Journal of Pathology Informatics 13 (2022) 100148

Contents lists available at ScienceDirect

Journal of Pathology Informatics

journal homepage: www.elsevier.com/locate/jpi

Impact of color augmentation and tissue type in deep learning for hematoxylin and eosin image super resolution

Cyrus Manuel, Philip Zehnder, Sertan Kaya, Ruth Sullivan, Fangyao Hu *

Genentech, South San Francisco, CA, USA

ARTICLE INFO

Keywords:
Artificial intelligence
Deep learning
Digital pathology
Generative adversarial networks
Histopathology
Image processing
Whole slide imaging

ABSTRACT

Single image super-resolution is an important computer vision task with applications including remote sensing, medical imaging, and surveillance. Modern work on super-resolution utilizes deep learning to synthesize high resolution (HR) images from low resolution images (LR). With the increased utilization of digitized whole slide images (WSI) in pathology workflows, digital pathology has emerged as a promising domain for super-resolution. Despite extensive existing research into super-resolution, there remain challenges specific to digital pathology. Here, we investigated image augmentation techniques for hematoxylin and eosin (H&E) WSI super-resolution and model generalizability across diverse tissue types. In addition, we investigated shortcomings with common quality metrics (peak signal-to-noise ratio (PSNR), structure similarity index (SSIM)) by conducting a perceptual quality survey for super-resolved pathology images. High performing deep super-resolution models were used to generate 20X HR images from LR images (5X or 10X equivalent) for 11 different tissues and 30 human evaluators were asked to score the quality of the generated versus the ground truth 20X HR images. The scores given by a human rater and the PSNR or the SSIM were compared to investigate the correlation between model training parameters. We found that models trained on multiple tissues generalized better than those trained on a single tissue type. We also found that PSNR correlated with perceptual quality ($R = 0.26$) less accurately than did SSIM ($R = 0.64$), suggesting that the SSIM quality metric is insufficient. The methods proposed in this study can be used to virtually magnify H&E images with better perceptual quality than interpolation methods (i.e., bicubic interpolation) commonly implemented in digital pathology software. The impact of deep SISR methods is more notable when scaling to 4X is needed, such as in the case of super-resolving a low magnification WSI from 10X to 40X.

Background

Single image super-resolution (SISR) is the task of reconstructing a high resolution (HR) image from its low resolution (LR) counterpart. Super-resolution poses a particularly difficult challenge since every LR image has infinite possible corresponding HR images. Several methods have been developed for a variety of computer vision applications in which resolution of an image must be improved, but image acquisition methods are unavailable or impractical, and image-processing techniques must be employed.[1,2] Such applications include remote sensing,[3,4] medical imaging,[5,6] and surveillance.[7] Conventional SISR approaches include example-based[8] and regression-based approaches.[9] Recently, deep learning has been applied to the SISR problem, achieving state-of-the-art performance on common super-resolution datasets.[1,10,11]

Pathologists interpret histopathology slides based on features observed at different resolutions. In a conventional pathology workflow, histopathology slides are examined under optical microscopes and the pathologist can increase the resolution by using a higher magnification objective. In contrast, in a digital pathology workflow, the resolution is constrained by the magnification of the scanning objective used to digitize the slides typically with 40X magnification (Hamamatsu). Evaluation and interpretation of histopathology slides by pathologists might be suboptimal if the desired resolution of the digital slides is not available. Deep learning-based super-resolution methods such as generative adversarial networks (GAN) have the potential to render structures that are present but difficult to see in LR images interpretable, aiding both human- and machine-based evaluation of WSI.[12,13] For example, leukocyte cytoplasmic granules provide critical information for cell subtype assessment and are optically present using 40X objectives, but are sometimes hard to clearly visualize and may require higher resolution for interpretation.[14]

Many recent efforts have applied super-resolution techniques to digital pathology[1,11,14–18] for a variety of purposes including increasing scanning throughput, reducing file storage costs, and improving downstream computer vision tasks. The data strategy for training these models has primarily focused on using a specific tissue for model development. Mukherjee et al. developed their SISR methods on renal, pancreatic, or breast cancer tissue

* Corresponding author at: Genentech, 1 DNA Way, South San Francisco, CA 94080, USA.
E-mail address: hu.fangyao@gene.com (F. Hu).

http://dx.doi.org/10.1016/j.jpi.2022.100148
Received 24 June 2022; Received in revised form 23 September 2022; Accepted 23 September 2022
Available online 01 October 2022

in order to improve tumor segmentation results.[17] Despite recent advances in the field, there remain questions surrounding how super-resolution models can be applied to other contexts. In this regard, it is not clear how single-tissue models generalize when applied to multiple tissue types. Notably, comparisons between specialist single-tissue models and generalist multi-tissue models have not been conducted. Thus, the generalizability of SISR models across tissue types warrants further investigation. Furthermore, studies using datasets with hematoxylin and eosin (H&E) WSI of different tissue types from research animals are uncommon.

In addition, super-resolution approaches are challenged by the variability of the color space of WSI in digital pathology. H&E stains are commonly used in digital pathology but the color in H&E slides varies due to inconsistencies in staining protocols and different scanning machines. While self-supervised stain normalization[15] and more general augmentation techniques have been investigated for super-resolution,[19,20] the research on the impact of color augmentation to SISR in WSI is limited. To make SISR models more robust to a variety of scanners and subtle stain variations, we evaluated the impact of color-based augmentation strategies on super-resolution performance.

Many quality metrics can be applied to super-resolved images as a way to evaluate SISR model performance. Common metrics for objective image quality include peak signal-to-noise ratio (PSNR)[21] and structural similarity index (SSIM).[22] While PSNR and SSIM continue to be valuable in assessing SISR model performance in terms of discrete pixel differences between the ground truth and super-resolved image, they cannot fully capture human perceptual image quality which is just as important, if not more important, for pathologists reviewing these fields-of-view in the slide. Importantly, the correlation between these metrics and perceived image quality in digital pathology is questionable.[1,23,24] Recent efforts have attempted to model perception scores with deep learning,[25,26] but have failed to achieve human performance. Thus, human scoring is often used as the gold-standard for evaluating perceptual photorealism.

We sought to optimize deep learning models for single image super-resolution of H&E images. In this study, we extend existing work[11,15,27] employing deep learning approaches to super-resolve H&E images at scaling factors up to 4X. We demonstrate the consequences of employing data augmentation and adversarial training on deep SR models in order to perform well in improving PSNR and SSIM of image reconstructions. We show how training Deep Back-Projection Network (DBPN)[10] with a GAN[13] improves perceived image quality. We highlight the disconnect between human perception of image quality and SSIM scores, and how SSIM failed as a performance metric for a GAN-based deep super-resolution model. We perform an initial exploration of single-tissue versus mixed-tissue training and the impact of this on the eventual goal of producing a super-resolution method that is applicable across a wide range of histologic sample types.

Methods

This study consists of 3 parts: (1) improving model performance by augmenting default training settings, (2) human rater scoring of super-resolved images, and (3) comparison of models trained on different tissue datasets. First, several training techniques were employed with 2 super-resolution models, Deep Back-Projection Networks For Super-Resolution (DBPN)[10] and Lightweight Image Super-Resolution with Information Multi-distillation Network (IMDN),[28] in the kidney dataset for hyperparameter tuning purposes. Second, high performing models were used to generate 20X, HR images from LR images at lower magnifications for 11 different tissues and 30 human raters were asked to score the quality of the generated versus the ground truth 20X images. In addition, the correlation between the scores from human rater and the PSNR or SSIM were computed to investigate the relation between the parameters. Last, to explore the generalizability of SISR models reconstructing unseen tissue types, DBPN and IMDN were trained with single tissues (lung, brain, or kidney) and mixed tissues separately using the optimal hyperparameters. Mixed-tissue and single-tissue model performances were examined.

