To be published in Biomedical Optics Express:

OPTICA

PUBLISHING GROUP

Formerly OSA

Super-resolution and Segmentation Deep Learning for Breast Cancer Histopathology Image Analysis.

Aniwat Juhong[1,7], Bo Li[1,7], Cheng-You Yao[2,7], Chia-Wei Yang[3,7], Dalen W. Agnew[4], Yu Leo Lei[5], Xuefei Huang[2,3,7], Wibool Piyawattanametha,[6,7] and Zhen Qiu[1,2,7]*

[1] Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI 48823, USA
[2] Department of Biomedical Engineering, Michigan State University, East Lansing, MI 48824, USA
[3] Department of Chemistry, Michigan State University, East Lansing, MI 48824, USA
[4] College of Veterinary Medicine, Michigan State University, East Lansing, MI 48824, USA
[5] Department of Periodontics Oral Medicine, University of Michigan, Ann Arbor, MI 48104, USA
[6] Department of Biomedical Engineering, School of Engineering, King Mongkut's Institute of Technology Ladkrabang (KMITL), Bangkok 10520, Thailand
[7] Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48823, USA
*qiuzhen@msu.edu

Abstract: 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.

1. Introduction

Pathology diagnosis is routine work usually performed by a skilled pathologist or cytologist. The diagnosis 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 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 [1, 2].

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 [3]. Laboratory infrastructure is typically clustered in urban settings and is relatively inaccessible in regions where significant portions of the affected population reside [4]. In particular, many neglected diseases are more prevalent in rural areas, far from these diagnostic centers [5]. 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 [6-10]. Those methods can surpass the traditional methods such as Otsu segmentation [11], Watershed method [12], and K-mean clustering [13] 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 [14] 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 [15] 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 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 [16] 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 [17]. 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 different 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 [18-24] such as fluorescence imaging, light-sheet imaging, and color imaging of pathological slides. However, those works employed the same concept of SRGAN [25] that the generator is built using the ResNet architecture or residual structure [26]. 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 [27] network is applied as the backbone for SGAN namely ESRGAN [28] 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 [29], 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 [30] have emerged as an alternative to CNNs architectures since they can provide better results on large datasets. However, 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 [31] (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) were 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 [32], 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 [33] to train our Inception U-net Model. Our Inception U-net Model's Intersection over Union (IoU) and Dice Similarity Coefficient (DSC) are 0.869 and 0.893, respectively. Since the SRGAN-ResNeXt and Inception U-net Models 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.
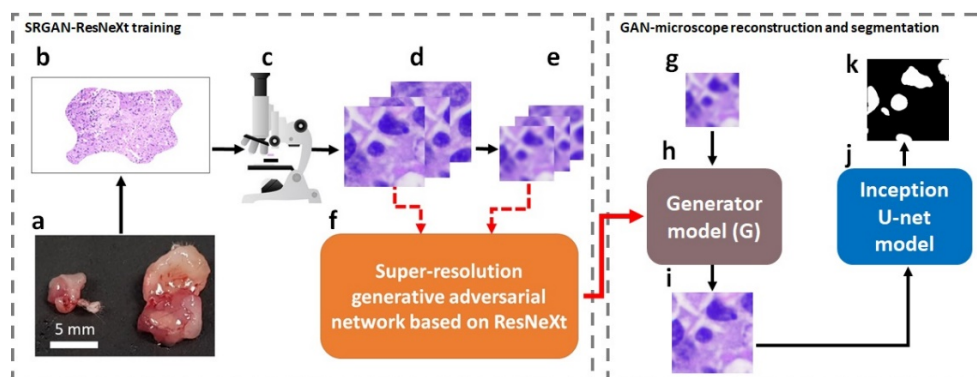

Fig. 1. The workflow of super high resolution and segmentation deep learning. (a) Fresh breast tumor tissues. (b) The corresponding H&E stained tissue slides. (c) A commercial microscope (Nikon Eclipse Ci) for capturing the H&E stained tissue slide images. (d) High-resolution images acquired by the microscope. (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.

Fig. 1 shows the overall workflow of the models. First, the breast tumor H&E slides were prepared on biopsy slides (Fig. 1(a)-(b)) to be imaged with a 40x magnification (Fig. 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 (Fig. 1(d)-(f)). Eventually, the well-trained generator model from the SRGAN-ResNeXt (Fig. 1(h)) was applied to the unseen low-resolution image (Fig. 1(g)) to enhance its quality by generating the high-resolution image (Fig. 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 (Fig. 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.

