scanned by the inVision MSOT system with the excitation pulse at wavelengths from 800 nm to 1000 nm (a comprehensive range of the NWs-ICG study). Since the inVision MSOT system can provide corresponding ultrasound images, NWs-ICG optoacoustic images obtained through linear spectral unmixing algorithm [87], and each single-wavelength optoacoustic image, these three imaging modalities were simultaneously acquired in every scanning position. Figure 11(d1-d4) shows the ultrasound images of the breast tumor with different scanning positions, Figure 11(e1-e4) shows the corresponding NWs-ICG optoacoustic images reconstructed from multispectral optoacoustic imaging with the excitation pulse at wavelengths from 800 to 1,000 nm by using the multispectral unmixing algorithm; Figure 11(f1-f4) shows the corresponding single-wave optoacoustic image at 800 nm excitation; and Figure 11(g1-g4) shows the corresponding overlaid images of these three imaging modalities.

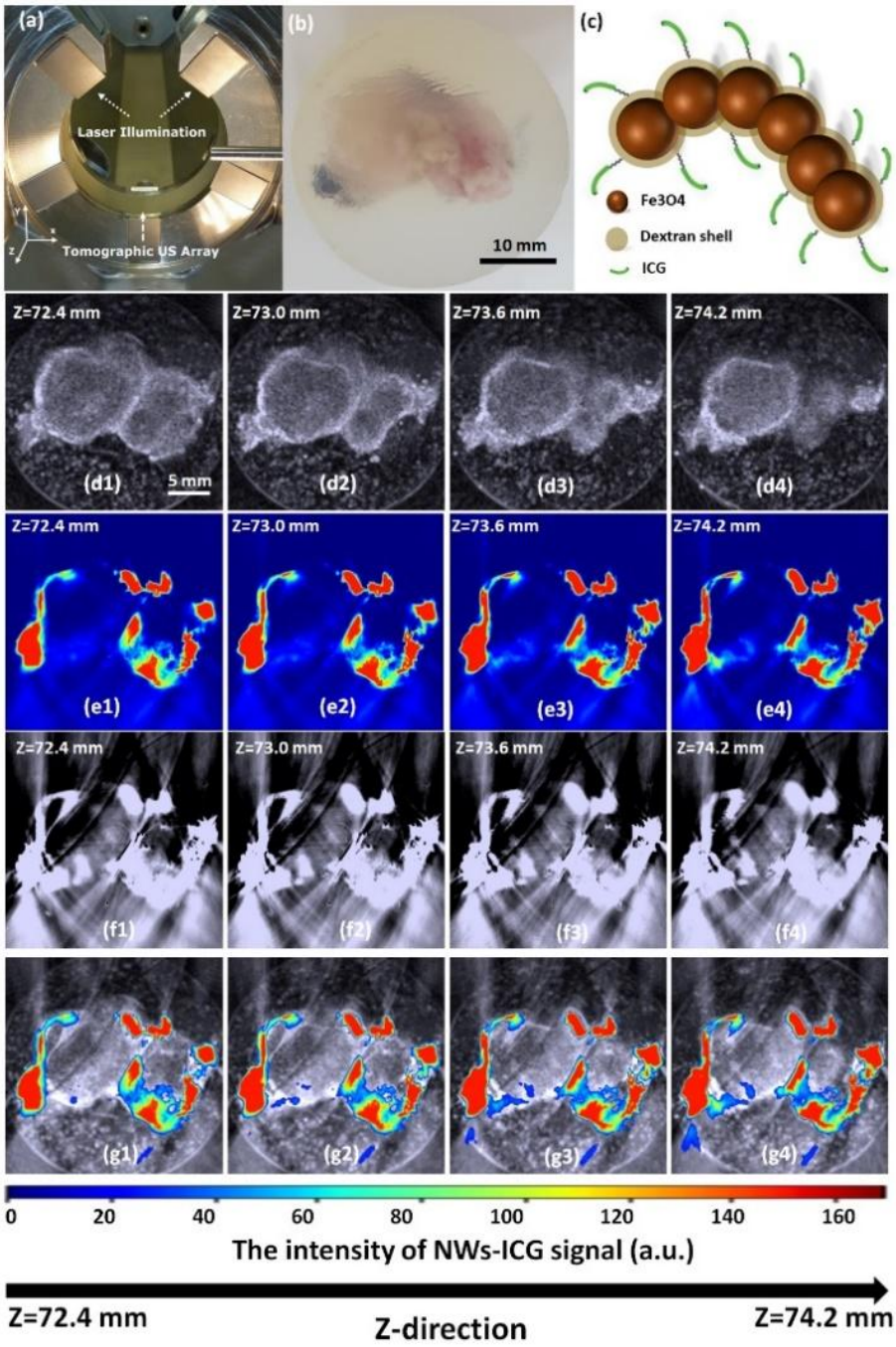


Figure 11. Ultrasound, NWs-ICG optoacoustic obtained through multispectral unmixing, and optoacoustic at 800 nm excitation imaging of an ex vivo breast tumor from a mouse intravenously injected with NWs-ICG. (a) The detection and illumination geometry in the imaging chamber of the MSOT system. (b) The breast tumor is embedded in agarose. (c) NWs-ICG structure. (d1-d4) Ultrasound images of the breast tumor with different step sizes. (e1-e4) The corresponding NWs-ICG optoacoustic images were obtained through multispectral unmixing. (f1-f4) The corresponding single-wavelength (λ_{ex} = 800 nm) optoacoustic images. (g1-g4) with an overlay of the ultrasound, the NWs-ICG optoacoustic(colormap), and the single-wavelength optoacoustic images.

3.2.2 I-Gen-LSTM and discriminator models

The I-Gen-LSTM model comprises three main neural networks depicted in Figure 12(a-c). The first neural network is the Inception encoder & decoder network based on Inception U-net architecture. The original U-net architect employs simple convolution blocks with the skip connection of encoders and decoders at the same dimension helping the model to circumvent the vanishing and exploding gradients problems. However, the simple convolution blocks might be insufficient to extract all crucial information comprehensively. Inception architecture is one of the effective CNNs architectures since it applies a wide range of kernel sizes to extract global and local features. A large and a small kernel size are tailored to extract information distributed globally and locally, respectively. With this attribute, the encoder & decoder network was designed using Inception U-net as its backbone as shown in Figure 2(a), for improving the model capability. This network takes two 2D images, acquired from an arbitrary consecutive position with a step size of 0.6 mm, as its inputs (input 1 and input 2, as shown in Figure 12(a)). The encoder shown on the left side of Figure 12(a) generates encoder outputs (E1n -E5n, where n is the input image number, i.e., 1 and 2). Inception architecture in the encoder with three different kernel sizes (1x1, 3x3, and 5x5) assembled as the parallel filters are used to extract features from the tensors followed by a rectified linear unit (ReLU) and a 2x2 max pooling with the stride of 2 steps for downsampling, respectively. Similarly, Inception architecture is also used in the decoder blocks. The encoder blocks are used to generate decoder outputs (D1n-D5n, where n is the input image number, i.e., 1 and 2) as shown in the right side of Figure 12(a) followed by a feature map upsampling, a 2x2 up-convolution (halving the number of feature channels), and a corresponding concatenation from the encoder part.

The second neural network is the convolutional LSTM network (ConvLSTM) [88], a recurrent neural network for spatio-temporal prediction. It has a convolutional structure in both the input-to-state and state-to-state transitions as shown in the bottom right of Figure 12(b). In other words, internal matrix multiplications are exchanged with convolution operations. Consequently, the data flowing through the ConvLSTM cells keeps the input dimension instead of being a 1D vector with features. The main equations of ConvLSTM are expressed in Equations (12-16) below, where '*' and '∘' represent the convolution operator and the Hadamard product (element-wise matrix multiplication), respectively. All variables in Equations (12-16) were shown in the "ConvLSTM block" in Figure 12(b).

$$i_t = \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} \circ C_{t-1} + b_i) \quad (12)$$

$$f_t = \sigma(W_{xf} * X_t + W_{ht} * H_{t-1} + W_{ct} \circ C_{t-1} + b_f) \quad (13)$$

$$c_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c) \quad (14)$$

$$o_t = \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_c) \quad (15)$$

$$H_t = o_t \circ \tanh(C_t) \quad (16).$$

The ConvLSTM takes the outputs of the Inception encoder from both input images (E11-E51 and E12-E52) as its inputs to generate five sequential blocks (Recurrent Conv1 to Recurrent Conv5) as shown in Figure 12(b). Recurrent Conv 1, 2, 3, 4, and 5 have dimensions of (5x128x128x512), (5x64x64x512), (5x32x32x512), (5x16x16x512), and (5x8x8x512), respectively. The first dimension represents the number of output images (five sequential output images). Lastly, it is the sequential image generator network inspired by U-net architecture. The model takes Recurrent Conv 1-5, two input images, encoder outputs (E11-E51 and E12-E52), and decoder outputs (D11-D41 and D12-D42) to reconstruct five sequential images of different scanning positions as shown in Figure 12(c). The left side of Figure 12(c) shows the concatenated encoder and decoder outputs

generated by the Inception encoder &decoder (Figure 12(a)). The right side of Figure 12(c) shows Conv2D transpose and Conv2D operations for the Recurrent Conv 1-5 generated by the ConvLSTM blocks (Figure 12(b)) and the concatenated encoder & decoder outputs.

All Conv2D transpose, Conv2D blocks utilize ReLU as their activation function except the last Conv2D* that applies hyperbolic tangent or tanh as its activation function. Indeed, the Recurrent Conv blocks regulate the gradual change in the sequential output images. In short, the I-Gen-LSTM model takes two images acquired by consecutive positions with 0.6 mm steps size and generates the five sequential images between these two images with gradual change following the scanning positions (step sizes of 0.1 – 0.5 mm). The ground truth images acquired using a small step size (0.1-0.5 mm) were used to determine the loss value from these five generated images. The loss functions will be elucidated in Section 3.2.3.

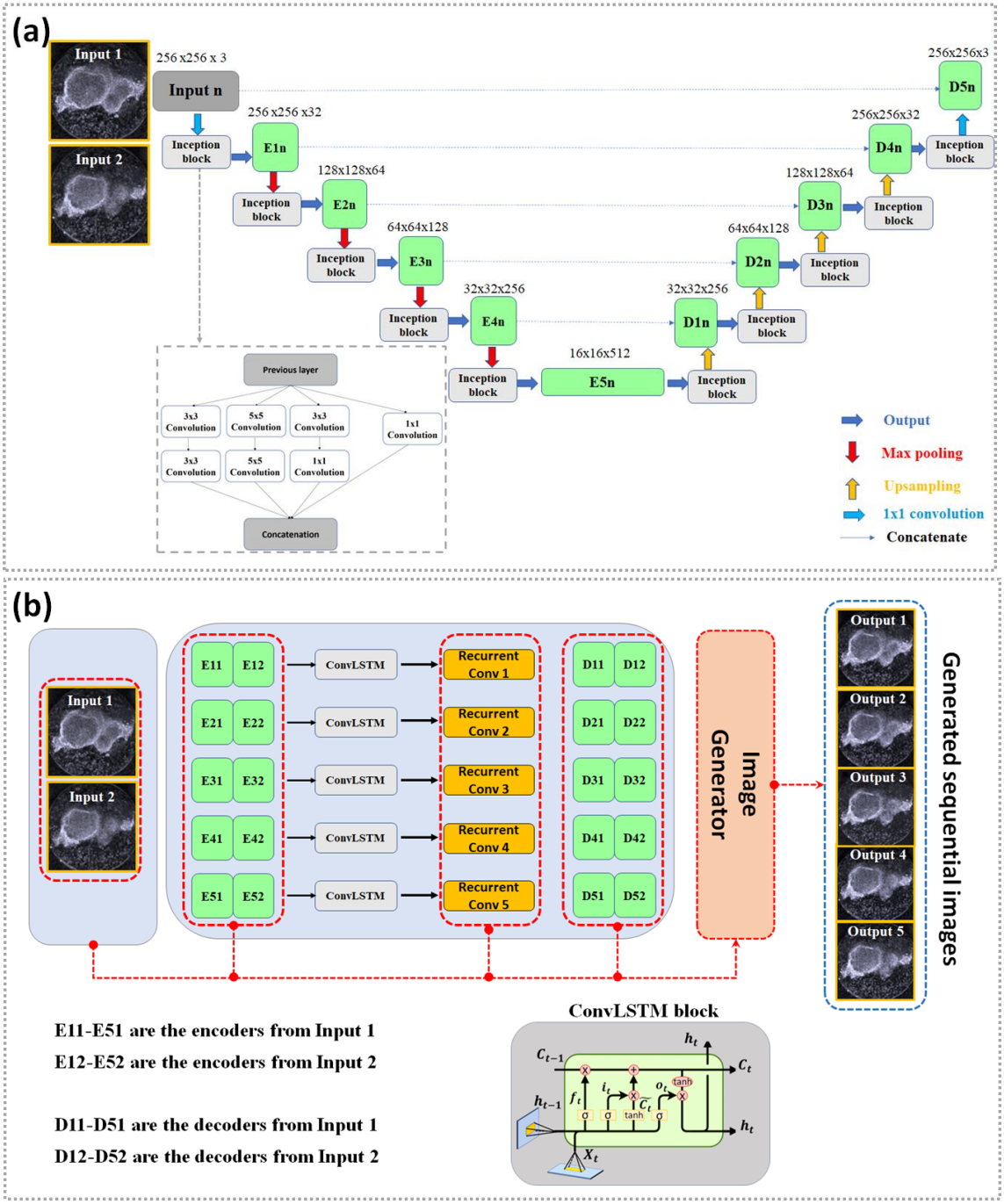
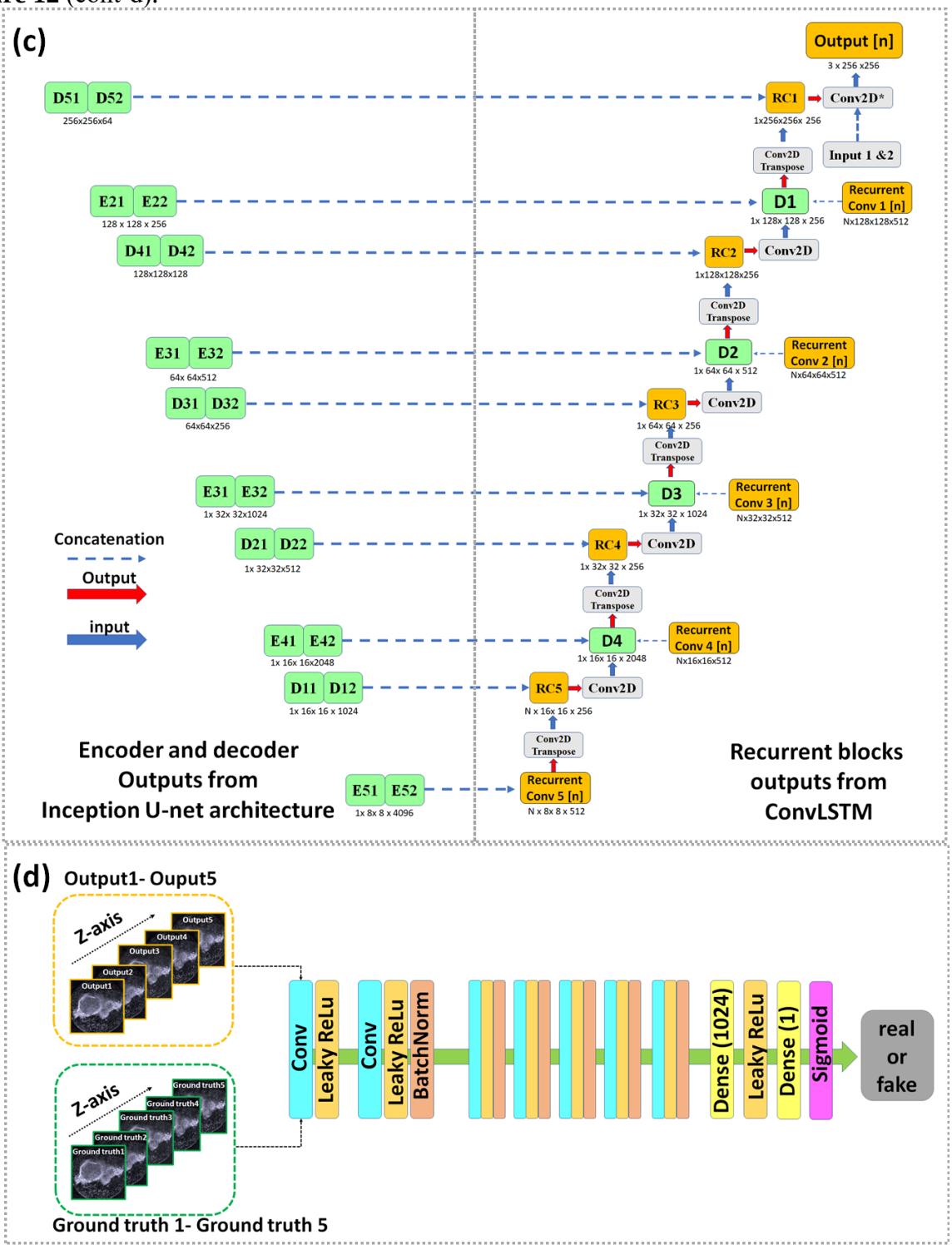


Figure 12. I-Gen-LSTM and discriminator architectures. (a) Inception encoder and decoder network were applied to both images (input1 and input2). (b) ConvLSTM network for generating the sequential blocks (Recurrent Conv 1-5) fed to the sequential image generator network for reconstructing the sequential output images. (c) The sequential image generator network. (d) The discriminator network.

Figure 12 (cont'd).



The discriminator network shown in Figure 12(d) is a simple convolution network designed to evaluate the similarity between the ground truths and generated images. The model comprises

eight convolutional layers and two fully connected layers. After each convolution block, a batch normalization layer is used, followed by an activation function named the Leaky ReLU function (α=0.2). The number of 3x3 filter kernels increases by a factor of 2 from 64 (the first layer) to 512 (the eighth layer) kernels. The last two layers are dense layers working as a classification block, predicting the probability of an image being either real or fake. To train the I-Gen-LSTM model, we assemble the models as a generative adversarial network (GAN) [89] shown in Figure 13 below.

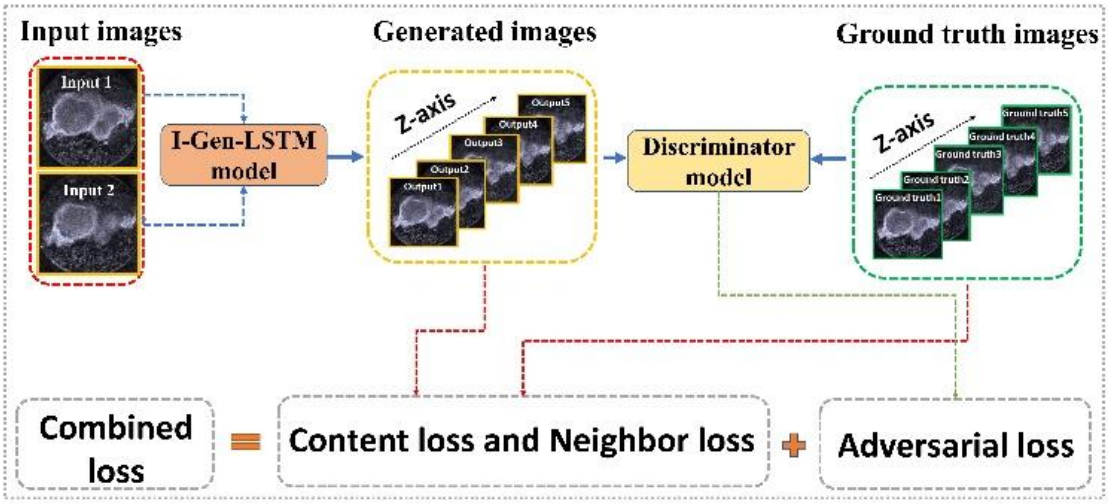


Figure 13. GAN with the combination of three loss functions (the content loss, the neighbor loss, and the adversarial loss functions) for training the I-Gen-LSTM model.

3.2.3 Loss functions

To optimize the I-Gen-LSTM model, we designed custom-made loss functions, namely the content loss (VGG19 loss, I^{SS}_{VGG}) [35], adversarial loss (Discriminator loss, I^{SS}_{Gen}), and neighbor loss (I^{SS}_{N}) as shown in Equation (17). Where C_{w1}, C_{w2}, and C_{w3} are the hyper-parameters set as 0.7, 0.1, and 0.2, respectively.

$$I^{SS} = C_{w1} I^{SS}_{VGG} + C_{w2} I^{SS}_{Gen} + C_{w3} I^{SS}_{N} \quad (17)$$

The content loss or VGG loss (I^{SS}_{VGG}), which is defined as the Euclidean distance between the feature map of the generated image ($G_{\theta G}(I^{LS})$) and the ground truth (I^{SS}), can extract high

dimensional features helping the model to generate the image with perceptually satisfying solutions without excessively smooth textures. The I_{VGG}^{SS} loss is based on the ReLU activation layers of the pre-train 19-layer VGG network and it can be calculated following Equation (18) as shown as

$$I_{VGG}^{SR} = \frac{1}{W_{i,j}H_{i,j}}\sum_{x=1}^{W_{i,j}}\sum_{y=1}^{H_{i,j}}\left(\emptyset_{i,j}(I^{SS})_{x,y} - \emptyset_{i,j}\left(G_{\theta G}(I^{LS})_{x,y}\right)^2 \right. \tag{18}$$

where $W_{i,j}\ and\ H_{i,j}$ describe the dimensions of the respective feature maps within the VGG network. The features map ($\emptyset_{i,j}$) can be obtained by the j-th convolution before the i^{th} maxpooling layer within the VGG19 network.

Moreover, the adversarial loss (I_{Gen}^{SS}) is also employed to distinguish the similarity of the two images. It is defined as the probabilities, varying from 0 to 1, which are the result of the discriminator model ($D_{\theta_D}(G_{\theta_G}(I^{LS}))$) as shown in Equation (19). Where I^{LS} is the input images, $G_{\theta_G}$is the generator model, and D_{θ_D} is the discriminator model.

$$I_{Gen}^{SS} = \sum_{n=1}^{N} -logD_{\theta_D}(G_{\theta_G}(I^{LS})) \tag{19}$$

Apart from using the content and adversarial losses, the neighbor loss is also applied to optimize the model. Since the I-Gen-LSTM model generates sequential images, the neighbor loss is essential to regulate the change of each generated image in the sequence. The concept of the neighbor loss function is to differentiate between the current generated image and the neighbor images in the same sequence as expressed in Equation (20) below as

$$I_N^{SS} = \sum_{n=1}^{N}(mse(I_n, I_{n-1}) + mse(I_n, I_{n+1})) \tag{20}$$

The custom-made loss function effectively leverages the combination of these three loss functions to train the I-Gen-LSTM model that can generate high-quality sequential images.