Whole slide image dataset

This study included a total of 567 WSI of H&E-stained histology slides of rat tissues which were collected from our in-house archive. Most scanned slides contained a single tissue type. The entire dataset consists of 12 tissue types: brain, ear, eye, kidney, lung, lymph node, ovary, sciatic nerve, skin, spinal cord, testis, and urinary bladder. Slides that contained tissues from the GI tract (colon, duodenum, ileum, and stomach) were also included and counted as the 13th tissue type. The slides were scanned using Hamamatsu NanoZoomer-XR (Hamamatsu Photonics, Hamamatsu City, Japan) at 40X magnification (0.226 μm/pixel) or 20X magnification (0.452 μm/pixel) and were manually inspected to be free of obvious blurring artifacts. The images at 20X magnification serve as the high resolution (HR) dataset (ground truth). WSI in the HR dataset were cropped to 1024 x 1024 image tiles without any overlap between image tiles. Tiles over 50 KB were kept, a file size which was derived empirically to filter out tiles lacking tissue. The LR image tiles were generated by down sampling with a scaling factor of 2X or 4X using bicubic interpolation from 20X images, yielding images with equivalent resolution to 10X magnification (0.904 μm/pixel) and 5X magnification (1.808 μm/pixel), respectively.

Data splitting

The WSIs was tiled and roughly split into training (70%), validation (20%), and testing (10%) datasets. (Table 1). The image tile sets were derived from sampling image tiles from the tissue WSI splits. Since tissues naturally varied in size, certain tissue types produced more image tiles than others. Thus, the training, validation and testing sets were limited to 15 000, 1950, and 1690 image tiles, respectively for data balancing. The single-tissue training dataset consisted of tiles derived from the same tissue type, while the mixed-tissue training dataset consisted of tiles from all tissue types. For single-tissue models, we made lung, kidney, and brain training sets. In contrast, mixed-tissue models had a training set consisting of an equal contribution ($\sim$1154 tiles) from all tissue types, with the exception of ovary ($n = 642$) and urinary bladder ($n = 1036$) which are small tissues and required tile resampling to attain the target sample size. Both single-tissue models and the mixed-tissue models were evaluated in the same validation and test set. The validation tile set was derived from the WSI validation split and was comprised of all 13 tissue types (Table 1). The test set was comprised of all 13 tissue types and 1690 image tiles in total (Table 1). The test set was used for hyperparameter tuning and evaluation of all models in this study.

Table 1
Whole slide image dataset splits and corresponding image patch breakdown.

Tissue type	# WSI train	# WSI validation	# Tiles in validation	# WSI test	# Tiles in test
Brain	21	6	150	3	130
Ear	21	6	150	3	130
Eye	21	6	150	3	130
Kidney	20	2	150	1	130
Lung	117	33	150	18	130
Lymph node	21	6	150	3	130
Ovary	10	3	150	2	130
Sciatic nerve	42	12	150	7	130
Skin	21	6	150	3	130
Spinal cord	63	18	150	9	130
GI tract	20	2	150	1	130
Testis	10	3	150	2	130
Urinary bladder	20	5	150	4	130

The training tile sets were limited to 15 000 tiles. The validation tile set was limited to 1950 tiles. The test tile set used to evaluate all models in this study was limited to 1690 tiles.

Model training

Two deep SR models were primarily investigated in this study: Deep Back-Projection Networks For Super-Resolution (DBPN)[10] and Lightweight Image Super-Resolution with Information Multi-distillation Network (IMDN).[28] Model architectures and implementation details for both networks were described by Haris et al[10] and Hui et al,[28] respectively. Models were implemented in PyTorch 1.5.1, an open source deep learning python library. Model training was done in 3 phases in alignment with the 3 parts of the study. The first phase focused on observing the effect of training parameters such as color and data augmentation to select the suitable models for the next phase. In the second phase, we focused on improving the image quality output of the selected models from the prior phase by training them in a GAN. In the third phase, we trained DPBN and IMDN models with different training datasets.

In the first phase, hyperparameters were found experimentally by training on the kidney dataset at 2X and 4X super-resolution scales. Following the training procedures in[10,28] DBPN and IMDN were initially trained with mean squared (Eq. 1) and mean absolute error (Eq. 2), respectively, to serve as the base or null model:

$$\mathcal{L}_{mse} = \mathcal{L}_{DBPN}\left(X_{HR}, \widehat{X}_{HR}\right) = \|X_{HR} - \widehat{X}_{HR}\|_2^2$$

Equation 1. Mean squared error loss or DBPN Loss function

$$\mathcal{L}_{mae} = \mathcal{L}_{IMDN}\left(X_{HR}, \widehat{X}_{HR}\right) = \|\widehat{X}_{HR} - X_{HR}\|_1$$

Equation 2. Mean absolute error loss or IMDN Loss function

where X_HR and Xhat_HR are the ground truth and reconstructed images, respectively.

To improve image quality generated by these models, we incorporated color augmentation and mix-up,[19] whereby we composed two different image tiles of the same tissue type into one image patch (Fig. 1a). When color augmentation was included in the training process, the HR image patch was randomly color augmented before downsampling. Similarly, when mix-up was included as a parameter in the model training, the composed image was first created with the HR image tiles and then downsampled to the LR image patch to be fed as input to the super-resolution model. Augmentations were applied during training. IMDN and DBPN models were trained with experimentally derived learning rates of 2×10^{-6} and 1×10^{-4}, respectively. All models were trained with a batch size of 16 on two Nvidia (Santa Clara, CA, USA) P6000 GPUs for 60 epochs using the ADAM optimizer.[29]

Three main training parameters were explored, producing 4 variations of DBPN and IMDN training protocols. These models were trained with either: (1) color jitter augmentation, (2) color jitter augmentation with 1 - SSIM added to the batch loss, (3) 1 - SSIM added to the batch loss, or with (4) color and mix-up augmentation with 1 - SSIM added to the batch loss.

In the second phase, we fine-tuned the IMDN and DBPN models from the previous phase with a GAN[13] by applying a discriminator network (Fig. 1b) to training process. We incorporated 3 additional loss terms as described in Haris et al[10]: adversarial, VGG, and style loss. In a GAN configuration, an adversarial loss (Eq. 3) is used to encourage a generator network G to compete with a discriminator network D. The generator network G is either the IMDN or DBPN model trained on kidney image patches with supervision using Eqs (2) and (1), respectively. The discriminator is a separate neural network classifier that attempts to determine which images are true HR images and which images are generated reconstructions.