2. Methods

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 Fig. 2(a) and Fig. 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, residual, 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 (Fig. 2(c)) is the combined models, which are the generator model, discriminator model, and VGG19-the latter works as the feature extractor [34].

2.1.1 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 Fig. 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 [Fig. 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 [35], 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 [Fig. 2(c)]. The discriminator and VGG19 models are untrainable during training the generator model.

2.1.2 Discriminator model

The discriminator network [36] 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.
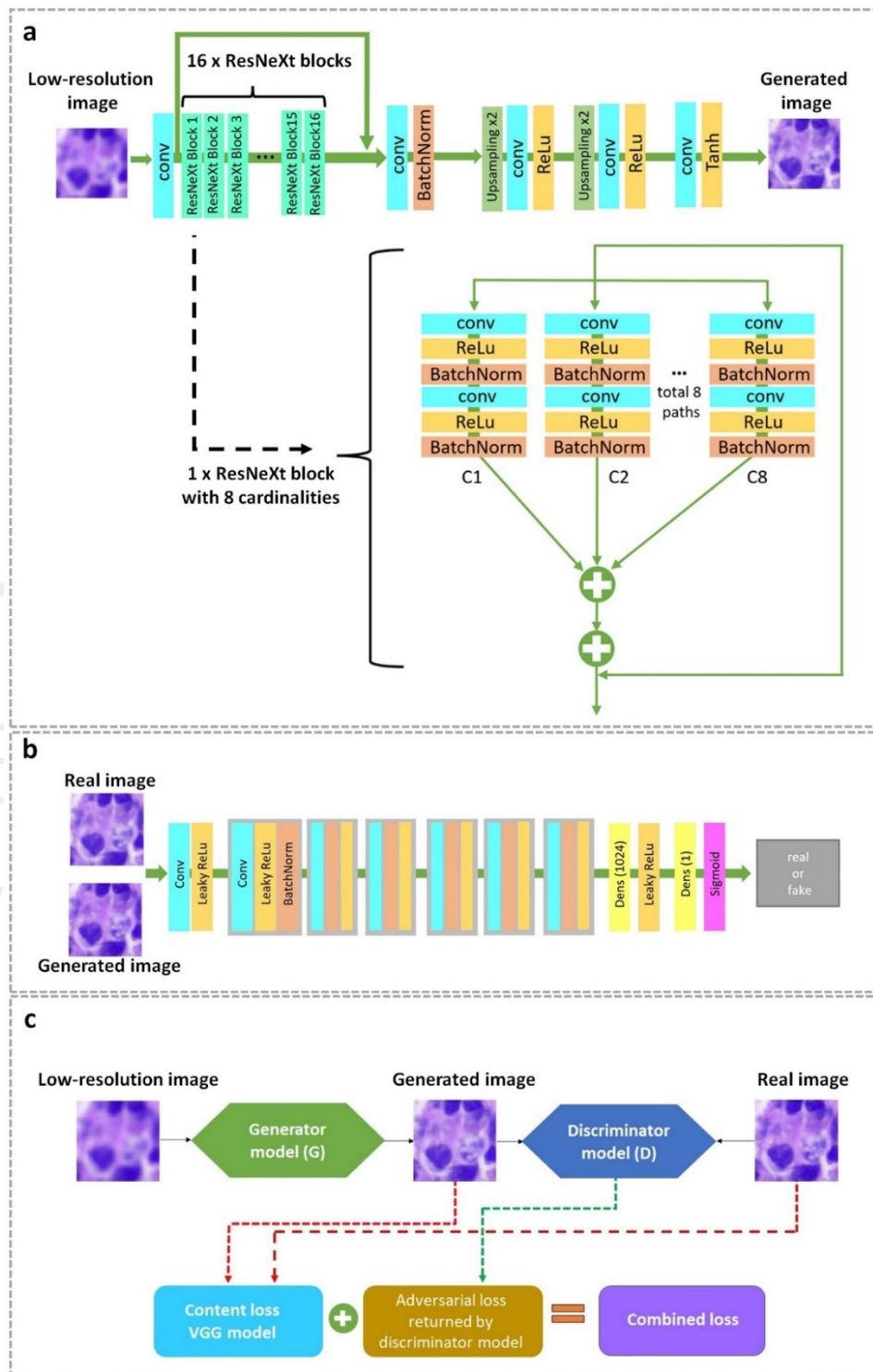
Fig. 2. Super-resolution generative adversarial network-based on SRGAN-ResNeXt. (a) Generator model. (b) Discriminator model. (c) The combined models so-called adversarial model for training Generator model.