3.2.4 I-Gen-LSTM model for Volumetric Imaging

To collect the database for training the model, 16 breast tumors from mice intravenously injected with NWs-ICG were acquired by the MSOT system. The data from these tumors were allocated for training (11 tumors), validation (3 tumors), and testing (2 tumors) datasets. The training time on Google Colaboratory (CoLab) Pro is approximately 40 hours. After initializing and importing the model, the I-Gen-LSTM can generate five sequential images by taking less than 1 second for the five output images on a personal computer (PC) with an 11^{th} Gen Intel core i7-11700k CPU, 16 GB RAM, and an NVIDIA RTX 3090 graphic card.

3.3 Results and discussion

3.3.1 Sequential NWs-ICG optoacoustic, ultrasound, and optoacoustic (λex = 800 nm) image reconstruction.

The breast tumor dissected from an NWs-ICG-injected mouse was scanned under the MSOT system. Figure 14 shows the generated sequential images generated by the I-Gen-LSTM model. Two input images of each modality, acquired from consecutive stage positions with a step size of 0.6 mm, are used as the inputs for the I-Gen-LSTM model. Here, we demonstrate a z-scanning range from 9.7 mm-10.3 mm with a step size of 0.1 mm as a representative.

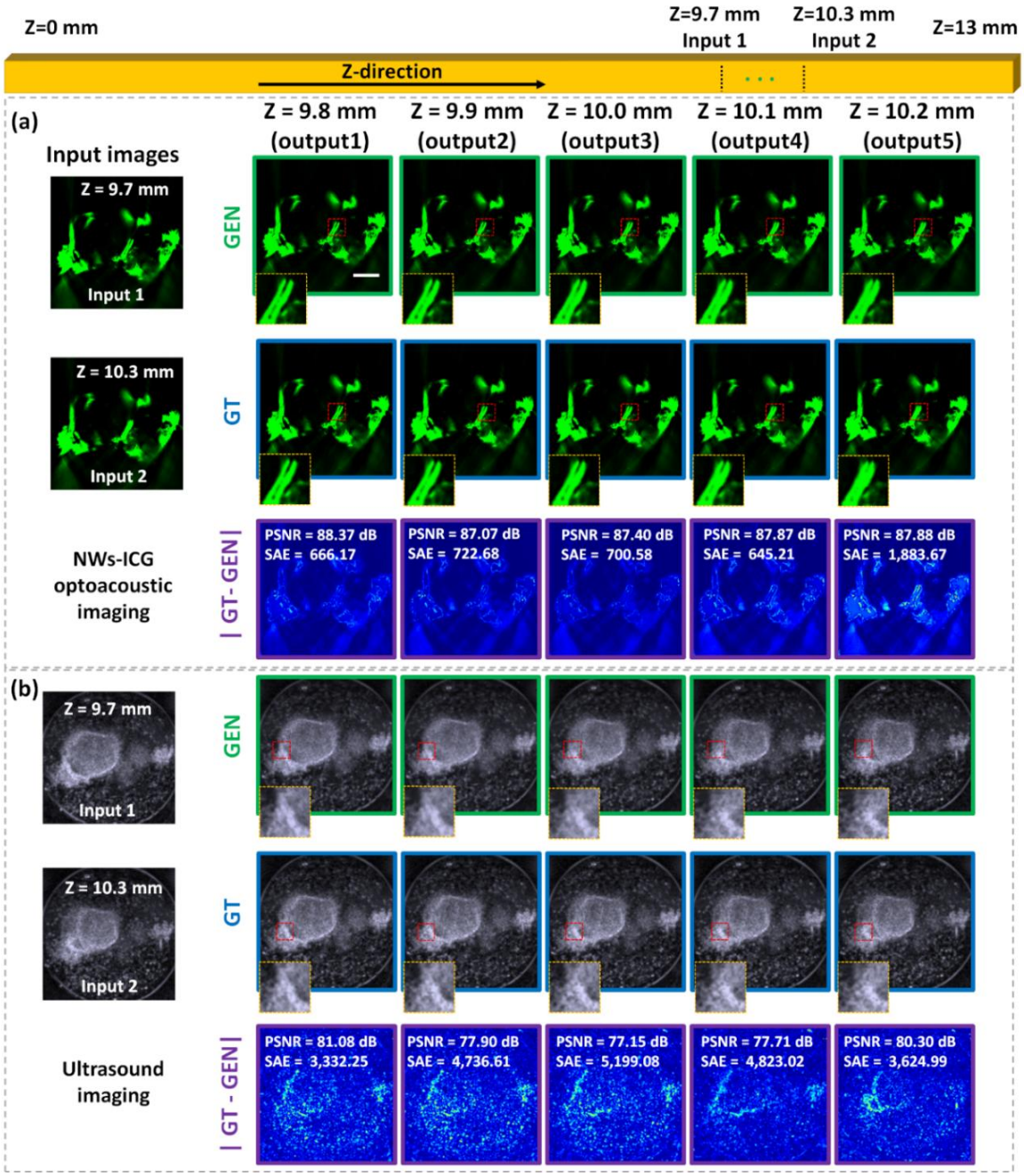
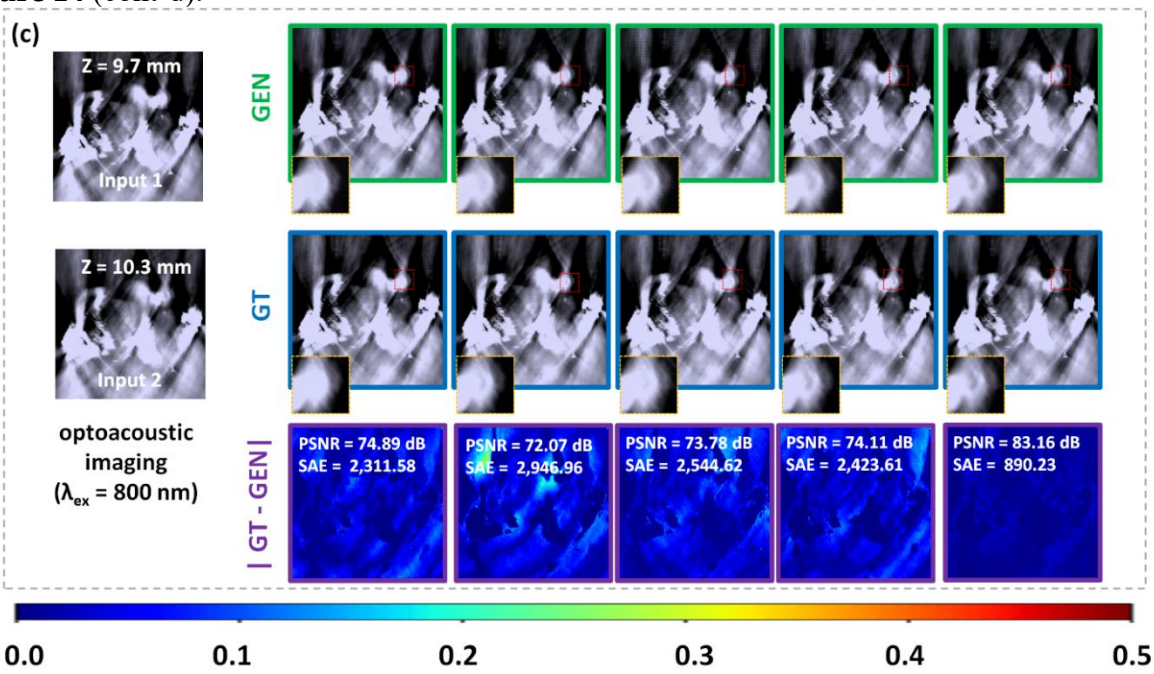


Figure 14. Results of sequential image reconstruction generated by the I-Gen-LSTM model. The two input images for each modality simultaneously acquired with a step size of 0.6 mm were fed into the I-Gen-LSTM model. The green, blue, and violet boxes show generated images (GEN), ground truth (GT), and the absolute error between GEN and GT images (|GT-GEN|) represented as color map images. The red-dashed boxes show the local features fairly change along the z-scanning position and the yellow-dashed boxes are the corresponding enlarged images of the red-dashed boxes. The scale bar is 5 mm. (a) NWs-ICG optoacoustic sequential image reconstruction result. (b) Ultrasound sequential image reconstruction result. (c) Single-wavelength optoacoustic (λ_{ex} = 800 nm) reconstruction result.

Figure 14 (cont'd).



The red-dashed boxes in Figure 14 show local features, which are fairly changing along the z-scanning position and are somewhat straightforward to observe. The orange-dashed boxes are the corresponding enlarged images of the red-dashed boxes. Figure 14(a) shows the sequential image reconstruction result of NWs-ICG optoacoustic imaging, Figure 14(b) shows the result of ultrasound imaging, and Figure 14(c) shows the result of single-wavelength optoacoustic (λ_{ex} = 800 nm) imaging. The average Peak-signal-to-noise ratio (PSNR) dB/ the average summation of absolute errors (SAE) between the ground truths (GT) and generated images (GEN) for this scanning range of NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) imaging are 87.72 dB/923.66 ,78.83 dB/4,323.19, 75.60 dB/2,223.40, respectively.

3.3.2 Three-dimensional reconstruction of the stack 2D NWs-ICG optoacoustic, ultrasound, and optoacoustic (λex = 800 nm) images

Since the MSOT system and our deep learning model provide the stack of multiple cross-sectional images for NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) images, we can use these images to reconstruct three-dimensional (3D) images by using Amira (Mercury Computer system, Berlin, Germany) software. Figure 15 shows the 3D reconstruction results of the ground truth and the generated images. Figure 15(a) demonstrates the 3D reconstruction of generated images from the I-Gen-LSTM model and Figure 15(b) shows the reconstruction of the ground truths acquired by mechanical scanning. After finished the experiment, the tumor was removed from the agarose and sent to the histopathology lab (MSU-IHPL Research facility) to prepare a Hematoxylin-and-Eosin (H&E) stained breast tumor slide shown in Figure 15(c).

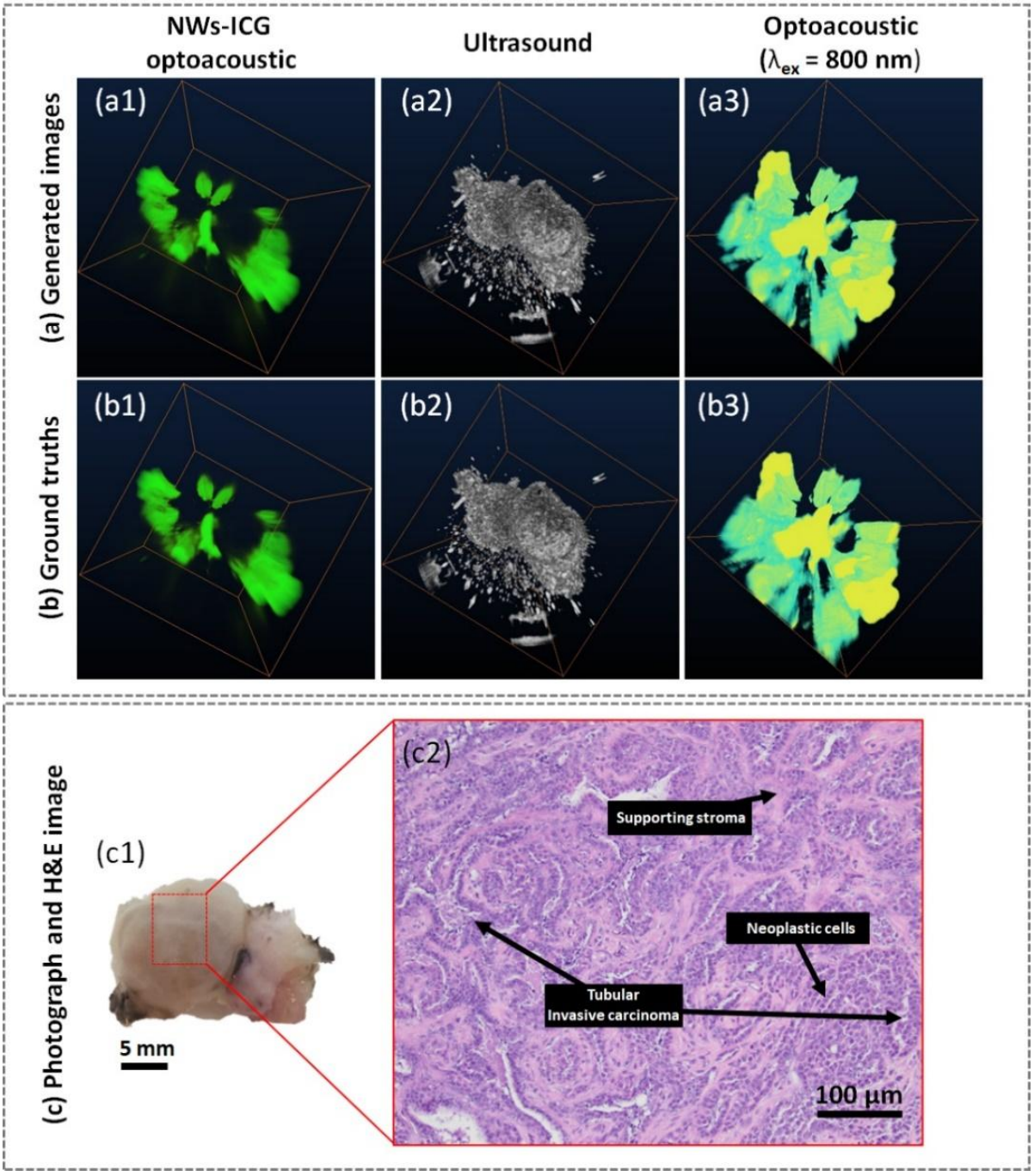


Figure 15. 3D image reconstruction of the breast tumor using cross-sectional NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) stacked images. (a) The 3D reconstruction result of the NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) images generated by the I-Gen-LSTM model with a step size of 0.1 mm. (b) The 3D reconstruction result acquired by mechanical scanning with a step size of 0.1 mm. (c) The photograph of the corresponding tumor and its H&E slide image.

3.3.3 Evaluations

The NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex} = 800 nm) images from two tumors not used for training the model were utilized for the model evaluation. Each tumor was scanned with a step size of 0.1 mm. Every two-image (with a 0.6 mm scanning step in between) was assigned as the input for the I-Gen-LSTM model to generate five sequential images with a step size of 0.1 mm. Here, the model was evaluated using four quantitative metrics: the average PSNR, SAE (GEN, GT), SAE ($Input_1, GT$),and SAE ($Input_2, GT$). They were applied to the testing dataset acquired from the tumors for all scanning positions. A large PSNR and a small SAE (GEN, GT) imply high-quality generated images. Indeed, if the SAE (GEN, GT) can perform better than SAE (Input1-GT) and SAE (Input2-GT), it also means that the model can effectively generate sequential images. All average evaluation metrics can be calculated following Equation (21-23).

$$\text{Average PSNR} = \frac{\sum_j^N \sum_i^5 PSNR_j(GEN_i\ , GT_i)}{5 \times N} \quad (21)$$

$$\text{Average SAE (GEN, GT)} = \frac{\sum_j^N \sum_i^5 SAE_j(GEN_i\ , GT_i)}{5 \times N} \quad (22)$$

$$\text{Average SAE } (Input_k\ \text{,GT}) = \frac{\sum_j^N \sum_i^5 SAE_j(Input_k\ , GT_i)}{5 \times N} \quad (23)$$

Where,

N is the number of scanning positions with a step size of 0.6 mm,

GEN_i is the generated image at “i” scanning position in between two input images (acquired with a step size of 0.6 mm),

GT_i is the corresponding ground truth,

$Input_k$ images are the two input images (k=1 and 2) acquired from arbitrary consecutive positions with a step of 0.6 mm.

Figure 16 shows the representative result from one of the evaluated tumors as the graph of the average PSNR and SAE (GEN, GT) vs. scanning positions. Table 5 shows the average evaluation metrics of the generated sequential NWs-ICG optoacoustic, ultrasound, and optoacoustic (λ_{ex}= 800 nm) images for all testing datasets. Overall, the average PSNR and SAE between generated images and ground truths of all modalities are greater than 75 dB and less than 2,000, respectively.

This indicates that the I-Gen-LSTM model can generate sequential images with promising results. To comprehensively evaluate the model performance, we also compared SAE (GEN, GT) to SAE ($Input_1, GT$) and SAE($Input_2, GT$) as the baseline for comparison. The average SAE (GEN, GT) of optoacoustic (λ= 800 nm) and ultrasound imaging performs better than the average SAE($Input_1, GT$) and SAE($Input_2, GT$), but the NWs-ICG optoacoustic imaging does not (the average SAE (GEN, GT) is slightly higher than the average of SAE($Input_1, GT$) and SAE($Input_2, GT$)) due to the tiny changing features in the sequential NWs-ICG optoacoustic imaging and the limited number of the training dataset. Although the overall result is favorable and encouraging, the deep learning model could be improved in future work. We will use a larger dataset with a larger image size to train the deep learning model so that the convolution/LSTM blocks can efficiently capture more sequential features, especially in a tiny changing feature modality such as NWs-ICG optoacoustic imaging.

Table 5. Average quantitative metrics of optoacoustic (λ_{ex} = 800 nm), NWs-ICG optoacoustic, and ultrasound images generated by the proposed deep learning model.

Average quantitative metrics	Optoacoustic (λ_{ex} = 800 nm)	NWs-ICG optoacoustic	Ultrasound
PSNR (dB)	76.53	83.75	80.44
SAE (GEN, GT)	**1,706.12**	858.54	**1,265.87**
SAE ($Input_1, GT$)	6,812.92	406.59	6,695.71
SAE ($Input_2, GT$)	5,294.94	**284.02**	4,902.67

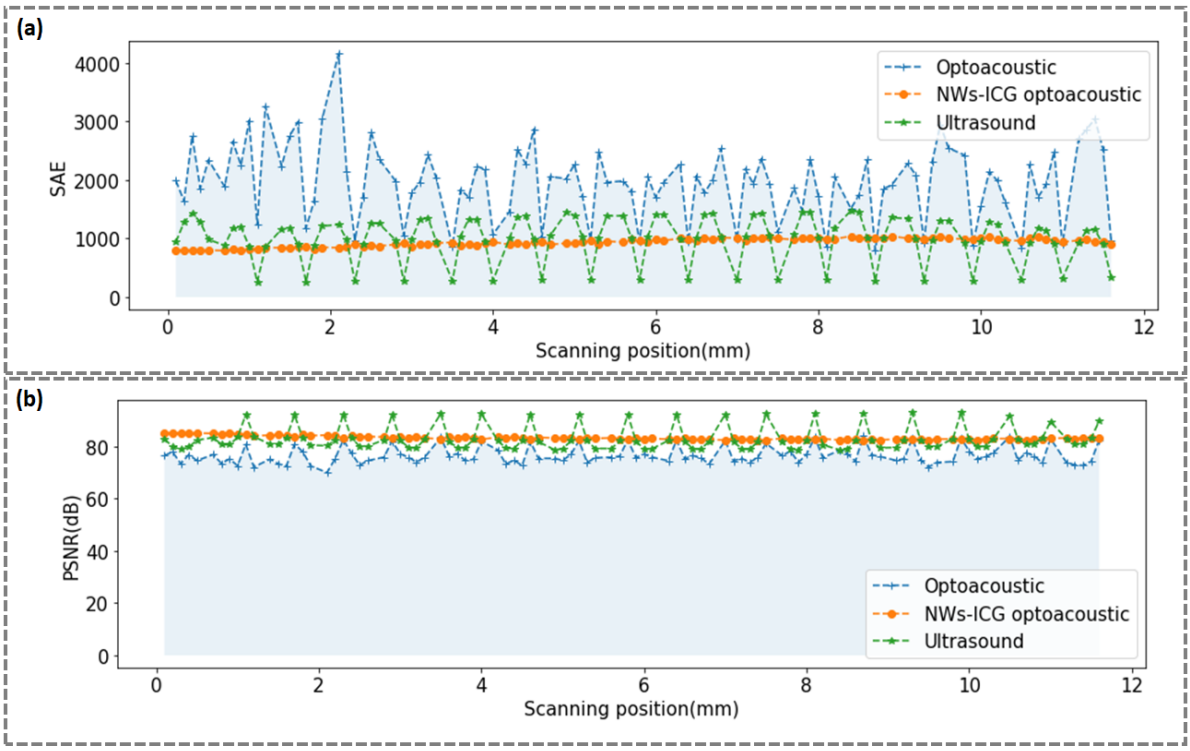


Figure 16. The PSNR and SAE (GEN, GT) evaluation in one of the testing tumors. (a-b) The graph between the PSNR and SAE (GEN, GT) values vs. scanning positions for all generated OPUS, NWs-ICG optoacoustic, and optoacoustic (λ_{ex} = 800 nm) images, respectively.

3.4 Conclusion

This work demonstrates a deep learning technique based on recurrent and convolution neural networks for generating sequential NWs-ICG optoacoustic (multispectral unmixing), ultrasound, and optoacoustic images. It has shown robust and promising performance in the accurate reconstruction of the sequential images for all modalities, according to the quantitative evaluation of model performance using the PSNR and SAE for all scanning positions of the generated images (reconstructed by the deep learning model) and ground truth (acquired by mechanical scanning). The architecture of our model is versatile since it can promisingly generate sequential cross-sectional images of three modalities from the commercial MSOT system. Using our deep learning can substantially reduce acquisition time. However, all the training data were acquired from *ex vivo* tissues completely fixed in agarose. Model performance with images acquired *in vivo* may be

affected by cardiac and respiratory motion. In the future, we will explore the possibility of optimizing and applying the model to generate sequential images of *in vivo* samples with motion artifacts.