$$\mathcal{L}_{adv} = log\ (D(X_{HR})) + log\ (1 - D(G(X_{LR}))$$

Equation 3. Adversarial loss

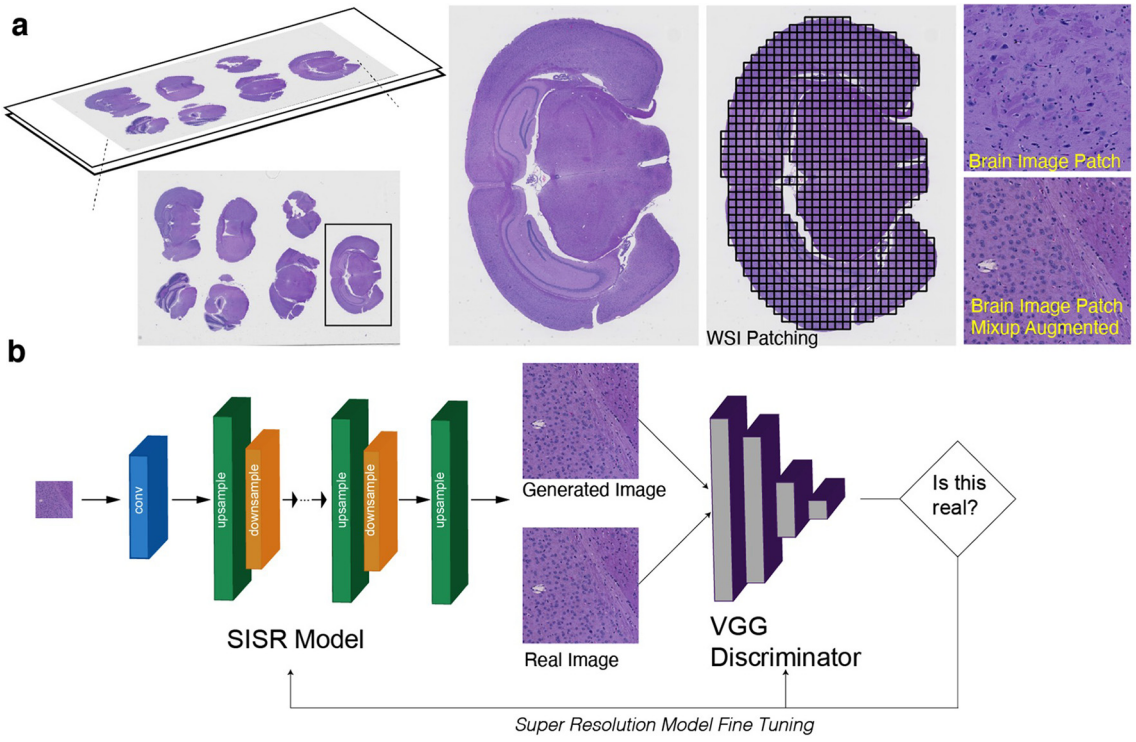

Fig. 1. Data preprocessing and model architecture overview. a. HR image patches are first extracted from tissue regions of the WSI and then are augmented before downscaling and feeding into a deep super-resolution model. b. Deep super-resolution models were trained to generate HR image patches from LR images. Additional model fine tuning was performed on select models by including a VGG discriminator to create a generative adversarial network. (WSI: Whole slide image, HR: High resolution image, SR: Super-resolution).

We used a VGG19[30] convolutional neural network pre-trained on the ImageNet dataset[31] as our discriminator. We employed VGG loss[32] (Eq. 4) to ensure consistent discriminator feature representations between the reconstruction and ground truth where f denotes the feature maps of the VGG network from multiple max-pool layers near the beginning of the network ($i = 2,3,4,5$).

$$\mathcal{L}_{vgg} = \sum_{i=2}^{5} \|f_i(X_{HR}) - f_i(\widehat{X}_{HR})\|_2^2$$

Equation 4. VGG loss

As per Haris et al,[10] style loss was used to facilitate the generation of high quality textures and was originally proposed by Gatys et al.[33] Style loss uses the same convolutional feature maps as in the VGG loss (Eq. 4) but are parameterized by a function, phi.

$$\mathcal{L}_{style} = \sum_{i=2}^{5} \|\phi(f_i(X_{HR})) - \phi(f_i(\widehat{X}_{HR}))\|_2^2$$

Equation 5. Style loss

To compute the style loss requires a Gram matrix where F denotes the feature maps as inputs to phi from Eq. (5). In Eq. (5), taking the dot product of the flattened image features (ground truth or super-resolved) with the convolutional feature map of the VGG network aims to adopt some of the visual style learned from the earlier layers of VGG network. Early layers

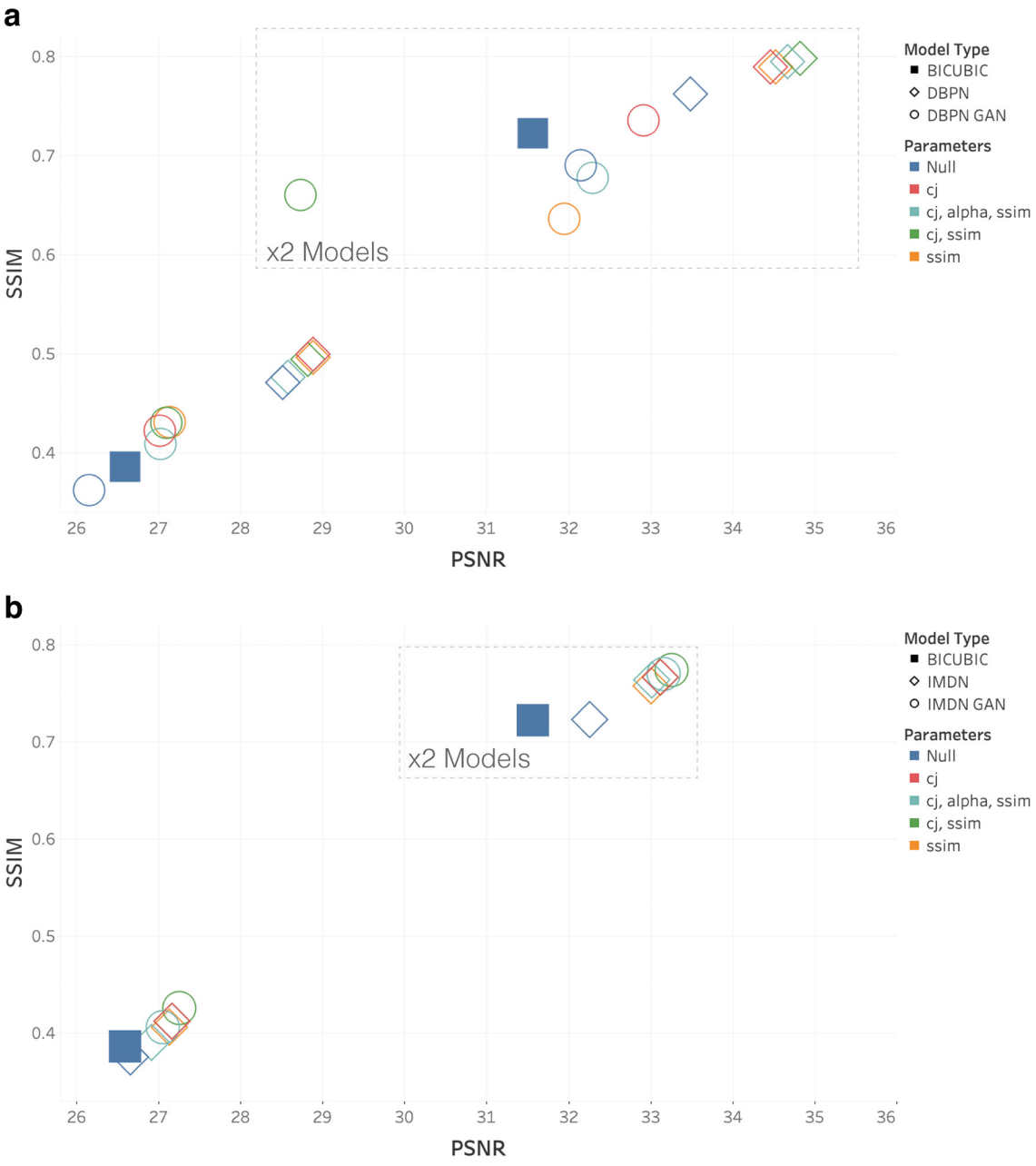

Fig. 2. PSNR and SSIM performance increase with data augmentation. Deep super-resolution models trained to upscale LR images by 2X or 4X. Hyperparameter adjustments to the DBPN and IMDN models increased metrics. DPBN models showed the highest PSNR and SSIM. a. DPBN & DBPN GAN b. IMDN & IMDN GAN (cj: color jitter augmentation, alpha: alpha mix-up augmentation, ssim: 1-SSIM added to the batch loss).