2.1.3 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, I_X^{SR}) and adversarial loss (Discriminator loss, I_{Gen}^{SR}) as shown in equation (1) as

$$I^{SR} = I_X^{SR} + C_w I_{Gen}^{SR}. \tag{1}$$

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.

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

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

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 ith 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

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

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 Dataset for training SRGAN-ResNeXt Model

To obtain breast cancer H&E images, the female MUC1 double-transgenic mice with breast tumors [37] 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.

Fig. 3(a) shows the cropping area from the large FOV H&E images. Fig. 3(b) are the small patches that were cropped from the large FOV image.
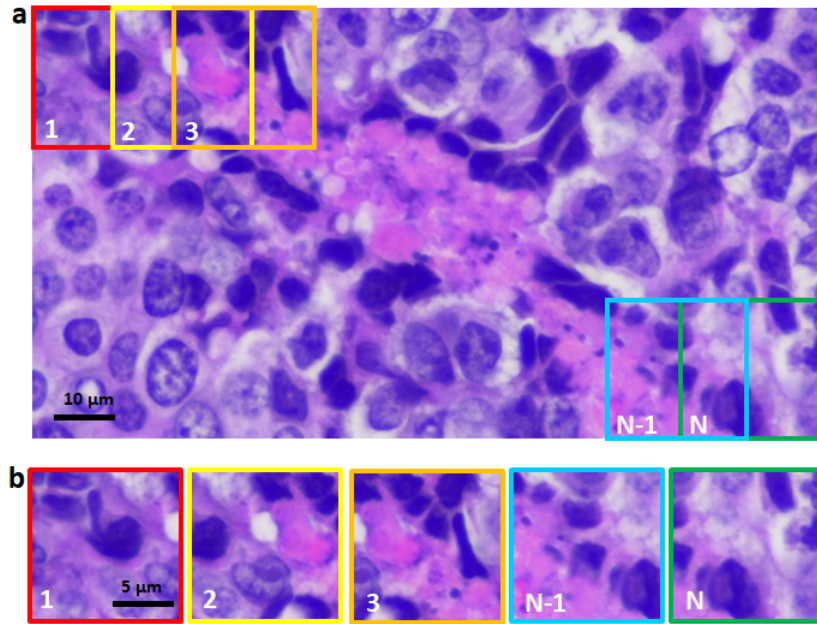

Fig. 3. Data set preparation for training SRGAN-ResNeXt, cropped image with 50% overlapping area. (a) Large field of view H&E image, (b) The small patches of the large image (a) with 50% overlapping area.

2.3 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 [38] 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 Fig. 4 below by replacing each convolution block in the original U-net architecture with the Inception blocks.
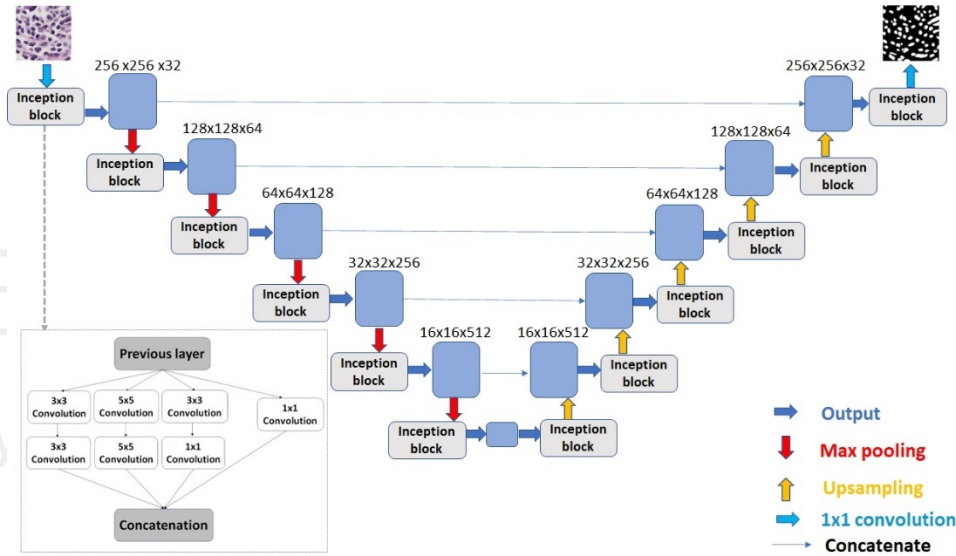