CHAPTER 4: Multi-head Attention U-Net for MPI-CT Image Segmentation

Magnetic particle imaging (MPI) is an emerging non-invasive molecular imaging modality with high sensitivity and specificity, exceptional linear quantitative ability, and potential for successful applications in clinical settings. Computed tomography (CT) is typically combined with the MPI image to obtain more anatomical information. Herein, we present a deep learning-based approach for MPI-CT image segmentation. The dataset utilized in training the proposed deep learning model is obtained from a transgenic mouse model of breast cancer following administration of indocyanine green (ICG)-conjugated superparamagnetic iron oxide nanoworms (NWs-ICG) as the tracer. The NWs-ICG particles progressively accumulate in tumors due to the enhanced permeability and retention (EPR) effect. The proposed deep learning model exploits the advantages of the multi-head attention mechanism and the U-Net model to perform segmentation on the MPI-CT images, showing superb results. In addition, we characterized the model with the different number of attention heads to explore the optimal number for our custom MPI-CT dataset.

4.1 Introduction

MPI is a highly sensitive imaging modality initially introduced in 2005 [91-93]. Unlike traditional imaging techniques such as magnetic resonance imaging (MRI), sonography, computed tomography (CT), and X-ray, MPI is not employed for structural imaging purposes. Nevertheless, it is a tracer imaging modality akin to positron emission tomography (PET) and single photon emission computed tomography (SPECT). The concept of MPI is to detect the three-dimensional

distribution of superparamagnetic iron-oxide nanoparticles (SPIONs) with extraordinary contrast and sensitivity, allowing us to track and quantify the tracer materials effectively. In addition, MPI signal can only be detected from the administered tracer providing an image without background as well as improving signal-to-noise ratios. Indeed, the development of MPI involved strengthening the existing imaging modalities (MRI, PET, SPECT, etc.). For instance, PET and SPECT tracers typically have half-lives in a range of minutes to hours, whereas the MPI tracer can last for several days to weeks [94]. Therefore, MPI is more eminently suitable for dynamic imaging applications than traditional tracer imaging methods. Numerous prototypes and commercial MPI scanners have demonstrated impressive results in *in-vivo* studies for vascular imaging [95-97], oncology [98-100], and cell tracking [101, 102]. The MPI system for humans is under development and may become available in the near future [103]. Like PET, an MPI image is frequently combined with a CT image for registering the particle signal (the MPI image) and the anatomical information (the CT image). This will enhance the diagnostic potential by identifying the precise location of functional events in the body [104].

Biocompatibility is one of the essential features for using biomaterials, particularly MPI tracers (iron oxide particles), for *in-vivo* applications and clinical trials. Nanoworms (NWs) are biocompatible iron oxide particles widely used for biomedical applications. NWs include a considerably lower inflammatory response than spherical iron oxide nanoparticles [105]. NWs are a nanostructure with an elongated assembly of iron oxide (IO) [106]. This structure can potentially augment the nanoparticles' capability for circulation and tumor targeting. Due to their nanoscale dimensions, NWs can remain in tumors longer than pure fluorescence contrast agents, also recognized as the enhanced permeability and retention (EPR) effect [107, 108].

Recently, image processing based on deep learning has become a promising approach for medical applications due to the rapid development of computation technologies for image classification [109-111], regression [112-114], reconstruction [115-117], and segmentation [118-121]. Deep learning models contain a large number of function approximators. As a result, the models without further modifications tend to neglect essential parts of the input and focus on others. The use of the attention mechanism [122] is one of the practical approaches to remedy this problem. The attention mechanism is an ingenious and powerful technique allowing neural networks to focus on meaningful parts of an input tensor. This mechanism is the key innovation behind numerous successful deep learning architectures such as TransUnet [123], BRET [124], and Swin transformer [125]. Multiplicative attention (Luong attention) [126] and additive attention (Bahanau attention) [127] are two initial instances of attention sparking the revolution. Since multiplicative attention implements matrix multiplication for calculating the output, it is more memory-efficient in practice and faster than additive attention. However, the additive attention can be superior to the multiplicative attention for large dimensional input features [128]. The U-Net architecture [129] is a widely recognized convolutional neural network (CNN) that has achieved prominence in the field of medical image segmentation due to its simplicity and remarkable performance. The original U-Net architecture contains two main components: an encoder and a decoder. The skip connection mechanism is added to the same dimensional encoder and decoder. Essentially, it combines spatial information from the down-sampling path (encoder) with the up-sampling path (decoder) to retain marvelous spatial information. In addition, the skip connection mechanism allows the gradient descent to readily propagate back to update the weights (learnable parameters). However, the skip connection mechanism brings along the poor feature representation from the encoder path. The attention U-Net architecture [40] can tackle this problem by

implementing the attention mechanism at the skip connection, allowing the model to actively suppress actions at irrelevant features. This reduces the computational resources wasted on irrelevant activations and provides superior network generalization. The attention mechanism applied in the attention U-Net is called the attention gates (AGs) [130] based on additive attention. The CNN model with AGs can be easily trained from scratch and boost the model's performance by automatically learning to focus on some crucial features without additional supervision. Available MPI data are remarkably limited for a computational study of robust MPI image quantification. Herein, we propose a multi-head attention U-Net model for the MPI-CT image segmentation. The MPI-CT images acquired from mice with breast tumors were manually labeled as the ground truths for training the model. The attention U-Net model [131] inspires the proposed model. Still, we apply the attention mechanism in parallel (multi-head attention) to step up the model capability for focusing on noteworthy features.

4.2 Methods

An extensive overview of the workflow involved in training the proposed multi-head attention U-Net model is shown in Figure 17 below. First, NWs were synthesized by the co-precipitation method of Fe^{2+} and Fe^{3+} salts with the polysaccharide dextran coating, as depicted in Figure 17(a1), the particles were then conjugated with ICG, resulting in the formation of conjugated superparamagnetic iron oxide nanoworms referred to as NWs-ICG [85]. In addition, we also acquired a transmission electron microscopy (TEM) image of NWs-ICG particles as shown in Figure 17(a2). With this structure, the detection of NWs-ICG can be achieved by fluorescence imaging and optoacoustic imaging, in addition to the use of MPI as shown in Figure 17(a3). Thus, this offers captivating prospects for a multimodal imaging study. However, this paper mainly focuses on MPI. A mouse with breast tumors was injected with NWs-ICG through the intravenous

administration injection method, followed by MPI-CT image acquisition. Figure 17(b1) shows the MPI and micro-CT image systems used in this work. The fundamental concept of MPI is illustrated in Figure 17(b2). In short, an intense magnetic field is generated by two permanent magnets, and the inside of this magnetic field contains a small area with low magnetic field intensity known as the field-free region (FFR). By rapidly moving the FFR across the imaging volume, the magnetization of SPIONs passing through the FFR induces a signal (oscillating changes in magnetization) in the imager's receive coil. In other words, SPIONs not passing not passing through the FFR do not generate a signal in the receiver coil due to a strong magnetic field outside the FFR inhibiting SPIONs from rotating. Lastly, the MPI-CT images were manually labeled as the ground truths for training the deep learning model as shown in Figure 17(c).

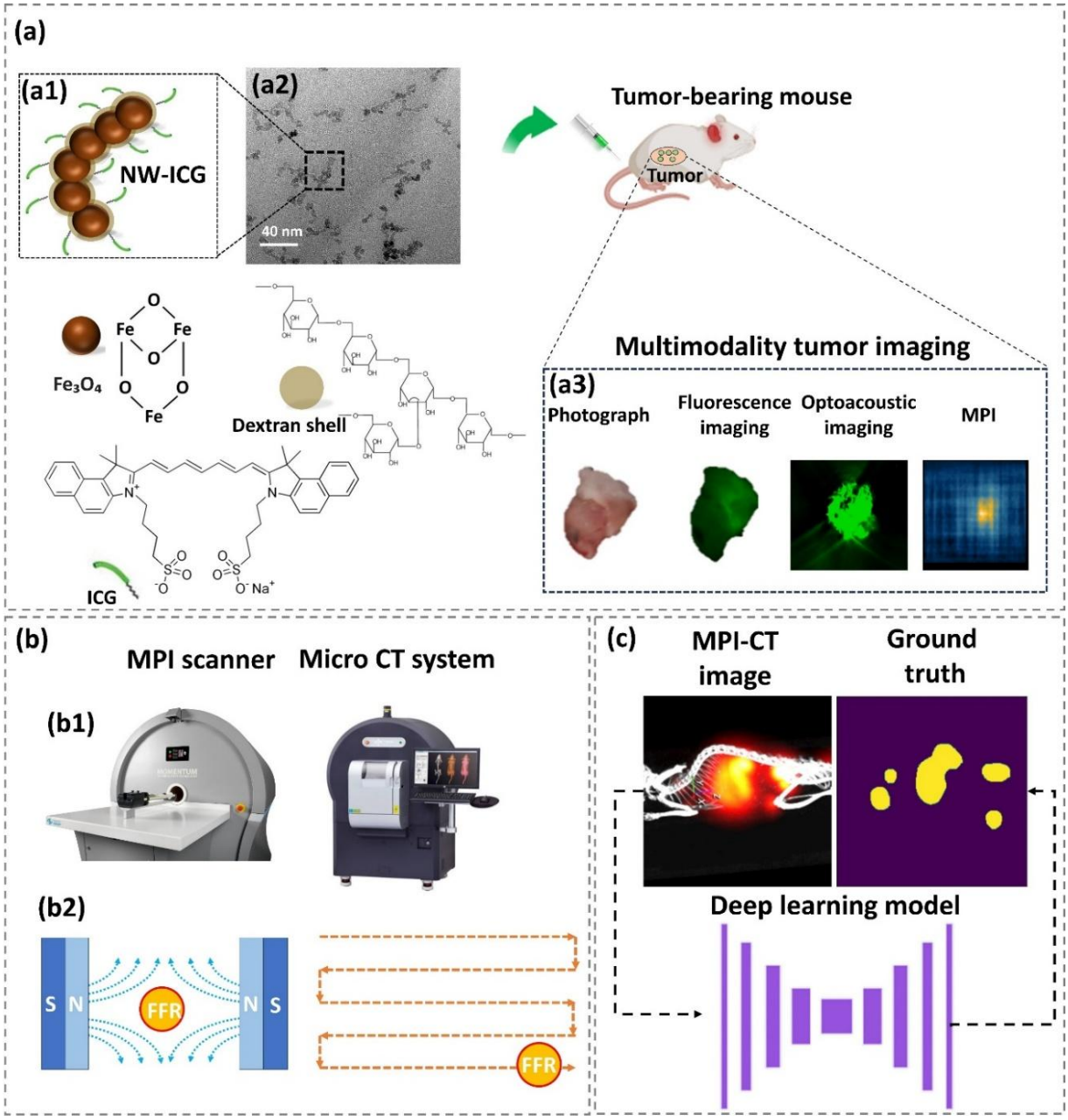


Figure 17. Overview of MPI-CT image segmentation using the custom dataset. (a) An injected-NWs-ICG breast tumor mouse; (a1) the chemical structure of NWs-ICG; (a2) TEM image of NWs-ICG particles with a scale bar of 40 nm; (a3) the multimodality imaging (fluorescence, optoacoustic, and MPI) of the tumor dissected from the NWs-ICG injected mouse. (b) MPI-CT image acquisition; (b1) MPI scanner and Micro-CT imaging system; (b2) illustration of the MPI principle. (c) Ground truth labeling in MPI-CT image segmentation.

4.2.1 Dataset preparation

To acquire a custom MPI-CT image dataset, MMTV-PyMT transgenic mice with breast cancer were intravenously injected with NWs-ICG at the concentration and volume of 2 mg/mL and 400 µL, respectively. All procedures used in experiments conducted on animals were approved by the Institutional Animal Care & Use Committee (IACUC) of Michigan State University. The Momentum MPI scanner (Magnetic Insight, Inc., Alameda, CA, USA) was employed to acquire the 3D MPI images of the NWs-ICG injected mice. The scanner was configured with the following parameters: 3D scan mode, Z FOV 10.0 cm, number of projections 21, and selection field gradient 5.7 T/m. The Micro CT system (PerkinElmer, Inc., Hopkinton, MA, USA) with the following parameters: speed scan mode and voltage of 90 kV was then used to acquire the corresponding CT images. Finally, 3D MPI-CT images were reconstructed using VivoQuant software (Magnetic Insight, Inc., Alameda, CA). The imaging was performed at four different time points: 1 hour, 24 hours, 48 hours, and 72 hours after injection. Therefore, with one mouse, we can obtain 3D datasets at these four different time points. However, we only focus on 2D images in this work. To obtain the 2D image dataset, the 3D images were rotated with random angles for capturing the 2D images, and we had to ensure that the perspectives or rotation angles were not the same (0 or 180 degrees from the existing images) for the data cleaning purpose. Figure 18(a) shows the MPI-CT images of the NWs-ICG injected mouse 1-72 hours post injection. MPI signal areas from MPI-CT images were manually labeled as the ground truth for training the segmentation deep learning model. There are 104 2D MPI-CT images and their corresponding ground truths from four different mice used for this study (91 images for a training dataset, 4 images for a validation dataset, and 9 images for a testing dataset). To affirm that there were NWs-ICG particles in the tumor tissues, after acquiring MPI-CT images, the tissues were dissected from the mice and preserved in a solution of

10% neutral buffered formalin (NBF). These NBF-fixed tissues were embedded in paraffin, followed by sectioning with a thickness of 5 μm and staining with Prussian Blue to detect ferric from iron and hematoxylin and eosin (H&E). All histological procedures were carried out by the Michigan State University investigative histopathology laboratory. Figure 18(c-d) show the Prussian blue stained histology image of one of the dissected tumors from NWs-ICG injected mice acquired by a commercially available microscope (Nikon Eclipse Ci, Nikon Inc, Tokyo, Japan).

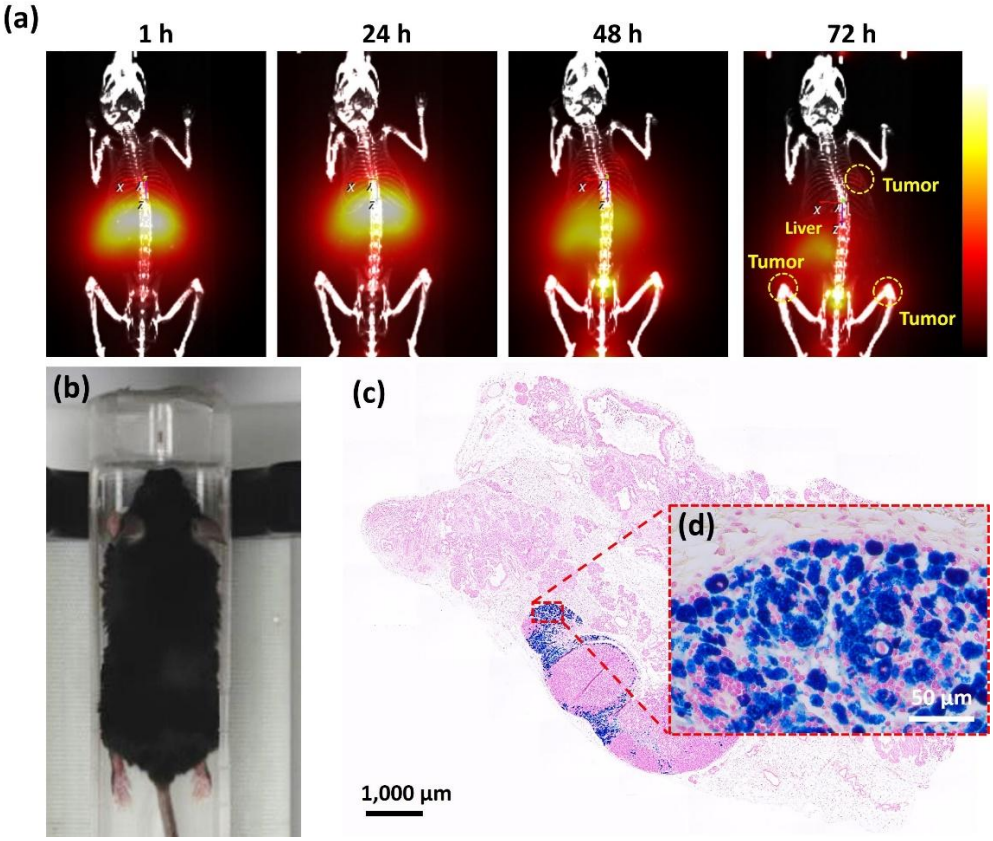


Figure 18. (a) MPI-CT images of the NWs-ICG injected mouse acquired from 1 – 72 hours post-injection. The yellow-dashed circles (MPI-CT image at 72 h) show the MPI signal of NWs-ICG from the tumors. (b) Photograph of the NWs-ICG injected mouse. (c-d) Prussian blue stained histological image of the breast tumor dissected from the NWs-ICG injected mouse acquired by 10x and 40x magnifications, respectively.

4.2.2. Multi-head attention U-Net

The main structure of the multi-head attention U-Net model is somewhat similar to the original attention U-Net model, which consists of the encoder, bottleneck, decoder, and single-head attention layers. However, the proposed model applies parallel attention gates (AGs) in each skip connection from encoder to decoder instead of a single attention head. This modification allows the model to collect and incorporate more salient information effectively. In addition, employing parallel AGs enables the model to simultaneously process input from distinct representation subspaces at numerous locations [31]. Figure 19**(**a) illustrates the multi-head attention U-Net architecture. The first part is the encoder (the left side of Figure 19(a)). The input image is progressively filtered and down-sampled by applying a convolution block, then a rectified linear unit (ReLU), and max-pooling 2x2 filters with a stride of 2. Furthermore, the number of feature channels is doubled at each downsampling step. The second part is multi-head attention gates (MH-AGs). The features propagated through the skip connections are filtered by exploiting these MH-AGs, which can help the model localize and focus on relevant features without cropping regions of interest. The third part is the decoder (the right side of Figure 19(a)). It consists of a concatenation of the attention weights from the MH-AG layer, a convolution block with the ReLU activation function, and a feature map upsampling followed by a 2x2 up-convolution resulting in a reduction of the number of feature channels by half. Figure 19(b) shows the MH-AG architecture employed between the encoder and decoder of the U-Net in Figure 19(a). MH-AG is a parallel mechanism block that minimizes the need for training a significant number of weights (learnable parameters) to enhance the performance of the U-Net model further. Moreover, the MH-AG adopts the same transformation in all branches to minimize the need to adjust hyperparameters in each branch manually. The output of each branch in MH-AG is obtained by performing element-wise

multiplication between the input feature maps and attention coefficients ($\hat{x}_n^l = x_i^l \cdot \alpha_i^l$) allowing the model to identify salient information. To identify focus areas, a gating vector (g_i) is assigned to each pixel. The gating vector encompasses contextual information utilized to suppress lower-level feature responses selectively. The gating coefficient is derived through the utilization of additive attention mathematically represented as follows:

$$q_{att}^l = \psi^T\left(\sigma_1\left(W_x^T x_i^l + W_g^T g_i + b_g\right) + b_\psi\right), \tag{24}$$

$$\sigma_i^l = \sigma_2\left(q_{att}^l\left(x_i^l, g_i; \Theta_{att}\right)\right), \tag{25}$$

"Where $\sigma_2(x_i) = \frac{1}{1+exp\,(-x_i)}$ represents the sigmoid activation function, Θ_{att} represents a group of parameters that comprises linear transformation $w_x \in R^{F_l \times F_{int}}$, $w_g \in R^{F_g \times F_{int}}$, $\psi \in R^{F_l \times F_{int}}$, and bias terms $b_\psi \in R$, and $b_g \in R^{F_g \times F_{int}}$. Channel-wise 1 x 1 x 1 convolutions for the input tensor are employed for computing the linear transformations.

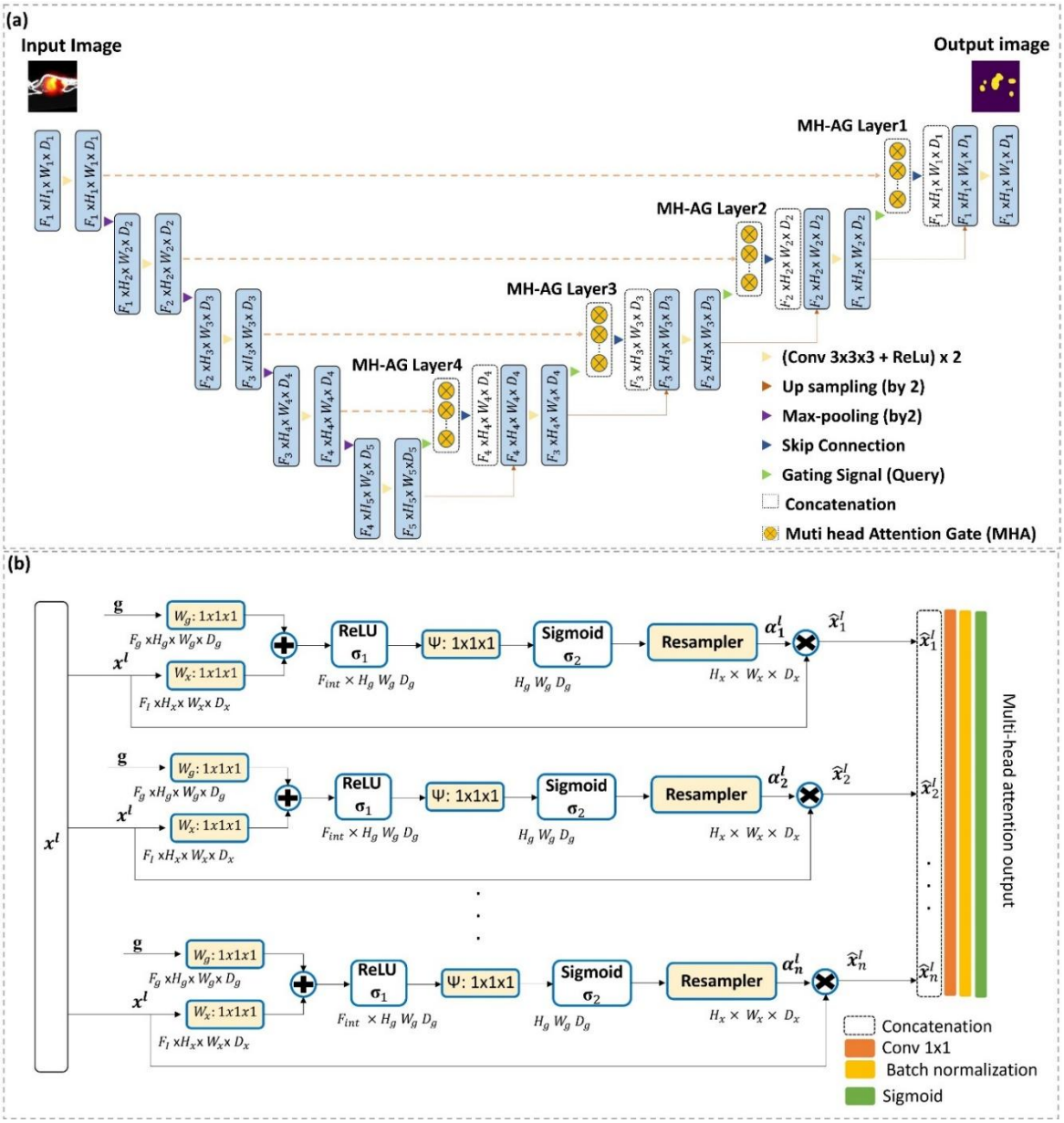


Figure 19. Schematic of the multi-head attention U-Net (the proposed model) for MPI-CT image segmentation. (a) The left side of the schematic represents the encoder blocks; the tensor is progressively down-sampled by a factor of 2 (e.g., $H_1 = H_5/16$); the right side represents the decoder blocks, the tensor is up-sampled gradually by a factor of 2. The muti-head attention gates (MH-AGs) are applied between the encoder and decoder to assign weights (learnable parameters) to noteworthy features. (b) Multi-head attention gate (MH-AG) architecture (n is the number of attention heads). Input features (x_n^l) are scaled with attention coefficients (α_n^l) computed in each branch of MH-AG. The gating signal (g) collected from a coarser scale provides activations and contextual information, which is applied to determine spatial regions. The output of each branch is then concatenated before feeding to the convolution layer, batch normalization, and sigmoid function to compute the final result of MH-AG.