of a convolutional neural network like VGG represent low-level features such as corners, blobs, and edges.

$$\phi(F) = FF^T \epsilon R^{nxn}$$

Equation 6. Gram matrix

The DBPN GAN and IMDN GAN models were trained on kidney image patches with the aforementioned VGG discriminator. During training, the generator and discriminator jointly optimize the following loss function:

$$\mathcal{L}_{GAN} = \mathcal{L}_{MSE} + \mathcal{L}_{VGG} + \mathcal{L}_{adv} + \mathcal{L}_{style}$$

Equation 7. GAN loss

Both IMDN GAN and DBPN GAN models use the same GAN loss (Eq. 7), which uses the aforementioned losses: adversarial loss (Eq. 3), VGG loss (Eq. 4), style loss (Eq. 5), and mean squared error loss (Eq. 1). The learning rates for IMDN GAN and DBPN GAN models were trained with the same rates in the first phase 2 x 10^{-6} and 1 x 10^{-4}, respectively. All GAN models were trained with a batch size of 16 on two Nvidia (Santa Clara, CA, USA) P6000 GPUs for 60 epochs using the ADAM optimizer.[29]

Lastly, the third phase focused on training IMDN and DBPN models with color jitter augmentation on different datasets: brain, kidney, lung, or a mixed tissue set. The DBPN and IMDN models were optimized according to Eq. (1) or Eq. (2), respectively. The number of epochs, learning rate, batch size, optimizer, and GPU hardware used to train these IMDN and DBPN models were the same as above.

Model evaluation

Model performance was evaluated using the PSNR and SSIM of the test dataset across all tissue types. To calculate the PSNR and SSIM of each tile, the original tile from the 20X WSI was first downscaled to the 5X or 10X WSI by bicubic interpolation and then reconstructed back to a 20X WSI equivalent with tile size (1024 x 1024 pixels).

PSNR and SSIM provide valuable quantitative metrics for image quality evaluation and are widely used in image analysis. However, PSNR and SSIM cannot fully capture human perception of image quality. To investigate the correlation between PSNR, SSIM, and the human perception of

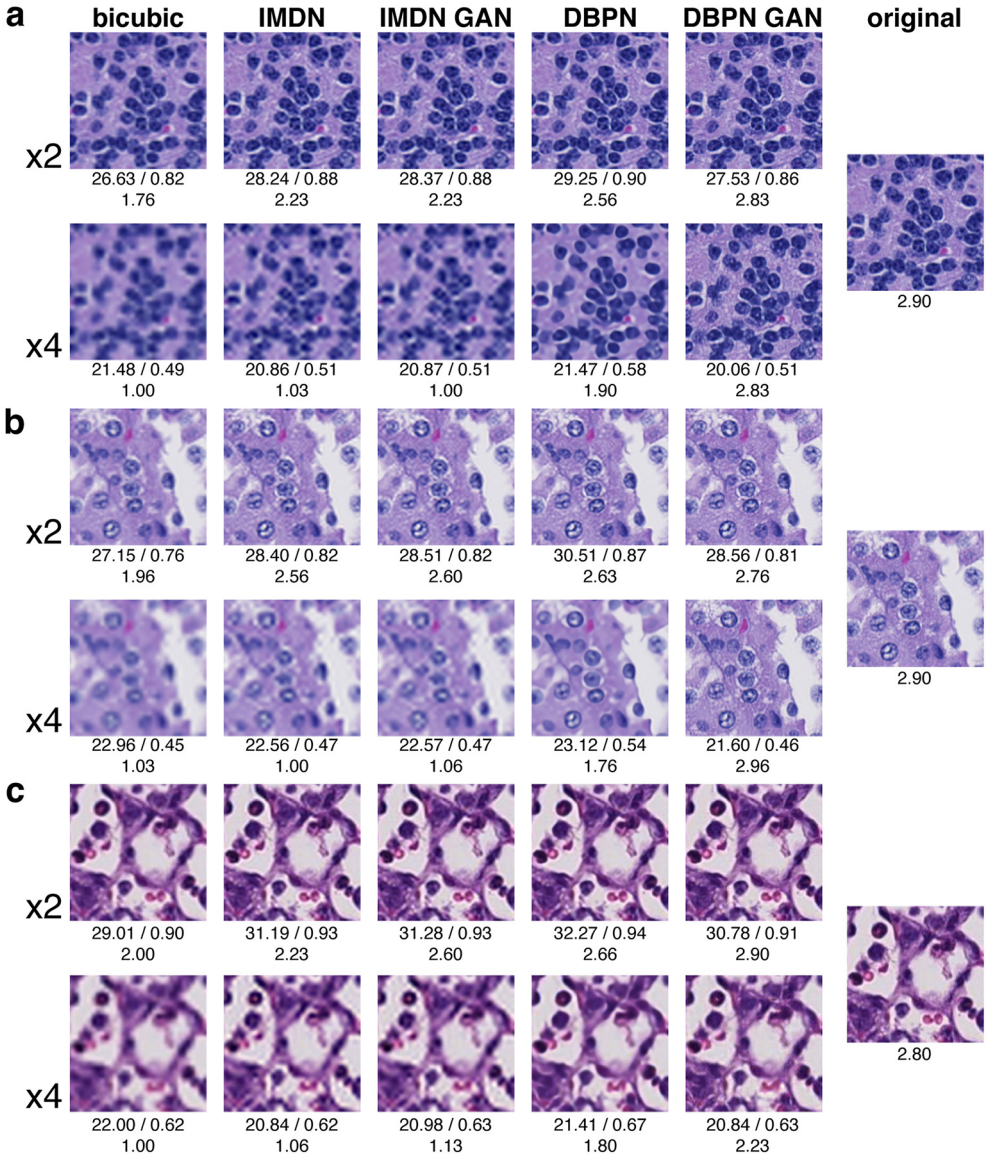

Fig. 3. Metrics of brain, kidney, and lung reconstructions in the image survey. DBPN GAN model generated images were rated the highest, despite the attributed lower PSNR and SSIM values. Captions under each tile denote PSNR, SSIM, mean rater score as PSNR / SSIM / mean rater score. Captions under the original column denote the mean rater score. a. Brain reconstructions. b. Kidney reconstructions. c. Lung reconstructions.

image quality in the digital pathology domain, an image quality survey of 31 human raters was conducted. Survey participants with varying experiences with H&E images from researcher with no pathology background to board-certified pathologists ($n = 5$) were asked to score perceived image quality on a scale from 1 to 3. The survey consisted of kidney models applied to all 11 tissue types. In addition, the survey included HR ground truths and LR image up scaled by bicubic interpolation. In total, each survey participant evaluated 121 tissue tiles. Spearman correlation was computed between the perceived image quality scores and the objective quality measures.