Fig. 4. Inception U-net architecture for H&E image segmentation. Every single blue box corresponds to a multi-channel feature map. The value over the boxes represents the number of channels.

Fig. 4 illustrates the Inception U-net architecture. The first part is the encoder (the left side of Fig. 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 Fig. 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

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

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.

2.4 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[33]. 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).

2.5 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. Fig. 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 Fig. 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

$$l^{JG} = l_X^{SR} + C_w l_{Gen}^{SR} + C_{w2} l_{GenS}^{SRS}, \tag{6}$$

Where l^{JG} is the combined loss of the jointly trained generator model, l_X^{SR} is the content loss (VGG19 loss), l_X^{SR} is the adversarial loss (Discriminator loss), l_{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.
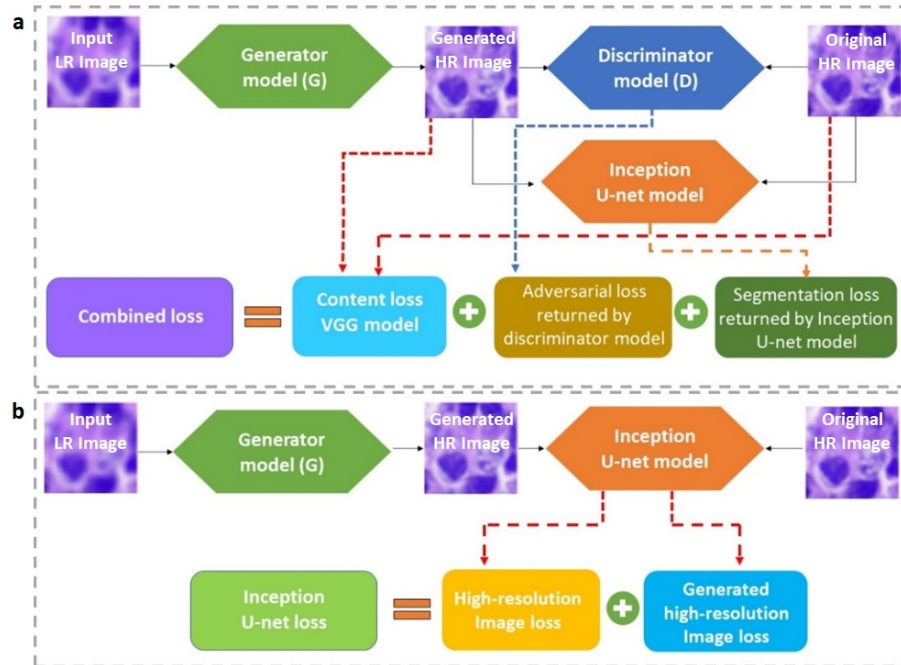
Fig. 5. Jointly trained SRGAN-ResNeXt Model and Inception U-net Model. (a) The assembled models for the jointly trained generator (JTG) Model. (b) The assembled models for the jointly trained Inception U-net (JTIU) Model.

2.6 Data set for the jointly trained SRGAN-ResNeXt and Inception U-net 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.

2.7 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 9th Gen Intel Core i7-9750H CPU, 16 GB RAM, and 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 the Google Colab-pro.