4.2.3 Loss function

Dice loss is widely used for medical image segmentation by comparing the similarity of two binary images (ground truth segmentation and predicted segmentation). Since our custom MPI-CT image dataset is limited and we want to prove the concept that multi-head attention can potentially enhance the model performance for MPI-CT image segmentation, the dice loss is simply used to train all models for a purpose of performance comparison. Equation 26 shows the dice loss function.

$$DiceLoss(y, \overline{y}) = 1 - \frac{(2y\overline{y}+1)}{(y+\overline{y}+1)}, \tag{26}$$

Where y represents the ground truth and $\overline{y}$ represents the predicted segmentation generated by a deep learning model. After assembling all the parts for building the models, the MPI-CT images and their corresponding segmentation masks were then utilized to train the models as inputs and ground truths, respectively with the following hyperparameters: an Adam optimizer [132] with an intimal rate of $5x10^{-4}$, a batch size of 8, and 60 epochs. All the models in this study were trained on a personal computer equipped with an 11^{th} Gen Intel core i7-11700k CPU, 64 GB of RAM, and an NVIDIA RTX 3090 graphic card.

4.3 Results and discussion

4.3.1 Gradient-weighted class activation maps (Grad-CAM)

Gradient-weighted class activation mapping (Grad-CAM) [133] is a class-discriminative localization technique. It can generate a visual representation of any CNN-based model without altering the model itself. Grad-CAM leverages the gradient information flowing through a specific convolutional layer to assign crucial weights to each neuron to determine a particular decision of interest. This gradient information is then used to calculate the localization map visualized as a heat map image. In short, the intuitive interpretation of Grad-CAM is based on the concept that

the model must observe some pixels and decide what object is present in the image, which can be interpreted as a gradient in mathematical terms. To compute Grad-CAM, the equations below are applied. Equation 27 is used to calculate the neuron's important weight (α_k^c) by calculating the global average pooling of the gradient from backpropagation. α_k^c is then employed to calculate the localization map Grad-CAM as shown in Equation 27 and 28.

$$\alpha_k^c = \frac{1}{Z}(\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}), \tag{27}$$

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k), \tag{28}$$

Where $\frac{\partial y^c}{\partial A_{ij}^k}$ is the gradient from backpropagation, A^k is feature map activation of a convolutional layer, α_k^c is neuron import weight, $L_{Grad-CAM}^c$ is localization map Grad-CAM (coarse heat map). Grad-CAM is applied to each multi-head attention layer (MH-AG layer 1-4) output in order to characterize and understand the multi-head attention U-Net model behavior. The attention weights of different MH-AG layers are visualized as shown in Figure 20. Figure 20(a) shows the input image, ground truth, and the segmentation outputs of 6-head, 4-head, and 2-head attention U-Net models. Figure 20(b) shows the Grad-CAM results of the corresponding attention U-Net models. According to these Grad-CAM results and final segmentation outputs, the 4-head attention U-Net model can exceptionally perform MPI-CT image segmentation and surpass 6-head and 2-head attention U-Net models since it can focus on more meaningful features and predict a more accurate result. It is interesting to note that each MH-AG layer output of the 4-head attention U-Net model pays attention to different meaningful features, the MH-AG layer 4 pays attention to the overall boundary of the MPI signal, the MH-AG layer 3 focuses on the increasingly precise boundary of the MPI signal, the MH-AG layer 2 changes the focus from the boundary of the MPI signal to the skeleton (bone structure, i.e., CT image), and the MH-AG layer 1 entirely focuses on the real target

MPI signal. With these different meaningful features, the learnable parameters of the model can be assigned to pay attention to the relevant features and circumvent irrelevant features for the final prediction. However, the 2-head and 6-head attention U-Net models behave in different ways. The MH-AG layers 4 and 3 of the 2-head attention U-Net poorly estimate the boundary of the MPI signal, and the MH-AG layers 2 and 1 focus on somewhat the same features (MPI signal areas). Although the MH-AG layers 4 and 3 of the 6-head attention U-Net can perform better than the 2-head attention model, the MH-AG layers 2 and 1 also pay attention to relatively the same features (MPI signal areas). Indeed, the optimal number of attention heads depends on the tasks we desire to train the deep learning model and the data features. If there are a larger number of important features, the higher number of attention heads could potentially help the model perform better by capturing more essential information. Nevertheless, the excessive number of attention heads could lead to less impressive performance, according to the Grad-CAM results illustrated in Figure 20 and our quantitative experiment discussed in the next section.

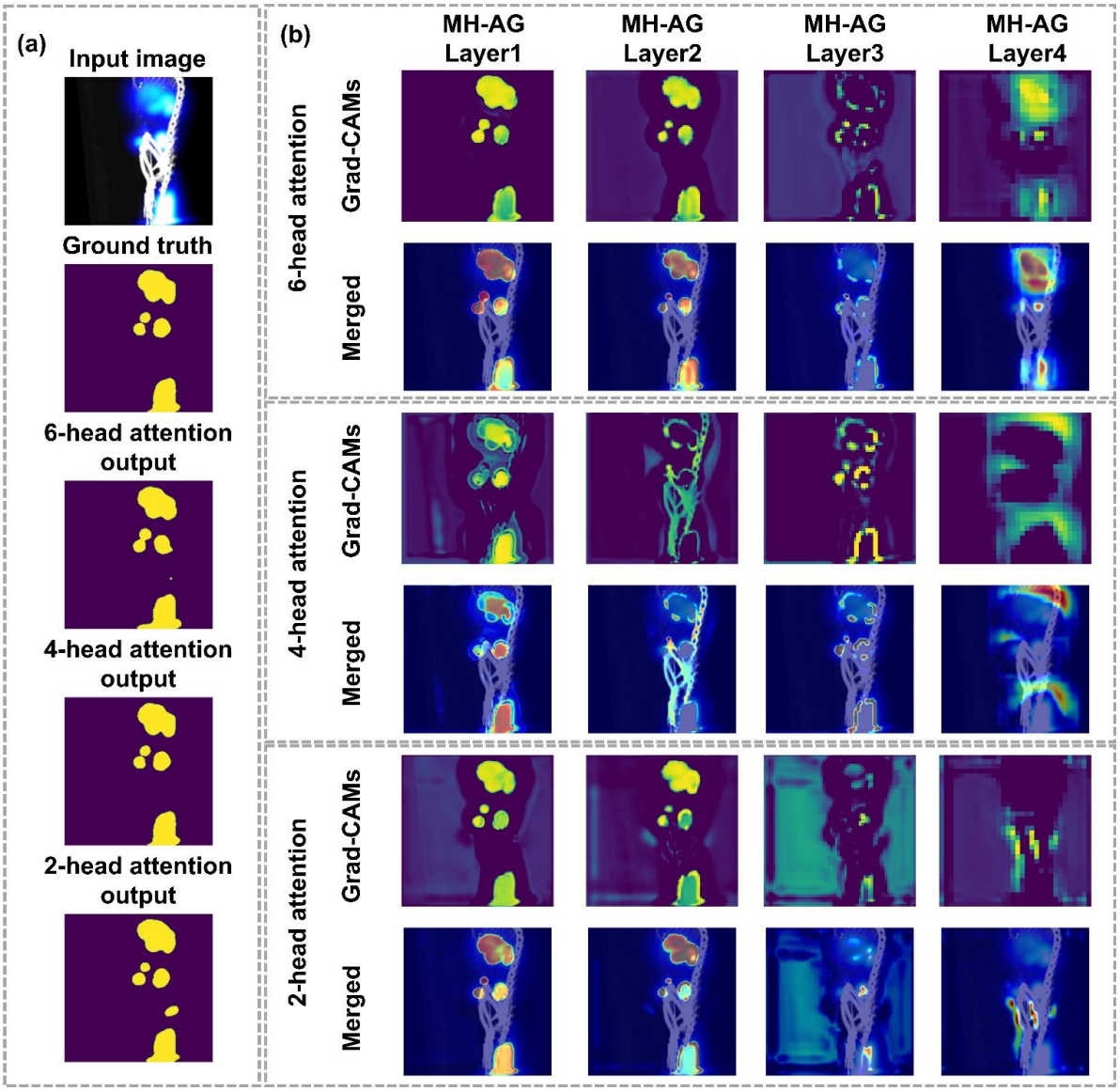


Figure 20. A comparison of Grad-CAMs results of 2-head attention, 4-head attention, and 6-head attention U-Net architectures. (a) Input MPI-CT image, segmentation ground truth and outputs of each attention architecture. (b) The Grad-CAM results of the attention architectures at different MH-AG layers (MH-AG layer (1-4)).

4.3.2 Implementation and evaluation metrics

Intersection over Union (IoU) is commonly used to evaluate the similarity between a predicted segmentation area and its ground truth [121]. The concept of IoU is to quantify the common area of the ground truth and prediction mask (intersection) divided by the entire number of pixels present across both the prediction mask and ground truth (union) as shown in the equation below.

$$IoU = \frac{ground\ truth \cap prediction}{ground\ truth \cup prediction} \tag{29}$$

The IoU ranges from 0 -1 (0-100%), with 0 indicating no overlapping area, whereas 1 indicates impeccably overlapping area.

The dice similarity coefficient (DSC) is another well-known parameter used to evaluate the similarity between the predicted area (our output) and ground truth [32]. The DSC can be calculated following the equation below.

$$DSC = \frac{2|ground\ truth\ \cap prediction|}{|ground\ truth| + |prediction|} \tag{30}$$

Precision is defined as the ratio of true positive results to the total number of positive results, which is the summation of true positive and false positive as shown in Equation 31.

$$Precision = \frac{TP}{TP+FP}, \tag{31}$$

Sensitivity, also known as Recall, is the number of true positive results over the summation of true positive and false negative results as shown in Equation 32.

$$Recall = \frac{TP}{TP+FN}, \tag{32}$$

Accuracy, also known as the Rand index, is the number of correct predictions divided by the total number of predictions as shown in Equation 33.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP}, \tag{33}$$

Where TP is true positive, TN is true negative, FP is false positive, and FN is false negative. As previously stated, if the number of attention heads is excessive, the performance of a deep learning model based on the attention heads could deteriorate. Thus, we characterized the number of attention heads and employed Dice and IoU as the representative benchmarks. Figure 21 illustrates the characterization results of the U-Net based on the different number of attention heads. With regards to the plot of Dice/IoU scores vs the number of attention heads, it begins at 0.889/0.804

with the 1-head attention architecture, it gradually increases and then reaches the highest score at 0.909/0.835 with the 4-head attention architecture before declining progressively to 0.906/0.829 and 0.901/0.822 with 5 and 6 attention heads, respectively. Therefore, the multi-head attention U-Net with 4 heads is the optimal model providing the best result for the MPI-CT image segmentation.

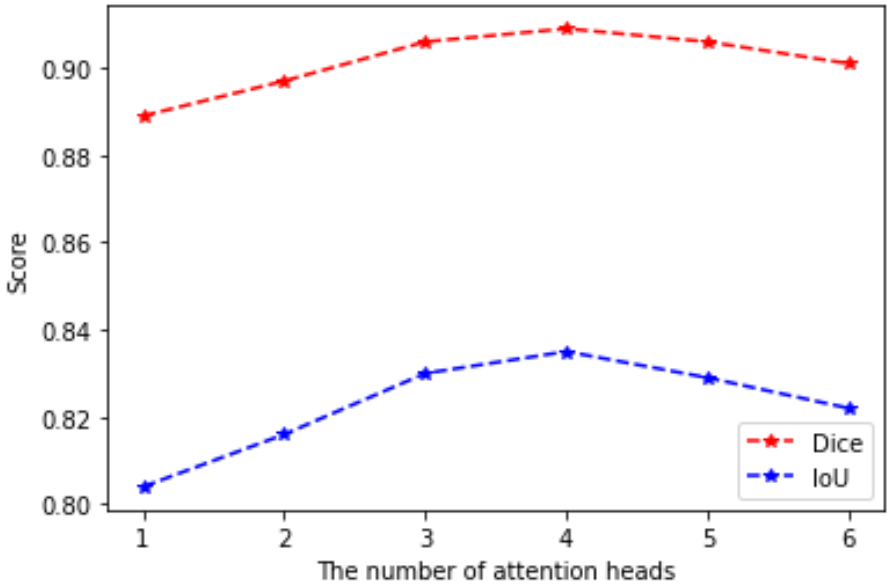


Figure 21. The performance of the multi-head attention U-Net models with the different number of attention heads (Dice/IoU scores vs the number of attention heads plot).

Table 6 shows the comprehensive characterization results of MPI-CT image segmentation of deep learning models with different architectures. Apart from using Dice and IoU scores as model evaluation metrics, we also characterized the performance of each model using accuracy, precision, and recall. Overall, the 4-head attention U-Net model can outperform other multi-head attention U-Net models including the original U-Net model as well as the state-of-the-art Transformer U-Net model. The representative visualization MPI-CT image segmentation results, together with the corresponding input images and ground truths of each architecture are illustrated in Figure 22**.**

Table 6. Quantitative evaluation (average ± standard deviation of each metric) of the different deep learning architectures for MPI-CT image segmentation.

Methods	Accuracy	Precision	Recall	Dice	IoU
U-Net	0.983± 0.004	0.891± 0.074	0.879± 0.076	0.883± 0.059	0.794± 0.089
Transformer U-Net	0.985± 0.005	0.909± 0.057	0.878± 0.069	0.892± 0.053	0.809± 0.083
1-head Attention U-Net	0.984± 0.005	0.892± 0.068	0.891± 0.069	0.889± 0.052	0.804± 0.083
2-head Attention U-Net	0.985± 0.004	0.888± 0.063	0.911± 0.057	0.897± 0.041	0.816± 0.052
3-head Attention U-Net	**0.987± 0.005**	**0.926± 0.038**	0.890± 0.065	0.906± 0.039	0.830± 0.063
4-head Attention U-Net (the proposed model)	**0.987± 0.005**	0.920± 0.040	0.902± 0.058	**0.909± 0.036**	**0.835± 0.060**
5-head Attention U-Net	0.986± 0.004	0.913± 0.049	0.903± 0.060	0.906± 0.030	0.830± 0.050
6-head Attention U-Net	0.985± 0.005	0.894± 0.074	**0.912± 0.053**	0.901± 0.043	0.822± 0.070

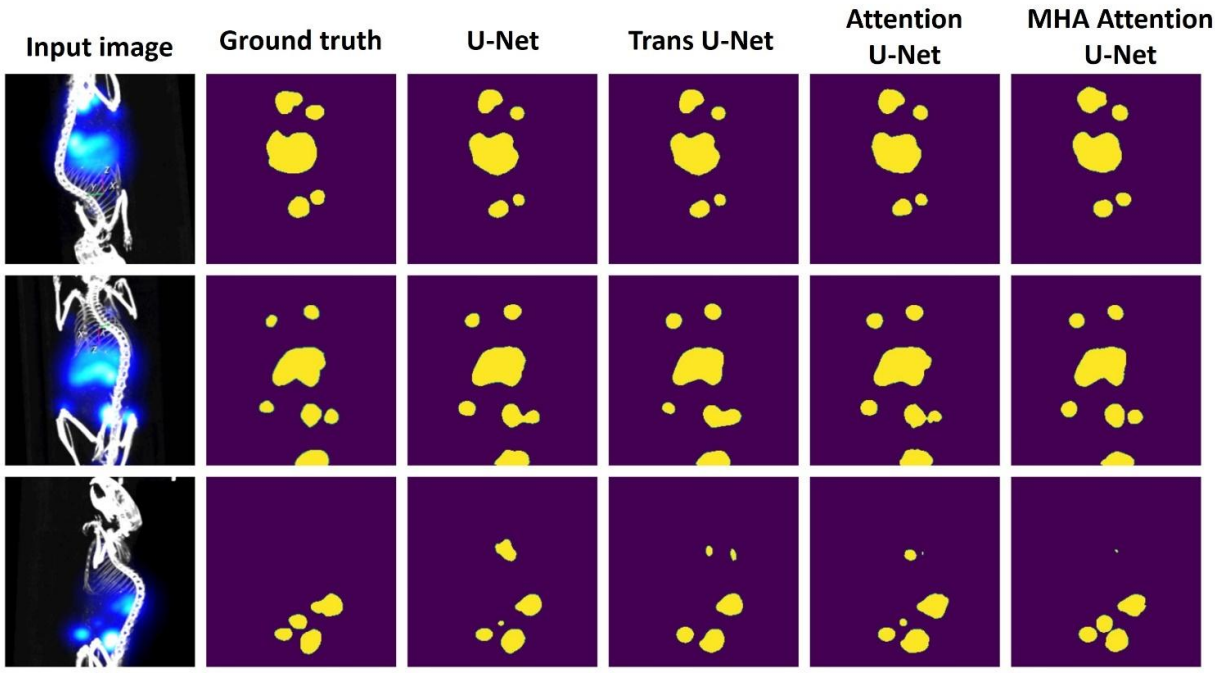


Figure 22. Visualization semantic segmentation results of the proposed model compared to other traditional U-Net models. From left to right, input MPI-CT images, the ground truth images, the segmentation results generated by U-Net, Trans-U-Net, Attention U-Net, and our proposed model (4-head attention), respectively.

4.4 Conclusion

Since MPI is a novel medical imaging technology, the data are strictly limited for a robust computation study. This work demonstrates the multi-head attention U-Net model, an efficient end-to-end deep learning based on U-Net architecture and multi-head attention mechanism, for

MPI-CT image segmentation. The proposed model was trained using a custom MPI-CT image dataset collected from transgenic mice with breast tumors injected with a promising MPI tracer for tumor imaging, namely NWs-ICG. To examine the concept of multi-head attention, a simple convolution block is employed as the backbone structure of the U-Net architecture to minimize the influence of other factors. Genuinely, the performance of the U-Net architecture can also be improved by using more efficient convolution blocks as the backbone. The optimal number of attention heads was experimentally observed in this study. Although an increase in the number of attention heads can potentially boost the model's capability, the excessive number of attention heads results in a decline in capability. Our study shows that the attention U-Net with 4 heads is the most favorable architecture for MPI-CT image segmentation. In future work, in addition to improving the model's performance, we would like to explore the possibility of exploiting deep learning for 3D MPI segmentation and MPI intensity segmentation. We anticipate this work to embark on an intensive study for MPI image analysis and implement it on humans in the near future.

CHAPTER 5: Monocular Depth Estimation Based on Deep Learning for Intraoperative Guidance Using Surface-enhanced Raman Scattering (SERS) Imaging

Imaging of surface-enhanced Raman scattering (SERS) nanoparticles (NPs) has been intensively studied for cancer detection due to its high sensitivity, unconstrained low signal-to-noise ratios, and multiplexing detection capability. Furthermore, conjugating SERS NPs with various biomarkers is straightforward, resulting in numerous successful studies on cancer detection and diagnosis. However, Raman spectroscopy only provides the spectral data from an imaging area without co-registered anatomic context. This is not practical and suitable for clinical applications. Here, we propose a custom-made Raman spectrometer together with computer vision-based positional tracking and monocular depth estimation using deep learning (DL) for the visualization of 2D and 3D SERS NPs imaging, respectively. In addition, the SERS NPs used in this study (hyaluronic acid (HA)-conjugated SERS NPs) showed clear tumor targeting capabilities (target CD44 typically overexpressed in tumors) by an *ex vivo* experiment and immunohistochemistry. The combination of Raman spectroscopy, image processing, and SERS molecular imaging, therefore, offers a robust and feasible potential for clinical applications.

5.1 Introduction

Surgical resection of a tumor is a standard of care therapy for most solid tumors. The ultimate goal of surgical resection is to remove the entire tumor with minimal damage to adjacent tissue, an outcome that strongly correlates with reduced tumor recurrence and improved survival [135, 136].

Tumor margins in numerous aggressive cancers are typically indistinct due to the primary tumor's propensity to invade into adjacent healthy tissue areas. As a result, defining appropriate margins for surgical resection remains challenging [137]. There are several modalities used in the clinic to visualize tumors and facilitate tumor removal such as magnetic resonance imaging (MRI), positron emission tomography (PET), and computed tomography (CT) [138-141]. However, these imaging modalities lack sufficient resolution needed to identify and remove microscopic sites of cancer invasion from the main tumor mass. To achieve precise tumor delineation and complete resection, a suitable intraoperative tool should meet the following requirements: high sensitivity and specificity, short acquisition time for real-time or near-real-time intraoperative detection, and high spatial resolution. With regards to imaging modalities, optical imaging exhibits distinct advantages compared to the previously mentioned non-optical imaging modalities in several aspects, such as lack of ionizing radiation, high sensitivity, and excellent spatiotemporal resolution [142-145].