Single-tissue DBPN and IMDN models were trained with a combination of losses and color augmentation techniques using the kidney data and evaluated on the test set to determine the best training strategy and the best performing model. To investigate the utility of mixed-tissue versus single-tissue type training, the best performing models were trained with either multiple-tissue or single-tissue type datasets using the training strategy described above. Single-tissue models were trained on brain, kidney, and lung tissue. Multiple-tissue models were trained on all tissue types simultaneously. The single-tissue models were evaluated in 2 analyses. In the first analysis, the single-tissue model was evaluated on a test subset of the same tissue type. For example, the trained model which was only exposed with kidney data was used to generate HR images for only kidney data in the test set and the trained brain model was used to generate HR images for only brain data. The evaluation metrics were computed separately for individual tissues. In the second, converse analysis, the single-tissue model was evaluated on the tissue types it was not trained on. In other words, the model trained with kidney data was used to generate HR images from non-kidney LR images in the test sets. The evaluation metrics were computed for the 3 single tissue models for the converse analysis. Both analyses were conducted at 2X and 4X super-resolution factors.

Results

SISR model performance

Fig. 2 shows the mean PSNR and SSIM scores in the test set for 2X and 4X kidney models trained with different parameters and illustrates the effects of augmentation and adversarial loss to the base (null) DBPN and

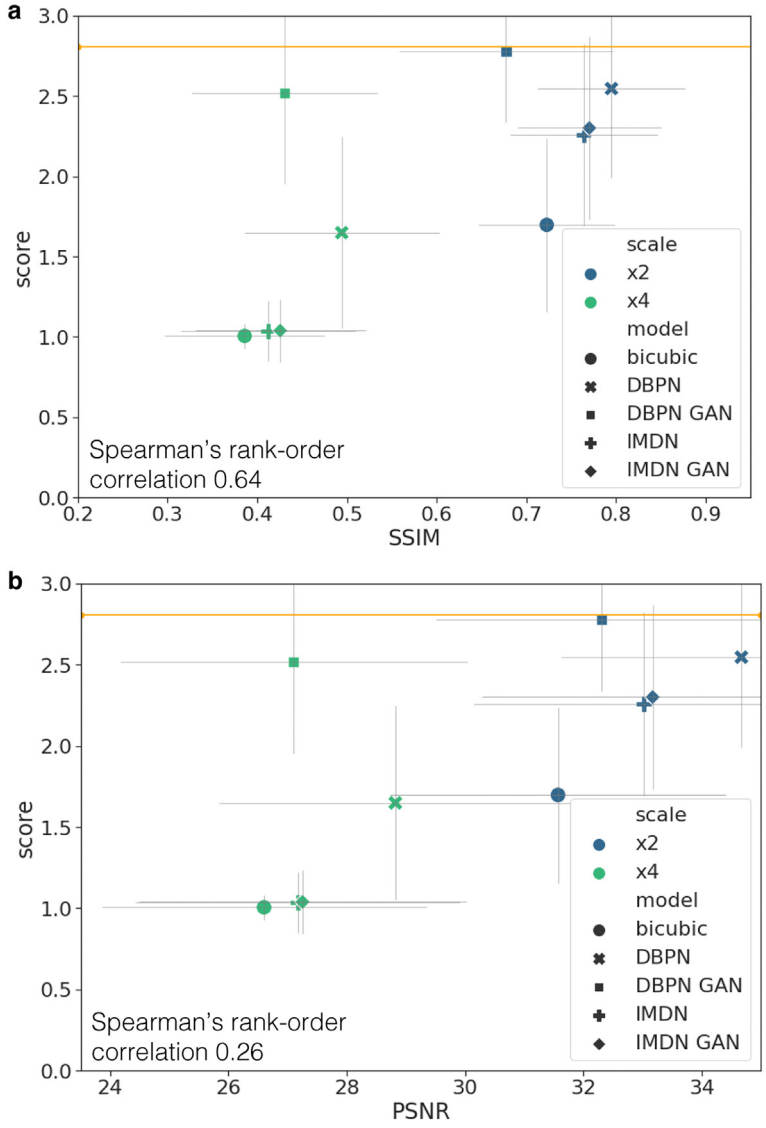

Fig. 4. Correlation of rater scores and model performance. The y-axis denotes the mean rater score and standard deviation of all the images reconstructed by the model used in the image survey. The x-axis denotes the model's mean and standard deviation SSIM and PSNR values derived from evaluation of the test set. Image reconstruction scores increase with the model's attributed SSIM and PSNR values. a. Mean rater score versus model SSIM. b. Mean rater score versus model PSNR. Orange horizontal lines show the mean rater scores of the ground truth images.

IMDN models. In this regard, all DBPN and IMDN models performed better than the bicubic interpolation baseline and the highest performing models for DBPN and IMDN at both scales were statistically significant for PSNR and SSIM (p-value <0.005) compared to their base (null) models (Supplemental Table S1).

Both DBPN and IMDN trained to upscale by 2X showed significant improvement to the base models in SSIM and PSNR when color jitter and mix-up augmentation was used (Supplemental Table S1). On average, the calculated PSNR/SSIM values of DBPN and IMDN 2X models for color jitter augmentation increased by 0.975/0.027 and 0.858/0.044, respectively and mix-up augmentations increased by 1.18/0.033 and 0.757/0.042, respectively.

Similarly, we observed an improvement in PSNR and SSIM compared to the base models trained to upscale by 4X. Color jitter augmentation resulted in the highest scores for both models and was significant (Supplemental Table S1). On average, the calculated PSNR/SSIM values of DBPN and IMDN 4X models for color jitter augmentation increased by 0.37/0.029 and 0.507/0.037, respectively, and mix-up augmentations increased by 0.07/0.006 and 0.256/0.015, respectively. The PSNR and SSIM metrics of the 4X models were lower than the 2X models, as expected. We conclude that the augmentation techniques explored for DBPN and IMDN are beneficial for super-resolution in terms of PSNR and SSIM metrics.

Additional fine tuning with a pre-trained VGG network as a preceptor (Fig. 1b) resulted in increased or decreased PSNR and SSIM scores that were dependent upon the neural network architecture (whether a pre-trained IMDN or DBPN was used). PSNR and SSIM decreased for all DBPN GAN models (Fig. 2). For example, DBPN GAN with color jitter augmentation decreased by -1.549 PSNR and -0.054 SSIM. In contrast, the IMDN GAN model showed an increase in these metrics for some cases such as when mix-up and color jitter augmentation were used during training.

Perceptual image quality on SSIM and PSNR

When super-resolved image quality was subjectively evaluated by human observers, the DBPN models were rated better than the IMDN models in the image survey. This agreed with the quantitative results as the DBPN models demonstrated higher PSNR and SSIM scores. Fig. 3 shows examples of the 5 reconstruction methods used to create image reconstructions for this survey across 3 tissue types (brain, kidney, and lung) along with the PSNR, SSIM, and averaged scores from raters. Across all tissue types, models trained with adversarial loss are preferred by human raters. In this set of image reconstructions, the average PSNR and SSIM across the three tissues for DBPN models were 26.34 and 0.75, respectively, which were higher than the DBPN GAN scores. Of note, the human rater scores suggested that the DBPN GAN model could super-resolve images with higher quality than the DBPN model (DBPN GAN average rater score: 2.75; DBPN average rater score: 2.21). These results suggest that PSNR and SSIM may not fully capture image quality preference by human evaluators.