3. Results and discussion

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. Fig. 6 shows the results of applying both the SRGAN-ResNeXt and the Inception U-net Models to a breast tumor H&E image. Fig. 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 Fig. 6(a2), 6(b2), and 6(c2). The SRGAN-ResNeXt Model was employed to enhance these low-resolution images by synthesizing high-resolution images (Fig. 6(a3), 6(b3), and 6(c3)). The Inception U-net was then applied to these generated high-resolution images for segmentation (Fig. 6(a4), 6(b4), and 6(c4)).
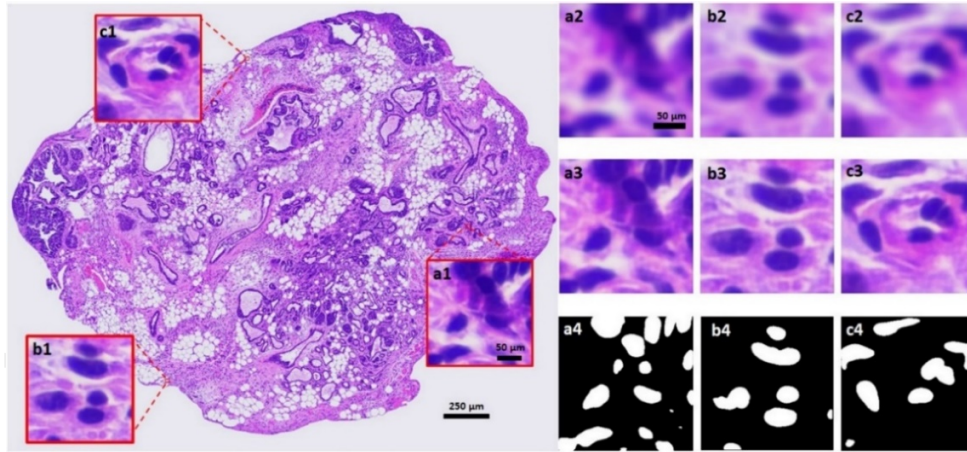

Fig. 6. The whole slide image (WSI) of a breast tumor H&E slide and the result of our deep learning model. (a1, b1, and c1) The high-resolution images of the WSI from different areas. (a2, b2, and c2) The low-resolution images. (a3, b3, and c3) The reconstructed high-resolution images using our deep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

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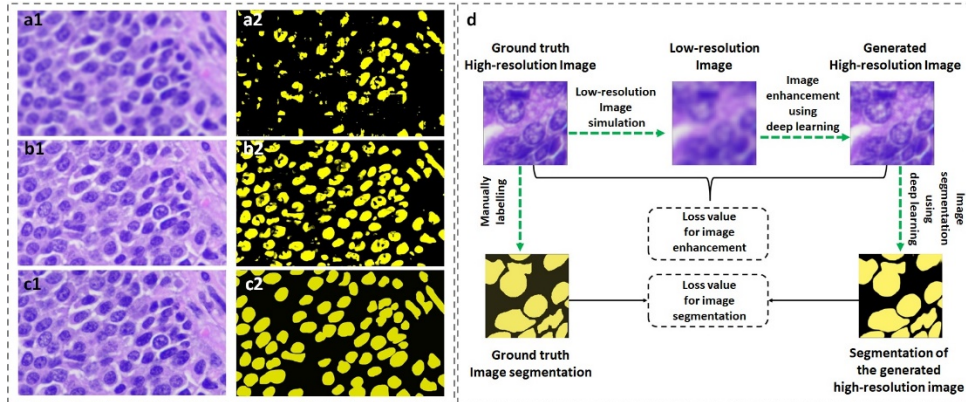
Fig.7. The H&E image segmentation of the low-resolution image and the enhanced-resolution image. (a1-a2) The low-resolution image and its segmentation image (output of the Inception U-net). (b1-b2) The enhanced-resolution image (output of the SRGAN-ResNeXt) and its segmentation image (output of the Inception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

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

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

The MSE is defined as the following

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

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) [39].

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

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^2 denotes 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 high-resolution generated and the ground truth (realistic high-resolution image) 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