Recently, surface-enhanced Raman spectroscopy (SERS) nanoparticles (NPs) imaging has increasingly been recognized as a promising molecular imaging technique for clear delineation of tumor margins and tumor surgical resection due to its exceptional sensitivity, distinctive Raman signature (fingerprint), multiplexing detection capability [146-152], and lack of autofluorescence and photobleaching problems associated with fluorescence imaging. SERS NPs are composed of a gold core, Raman active dye, and silica shell, which have been developed to function as tumor-targeting beacons showing substantially strong signals due to the surface plasmon resonance (SPR) effect [153] of the metallic core (gold). In addition, they can be effortlessly conjugated with various tumor-targeting ligands as well as fabricated with different Raman-active dyes. Each Raman dye emits a unique Raman spectrum, called "flavor", facilitating multiplexing. Several research groups, as well as our group, have demonstrated encouraging results of SERS NPs imaging for *ex*

vivo, *in vivo*, and image-guided surgery experiments [154-158]. However, Raman spectroscopy predominantly provides spectral data, lacking the capability to co-register and visually represent anatomic features, limiting applications for image-guided surgery.

To overcome this problem, we propose a custom-made Raman spectroscopy system together with computer vision-based positional tracking and DL-based techniques to visualize 2D and 3D SERS NPs imaging, respectively. Specifically, the traditional template matching algorithm [159] is employed for probe tracking, and the affine transformation [160] is then used to co-register a 2D SERS image (reconstructed by using the multiplexing algorithm [161, 162]) and a sample photograph. For 3D imaging, the image is reconstructed based on a deep-learning monocular depth estimation (distance relative to the camera) of each given pixel in the input image. Multiple Depth Estimation Accuracy with Single Network (MiDaS) is a promising DL technique that estimates depth from an arbitrary input image. MiDaS utilizes a conventional encoder-decoder structure to generate the depth map images. The legacy MiDaS V2.1 model [163] uses a residual network as the backbone for feature extraction as this network structure is invulnerable to vanishing gradients and allows MiDaS to extract multi-channel feature maps from input tensors. The vision transformer (ViT) [164] is the state-of-the-art model employed in computer vision tasks. It can surpass convolutional neural networks (CNNs)-based models across various domains and settings. Therefore, the latest MiDaS versions (3.0 [165] and 3.1 [166]) replace the CNNs backbone with vision transformer networks showing superior results. In this work, we directly utilized the pre-trained MiDaS 3.1 to reconstruct a 3D mouse image and co-register with the SERS image.

5.2 Methods

5.2.1 Raman spectrometer

A schematic of the proposed Raman system is illustrated in Figure 23. A 785-nm laser (iBeam Smart 785, Toptica Photonics, Munich, Germany) is employed for the excitation source, the custom-made fiber bundle Raman catheter (Fiber guide Industries, Caldwell, ID, USA) is used for the laser illumination and the Raman spectra collection. A proximal end of the probe is made up of one single mode fiber (780HP, 4.4 µm core diameter) for 785 nm laser illumination and 36 multimode fibers (AFS200/220T, 200-µm core) for the Raman spectra collection as shown in Figure 23(b). The single-mode fiber for illumination is centrally positioned with the probe and encompassed by the 36 multimode fibers for Raman spectra acquisition. In addition, a fused silica plano-convex lens (L1, f=6.83 mm, PLCS-4.0-3.1-UV, CVI Laser Optics, Albuquerque, NM, USA) is placed in front of the probe to collimate the 785 nm laser illumination with a beam diameter of 1 mm and power of 30 mW on the sample. For the distal end, it is arranged in a vertical array or linear array for effectively coupling the light to the spectrometer (Kymera 193i-A, Andor Technology, Belfast, UK) by using optical relay lenses (L2, f = 100 mm, AC254-100-B and L3, f=80 mm, AC254-080-B, Thorlabs Inc., Newton, NJ, USA). In addition, the Rayleigh scattering from the collected light is filtered out by a long-pass filter (LPF, λ_c = 830 nm; BLP01-830R-25, Semrock, Rochester, NY, USA), placed between the relay lenses. As a result, the light that traverses the spectrometer is solely subjected to Stokes-Raman scattering. The Stokes-Raman scattering light from the spectrometer is then collected by a cooled deep-depletion spectroscopic charge-coupled device (CCD) array (1024 x 256 pixels with a pixel size of 26 µm x 26 µm; DU920P Bx-DD, Andor technology, Belfast, UK) with a wavelength range of 835- 912 nm (Raman shift of 770 -1777 cm^{-1}). To achieve raster scanning, a two-axis translation stage is

constructed by joining two linear stages in an orthogonal manner (DDS050, Thorlabs Inc., Newton, NJ, USA). Furthermore, a color monocular camera (ELP 5-50mm, with Sony IMX323 chip, Shenzhen, China) is applied to track the Raman probe position and capture the sample photograph to reconstruct the 2D and 3D co-registered SERS images.

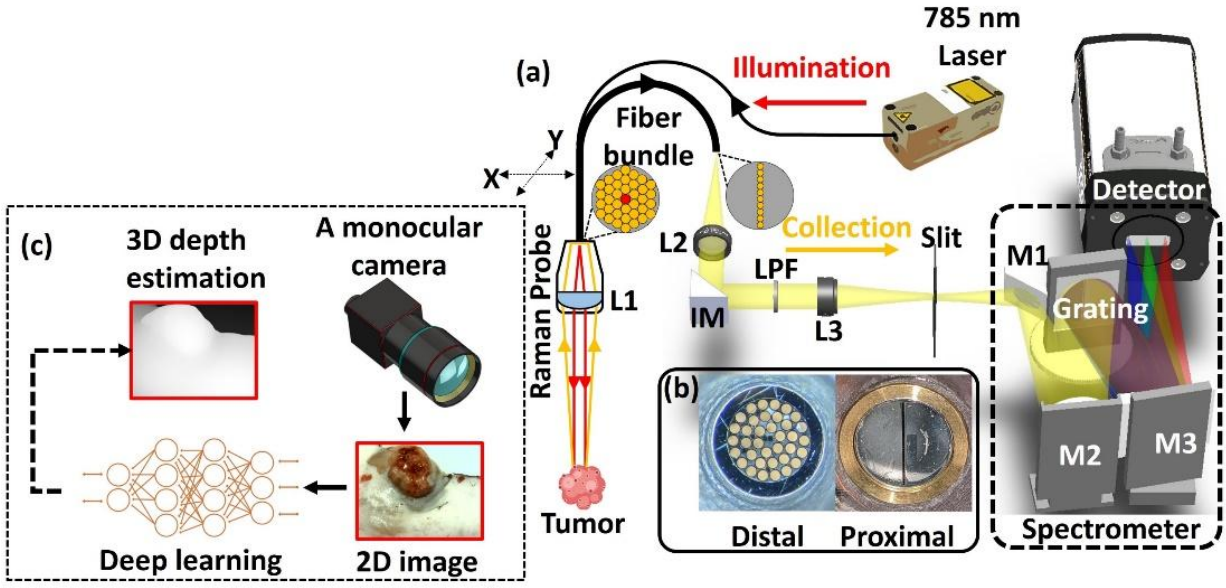


Figure 23. Schematic of the custom-made Raman imaging system together with the visualization system. (a) The optical diagram of the Raman spectroscopy system. A 785 nm laser is used to illuminate the sample through a single mode fiber and collimated by an L1 lens. The scattered light is then collected by the Raman probe, coupled into the spectrometer using the relay optics (L2 and L3 lenses) with an interchangeable mirror (IM) and a long pass filter (LPF) in between. The spectrometer consists of a rotatable grating, three mirrors (M1: reflection mirror, M2: collimating mirror, and M3: focusing mirror), and a back-illuminated deep-depletion CCD. To perform 2D Raman imaging, the Raman probe is translated by a two-axis motorized stage. (b) the photograph of the distal and proximal ends of the custom-made fiber bundle. (c) Schematic of the visualization system for generating the 2D and 3D co-registered SERS imaging.

5.2.2 SERS NPs synthesis

SERS NPs were synthesized using the tris-based assisted synthesis protocol with Au NPs formation at elevated temperature as shown in Figure 24(a). First, the sodium citrate reduction approach was employed to prepare 17 nm Au-NP seeds. The seeds were then mixed with tris at 98 °C, followed by adding gold chloride for seed-mediated growth to obtain 50 nm Au-NPs. The Raman dye was promptly added after the formation of 50 nm Au-NPs, and the solution was stirred

for one minute, followed by cooling in an ice bath. To functionalize SERS NPs with biomolecules, particularly hyaluronic acid (HA) and polyethylene glycol (PEG), thiol groups were employed for the attachment of these biomolecules and to Au NPs via gold-thiol interaction [167-171] . S420 SERS NPs were mixed with thiolated-HA and this mixture solution was then incubated at 4 °C overnight. After that, unbounded HA was removed by repeated centrifugation. Likewise, the procedure to conjugate PEG with S481 SERS NPs is the same as the HA conjugation. The size and shape of synthesized SERS NPs were characterized by a transmission electron microscope (TEM; 2200FS, JEOL Ltd., Tokyo, Japan) and a dynamic light scattering particle analyzer (DLS; Zetasizer Nano ZS, Malvern Panalytical Ltd., Malvern, England, UK). SERS NPs are homogenous spheres with approximately 50 nm in diameter as shown in Figure 24(b). The DLS result was also applied to validate the distribution size with a measurement of 56 nm as shown in Figure 24(c). The comprehensive synthesis protocol and characterization of SERS-NPs are demonstrated in our previous work [157]. The normalized Raman spectra (acquired by our custom-made Raman spectrometer) of S420 and S481 SERS NPs with a concentration of 500 pM are demonstrated in Figure 24(d).

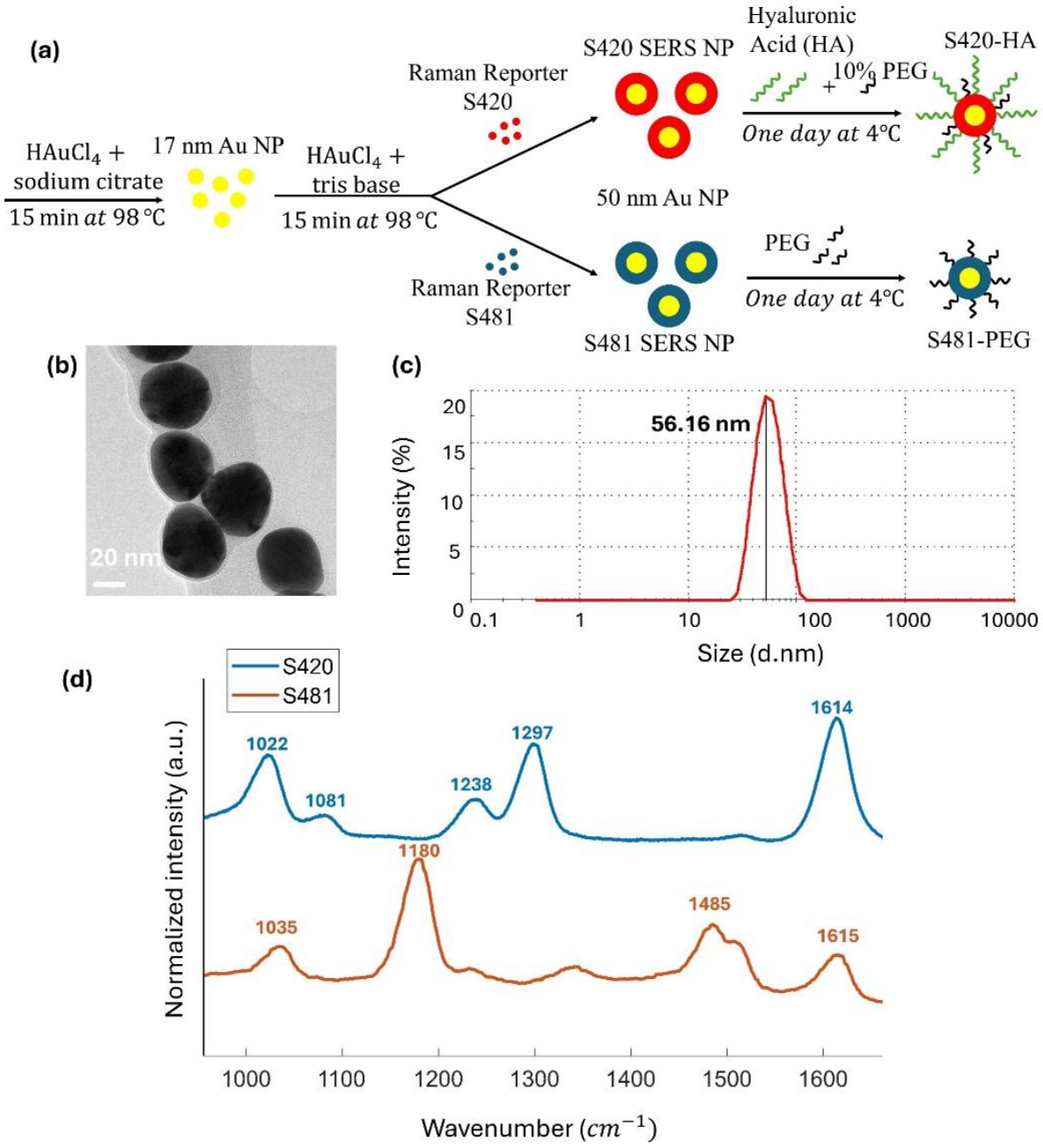


Figure 24. Synthesis of the SERS NPs. (a) SERS NPs synthesis and HA/PEG conjugation procedure. First, 17 nm gold seeds (Au NP) are formed. Second, the NPs further grow to 50 nm meanwhile different Raman reporters (S420 and S481) are attached to the gold surface. Lastly, the SERS NPs are functionalized with HA or PEG. (b) TEM image of the SERS NP with diameter of approximately 50 nm. (c) DLS result of the corresponding SERS NPs. The measured size is 56.16 nm in diameter. (d) Normalized Raman spectra of the stock SERS solution of both flavors (S420 and S481).

5.2.3 Position tracking and image co-registration algorithms

Before processing the data acquired by a low-cost camera, camera calibration [172, 173] was applied to correct the image distortion due to the lens quality and optical alignment. Template matching algorithm [174] is then used to determine the precise position of a Raman probe image (the template image) in a large surgery area image (the input image). The concept of this algorithm is to slide the template image over the input image, akin to a 2D convolutional operation, followed by a comparison of the template and the corresponding patch of the input image, which can be done by several methods. In this work, we employed a normalized cosine coefficient (TM_CCOEF_NORMED) implemented in Python using the OpenCV library [175] to calculate the template matching for the Raman probe detection. With the Raman probe position, the scanning position can be easily estimated during data acquisition. In addition, to accurately overlay the SERS image (X) and surgery area image (Y), an image co-registration algorithm is required by calculating the geometric transformation matrix (T) as shown in the equations below.

$$Y = T.X, \tag{34}$$

$$X = \begin{bmatrix} x'_1 & x'_2 & \cdots & x'_n \\ y'_1 & y'_2 & \dots & y'_n \\ 1 & 1 & \dots & 1 \end{bmatrix}, \tag{35}$$

$$Y = \begin{bmatrix} x_1 & x_2 & \dots & x_n \\ y_1 & y_2 & \dots & y_n \\ 1 & 1 & \dots & 1 \end{bmatrix}, \tag{36}$$

$$T = \begin{bmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ 0 & 0 & 1 \end{bmatrix}, \tag{37}$$

where (x'_n , y'_n) and (x_n, y_n) are the corresponding positions (n is the number of corresponding positions) in the input image X and the reference image (Y), respectively, and m_{ij} is the simplified transformation matrix parameters derived from the rotation, scaling, shearing, and translation matrices as shown in the equation below.

$$T = \begin{bmatrix} 1 & 0 & t_x \\ 0 & 1 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & sh_x & 0 \\ sh_y & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\theta) & \sin(\theta) & 0 \\ -\sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (38)$$

Where the translation matrix: t_x and t_y are the displacement along the x and y axes, respectively, the scaling matrix: s_x and s_y are the scale factors along the x and y axes, respectively, the shear matrix: sh_x and sh_y are the shear factors along the x and y axes, respectively, and the rotation matrix: θ is the angle of rotation. Indeed, T matrix can be estimated by using corresponding points together with the minimized least square error (ε^2) as shown below:

$$\varepsilon^2 = \|TX - Y\|^2, \quad (39)$$

$$\frac{d\varepsilon^2}{dT} = -2X^T(Y - TX) = 0\,, \quad (40)$$

$$X^TY = X^TTX\,, \quad (41)$$

$$T = (X^TX)^{-1}\ (X^T\mathrm{Y}). \quad (42)$$

To obtain a more accurate co-registration result (2D co-registered SERS image), the estimated transformation matrix (T) is then applied to the reconstructed SERS image (X) derived from the demultiplexing algorithm. In our case, the raster scan was applied to reconstruct the SERS image and the fiducial landmarks (four corners of the scanning area) were marked on the sample. Thus, the four corners of the SERS image were used as the corresponding points to the four fiducial points on the samples for the image co-registration.

5.2.4 Depth estimation using DL

MiDaS is considered as a promising model for performing monocular depth estimation, and the original MiDaS V 2.1 [163] is based on a CNN backbone, however the newer versions (MiDaS V 3.0 [165] and V 3.1 [166]) employ a transformer architectures as their backbones, which can significantly outperform the original version. The training protocol of the MiDaS V2.1, 3.0, and 3.1 models are analogous. Breifly, the MiDaS models were trained by using 12 mixing datasets,

multi-objective optimization [176] with Adam [177], and scale-and-shift-invariant loss [178]. The encoder and decoder weights were updated by applying the learning rates of 1e-5 and 1e-4, respectively. The models were initially pre-trained on a subset of the datasets for 60 epochs, followed by training for another 60 epochs on the full dataset. The complete training details are elucidated in the original MiDaS V 2.1 paper. All DL models demonstrated in this work were implemented on a personal computer equipped with an 11^{th} Gen Intel core i7-11700k CPU, 64 GB, and an NVIDIA RTX 3090 graphic processing unit (GPU). Indeed, all MiDas models are built using encoder and decoder structures. Each MiDaS model differs in the backbone of the encoder part (variant of CNNs and Transfomer architectures), while the rest of the model remains consistent. Since the latest MiDaS V 3.1 provides the best result compared to other versions, it is used in this study. Bi-direction Encoder repression from Image Tranfomers (BEiT) [179] is used as the backbone of MiDas V 3.1, as shown in Figure 25(a-b). BEiT is a state-of-the-art architecture that enables self-supervised pretraining of vision transformer (ViT) to surpass supervision pretraining. The pre-train task in BEiT is the masked image modeling (MIM) head, as shown in Figure 25(b). The concept of MIM is to recover the original visual tokens based on the corrupted image patches. In other words, MIM uses two views for each image to train the model. First, the 2D image with a size of HxWxC is divided into a sequence of HW/P^2 patches for each channel, where (H,W) is the image size, C is the number of channels, and (P,P) is the patch size. All the patches are then flatten into vectors and linearly projected. Second, an image tokenizer converts the image into a sequence of discrete tokens rather than using raw pixels. Discrete variational autoencoder (dVAE) [180, 181] is directly used to train this image tokenizer. Indeed, the image tokenizer is a readily trained token genertor for the input patches.

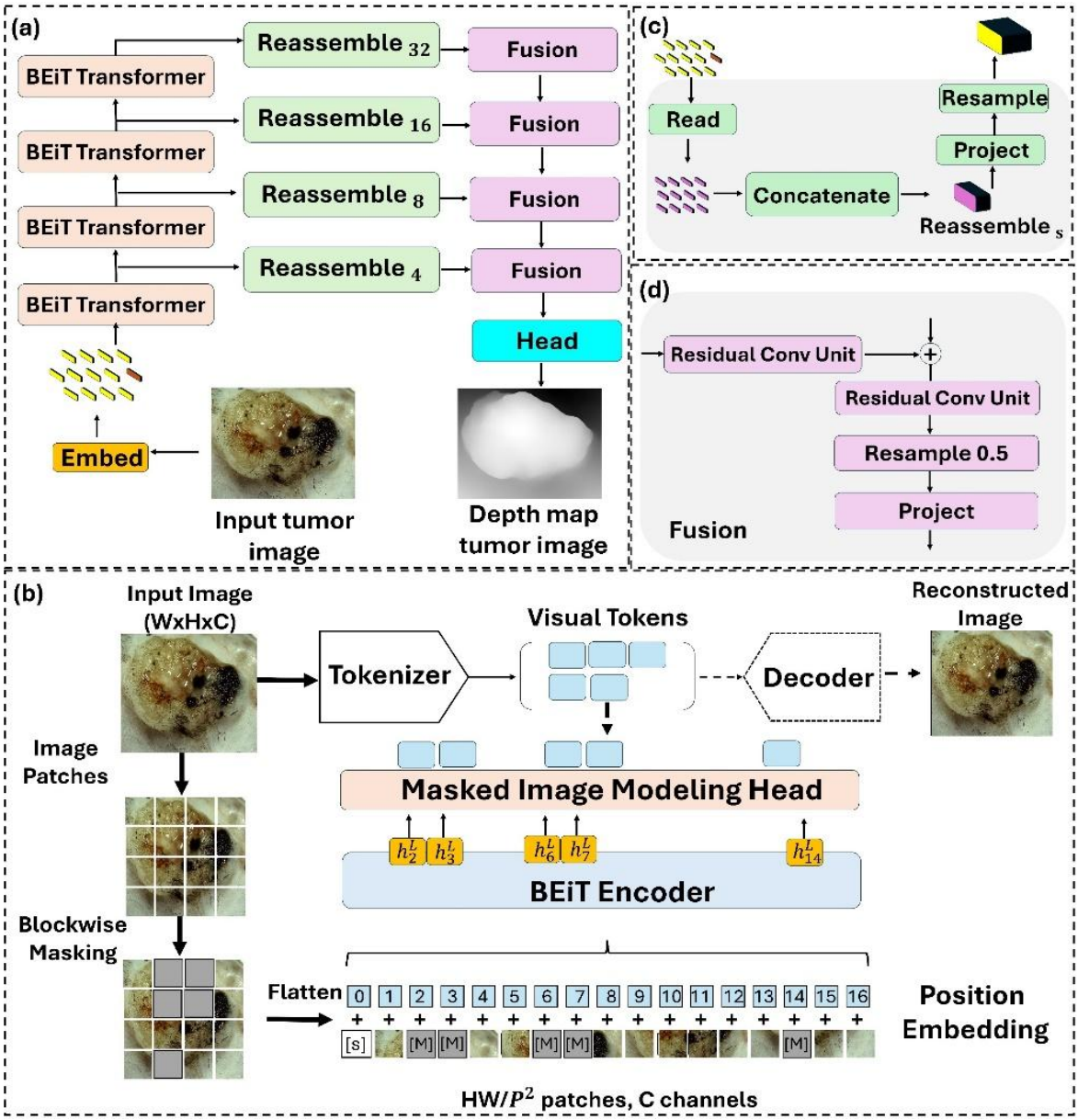


Figure 25. (a) Overview of the MiDaS V 3.1 architecture. The input image is embedded with a positional embedding and a patch-independent readout token (orange) is included. These patches are fed to four BEiT stages. At each BEiT, the output tensor is passed through the Reassemble and Fusion blocks to predict the encoder outputs for each stage. (b) BEiT transformer architecture used in the encoder part in (a). (c) Reassemble block applied to assemble the tokens into feature maps with 1/s the spatial resolution of the input image. (d) Fusion block used to combine the features and upsample the features maps by two times.