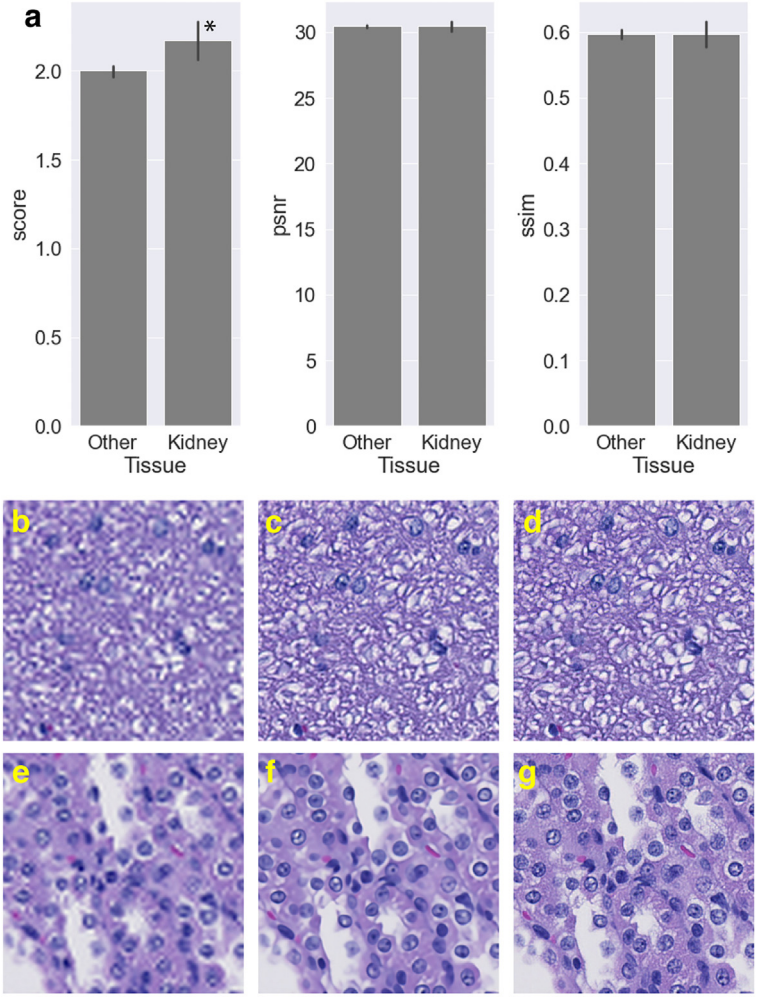

Fig. 5. Kidney compared to other tissues in the image survey. Kidney was rated the highest compared to other tissues in the image survey. a. Kidney compared to other tissues by rater score, PSNR, and SSIM. Shown as mean with 95% confidence interval. b–d. Example spinal cord image reconstructions (250 by 250 pixels), that were mostly rated as 1 (b), 2 (c), and 3 (d). e–g. Example kidney image reconstructions (250 by 250 pixels), that were mostly rated as 1 (e), 2 (f), and 3 (g).

From the image survey, SSIM correlated higher (R = 0.64) with image quality ratings given by human raters than PSNR (R = 0.26). Fig. 4 illustrates model PSNR and SSIM values plotted against rater scores (number of data points = 3630). Fig. 4a shows a plot of each model evaluated SSIM on the test set and the corresponding mean rater score. Spearman's rank-order correlation between rater score and model SSIM was strong with a significant *p*-value of 0.003. Fig. 4b shows a plot of each model PSNR and the corresponding mean rater score. Spearman's rank-order correlation between human rater score and model PSNR was significant ($p<<0.01$) with a *p*-value 2.1 x 10^{-56}, suggesting that the higher correlation value of SSIM to perceptual rating could be a better proxy measure for human preference than PSNR. This is likely due to SSIM captures local, structural information better than PSNR.

For simplicity, a single-tissue kidney model was used to generate all HR images for the images survey instead of using a model trained on multiple tissues. As expected, the kidney tissue reconstructed image was rated the highest (Fig. 5a, Supplemental Figure S1) since the models have been trained with those data. A Whitney–Mann rank-sum test was computed and kidney reconstructions were significantly preferred to non-kidney reconstructions with a *p*-value 0.002. However, the kidney model demonstrated promising reconstructions for other tissues like brain, lung, and urinary bladder with average score greater than 2 (Supplemental Figure S1). In Fig. 5, we illustrate low, medium, and high scoring super-resolved images of kidney (Fig. 5b–d) and spinal cord (Fig. 5e–g).

Single-tissue versus mixed-tissue models

Sample single-tissue models (brain, kidney, or lung) and a mixed-tissue model evaluated on example brain, kidney, lung, and urinary bladder tiles at 2X generated images with quality scores that differed by less than 15.6% (PSNR) and 12.4% (SSIM) from each other. Supplemental Figures 2 & 3 illustrate the DBPN image reconstruction improvements over the bicubic interpolation method for 2X and 4X scales on these example tissues. For instance, upscaling of a lung image patch by 2X resulted in PSNR and SSIM values that were higher in both the lung model (33.92/0.94) and the mixed-tissue model (33.61/0.93) compared to bicubic (29.68/0.88). At 4X a similar trend was observed with PSNR but SSIM had a larger spread; PSNR and SSIM of images reconstructed by single-tissue or mixed-tissue models were within 12.2% and 33.3% of each other on these example tissue tiles, respectively (see Fig. 6 for example lung and kidney tissue

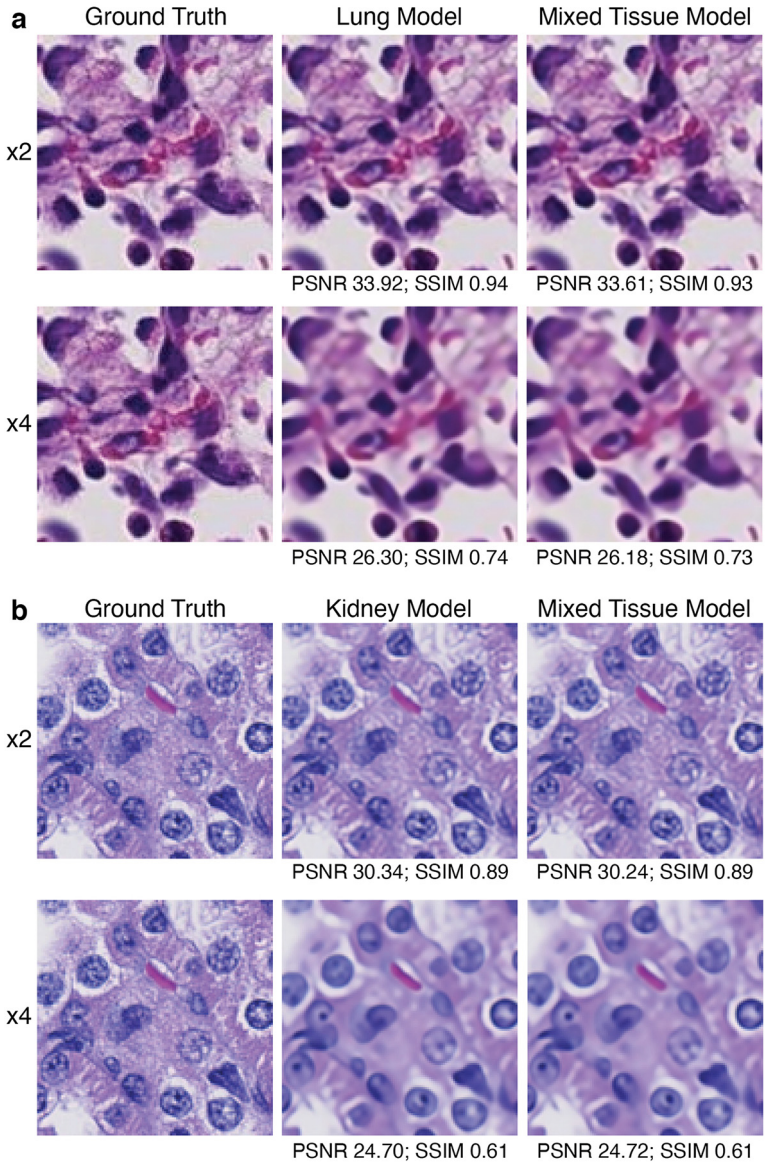

Fig. 6. Image reconstruction comparison between single-tissue and mixed-tissue DBPN model. Examples to illustrate the similarity of reconstructed images by either a single-tissue or mixed-tissue DPBN model. a. 20X lung image patches b. 20X kidney image patches.