Fig. 8 compares the reconstruction results of the typical SRGAN, SRGAN-Transformer, SRGAN-DenseNet, and our SRGAN-ResNeXt. Fig. 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. Fig. 8(c), 8(d), 8(e), and 8(f) show the generated high-resolution H&E image 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 Fig. 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.)
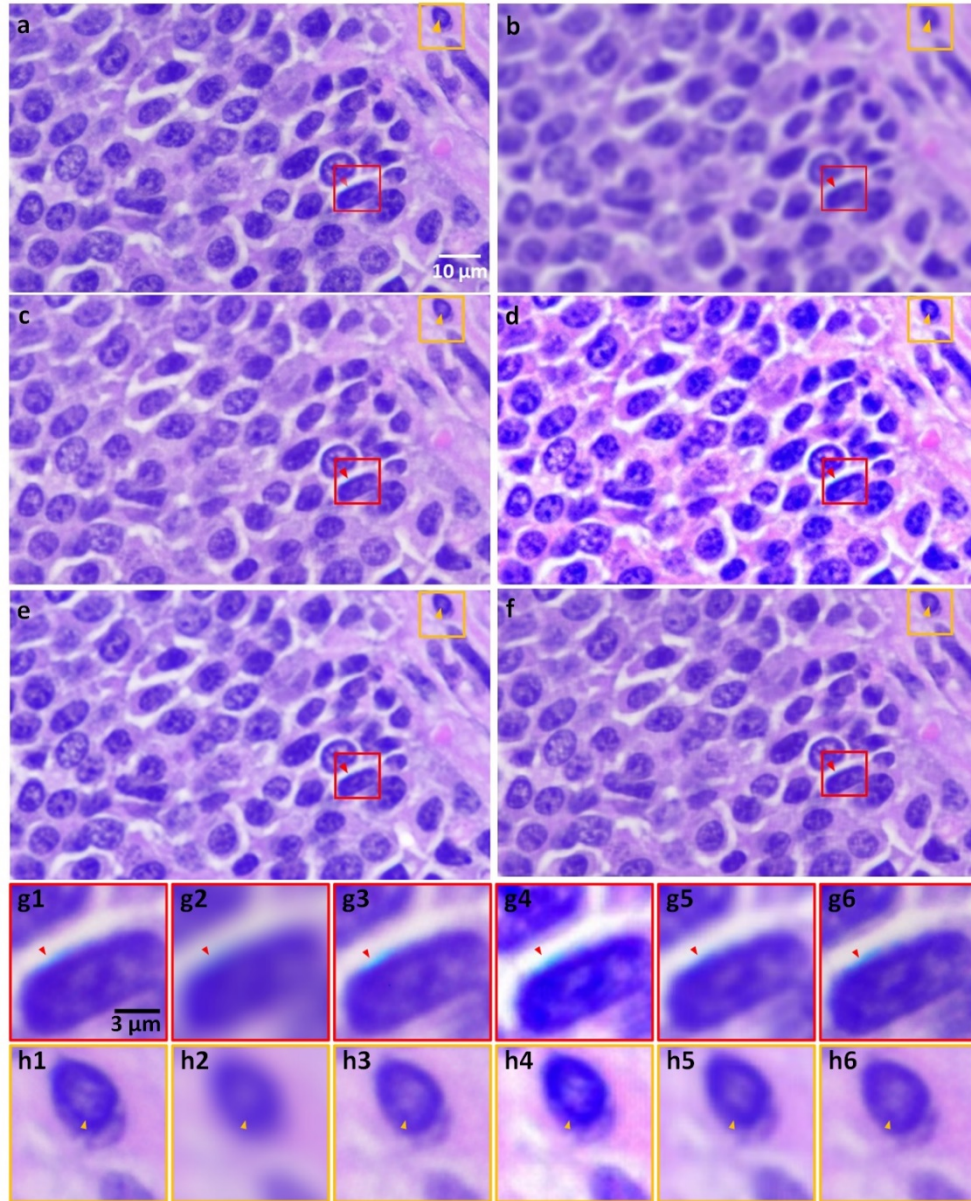
Fig. 8. Comparison of the results for our deep-learning model based on ResNeXt against bicubic interpolation of the low-resolution image, SRGAN, SRGAN-Transformer, and SRGAN-DenseNet. (a) The original ground truth image. (b) Bicubic interpolation of the low-resolution image. (c) The SRGAN result. (d) 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.

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) [40]. 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

$$IoU = \frac{target \cap prediction}{target \cup prediction}.\tag{10}$$

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

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

It is remarkably similar to the IoU. They are positively correlated.