The outputs from the tokenizer and MIM are used to determine the loss value to update the learnable parameters allowing the network to obtain a deep understanding of underlying image patterns without the explicit lables. It is important to note that the BEiT was initially designed for

an image classification problem and does not provide depth estimation functionality. To assemble MiDaS V 3.1, BEiT is used as a feature extractor and must be appropriately connected to the depth decoder. Regarding the encoder-decoder in MiDaS, the input is progessively processed for each encoder stage, similar to the decoder stage. Thus, the BEiT backbone can be integrated by placing appropriate hooks, meaning a tensor computed in the encoder is taken and available as input for the decoder at one of its stage. This requires a reassembling process to reshape the tensors to fit the decoder, as shown in Figure 25(c-d). Essentially, the input image is embedded as the tokens, which are passed through serval BEiT stages. At each stage, the tokens are ressemable into image-like represtanion with different resolutions. After that, the fusion module is employed to fuse and upsample these image-like represtanions in order to generate an exquisite prediction. The final prediction is then fed to a task-specific ouput head to generate the depthmap image. The depth map image generated by the MiDaS model is considered as a dispartily-like image (inversely propotional to the depth map intesnity), which is then projected into 3D space using the reprojectImageTo3D function in OpenCV [175]. Lastly, The color of each pixel in the 2D co-registered SERS image is mapped onto the corresponding positions (x-y plane) in the 3D space of depth map image to obtain the final 3D SERS image.

5.3 Results and discussion

5.3.1 Phantom characterizations

The step-wedge with a height of 9.5 mm of each step, which was constructed from the standard mounting bases (BA1S, Thorlabs Inc., Newton, NJ, USA), was used as a phantom to characterize the depth estimation DL models. The camera captured this phantom photograph and was used as the input for the three different MiDaS models (CNN, ViT, and BEiT) to estimate the depth and compare the performance of each model. To quantify the performance of each model, the depth

map intensities from step 4 to step 1 (along with the white-dashed line) were plotted as illustrated in Figure 26(a). The absolute errors were then calculated from the intensity profiles of each model and the ground truth (the black line). Table 7 shows the average absolute error ± standard deviation results of each model. It shows that the MiDaS model based on BEiT architecture can surpass other models with the lowest average absolute error of 0.0485 ± 0.1737.

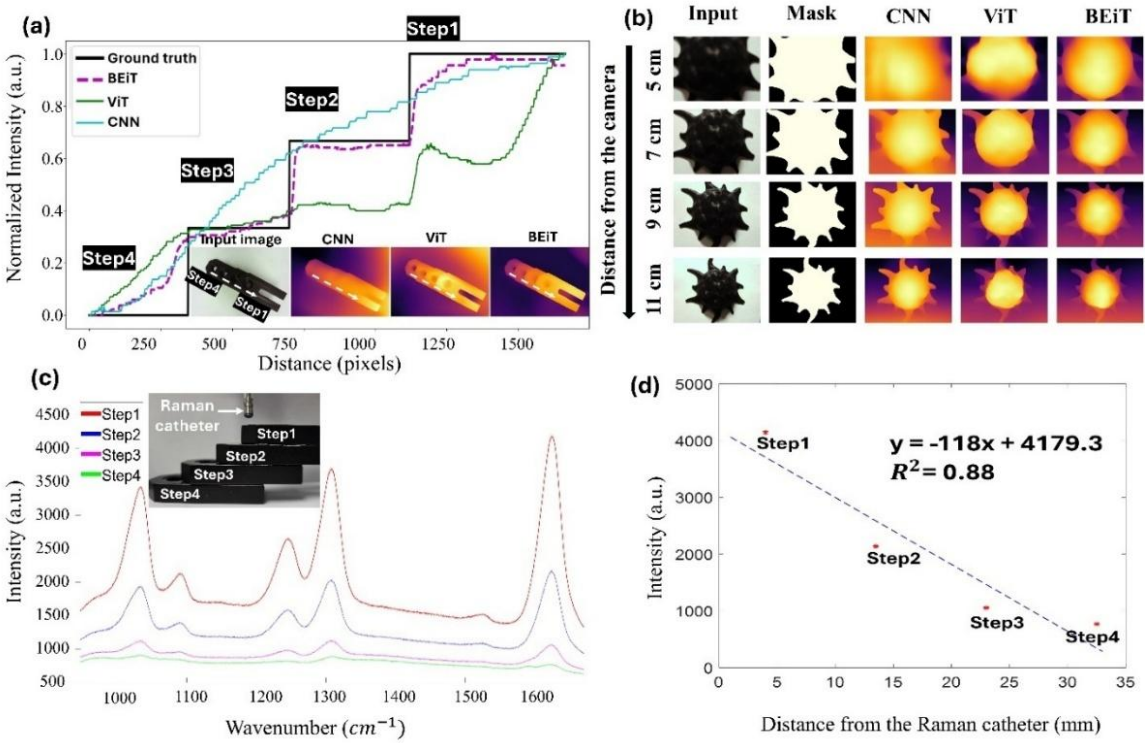


Figure 26. Validation of depth map imaging and Raman spectra at different distances from a camera and a Raman catheter, respectively. (a) Depth map imaging of a step-wedge phantom generated by MiDaS models based on three different backbones (CNN, ViT, and BEiT) and the comparison of the depth map intensity profiles of each model. (b) Depth map imaging of a tumor phantom with different distances from the camera. (c) The Raman spectra of S420 SERS NPs characterization at different distances from the Raman catheter by using the step-wedge phantom. (d) A linearity plot of the highest intensity of S420 (1614 cm^{-1}) versus the distances from the Raman catheter.

Table 7. Depth map intensity characterization results (Average absolute error ± Standard deviation) of MiDaS models with three different architectures: CNN, VIT, and BEiT.

Step number	**CNN**	**ViT**	**BEiT**
Step 1	0.074 ± 0.56	0.318 ± 0.14	**0.051** ± 0.56
Step 2	0.070± 0.046	0.252± 0.01	**0.032** ± 0.04
Step 3	0.135± 0.088	**0.018± 0.016**	0.024 ± 0.012
Step 4	0.092± 0.077	0.161± 0.10	**0.087± 0.083**

Furthermore, a 3D-printed tumor phantom was utilized for thorough characterization of the MiDaS models, as depicted in Figure 26(b). The distance between the phantom and camera varied from 5 cm to 11 cm with an increment of 2 cm. The phantom depth map images were then generated by the MiDaS models. The quality images captured at the out-of-focus distances (5 cm and 7 cm) are unsatisfactory, leading to deterioration of depth map quality, as the models cannot correctly recognize some poor resolution areas to generate the depth map image, especially the CNN MiDaS model. Nevertheless, the BEiT model can still generate somewhat decent quality depth map images. Table 8. shows four evaluation metrics (average value from all distances ± standard deviation): IoU, F1-score, Recall, and Precision, of the depth map images and their corresponding masks. This evaluation shows the overall performance of the MiDaS models for generating depth map images of the same object with different image quality (in-focus and out-of-focus images), particularly the BEiT MiDaS model can surpass other models with the promising scores of all evaluation metrics. In addition, the complexity and average execution time for one input image were evaluated to assess the feasibility for intraoperative guidance applications. Although we implemented MiDaS on a moderate-budget GPU (an NVIDIA RTX 3090 GPU), the execution time is feasible for intraoperative guidance applications. Indeed, the execution time can be improved by using more powerful GPUs currently available on the market.

Table 8. Tumor phantom characterization result of the three different MiDaS models

Evaluation	CNN	ViT	BEiT
IoU	0.139± 0.026	0.241± 0.018	0.272 ± 0.033
F1-score	0.244± 0.041	0.389± 0.024	0.426 ± 0.042
Recall	0.262 ± 0.024	0.370 ± 0.027	0.402 ± 0.029
Precision	0.234 ± 0.058	0.421 ± 0.074	0.466 ± 0.088
Execution time (second)	0.861	0.998	1.175
The number of parameters	105 M	334 M	345 M

In addition to the depth map image characterization, the intensity of Raman spectra of the same sample at various distances from the Raman catheter was also characterized by using the step-wedge phantom from Figure 26(a) and S420 SERS NPs solution with a concentration of 500 pM as shown in Figure 26(c). The SERS NPs solution was dropped on each step with a volume of 20 µL, followed by acquiring the Raman spectra using 30 mW laser power and 1 second exposure time. The linearity plot of the highest peak of S420 (1614 cm^{-1}) and the distance between the Raman catheter and sample is illustrated in Figure 26(d). The distance between the catheter and the sample is inversely proportional to the intensity of the Rama spectra. Thus, this has to be addressed to enhance the accuracy of clinical applications.

5.3.2 Ex-vivo experiment

To validate the targeting capability of the conjugated-HA SERS NPs, we performed an *ex-vivo* experiment on tumor tissue and spleen connective tissue (control) harvested from the MUC1 breast tumor mouse model [38]. All procedures used in experiments conducted on animals were approved by the Institutional Animal Care & Use Committee (IACUC) of Michigan State University. SERS-NPs used in this experiment were also published in our previous work [157]. First, we scanned the background signal from all the tissues. Second, all tissues were incubated with the mixture solution of S420-HA and S481-PEG SERS NPs with a concentration of 250 pM for 15 minutes. The S481-PEG was used as a control SERS NPs solution (non-targeting). In the next step, all the tissues were

rinsed by phosphate-buffered saline (PBS) 4-5 times, followed by acquiring the Raman spectra and reconstructing the image using the demultiplexing algorithm [161, 162]. This algorithm is based on the direct classical least squares (DCLS) method, using measured Raman spectra, reference spectra of SERS NPs of each flavor (spectra of a pure SERS NPs solution at a high concentration), and background spectra as inputs to estimate the weight of a specific flavor.

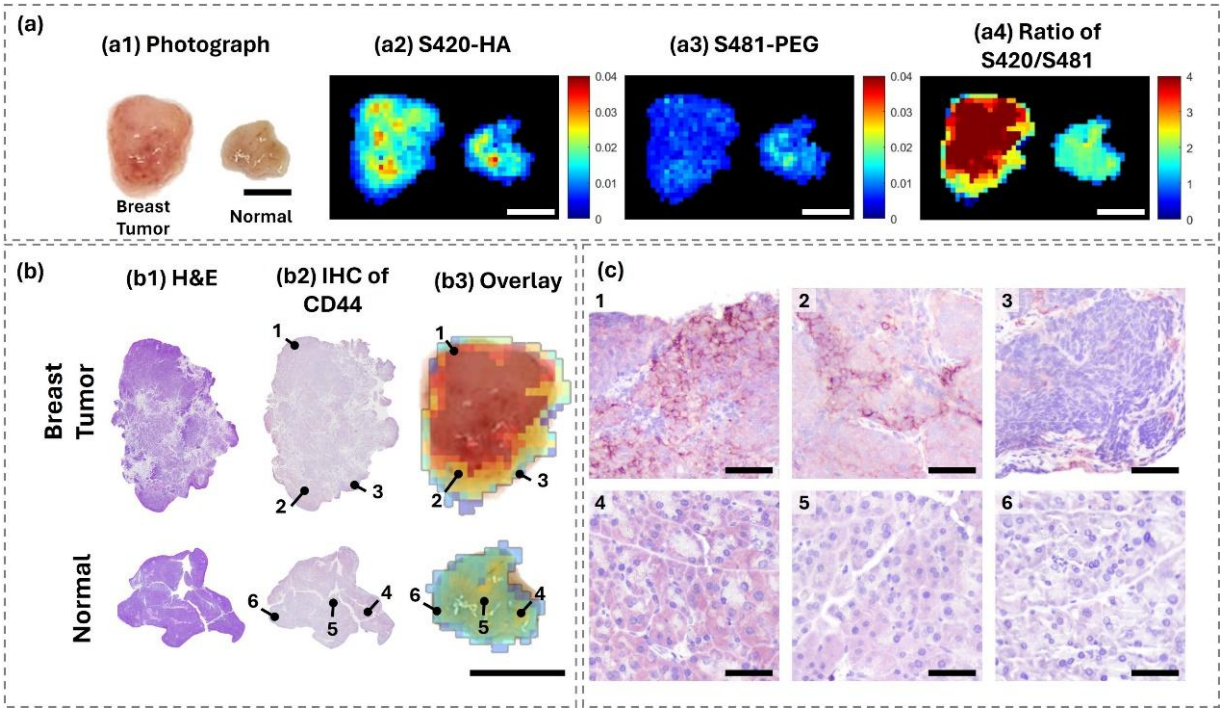


Figure 27. (a) Multiplexed Raman images of tissues topically stained with the mixture of SERS-HA (CD44 targeting) and SERS-PEG (control) solution, (a1) Photographs of the mouse tumor tissue and spleen connective tissue (control), and (a2-a4) Raman images of individual channels and ratiometric results. (b) H&E and IHC-CD44 images of the corresponding tissues. (c) Representative enlarged IHC images in (b) of the breast tumor and normal tissues. Scale bars in (a-b) and (c) are 5 mm and 50 μm, respectively.

Ideally, by rinsing tissues after incubation, the non-targeting NPs (S481-PEG) should be removed from the incubated tissues, and the majority of targeting NPs (S420-HA) should remain on the tumor with overexpressed CD44. However, in the practical experiment, we detected signals from both S420-HA and S481-PEG in both the tumor and normal tissues, as shown in Figure 27(a1-a3), due to tissue texture and non-specific binding. Therefore, the Raman ratiometric image of S420-

HA and S481-PEG was applied to evaluate the targeting of the NPs, as shown in Figure 27(a4). According to the ratiometric result, the ratio of targeting NPs (S420-HA) on the tumor tissue is significantly stronger than the ratio on the control tissue, which is encouraging and promising. Furthermore, the H&E and IHC of CD44 of the corresponding tissues were prepared, and the results are shown in Figure 27(b1-b2), respectively. CD44 is labeled as brown areas, and they are intense (overexpressed) in the tumor tissue as shown in Figure 27(c). This is also consistent with the ratiometric result.

5.3.3 Image-guided surgery experiment

In this experiment, we would like to validate the capability of the proposed Raman system and SERS NPs and closely replicate the clinical conditions of human surgery. A 5-month-old female C57BL6 double transgenic mouse with breast cancer was used for this experiment. First, the operative surgery area (tumor area) was defined, followed by acquiring the Raman signal as the background signal. The mouse was then intratumorally injected with the S420-HA solution with a concentration of 500 pM, a volume of 100 μL, and a depth of injection of approximately 2-3 mm. 42 hours after the injection, the mouse was euthanized by using a table-top research anesthesia machine (V300PS-PARKLAND SCIENTIFIC, USA) with 10 lpm of oxygen flow and 1.5% of anesthetic agent vapor in oxygen during the image-guided surgery imaging. The tumor skin was then cut open followed by rinsing the tumor area with PBS 4-5 times and acquiring Raman spectra. After that, the Raman image (weight of S420-HA) of the scanned area was reconstructed and the tumor was also gradually resected following the white boundaries as shown in Figure 28. It is important to note that the deeper the resection is performed, the weaker the signal of SERS NPs is. This is due to the effective working distance of the Raman probe. Therefore, the depth of information on the operative area is essential for providing additional insights and guidance for

more effective surgery, and we also demonstrate the concept of the 3D SERS NPs imaging in the next section.

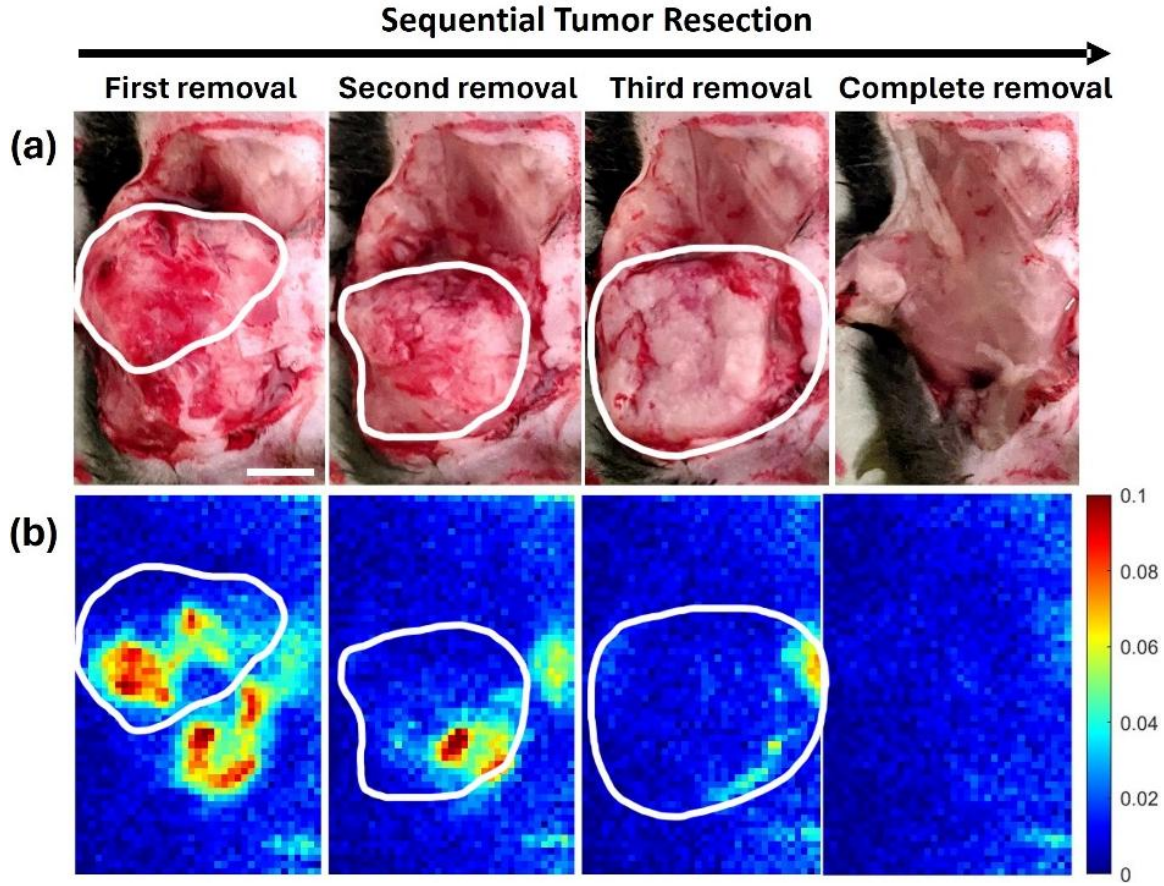


Figure 28. SERS image-guided surgery for resection of a mouse with a breast tumor. (a) Photographs of the tumor during the intraoperative SERS image-guided surgery from the first removal to the complete removal. (b) the corresponding SERS imaging (weight of S420-HA) reconstructed by the demultiplexing algorithm. The scale bar is 5 mm, and the white boundaries depict the resection regions.

5.3.4 2D tracking and 3D SERS imaging

In addition to the image-guided surgery and *ex-vivo* experiments, we demonstrate our custom-made Raman system and monocular depth estimation based on DL to visualize the SERS NPs signal on the sample in 2D and 3D surfaces in the physical world. To simplify the experiment, the S420-HA solution with a concentration of 500 pM was directly dropped on the cut-open tumor of another breast tumor mouse with an incubation time of 15 minutes followed by rinsing with PBS

4-5 times and acquiring Raman spectra, respectively. Before applying this S420-HA solution, the background Raman signal was also acquired as it is one of the input variables for the SERS image reconstruction. A color camera was used to record the video of the scanning area and capture the photograph of the sample to generate the 2D SERS mapping video and the 3D SERS image. To generate the 2D SERS mapping video, the template matching algorithm was applied to track the Raman catheter position to estimate the scanning positions. After that, the SERS signals (the weights of S420-HA) were then generated on these estimated scanning positions as shown in Figure 29 (a). After completing the scanning, the image co-registration algorithm was applied to co-register 2D SERS image with the sample photograph and the MiDaS DL based on BEiT was utilized to generate the depth map image. With these 2D co-registered SERS and depth map images, the 3D-coregiesterd SERS image was reconstructed and projected as point clouds in the 3D space as shown in Figure 29(b). Since Figure 29(a) shows the Raman catheter tracking with real-time 2D SERS image reconstruction, the large field of view (FOV) was needed to acquire the image for covering the catheter and scanning area images. Nevertheless, the smaller FOV was employed to illustrate greater detail in the 3D SERS image shown in Figure 29(b). According to these promising results, the proposed method can facilitate 2D and 3D SERS imaging through the utilization of a Raman catheter system and a simple camera, which can immeasurably improve the visualization and precision of SERS NPs distribution leading to more efficient clinical applications. Specifically, it is beneficial for image-guided surgery by assisting surgeons to locate solid tumors and achieve more precise resections. However, there is an obvious artifact pattern in 3D SERS imaging. It is caused by the large excitation laser (approximately 1 mm). This could be resolved by improving the optic design of the Raman system to reduce the beam size and adding a scanner to maintain the acquisition speed, which could be our future work.