Table 2
Summary of SSIM performance by models trained with single-tissue versus mixed-tissues.

	DBPN (x2 / x4)		IMDN (x2 / x4)	
	Single tissue	Mixed tissue	Single tissue	Mixed tissue
Lung	0.90 / 0.66	0.88 / 0.65	0.85 / 0.54	0.84 / 0.50
Brain	0.83 / 0.41	0.83 / 0.41	0.80 / 0.34	0.80 / 0.35
Kidney	0.87 / 0.60	0.87 / 0.60	0.83 / 0.50	0.82 / 0.49

The mean SSIM was calculated from brain, lung, or kidney tiles in the test set. The single tissue model column corresponds with either the lung, brain, or kidney trained model.

Table 3
Summary of PSNR performance by models trained with single versus mixed tissues.

	DBPN (x2 / x4)		IMDN (x2 / x4)	
	Single tissue	Mixed tissue	Single tissue	Mixed tissue
Lung	36.86 / 29.09	36.42 / 28.94	34.58 / 26.30	34.15 / 25.81
Brain	32.85 / 27.69	32.62 / 27.64	31.96 / 26.67	31.88 / 26.69
Kidney	33.94 / 28.19	33.87 / 28.18	31.94 / 26.18	31.59 / 26.04

The mean PSNR was calculated from brain, lung, or kidney tiles in the test set. The single tissue model column corresponds with either the lung, brain, or kidney trained model.

reconstructions to compare images generated using a single-tissue or mixed-tissue model).

A super-resolved urinary bladder tile had PSNR and SSIM within 4.9% and 4.7% of each other using single-tissue versus a mixed-tissue model at 2X (PSNR Mixed Model = 31.83 and PSNR Lung Model = 30.31; SSIM Mixed Model = 0.87 and SSIM Lung Model = 0.83). Similarly, at 4X, the PSNR and SSIM were within 5.3% and 9.2% of each other using a single-tissue versus mixed-tissue model (PSNR Mixed Model = 26.51 and PSNR Brain Model = 26.15; SSIM Mixed Model = 0.57 and SSIM Lung Model 0.52). These results suggest that both single-tissue and mixed-tissue training approaches are viable when developing SISR models for many tissue types.

Table 2 summarizes the calculated SSIM of brain, lung, and kidney image tiles in the test set. The average SSIM of the lung, brain, and kidney tiles reconstructed by either a single-tissue or a mixed-tissue model were within 8.1% for DPBN and 6% for IMDN at 2X and within 46.7% for DPBN and 45.4% for IMDN at 4X. In addition, single-tissue models evaluated on their respective image tiles (i.e., lung model evaluated on lung tiles) were within 2.2% or equal in SSIM compared to the mixed-tissue model.

Table 3 summarizes the calculated PSNR of brain, lung, and kidney image tiles in the test set. The average PSNR of the lung, brain, and kidney tiles reconstructed by either a single-tissue or a mixed-tissue model were within 12.2% for DBPN and 9% for IMDN at 2X and within 5% for DPBN and 3.4% for IMDN at 4X. In addition, single-tissue models evaluated on their respective image tiles were within 1.2% in PSNR compared to the mixed-tissue model at either scale or model architecture (DPBN or IMDN).

A detailed breakdown of the calculated SSIM and PSNR of the other tissues in the test set by these super-resolution models are summarized in Supplemental Tables S2 and S3 for 2X and 4X results, respectively. In general, both single-tissue and mixed-tissue training are viable when building SISR models.

Conclusions

Image resolution plays an important role in pathology but can be constrained in digital workflows. However, resolution of the digitized slides can be enhanced with computational methods such as SISR algorithms. The results from this work inform SISR model improvement for H&E WSI with supervised training strategies that involve data augmentation. Additionally, we demonstrate that the use of PSNR and SSIM metrics was not an ideal surrogate for human-perceived image quality. The proposed methods can be used to virtually magnify H&E images with better perceptual quality than interpolation methods (i.e., bicubic) commonly implemented in digital pathology viewers. The impact of deep SISR methods is more notable when scaling to 4X is needed, such as in the case of super-resolving a low magnification WSI from 10X to 40X.

In this study, DBPN GAN models achieved lower PSNR and SSIM values than their non-GAN counterparts (Fig. 2), but the DBPN GAN models reached higher perceived quality scores provided by human raters and the highest perceived image score was achieved by DBPN GAN (Fig. 4). These data suggest that the DBPN GAN models learned features in H&E WSI that humans use for image quality evaluation. These features cannot be entirely quantified by SSIM or PSNR. Consistent with our findings, Ledig et al[32]. performed a mean opinion score test to quantify the ability of various super-resolution algorithms (GAN-based and non-GAN-based) approaches to reconstruct perceptually convincing images and observed that GAN models achieved higher mean opinion scores while demonstrated lower PSNR and SSIM scores in various datasets.

Although an imperfect reflection of human perception scores, and in spite of the findings of Ledig et al. that SSIM and PSNR were not optimal for evaluating image quality,[32] SSIM was significantly correlated with perceptual image quality scored by human raters in our work (Fig. 4). Therefore, we used SSIM to evaluate the performance of single-tissue and mixed-tissue training in select examples (Tables 2 and 3). The 2 approaches resulted in image reconstructions that differed by less than 12.4% SSIM at 2X. Depending upon the goals and constraints of a particular project, our results suggest that either approach may be appropriate. For example, a single tissue DBPN GAN model could be suitable to reconstruct other tissue types (Fig. 3).

Two aspects of our study that may limit the generalizability of our results include: (1) limited training data, and (2) the 20X target magnification. In all model training, we limit the training to 15 000 WSI tiles to control across all training experiments. While we saw the best results with DBPN GAN, the image reconstructions could possibly be improved with additional training data as deep learning models are notoriously data hungry. While our target magnification was 20X, we cannot say with certainty that the same training principles detailed herein could be applied to higher magnifications. As 40X scanners become more common, more training data will be available to help answer this question. In the future, we hope to gather a large enough dataset of 40X magnification to determine if our method generalizes to higher target resolution.

We present a deep SISR framework to obtain high perceived image quality of WSI reconstructions. These data augmentation methods could be used to extend future model architectures that may emerge. This work adds to the limited research done on developing deep SISR methods in digital pathology by highlighting the limitations of using PSNR and SSIM as a proxy to perceived image quality. We demonstrate that optimizing deep SISR models with GAN loss is promising and additional controlled studies of increasing the training set are warranted. Future work will investigate augmenting the digital slide review process by a pathologist with SISR methods to bring more visual clarity to low magnification slides.

Authors' Contributions

CM - conception and design, data acquisition, analysis and interpretation of data, drafting the manuscript or revising it critically for important intellectual content

PZ - data preprocessing and augmentation code, preliminary experiments, conception and design, analysis and interpretation of data, drafting the manuscript or revising it critically for important intellectual content

SK - conception and design, execution of preliminary experiments

RS - conception and design, drafting the manuscript or revising it critically for important intellectual content, given final approval of the version to be published

FH - conception and design, drafting the manuscript or revising it critically for important intellectual content, given final approval of the version to be published

Funding

This work was supported by Genetech Inc.

Conflicting Interest

The authors declare no conflict of interest

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Fangyao Hu reports a relationship with Genentech Inc that includes: employment. Cyrus Manuel reports a relationship with Genentech Inc that includes: employment. Sertan Kaya reports a relationship with Genentech Inc that includes: employment. Ruth Sullivan reports a relationship with Genentech Inc that includes: employment.