The unseen H&E cancer images from the cancer imaging archive [33] 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

| H&E image | Ground truth | U-net | Inception U-net |

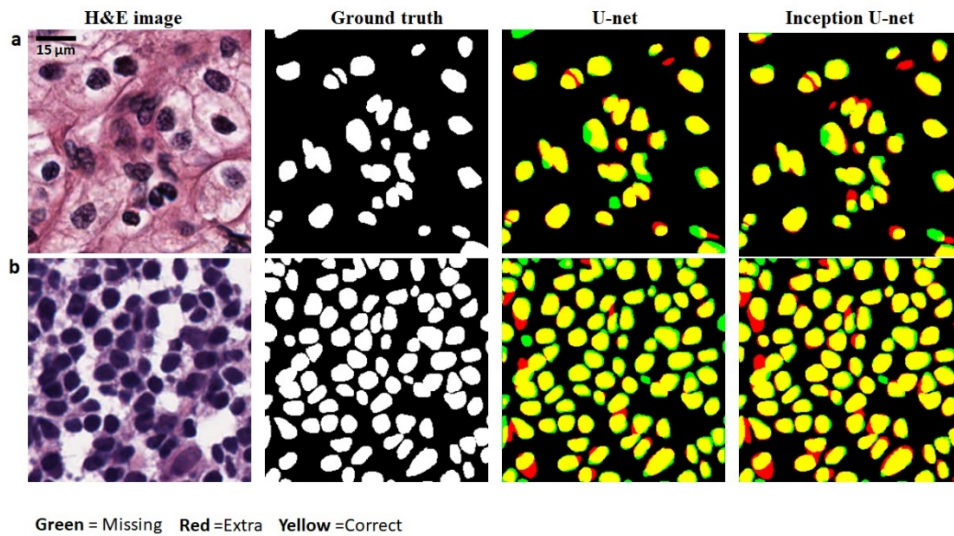Green = Missing   Red = Extra   Yellow = Correct

Fig. 9. Comparison results between the traditional U-net and Inception U-net by using H&E images and ground truth from the dataset [33]. (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.

3.2 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 Fig. 10, Table. 3, and Table. 4 below.
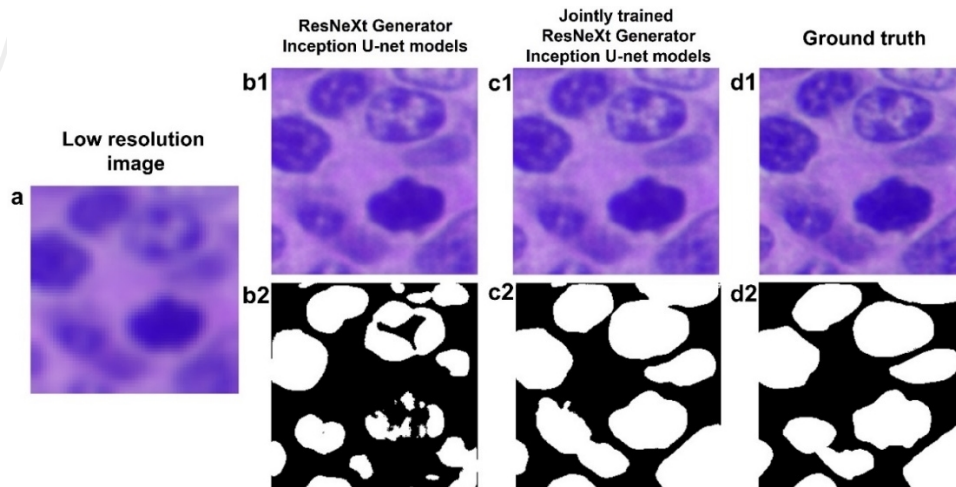

Fig. 10. The improvement of the SRGAN-ResNeXt and Inception U-net after training them jointly. (a) Low-resolution image input. (b1-b2) The ResNeXt generator and Inception U-net models' results. (c1-c2) The jointly trained models' results. (d1-d2) High-resolution and segmentation ground truth images.

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**

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.

Disclosures: The authors declare no conflicts of interest related to this article.

Data availability: Data underlying the results presented in this paper are not publicly available at this time but may be obtained from the authors upon reasonable request.