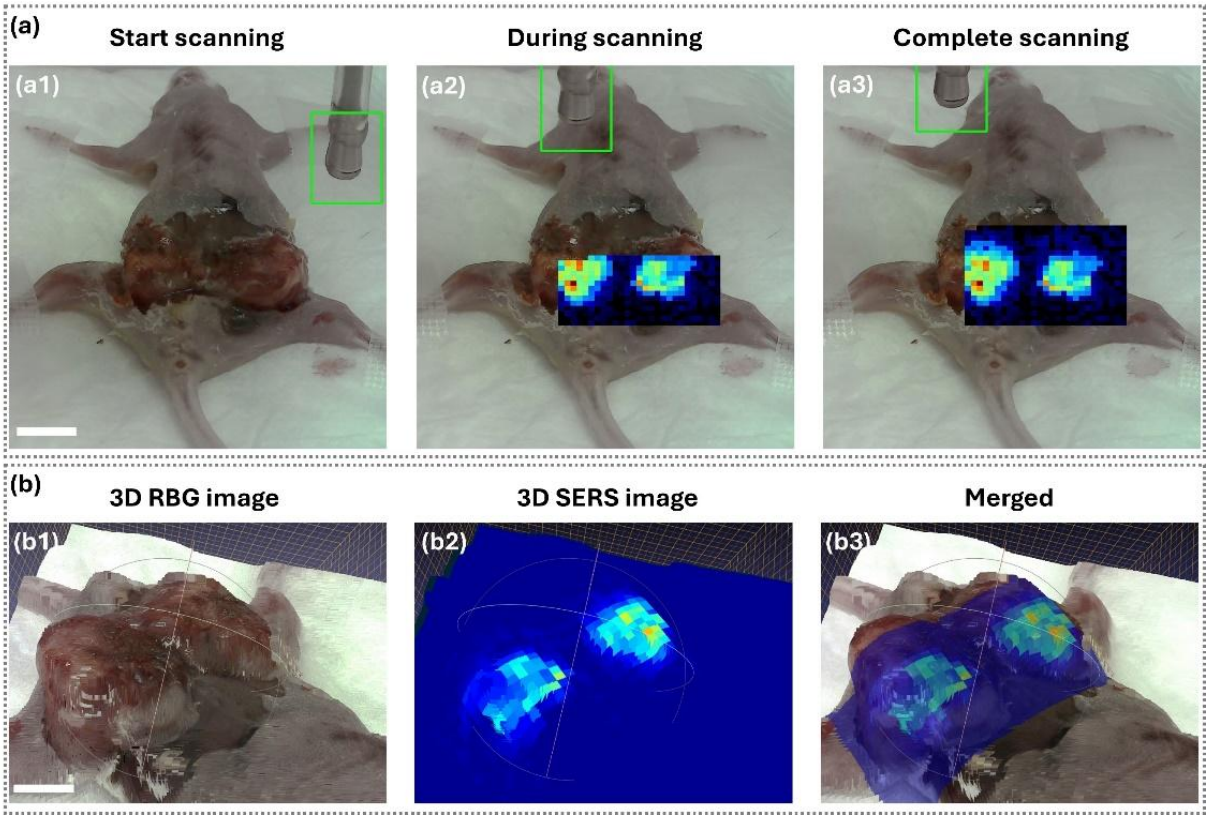


Figure 29. (a) 2D SERS image during Raman spectra acquisition, (a1) before scanning, (a2) during scanning, and (a3) complete scanning. (b) 3D image of the sample, SERS, and co-registered SERS reconstructed by using affine transformation and Midas 3.1 DL model with the BiET backbone architecture. The scale bars of (a1) and (b1) are 10 mm and 8 mm, respectively.

5.4 Conclusion

Intraoperative imaging systems, in tandem with exogenous contrast agents, play a crucial role in tumor resection by assisting a surgeon to identify tumor areas with a high degree of sensitivity and specificity. However, traditional imaging systems commonly encounter poor tumor margin visualization, particularly the weak signal of a tumor at deeper layers. Without depth information, these weak signals might be neglected, leading to ineffective tumor resection. Therefore, the whole tumor might not be completely removed, causing tumor recurrence. In recent years, SERS NPs imaging has been increasingly recognized as an encouraging molecular imaging technique due to its remarkable sensitivity, multiplexing detection capability, and photostability. In addition, it has

demonstrated significant potential in cancer detection and enhancing delineation of tumor margin, as SERS NPs can be easily conjugated with various biomarkers.

In this work, we propose an approach to visualize 2D and 3D SERS imaging. A step-wedge phantom and a tumor phantom were used to evaluate the depth map estimation performance of MiDaS models with three different back-bone architectures: CNN, ViT, and BEiT. MiDas based on BEiT can outperform other models; thus, it was employed for 3D visualization of SERS NPs. HA-conjugated SERS NPs were evaluated by *ex-vivo* and image-guided surgery experiments by using the traditional 2D SERS image reconstruction showing promising results. Nevertheless, it lacks the depth information for practical clinic applications, affecting surgery outcomes. Therefore, the proposed approach combines the use of a custom-made Raman spectrometer with computer vision-based positional tracking for 2D SERS imaging and monocular depth estimation based on the MiDaS model for 3D SERS imaging. This combination can overcome the disadvantage of the conventional Raman system, which only provides spectra information and is unsuitable for clinical applications. The 2D and 3D image co-registration between the Raman imaging and the sample photograph in the physical world enables better performance and efficiency of tumor resection, potentially leading to its implementation in human clinical trials in the near future. Essentially, the proposed method shows a proof-concept study of image-guided surgery by using 3D and 2D SERS imaging. However, there are some limitations that need to be improved in the future, particularly the resolution of SERS imaging. The excitation laser beam diameter in the proposed system is somewhat large (roughly 1 mm), causing the artifact in 3D and 2D image reconstruction, which is unsuitable for small tumor resection. Therefore, the optics part should be re-designed to obtain smaller beam size for enhanced resolution. In addition, the depth map estimation using MiDaS can be influenced by the resolution of an input image acquired at an

out-of-focus distance. Thus, auto-focus approaches, such as resolution enhancement deep learning or a hardware-based approach, should be considered to avoid this problem. The proposed method may be more feasible for future clinical applications as a result of these improvements.

CHAPTER 6: Summary and future work

In this dissertation, a wide range of biomedical applications based on different deep learning techniques have been presented. Firstly, a practical deep learning model for the resolution enhancement of H&E-stained images by using the state-of-the-art SRGAN-ResNeXt network has been demonstrated. The model can deeply learn how to map the low-resolution images to their corresponding high-resolution images. Even though cell images contain sophisticated patterns and structures, the SRGAN-ResNeXt model can still provide high-quality reconstruction results. Moreover, it can outperform the original SRGAN model. Therefore, we take these advantages to characterize and quantify the nuclei from the generated high-resolution images. Secondly, deep learning based on recurrent and convolutional neural networks has been demonstrated for generating sequential NWs-ICG optoacoustic (multispectral unmixing), ultrasound, and optoacoustic images. It has shown robust and promising performance in the accurate reconstruction of the sequential images for all modalities, according to the quantitative evaluation of model performance using the PSNR and SAE for all scanning positions of the generated images (reconstructed by the deep learning model) and ground truth (acquired by mechanical scanning). The architecture of our model is versatile since it can promisingly generate sequential cross-sectional images of three modalities from a commercial MSOT system. Using our deep learning can substantially reduce acquisition time. However, all the training data were acquired from *ex vivo* tissues completely fixed in agarose. Model performance with images acquired *in vivo* may be affected by cardiac and respiratory motion. Thirdly, the proposed multi-head attention U-Net model, an efficient end-to-end deep learning based on U-Net architecture and multi-head attention mechanism, was demonstrated for MPI-CT image segmentation. The proposed model was trained using a custom MPI-CT image dataset collected from transgenic mice with breast tumors injected

with a promising MPI tracer for tumor imaging, namely NWs-ICG. The optimal number of attention heads was experimentally observed in this study. Although an increase in the number of attention heads can potentially boost the model's capability, the excessive number of attention heads results in a decline in capability. Our study shows that the attention U-Net with four heads is the most favorable architecture for MPI-CT image segmentation. Lastly, we propose a method to generate 2D and 3D SERS imaging. The proposed method integrates the use of a custom-made Raman spectrometer with image processing and deep learning to generate 2D and 3D SERS image, which can overcome the drawback of the conventional Raman system, only providing spectra information. The 2D and 3D image co-registration between the Raman imaging and the sample photograph in the physical world enables better performance and efficiency of tumor resection, potentially leading to its implementation in human clinical trials in the near future.

In addition to the applications mentioned above, I am working on virtual H&E images using deep learning. In this work, the virtual H&E deep learning model is employed to transform auto-fluorescence images of unstained tissue slides to virtual H&E images. Another deep learning model is then applied to screening the cancer areas. With this concept, it could potentially shorten the standard cancer diagnosis and be useful for practical clinical applications. Furthermore, in my future work, I plan on developing a universal visual-language foundation deep learning model using a variety of pathology images and biomedical fundamental texts for cancer detection with several downstream tasks related to pathology images to achieve superb performance on pathology image classification, segmentation, and biomarker quantitative.

BIBLIOGRAPHY

1. Juhong, A., Li, B., Yao, C.-Y., Yang, C.-W., Agnew, D. W., Lei, Y. L., Huang, X., Piyawattanametha, W., and Qiu, Z. (2022). Super-resolution and segmentation deep learning for breast cancer histopathology image analysis. Biomedical Optics Express: 14, 18-36.

2. Litjens, G., Sánchez, C. I., Timofeeva, N., Hermsen, M., Nagtegaal, I., Kovacs, I., Hulsbergen-Van De Kaa, C., Bult, P., Van Ginneken, B., and Van Der Laak, J. (2016). Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Scientific reports: 6, 1-11.

3. Mendez, A. J., Tahoces, P. G., Lado, M. a. J., Souto, M., and Vidal, J. J. (1998). Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms. Medical Physics: 25, 957-964.

4. Bogoch, I. I., Koydemir, H. C., Tseng, D., Ephraim, R. K., Duah, E., Tee, J., Andrews, J. R., and Ozcan, A. (2017). Evaluation of a mobile phone-based microscope for screening of Schistosoma haematobium infection in rural Ghana. The American journal of tropical medicine and hygiene: 96, 1468.

5. Petti, C. A., Polage, C. R., Quinn, T. C., Ronald, A. R., and Sande, M. A. (2006). Laboratory medicine in Africa: a barrier to effective health care. Clinical Infectious Diseases: 42, 377-382.

6. Colley, D. G., Bustinduy, A. L., Secor, W. E., and King, C. H. (2014). Human schistosomiasis. The Lancet: 383, 2253-2264.

7. Irshad, H., Veillard, A., Roux, L., and Racoceanu, D. (2013). Methods for nuclei detection, segmentation, and classification in digital histopathology: a review—current status and future potential. IEEE reviews in biomedical engineering: 7, 97-114.

8. Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R., Cree, I. A., and Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. IEEE transactions on medical imaging: 35, 1196-1206.

9. Song, Y., Zhang, L., Chen, S., Ni, D., Lei, B., and Wang, T. (2015). Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning. IEEE Transactions on Biomedical Engineering: 62, 2421-2433.

10. Xing, F., Xie, Y., and Yang, L. (2015). An automatic learning-based framework for robust nucleus segmentation. IEEE transactions on medical imaging: 35, 550-566.

11. Xing, F. and Yang, L. (2016). Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. IEEE reviews in biomedical engineering: 9, 234-263.

12. Otsu, N. (1979). A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics: 9, 62-66.

13. Yang, X., Li, H., and Zhou, X. (2006). Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy. IEEE Transactions on Circuits and Systems I: Regular Papers: 53, 2405-2414.

14. Filipczuk, P., Kowal, M., and Obuchowicz, A. (2011) Automatic breast cancer diagnosis based on k-means clustering and adaptive thresholding hybrid segmentation. Image processing and communications challenges 3 (Springer), pp. 295-302.

15. Graham, S., Vu, Q. D., Raza, S. E. A., Azam, A., Tsang, Y. W., Kwak, J. T., and Rajpoot, N. (2019). Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images. Medical Image Analysis: 58, 101563.

16. Schmidt, U., Weigert, M., Broaddus, C., and Myers, G. (2018). Cell detection with star-convex polygons. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer: 265-273.

17. Chen, S., Ding, C., Liu, M., and Tao, D. (2021). CPP-net: Context-aware polygon proposal network for nucleus segmentation. arXiv preprint arXiv:2102.06867.

18. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, Springer: 234-241.

19. de Haan, K., Zhang, Y., Liu, T., Sisk, A. E., Diaz, M. F., Zuckerman, J. E., Rivenson, Y., Wallace, W. D., and Ozcan, A. (2020). Deep learning-based transformation of the H&E stain into special stains improves kidney disease diagnosis. arXiv preprint arXiv:2008.08871.

20. Liu, T., De Haan, K., Rivenson, Y., Wei, Z., Zeng, X., Zhang, Y., and Ozcan, A. (2019). Deep learning-based super-resolution in coherent imaging systems. Scientific reports: 9, 1-13.

21. Mukherjee, L., Keikhosravi, A., Bui, D., and Eliceiri, K. W. (2018). Convolutional neural networks for whole slide image superresolution. Biomedical optics express: 9, 5368-5386.

22. Rivenson, Y., Göröcs, Z., Günaydin, H., Zhang, Y., Wang, H., and Ozcan, A. (2017). Deep learning microscopy. Optica: 4, 1437-1443.

23. Wang, H., Rivenson, Y., Jin, Y., Wei, Z., Gao, R., Günaydın, H., Bentolila, L. A., Kural, C., and Ozcan, A. (2019). Deep learning enables cross-modality super-resolution in fluorescence microscopy. Nature methods: 16, 103-110.

24. Zhang, H., Fang, C., Xie, X., Yang, Y., Mei, W., Jin, D., and Fei, P. (2019). High-throughput, high-resolution deep learning microscopy based on registration-free generative adversarial network. Biomedical optics express: 10, 1044-1063.

25. Zheng, T., Oda, H., Moriya, T., Sugino, T., Nakamura, S., Oda, M., Mori, M., Takabatake, H., Natori, H., and Mori, K. (2020). Multi-modality super-resolution loss for GAN-based super-resolution of clinical CT images using micro CT image database. In *Medical Imaging 2020: Image Processing*, International Society for Optics and Photonics: 1131305.

26. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., and Wang, Z. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681-4690.

27. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

28. Iandola, F., Moskewicz, M., Karayev, S., Girshick, R., Darrell, T., and Keutzer, K. (2014). Densenet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869.

29. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., and Change Loy, C. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 0-0.

30. Bianco, S., Cadene, R., Celona, L., and Napoletano, P. (2018). Benchmark analysis of representative deep neural network architectures. IEEE access: 6, 64270-64277.

31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems: 30.

32. Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492-1500.

33. Delibasoglu, I. and Cetin, M. (2020). Improved U-Nets with inception blocks for building detection. Journal of Applied Remote Sensing: 14, 044512.

34. Hou, L., Gupta, R., Van Arnam, J. S., Zhang, Y., Sivalenka, K., Samaras, D., Kurc, T. M., and Saltz, J. H. (2020). Dataset of segmented nuclei in hematoxylin and eosin stained histopathology images of ten cancer types. Scientific data: 7, 1-12.

35. Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

36. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., and Wang, Z. (2016). Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1874-1883.

37. Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

38. Stergiou, N., Gaidzik, N., Heimes, A.-S., Dietzen, S., Besenius, P., Jäkel, J., Brenner, W., Schmidt, M., Kunz, H., and Schmitt, E. (2019). Reduced breast tumor growth after immunization with a tumor-restricted MUC1 glycopeptide conjugated to tetanus toxoid. Cancer Immunology Research: 7, 113-122.

39. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1-9.

40. Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing: 13, 600-612.

41. Chang, H.-H., Zhuang, A. H., Valentino, D. J., and Chu, W.-C. (2009). Performance measure characterization for evaluating neuroimage segmentation algorithms. Neuroimage: 47, 122-135.

42. Juhong, A., Li, B., Liu, Y., Yao, C. Y., Yang, C. W., Agnew, D. W., Lei, Y. L., Luker, G. D., Bumpers, H., and Huang, X. (2023). Recurrent and convolutional neural networks for sequential multispectral optoacoustic tomography (MSOT) imaging. Journal of Biophotonics: 16, e202300142.

43. Ntziachristos, V. and Razansky, D. (2010). Molecular imaging by means of multispectral optoacoustic tomography (MSOT). Chemical reviews: 110, 2783-2794.

44. Wang, L. V. and Hu, S. (2012). Photoacoustic tomography: in vivo imaging from organelles to organs. science: 335, 1458-1462.

45. Buehler, A., Kacprowicz, M., Taruttis, A., and Ntziachristos, V. (2013). Real-time handheld multispectral optoacoustic imaging. Optics letters: 38, 1404-1406.

46. Dima, A. and Ntziachristos, V. (2016). In-vivo handheld optoacoustic tomography of the human thyroid. Photoacoustics: 4, 65-69.

47. Tam, A. C. (1986). Applications of photoacoustic sensing techniques. Reviews of Modern Physics: 58, 381.

48. Razansky, D., Distel, M., Vinegoni, C., Ma, R., Perrimon, N., Köster, R. W., and Ntziachristos, V. (2009). Multispectral opto-acoustic tomography of deep-seated fluorescent proteins in vivo. Nature photonics: 3, 412-417.

49. Tzoumas, S., Deliolanis, N. C., Morscher, S., and Ntziachristos, V. (2013). Unmixing molecular agents from absorbing tissue in multispectral optoacoustic tomography. IEEE transactions on medical imaging: 33, 48-60.

50. Diot, G., Metz, S., Noske, A., Liapis, E., Schroeder, B., Ovsepian, S. V., Meier, R., Rummeny, E., and Ntziachristos, V. (2017). Multispectral optoacoustic tomography (MSOT) of human breast cancer. Clinical Cancer Research: 23, 6912-6922.

51. Quiros-Gonzalez, I., Tomaszewski, M. R., Aitken, S. J., Ansel-Bollepalli, L., McDuffus, L.-A., Gill, M., Hacker, L., Brunker, J., and Bohndiek, S. E. (2018). Optoacoustics delineates murine breast cancer models displaying angiogenesis and vascular mimicry. British journal of cancer: 118, 1098-1106.

52. Ron, A., Deán-Ben, X. L., Gottschalk, S., and Razansky, D. (2019). Volumetric optoacoustic imaging unveils high-resolution patterns of acute and cyclic hypoxia in a murine model of breast cancer. Cancer research: 79, 4767-4775.

53. Taruttis, A., van Dam, G. M., and Ntziachristos, V. (2015). Mesoscopic and macroscopic optoacoustic imaging of cancer. Cancer research: 75, 1548-1559.

54. Tomaszewski, M. R., Gehrung, M., Joseph, J., Quiros-Gonzalez, I., Disselhorst, J. A., and Bohndiek, S. E. (2018). Oxygen-enhanced and dynamic contrast-enhanced optoacoustic tomography provide surrogate biomarkers of tumor vascular function, hypoxia, and necrosis. Cancer research: 78, 5980-5991.

55. Regensburger, A. P., Fonteyne, L. M., Jüngert, J., Wagner, A. L., Gerhalter, T., Nagel, A. M., Heiss, R., Flenkenthaler, F., Qurashi, M., and Neurath, M. F. (2019). Detection of collagens by multispectral optoacoustic tomography as an imaging biomarker for Duchenne muscular dystrophy. Nature medicine: 25, 1905-1915.

56. Song, W., Tang, Z., Zhang, D., Burton, N., Driessen, W., and Chen, X. (2015). Comprehensive studies of pharmacokinetics and biodistribution of indocyanine green and liposomal indocyanine green by multispectral optoacoustic tomography. RSC advances: 5, 3807-3813.

57. Anani, T., Brannen, A., Panizzi, P., Duin, E. C., and David, A. E. (2020). Quantitative, real-time in vivo tracking of magnetic nanoparticles using multispectral optoacoustic tomography (MSOT) imaging. Journal of pharmaceutical and biomedical analysis: 178, 112951.

58. Gurka, M. K., Pender, D., Chuong, P., Fouts, B. L., Sobelov, A., McNally, M. W., Mezera, M., Woo, S. Y., and McNally, L. R. (2016). Identification of pancreatic tumors in vivo with ligand-targeted, pH responsive mesoporous silica nanoparticles by multispectral optoacoustic tomography. Journal of controlled release: 231, 60-67.

59. Li, D., Zhang, G., Xu, W., Wang, J., Wang, Y., Qiu, L., Ding, J., and Yang, X. (2017). Investigating the effect of chemical structure of semiconducting polymer nanoparticle on photothermal therapy and photoacoustic imaging. Theranostics: 7, 4029.

60. Wang, S., Zhang, L., Zhao, J., He, M., Huang, Y., and Zhao, S. (2021). A tumor microenvironment–induced absorption red-shifted polymer nanoparticle for simultaneously activated photoacoustic imaging and photothermal therapy. Science Advances: 7, eabe3588.

61. Gröhl, J., Schellenberg, M., Dreher, K., Holzwarth, N., Tizabi, M. D., Seitel, A., and Maier-Hein, L. (2021). Semantic segmentation of multispectral photoacoustic images using deep learning. arXiv preprint arXiv:2105.09624.

62. Yuan, A. Y., Gao, Y., Peng, L., Zhou, L., Liu, J., Zhu, S., and Song, W. (2020). Hybrid deep learning network for vascular segmentation in photoacoustic imaging. Biomedical Optics Express: 11, 6445-6457.

63. Luke, G. P., Hoffer-Hawlik, K., Van Namen, A. C., and Shang, R. (2019). O-Net: a convolutional neural network for quantitative photoacoustic image segmentation and oximetry. arXiv preprint arXiv:1911.01935.

64. Lan, H., Jiang, D., Yang, C., and Gao, F. (2019). Y-Net: a hybrid deep learning reconstruction framework for photoacoustic imaging in vivo. arXiv preprint arXiv:1908.00975.

65. Zhang, J., Chen, B., Zhou, M., Lan, H., and Gao, F. (2018). Photoacoustic image classification and segmentation of breast cancer: a feasibility study. IEEE Access: 7, 5457-5466.

66. Chen, T., Lu, T., Song, S., Miao, S., Gao, F., and Li, J. (2020). A deep learning method based on U-Net for quantitative photoacoustic imaging. In *Photons Plus Ultrasound: Imaging and Sensing 2020*, International Society for Optics and Photonics: 112403V.

67. Bench, C., Hauptmann, A., and Cox, B. T. (2020). Toward accurate quantitative photoacoustic imaging: learning vascular blood oxygen saturation in three dimensions. Journal of Biomedical Optics: 25, 085003.

68. Yang, C., Lan, H., Zhong, H., and Gao, F. (2019). Quantitative photoacoustic blood oxygenation imaging using deep residual and recurrent neural network. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, IEEE: 741-744.

69. Gröhl, J., Kirchner, T., Adler, T., and Maier-Hein, L. (2019). Estimation of blood oxygenation with learned spectral decoloring for quantitative photoacoustic imaging (LSD-qPAI). arXiv preprint arXiv:1902.05839.

70. Cai, C., Deng, K., Ma, C., and Luo, J. (2018). End-to-end deep neural network for optical inversion in quantitative photoacoustic imaging. Optics letters: 43, 2752-2755.

71. Allman, D., Reiter, A., and Bell, M. A. L. (2018). Photoacoustic source detection and reflection artifact removal enabled by deep learning. IEEE transactions on medical imaging: 37, 1464-1477.