Acknowledgments

Safety Assessment and Development Sciences Informatics project sponsors.

Our Roche and Genentech colleagues who participated in our perceptual image survey.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jpi.2022.100148.

References

1. Bashir SMA, Wang Y, Khan M, Niu Y. A comprehensive review of deep learning-based single image super-resolution. Peeri Comput Sci 2021:e621. https://doi.org/10.7717/peerj-cs.621.
2. Glasner D, Bagon S, Irani M. Super-resolution from a single image. IEEE 12th Int Co nf Comput Vis; 2009. p. 349–356. https://doi.org/10.1109/iccv.2009.5459271. Conference: Location.
3. Pashaei M, Starek MJ, Kamangir H, Berryhill J. Deep learning-based single image super-resolution: an investigation for dense scene reconstruction with UAS photogrammetry. Remote Sens-Basel 2020;12:1757. https://doi.org/10.3390/rs12111757.
4. Xu W, Xu G, Wang Y, Sun X, Lin D, Wu Y. High Quality remote sensing image super-resolution using deep memory connected network. Arxiv 2020. https://doi.org/10.1109/igarss.2018.8518855.
5. Chen Y, Shi F, Christodoulou AG, Zhou Z, Xie Y, Li D. Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network. ArXiv 2018. https://doi.org/10.48550/arXiv.1803.01417.
6. Mithra KKM, Ramanarayanan S, Ram K, Sivaprakasam M. Reference-based texture transfer for single image super-resolution o f magnetic resonance images. IEEE 18th Int. Symposium Biomed Imaging (ISBI); 2021. p. 579–583. https://doi.org/10.48550/arXiv.2102.05450. Conference: Location.
7. Khan M, Khan MA, Obaid F, Jadoon S, Khan MA, Sikandar M. A novel multi-frame super resolution algorithm for surveillance camera image reco nstruction. First International Conference on Anti-Cybercrime (ICACC); 2015. https://doi.org/10.1109/Anti-Cybercrime.2015.7351950. Conference: Location.
8. Zhu Y, Zhang Y, Yuille AL. Single image super-resolution using deformble patches. IEEE Co nf Comput Vis Pattern Recognit; 2014. p. 2917–2924. https://doi.org/10.1109/cvpr.2014.373. Conference: Location.
9. Yang J, Lin Z, Cohen S. Fast image super-resolution based on in-place example regressi on. IEEE Conf Comput Vis Pattern Recognit; 2013. p. 1059–1066. https://doi.org/10.1109/cvpr.2013.141. Conference: Location.
10. Haris M, Shakhanarovich G, Ukita N. Deep back-projection networks for single image super-resolution. IEEE T Pattern Anal Machine Intelligence 2021;43:4323–4337. https://doi.org/10.1109/TPAMI.2020.3002836.
11. Sun K, Gao Y, Xie T, Wang X, Yang Q, Chen L, et al. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. ArXiv 2021. https://doi.org/10.48550/arXiv.2105.07200.
12. Çelik G, Talu MF. Resizing and cleaning of histopathological images using generative adversarial networks. Phys A Stat Mech Appl 2020:554. https://doi.org/10.1016/j.physa.2019.1226.
13. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. ArXiv 2014. https://doi.org/10.48550/arXiv.1406.2661.
14. Hua X, Cai Y, Zhou Y, Yan F, Cao X. Leukocyte super-resolution via geometry prior and structural consistency. J Biomed Opt 2020:25. https://doi.org/10.1117/1.JBO.25.10.106501.
15. Li B, Keikhosravi A, Loeffler AG, Eliceiri KW. Single image super-resolution for whole slide image using convolutional neural networks and self-supervised color normalization. Med Image Anal 2021;68, 101938. https://doi.org/10.1016/j.media.2020.101938.
16. Ma J, Yu J, Liu S, Chen L, Li X, Feng J, et al. PathSRGAN: multi-supervised super-resolution for cytopathological images using generative adversarial network. IEEE Trans Med Imaging 2020;39:2920–2930. https://doi.org/10.1109/TMI.2020.2980839.
17. Mukherjee L, Bui HD, Keikhosravi A, Loeffler A, Eliceiri K. Super-resolution recurrent convolutional neural networks for learning with multi-resolution whole slide images. J Biomed Opt 2019;24:1-15. https://doi.org/10.1117/1.JBO.24.12.126003.
18. Singh A, Ohgami RS. Super-resolution digital pathology image processing of bone marrow aspirate and cytology smears and tissue sections. J Pathol Inform 2018;9:48. https://doi.org/10.4103/jpi.jpi_56_18.
19. Feng R, Gu J, Qiao Y, Dong C. Suppressing Model Overfitting for Image Super-Resolution Networks 2019. https://doi.org/10.48550/arXiv.1906.04809.
20. Yoo J, Ahn N, Sohn K-A. Rethinking data augmentation for image super-resolution: a comprehensive analysis and a new strategy. ArXiv 2020. https://doi.org/10.48550/arXiv.2004.00448.
21. Wang Z, Bovik AC. Mean squared error: Love it or leave it? A new look at signal fidelity measures. IEEE Signal Proc Mag 2009;26:98-117. https://doi.org/10.1109/MSP.2008.930649.
22. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE T Image Process 2004;13:600–612. https://doi.org/10.1109/TIP.2003.819861.
23. Mason A, Rioux J, Clarke SE, Costa A, Schmidt M, Keough V, et al. Comparison of objective image quality metrics to expert radiologists' scoring of diagnostic quality of MR images. IEEE Trans Med Imaging 2020;39:1064–1072. https://doi.org/10.1109/TMI.2019.2930338.
24. Wan W, Wu J, Shi G, Li Y, Dong W. Super-resolution quality assessment subjective evaluation database and quality index based on perpetual structure measurement. 2018 IEEE Int Conf Multimedia Expo Icme (ICME); 2018. p. 1–6. https://doi.org/10.1109/ICME.2018.8486519. Conference: Location.
25. Cheng Z, Akyazi P, Sun H, Katto J, Ebrahimi T. Perceptual quality study on deep learning based image compression. ArXiv 2019. https://doi.org/10.48550/arXiv.1905.03951.
26. Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. ArXiv 2018. https://doi.org/10.48550/arXiv.1801.03924.
27. Deng Y, Feng M, Jiang Y, Zhou Y, Qin H, Xiang F, et al. Development of pathological reconstructed high-resolution images using artificial intelligence based on whole slide image. MedComm 2020;1:410–417. https://doi.org/10.1002/mco2.39.
28. Hui Z, Gao X, Yang Y, Wang X. Lightweight image super-resolution with information multi-distillation network. Proceedings of the 27th ACM Inter Conf on Multimedia; 2019. p. 2024–2032. https://doi.org/10.1145/3343031.3351084.Conference: Location.
29. Kingma DP, Ba J. Adam: a method for stochastic optimization. ArXiv 2014. https://doi.org/10.48550/arXiv.1412.6980.
30. Simoyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv 2014. https://doi.org/10.48550/arXiv.1409.1556.
31. Deng J, Dong W, Socher R, Li L-J, Li KF-FL. ImageNet: a large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition; 2009. https://doi.org/10.1109/CVPR.2009.5206848. Conference: Location.
32. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, et al. Photo-realistic single image super-resolution using a generative adversarial network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Conference; 2017. https://doi.org/10.1109/CVPR.2017.19. Conference: Location.
33. Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016. p. 2414–2423. https://doi.org/10.1109/CVPR.2016.265. Conference: Location.