References

1. G. Litjens, C. I. Sánchez, N. Timofeeva, M. Hermsen, I. Nagtegaal, I. Kovacs, C. Hulsbergen-Van De Kaa, P. Bult, B. Van Ginneken, and J. Van Der Laak, "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," Scientific reports **6**, 1-11 (2016).
2. A. J. Mendez, P. G. Tahoces, M. a. J. Lado, M. Souto, and J. J. Vidal, "Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms," Medical Physics **25**, 957-964 (1998).
3. I. I. Bogoch, H. C. Koydemir, D. Tseng, R. K. Ephraim, E. Duah, J. Tee, J. R. Andrews, and A. Ozcan, "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 (2017).
4. C. A. Petti, C. R. Polage, T. C. Quinn, A. R. Ronald, and M. A. Sande, "Laboratory medicine in Africa: a barrier to effective health care," Clinical Infectious Diseases **42**, 377-382 (2006).
5. D. G. Colley, A. L. Bustinduy, W. E. Secor, and C. H. King, "Human schistosomiasis," The Lancet **383**, 2253-2264 (2014).
6. H. Irshad, A. Veillard, L. Roux, and D. Racoceanu, "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 (2013).
7. K. Sirinukunwattana, S. E. A. Raza, Y.-W. Tsang, D. R. Snead, I. A. Cree, and N. M. Rajpoot, "Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images," IEEE transactions on medical imaging **35**, 1196-1206 (2016).
8. Y. Song, L. Zhang, S. Chen, D. Ni, B. Lei, and T. Wang, "Accurate segmentation of cervical cytoplasm and nuclei based on multiscale convolutional network and graph partitioning," IEEE Transactions on Biomedical Engineering **62**, 2421-2433 (2015).
9. F. Xing, Y. Xie, and L. Yang, "An automatic learning-based framework for robust nucleus segmentation," IEEE transactions on medical imaging **35**, 550-566 (2015).
10. F. Xing and L. Yang, "Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review," IEEE reviews in biomedical engineering **9**, 234-263 (2016).
11. N. Otsu, "A threshold selection method from gray-level histograms," IEEE transactions on systems, man, and cybernetics **9**, 62-66 (1979).
12. X. Yang, H. Li, and X. Zhou, "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 (2006).
13. P. Filipczuk, M. Kowal, and A. Obuchowicz, "Automatic breast cancer diagnosis based on k-means clustering and adaptive thresholding hybrid segmentation," in *Image processing and communications challenges 3* (Springer, 2011), pp. 295-302.
14. S. Graham, Q. D. Vu, S. E. A. Raza, A. Azam, Y. W. Tsang, J. T. Kwak, and N. Rajpoot, "Hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images," Medical Image Analysis **58**, 101563 (2019).
15. U. Schmidt, M. Weigert, C. Broaddus, and G. Myers, "Cell detection with star-convex polygons," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, (Springer, 2018), 265-273.
16. S. Chen, C. Ding, M. Liu, and D. Tao, "CPP-net: Context-aware polygon proposal network for nucleus segmentation," arXiv preprint arXiv:2102.06867 (2021).
17. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, (Springer, 2015), 234-241.
18. K. de Haan, Y. Zhang, T. Liu, A. E. Sisk, M. F. Diaz, J. E. Zuckerman, Y. Rivenson, W. D. Wallace, and A. Ozcan, "Deep learning-based transformation of the H&E stain into special stains improves kidney disease diagnosis," arXiv preprint arXiv:2008.08871 (2020).
19. T. Liu, K. De Haan, Y. Rivenson, Z. Wei, X. Zeng, Y. Zhang, and A. Ozcan, "Deep learning-based super-resolution in coherent imaging systems," Scientific reports **9**, 1-13 (2019).
20. L. Mukherjee, A. Keikhosravi, D. Bui, and K. W. Eliceiri, "Convolutional neural networks for whole slide image superresolution," Biomedical optics express **9**, 5368-5386 (2018).
21. Y. Rivenson, Z. Göröcs, H. Günaydin, Y. Zhang, H. Wang, and A. Ozcan, "Deep learning microscopy," Optica **4**, 1437-1443 (2017).
22. H. Wang, Y. Rivenson, Y. Jin, Z. Wei, R. Gao, H. Günaydın, L. A. Bentolila, C. Kural, and A. Ozcan, "Deep learning enables cross-modality super-resolution in fluorescence microscopy," Nature methods **16**, 103-110 (2019).

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

24. T. Zheng, H. Oda, T. Moriya, T. Sugino, S. Nakamura, M. Oda, M. Mori, H. Takabatake, H. Natori, and K. Mori, "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, 2020), 1131305.

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

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

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

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

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

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

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

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

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

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

35. W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "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*, 2016), 1874-1883.

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

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

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

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

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