72. Davoudi, N., Deán-Ben, X. L., and Razansky, D. (2019). Deep learning optoacoustic tomography with sparse data. Nature Machine Intelligence: 1, 453-460.

73. Hariri, A., Alipour, K., Mantri, Y., Schulze, J. P., and Jokerst, J. V. (2020). Deep learning improves contrast in low-fluence photoacoustic imaging. Biomedical optics express: 11, 3360-3373.

74. Lu, T., Chen, T., Gao, F., Sun, B., Ntziachristos, V., and Li, J. (2021). LV-GAN: A deep learning approach for limited-view optoacoustic imaging based on hybrid datasets. Journal of biophotonics: 14, e202000325.

75. Sivasubramanian, K. and Xing, L. (2020). Deep learning for image processing and reconstruction to enhance led-based photoacoustic imaging. LED-Based Photoacoustic Imaging: From Bench to Bedside, 203-241.

76. Lafci, B., Merčep, E., Morscher, S., Deán-Ben, X. L., and Razansky, D. (2020). Deep learning for automatic segmentation of hybrid optoacoustic ultrasound (OPUS) images. IEEE transactions on ultrasonics, ferroelectrics, and frequency control: 68, 688-696.

77. Aydın, M., Kiraz, B., Eren, F., Uysallı, Y., Morova, B., Ozcan, S. C., Acilan, C., and Kiraz, A. (2022). A Deep Learning Model for Automated Segmentation of Fluorescence Cell images. In *Journal of Physics: Conference Series*, IOP Publishing: 012003.

78. de Haan, K., Ceylan Koydemir, H., Rivenson, Y., Tseng, D., Van Dyne, E., Bakic, L., Karinca, D., Liang, K., Ilango, M., and Gumustekin, E. (2020). Automated screening of sickle cells using a smartphone-based microscope and deep learning. NPJ digital medicine: 3, 76.

79. Ibtehaz, N. and Rahman, M. S. (2020). MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation. Neural networks: 121, 74-87.

80. Punn, N. S. and Agarwal, S. (2022). Modality specific U-Net variants for biomedical image segmentation: a survey. Artificial Intelligence Review: 55, 5845-5889.

81. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818-2826.

82. Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural computation: 9, 1735-1780.

83. Xu, M. and Wang, L. V. (2005). Universal back-projection algorithm for photoacoustic computed tomography. Physical Review E: 71, 016706.

84. Stergiou, N., Gaidzik, N., Heimes, A.-S., Dietzen, S., Besenius, P., Jäkel, J., Brenner, W., Schmidt, M., Kunz, H., and Schmitt, E. (2019). Reduced Breast Tumor Growth after Immunization with a Tumor-Restricted MUC1 Glycopeptide Conjugated to Tetanus ToxoidImmunization against Tumor-Restricted MUC1 in Breast Cancer. Cancer Immunology Research: 7, 113-122.

85. Yang, C.-W., Liu, K., Yao, C.-Y., Li, B., Juhong, A., Qiu, Z., and Huang, X. (2022). Indocyanine Green-Conjugated Superparamagnetic Iron Oxide Nanoworm for Multimodality Breast Cancer Imaging. ACS Applied Nano Materials: 5, 18912-18920.

86. Greish, K. (2010). Enhanced permeability and retention (EPR) effect for anticancer nanomedicine drug targeting. Cancer nanotechnology: Methods and protocols, 25-37.

87. Keshava, N. and Mustard, J. F. (2002). Spectral unmixing. IEEE signal processing magazine: 19, 44-57.

88. Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, 802-810.

89. Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. IEEE signal processing magazine: 35, 53-65.

90. Juhong, A., Li, B., Liu, Y., Yang, C. W., Yao, C. Y., Agnew, D. W., Lei, Y. L., Luker, G. D., Bumpers, H., and Huang, X. (2024). Multihead Attention U-Net for Magnetic Particle Imaging–Computed Tomography Image Segmentation. Advanced Intelligent Systems: 6, 2400007.

91. Bulte, J. W. (2019). Superparamagnetic iron oxides as MPI tracers: A primer and review of early applications. Advanced drug delivery reviews: 138, 293-301.

92. Gleich, B. and Weizenecker, J. (2005). Tomographic imaging using the nonlinear response of magnetic particles. Nature: 435, 1214-1217.

93. Scarfe, L., Brillant, N., Kumar, J. D., Ali, N., Alrumayh, A., Amali, M., Barbellion, S., Jones, V., Niemeijer, M., and Potdevin, S. (2017). Preclinical imaging methods for assessing the safety and efficacy of regenerative medicine therapies. NPJ Regenerative medicine: 2, 28.

94. Zheng, B., Vazin, T., Goodwill, P. W., Conway, A., Verma, A., Ulku Saritas, E., Schaffer, D., and Conolly, S. M. (2015). Magnetic particle imaging tracks the long-term fate of in vivo neural cell implants with high image contrast. Scientific reports: 5, 14055.

95. Rahmer, J., Gleich, B., Weizenecker, J., and Borgert, J. (2010). 3D real-time magnetic particle imaging of cerebral blood flow in living mice. In *Proceedings of the International Society for Magnetic Resonance in Medicine*, 714.

96. Ludewig, P., Gdaniec, N., Sedlacik, J., Forkert, N. D., Szwargulski, P., Graeser, M., Adam, G., Kaul, M. G., Krishnan, K. M., and Ferguson, R. M. (2017). Magnetic particle imaging for real-time perfusion imaging in acute stroke. ACS nano: 11, 10480-10488.

97. Orendorff, R., Keselman, K., and Conolly, S. (2018). Quantitative cerebral blood flow and volume measurements by magnetic particle imaging. In *13th European Molecular Imaging Meeting*, 20-23.

98. Fu, A., Wilson, R. J., Smith, B. R., Mullenix, J., Earhart, C., Akin, D., Guccione, S., Wang, S. X., and Gambhir, S. S. (2012). Fluorescent magnetic nanoparticles for magnetically enhanced cancer imaging and targeting in living subjects. ACS nano: 6, 6862-6869.

99. Tomitaka, A., Arami, H., Gandhi, S., and Krishnan, K. M. (2015). Lactoferrin conjugated iron oxide nanoparticles for targeting brain glioma cells in magnetic particle imaging. Nanoscale: 7, 16890-16898.

100. Finas, D., Baumann, K., Sydow, L., Heinrich, K., Gräfe, K., Rody, A., Lüdtke-Buzug, K., and Buzug, T. (2013). Lymphatic tissue and superparamagnetic nanoparticles-magnetic particle imaging for detection and distribution in a breast cancer model. Biomedical Engineering/Biomedizinische Technik: 58, 000010151520134262.

101. Song, G., Chen, M., Zhang, Y., Cui, L., Qu, H., Zheng, X., Wintermark, M., Liu, Z., and Rao, J. (2018). Janus iron oxides@ semiconducting polymer nanoparticle tracer for cell tracking by magnetic particle imaging. Nano letters: 18, 182-189.

102. Zheng, B., von See, M. P., Yu, E., Gunel, B., Lu, K., Vazin, T., Schaffer, D. V., Goodwill, P. W., and Conolly, S. M. (2016). Quantitative magnetic particle imaging monitors the transplantation, biodistribution, and clearance of stem cells in vivo. Theranostics: 6, 291.

103. Wu, L. C., Zhang, Y., Steinberg, G., Qu, H., Huang, S., Cheng, M., Bliss, T., Du, F., Rao, J., and Song, G. (2019). A review of magnetic particle imaging and perspectives on neuroimaging. American Journal of Neuroradiology: 40, 206-212.

104. Herz, S., Vogel, P., Dietrich, P., Kampf, T., Rückert, M. A., Kickuth, R., Behr, V. C., and Bley, T. A. (2018). Magnetic particle imaging guided real-time percutaneous transluminal angioplasty in a phantom model. Cardiovascular and interventional radiology: 41, 1100-1105.

105. Hossaini Nasr, S., Tonson, A., El-Dakdouki, M. H., Zhu, D. C., Agnew, D., Wiseman, R., Qian, C., and Huang, X. (2018). Effects of nanoprobe morphology on cellular binding and inflammatory responses: hyaluronan-conjugated magnetic nanoworms for magnetic resonance imaging of atherosclerotic plaques. ACS applied materials & interfaces: 10, 11495-11507.

106. Park, J. H., von Maltzahn, G., Zhang, L., Schwartz, M. P., Ruoslahti, E., Bhatia, S. N., and Sailor, M. J. (2008). Magnetic iron oxide nanoworms for tumor targeting and imaging. Advanced materials: 20, 1630-1635.

107. Iyer, A. K., Khaled, G., Fang, J., and Maeda, H. (2006). Exploiting the enhanced permeability and retention effect for tumor targeting. Drug discovery today: 11, 812-818.

108. Kobayashi, H., Watanabe, R., and Choyke, P. L. (2014). Improving conventional enhanced permeability and retention (EPR) effects; what is the appropriate target? Theranostics: 4, 81.

109. Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D. D., and Chen, M. (2014). Medical image classification with convolutional neural network. In *2014 13th international conference on control automation robotics & vision (ICARCV)*, IEEE: 844-848.

110. Liu, Q., Yu, L., Luo, L., Dou, Q., and Heng, P. A. (2020). Semi-supervised medical image classification with relation-driven self-ensembling model. IEEE transactions on medical imaging: 39, 3429-3440.

111. Deepa, S. and Devi, B. A. (2011). A survey on artificial intelligence approaches for medical image classification. Indian Journal of Science and Technology: 4, 1583-1595.

112. Goldstein, B. A., Navar, A. M., and Carter, R. E. (2017). Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges. European heart journal: 38, 1805-1814.

113. Maulud, D. and Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends: 1, 140-147.

114. Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., and Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of clinical epidemiology: 110, 12-22.

115. Zhang, S., Liang, G., Pan, S., and Zheng, L. (2018). A fast medical image super resolution method based on deep learning network. IEEE Access: 7, 12319-12327.

116. Wang, G., Ye, J. C., Mueller, K., and Fessler, J. A. (2018). Image reconstruction is a new frontier of machine learning. IEEE transactions on medical imaging: 37, 1289-1296.

117. Lundervold, A. S. and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Zeitschrift für Medizinische Physik: 29, 102-127.

118. Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. Journal of digital imaging: 32, 582-596.

119. Tajbakhsh, N., Jeyaseelan, L., Li, Q., Chiang, J. N., Wu, Z., and Ding, X. (2020). Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. Medical Image Analysis: 63, 101693.

120. Maier, A., Syben, C., Lasser, T., and Riess, C. (2019). A gentle introduction to deep learning in medical image processing. Zeitschrift für Medizinische Physik: 29, 86-101.

121. Wang, R., Lei, T., Cui, R., Zhang, B., Meng, H., and Nandi, A. K. (2022). Medical image segmentation using deep learning: A survey. IET Image Processing: 16, 1243-1267.

122. Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. Neurocomputing: 452, 48-62.

123. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

124. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

125. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012-10022.

126. Luong, M.-T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025.

127. Chorowski, J. K., Bahdanau, D., Serdyuk, D., Cho, K., and Bengio, Y. (2015). Attention-based models for speech recognition. Advances in neural information processing systems: 28.

128. Britz, D., Goldie, A., Luong, M.-T., and Le, Q. (2017). Massive exploration of neural machine translation architectures. arXiv preprint arXiv:1703.03906.

129. Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer: 234-241.

130. Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., and Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis: 53, 197-207.

131. Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., and Kainz, B. (2018). Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999.

132. Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

133. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618-626.

134. Juhong, A., Li, B., Liu, Y., Yao, C.-Y., Yang, C.-W., Atique Ullah, A., Liu, K., Lewandowski, R. P., Harkema, J. R., and Agnew, D. W. (2025). Monocular depth estimation based on deep learning for intraoperative guidance using surface-enhanced Raman scattering imaging. Photonics Research: 13, 550-560.

135. Lukianova-Hleb, E. Y., Kim, Y.-S., Belatsarkouski, I., Gillenwater, A. M., O'Neill, B. E., and Lapotko, D. O. (2016). Intraoperative diagnostics and elimination of residual microtumours with plasmonic nanobubbles. Nature Nanotechnology: 11, 525-532.

136. Wang, T., Wang, D., Yu, H., Feng, B., Zhou, F., Zhang, H., Zhou, L., Jiao, S., and Li, Y. (2018). A cancer vaccine-mediated postoperative immunotherapy for recurrent and metastatic tumors. Nature communications: 9, 1532.

137. Anup, N., Gadeval, A., and Tekade, R. K. (2023). A 3D-printed graphene BioFuse implant for postsurgical adjuvant therapy of cancer: proof of concept in 2D-and 3D-spheroid tumor models. ACS Applied Bio Materials: 6, 1195-1212.

138. Aydın, H., Sillenberg, I., and von Lieven, H. (2001). Patterns of failure following CT-based 3-D irradiation for malignant glioma. Strahlentherapie und Onkologie: 177, 424-431.

139. Gao, R. W., Teraphongphom, N. T., van den Berg, N. S., Martin, B. A., Oberhelman, N. J., Divi, V., Kaplan, M. J., Hong, S. S., Lu, G., and Ertsey, R. (2018). Determination of tumor margins with surgical specimen mapping using near-infrared fluorescence. Cancer research: 78, 5144-5154.

140. Gao, X., Yue, Q., Liu, Z., Ke, M., Zhou, X., Li, S., Zhang, J., Zhang, R., Chen, L., and Mao, Y. (2017). Guiding brain-tumor surgery via blood–brain-barrier-permeable gold nanoprobes with acid-triggered MRI/SERRS signals. Advanced Materials: 29, 1603917.

141. Kunjachan, S., Ehling, J., Storm, G., Kiessling, F., and Lammers, T. (2015). Noninvasive imaging of nanomedicines and nanotheranostics: principles, progress, and prospects. Chemical reviews: 115, 10907-10937.

142. Kircher, M. F., Mahmood, U., King, R. S., Weissleder, R., and Josephson, L. (2003). A multimodal nanoparticle for preoperative magnetic resonance imaging and intraoperative optical brain tumor delineation. Cancer research: 63, 8122-8125.

143. Pal, S., Ray, A., Andreou, C., Zhou, Y., Rakshit, T., Wlodarczyk, M., Maeda, M., Toledo-Crow, R., Berisha, N., and Yang, J. (2019). DNA-enabled rational design of fluorescence-Raman bimodal nanoprobes for cancer imaging and therapy. Nature communications: 10, 1926.

144. Qi, J., Li, J., Liu, R., Li, Q., Zhang, H., Lam, J. W., Kwok, R. T., Liu, D., Ding, D., and Tang, B. Z. (2019). Boosting fluorescence-photoacoustic-Raman properties in one fluorophore for precise cancer surgery. Chem: 5, 2657-2677.

145. Zysk, A. M., Chen, K., Gabrielson, E., Tafra, L., May Gonzalez, E. A., Canner, J. K., Schneider, E. B., Cittadine, A. J., Scott Carney, P., and Boppart, S. A. (2015). Intraoperative assessment of final margins with a handheld optical imaging probe during breast-conserving surgery may reduce the reoperation rate: results of a multicenter study. Annals of surgical oncology: 22, 3356-3362.

146. Laing, S., Jamieson, L. E., Faulds, K., and Graham, D. (2017). Surface-enhanced Raman spectroscopy for in vivo biosensing. Nature Reviews Chemistry: 1, 0060.

147. Langer, J., Jimenez de Aberasturi, D., Aizpurua, J., Alvarez-Puebla, R. A., Auguié, B., Baumberg, J. J., Bazan, G. C., Bell, S. E., Boisen, A., and Brolo, A. G. (2019). Present and future of surface-enhanced Raman scattering. ACS nano: 14, 28-117.

148. Li, M., Cushing, S. K., and Wu, N. (2015). Plasmon-enhanced optical sensors: a review. Analyst: 140, 386-406.

149. Li, D., Hui, H., Zhang, Y., Tong, W., Tian, F., Yang, X., Liu, J., Chen, Y., and Tian, J. (2020). Deep learning for virtual histological staining of bright-field microscopic images of unlabeled carotid artery tissue. Molecular imaging and biology: 22, 1301-1309.

150. Pan, X., Li, L., Lin, H., Tan, J., Wang, H., Liao, M., Chen, C., Shan, B., Chen, Y., and Li, M. (2019). A graphene oxide-gold nanostar hybrid based-paper biosensor for label-free SERS detection of serum bilirubin for diagnosis of jaundice. Biosensors and Bioelectronics: 145, 111713.

151. Shan, B., Pu, Y., Chen, Y., Liao, M., and Li, M. (2018). Novel SERS labels: Rational design, functional integration and biomedical applications. Coordination Chemistry Reviews: 371, 11-37.

152. Wang, Y., Kang, S., Khan, A., Ruttner, G., Leigh, S. Y., Murray, M., Abeytunge, S., Peterson, G., Rajadhyaksha, M., and Dintzis, S. (2016). Quantitative molecular phenotyping with topically applied SERS nanoparticles for intraoperative guidance of breast cancer lumpectomy. Scientific reports: 6, 21242.

153. Liang, A., Liu, Q., Wen, G., and Jiang, Z. (2012). The surface-plasmon-resonance effect of nanogold/silver and its analytical applications. TrAC Trends in Analytical Chemistry: 37, 32-47.

154. Davis, R. M., Campbell, J. L., Burkitt, S., Qiu, Z., Kang, S., Mehraein, M., Miyasato, D., Salinas, H., Liu, J. T., and Zavaleta, C. (2018). A raman imaging approach using CD47 antibody-labeled SERS nanoparticles for identifying breast cancer and its potential to guide surgical resection. Nanomaterials: 8, 953.

155. Gao, H. (2016). Progress and perspectives on targeting nanoparticles for brain drug delivery. Acta Pharmaceutica Sinica B: 6, 268-286.

156. Huang, R., Harmsen, S., Samii, J. M., Karabeber, H., Pitter, K. L., Holland, E. C., and Kircher, M. F. (2016). High precision imaging of microscopic spread of glioblastoma with a targeted ultrasensitive SERRS molecular imaging probe. Theranostics: 6, 1075.

157. Liu, K., Ullah, A. A., Juhong, A., Yang, C. W., Yao, C. Y., Li, X., Bumpers, H. L., Qiu, Z., and Huang, X. (2024). Robust Synthesis of Targeting Glyco-Nanoparticles for Surface Enhanced Resonance Raman Based Image-Guided Tumor Surgery. Small Science, 2300154.

158. Zavaleta, C. L., Smith, B. R., Walton, I., Doering, W., Davis, G., Shojaei, B., Natan, M. J., and Gambhir, S. S. (2009). Multiplexed imaging of surface enhanced Raman scattering nanotags in living mice using noninvasive Raman spectroscopy. Proceedings of the National Academy of Sciences: 106, 13511-13516.

159. Brunelli, R. (2009). Template matching techniques in computer vision: theory and practice. (John Wiley & Sons).

160. Mikolajczyk, K. and Schmid, C. (2004). Scale & affine invariant interest point detectors. International journal of computer vision: 60, 63-86.

161. Garai, E., Sensarn, S., Zavaleta, C. L., Van de Sompel, D., Loewke, N. O., Mandella, M. J., Gambhir, S. S., and Contag, C. H. (2013). High-sensitivity, real-time, ratiometric imaging of surface-enhanced Raman scattering nanoparticles with a clinically translatable Raman endoscope device. Journal of biomedical optics: 18, 096008-096008.

162. Zavaleta, C. L., Garai, E., Liu, J. T., Sensarn, S., Mandella, M. J., Van de Sompel, D., Friedland, S., Van Dam, J., Contag, C. H., and Gambhir, S. S. (2013). A Raman-based endoscopic strategy for multiplexed molecular imaging. Proceedings of the National Academy of Sciences: 110, E2288-E2297.

163. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE transactions on pattern analysis and machine intelligence: 44, 1623-1637.

164. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., and Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

165. Ranftl, R., Bochkovskiy, A., and Koltun, V. (2021). Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12179-12188.

166. Birkl, R., Wofk, D., and Müller, M. (2023). Midas v3. 1--a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460.

167. Gotov, O., Battogtokh, G., Shin, D., and Ko, Y. T. (2018). Hyaluronic acid-coated cisplatin conjugated gold nanoparticles for combined cancer treatment. Journal of industrial and engineering chemistry: 65, 236-243.

168. Lee, H., Lee, K., Kim, I. K., and Park, T. G. (2008). Synthesis, characterization, and in vivo diagnostic applications of hyaluronic acid immobilized gold nanoprobes. Biomaterials: 29, 4709-4718.

169. Lee, M.-Y., Yang, J.-A., Jung, H. S., Beack, S., Choi, J. E., Hur, W., Koo, H., Kim, K., Yoon, S. K., and Hahn, S. K. (2012). Hyaluronic acid–gold nanoparticle/interferon α complex for targeted treatment of hepatitis C virus infection. ACS nano: 6, 9522-9531.

170. Li, X., Zhou, H., Yang, L., Du, G., Pai-Panandiker, A. S., Huang, X., and Yan, B. (2011). Enhancement of cell recognition in vitro by dual-ligand cancer targeting gold nanoparticles. Biomaterials: 32, 2540-2545.

171. Xue, Y., Li, X., Li, H., and Zhang, W. (2014). Quantifying thiol–gold interactions towards the efficient strength control. Nature communications: 5, 1-9.

172. Juhong, A., Li, B., Yao, C.-Y., Yang, C.-W., Liu, K., Agnew, D. W., Lei, Y. L., Luker, G. D., Bumpers, H., and Huang, X. (2023). Cost-Effective Near Infrared Fluorescence Wide-Field Camera for Breast Tumor Imaging. IEEE Photonics Technology Letters: 35, 813-816.

173. Zhang, Z. (2000). A flexible new technique for camera calibration. IEEE Transactions on pattern analysis and machine intelligence: 22, 1330-1334.

174. Jurie, F. and Dhome, M. (2001). A simple and efficient template matching algorithm. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, IEEE: 544-549.

175. Bradski, G. and Kaehler, A. (2000). OpenCV. Dr. Dobb's journal of software tools: 3.

176. Sener, O. and Koltun, V. (2018). Multi-task learning as multi-objective optimization. Advances in neural information processing systems: 31.

177. Diederik, P. K. (2014). Adam: A method for stochastic optimization.

178. Li, Z. and Snavely, N. (2018). Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2041-2050.

179. Bao, H., Dong, L., Piao, S., and Wei, F. (2021). Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254.

180. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. (2021). Zero-shot text-to-image generation. In *International conference on machine learning*, Pmlr: 8821-8831.

181. Rolfe, J. T. (2016). Discrete variational autoencoders. arXiv preprint arXiv:1609.02200.