Optics Letters

Fiber bundle image restoration using deep learning

Jianbo Shao,[1,2] Junchao Zhang,[1,3] Xiao Huang,[1,4] Rongguang Liang,[1,*] and Kobus Barnard[2,5]

[1]College of Optical Sciences, University of Arizona, Tucson, Arizona 85721, USA
[2]Department of Electrical and Computer Engineering, University of Arizona, Tucson, Arizona 85721, USA
[3]Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, Liaoning Province 110016, China
[4]College of Optical Science and Engineering, Zhejiang University, Hangzhou, Zhejiang Province 310027, China
[5]Department of Computer Science, University of Arizona, Tucson, Arizona 85721, USA
*Corresponding author: rliang@optics.arizona.edu

Received 15 November 2018; revised 16 January 2019; accepted 19 January 2019; posted 22 January 2019 (Doc. ID 352057); published 19 February 2019

We propose a deep learning-based restoration method to remove honeycomb patterns and improve resolution for fiber bundle (FB) images. By building and calibrating a dual-sensor imaging system, we capture FB images and corresponding ground truth data to train the network. Images without fiber bundle fixed patterns are restored from raw FB images as direct inputs, and spatial resolution is significantly enhanced for the trained sample type. We also construct the brightness mapping between the two image types for the effective use of all data, providing the ability to output images of the expected brightness. We evaluate our framework with data obtained from lens tissues and human histological specimens using both objective and subjective measures.

https://doi.org/10.1364/OL.44.001080

An imaging fiber bundle (FB) contains thousands of single fiber cores which can relay images from remote regions to digital sensors. Due to its unique flexible feature, the FB has been extensively used in medical endoscopy [1]. However, its geometric nature introduces honeycomb-like fixed patterns on its output images. In addition, effective spatial resolution of an FB imaging system is limited by the individual fiber core diameter and fiber density, rather than the optical system and camera sensor. Therefore, there is a critical need to remove honeycomb patterns and improve spatial resolution.

We classify existing methods based on whether a single FB image is used for input or multiple FB images are used for input. In the case of single FB image input, initial methods [2,3] include spatial and frequency domain filtering and interpolation, together with prior learning of the FB structure. These methods remove the fixed pattern, but spatial resolution is not substantively improved. Using multiple images as input, existing methods digitally register FB images first and reconstruct high-resolution images by calculating median images [4] or using maximum-a-posterior (MAP) estimation [5], leading to gains in resolution due to the additional information available in multiple images.

Recently, Ravi *et al.* [6] reported the first attempt to translate neural network methods into FB imaging. They first removed honeycomb patterns using interpolation and then applied neural networks to learn mappings from restored FB images to ground truth (GT) images for further resolution enhancement. Due to the lack of actual image data, they estimated pseudo GT data for FB images by registering multiple frames from a microendoscope, and then generated FB images from the pseudo GT data. The accuracy of the pseudo GT image generation depends heavily on image registration, which can be computationally costly and erroneous as pointed out by the authors. Thus, efficiently obtaining accurately matched pairs of FB images and their "real" GT data is critical for applying neural networks into FB imaging domain.

In this Letter, we propose a method for restoring FB images using deep learning. To acquire well-registered GT and FB images simultaneously, we built a dual-sensor imaging system. We propose a generative adversarial restoration neural network (GARNN) to learn a direct mapping from FB images to their corresponding GT data. We normalize brightness for training to make the best use of our data. To restore images, we similarly normalize network input and reverse the normalization on output to get approximately correct brightness. We report experiments using lens tissues and three types of human histological specimens. We find that we can remove the fixed-pattern noise completely, and that hidden details are also significantly recovered when the training and testing images are from the same type. We compare the GARNN with two state-of-the-art restoration neural networks and our MAP-based method [5] with single image input. Both objective and subjective image quality measures suggest that the GARNN reconstructs sharper images free of the fixed pattern. We also evaluate our system using cross-brightness experiments to show the robustness of our model under varying illuminations.

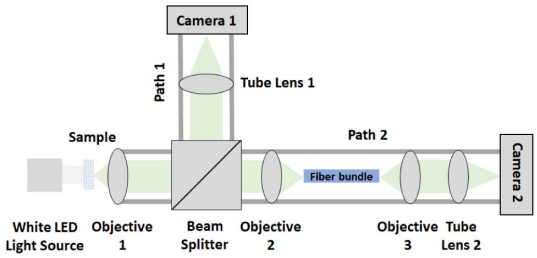

Fig. 1. Dual-sensor imaging system for capturing raw FB images and corresponding GT data for the training neural network.

Figure 1 shows our dual-sensor imaging system for capturing "one-to-one" pairs of FB and GT images. We used a silica-based FB (FIGH-30-650S, Fujikura) containing $30,000 \pm 3,000$ cores with 600 ± 30 μm diameter and three identical 0.25 NA microscope objectives. To obtain GT images and FB images simultaneously, we added a beam splitter to split the light from Objective 1 into two paths. Path 1 is a typical microscope configuration consisting of Objective 1, Tube Lens 1, and Camera 1, for HR imaging. In Path 2, Objective 2 focuses the light onto the FB, and Objective 2 and Tube lens 2 image the output surface of the FB to Camera 2 for FB imaging. Since Objectives 1 and 2 form a perfect 1:1 relay system imaging the input surface of the FB to the object directly, Path 2 is optically equivalent to the imaging condition where the FB contacts the object for FB imaging. We aligned the system using a 1951 USAF target by observing overlaid images from the two cameras. Specifically, we adjusted the relative position between the FB and Objective 2 until the two images were aligned to less than one pixel.

Figure 2 shows the architecture of the GARNN. It consists of a generative network and a discriminative network. We construct the generative network by following a state-of-the-art restoration neural network design [7]. Block 1 serves as the input layer of the FB image. It has 64 convolutional (Conv) filters with a kernel size of 3×3 and uses rectified linear units (ReLU) [8] for the activation functions. Blocks 2 to 16 contain 64 filters with a kernel size of $3 \times 3 \times 64$. Batch normalization (BN) is added between Conv and ReLU for faster convergence [9]. The last Block, 17, outputs the restored clean image, and it contains one Conv filter with a size of $3 \times 3 \times 64$.

The objective function of this generative network, $\mathcal{G}$, is the content loss between FB and GT images given by

$$L_{\text{Content}} = \frac{1}{N}\sum \|\boldsymbol{X} - \mathcal{G}(\boldsymbol{G})\|_2^2, \tag{1}$$

where $\|\cdot\|_2^2$ denotes L_2 norm, $\boldsymbol{X}$ is GT data, and $\mathcal{G}(\boldsymbol{G})$ represents the generative network output restored image with FB image $\boldsymbol{G}$ as input. N is the number of image pairs for one training iteration. However, by using content loss alone, this approach usually leads to unwanted over-smoothing, where output images lack high frequency information [10].

Thus, to obtain sharper and more realistic restored images, we add a discriminative network by following the design proposed by Ledig *et al.* [10] to perform adversarial learning. This discriminative network is trained to differentiate between restored and GT images, and helps the generative network learn to output images that are more similar to GT data. In this discriminative network, Block 1 is the input layer, which contains 64 Conv filters and uses Leaky ReLU [11] as its activation function. The layers in Blocks 2 to 8 have increasing numbers of filters (kernel size 3×3) with strided convolution. BN is added between Conv and leaky ReLu to increase training speed. Block 9 is the output block, consisting of two dense layers with Leaky ReLU and a sigmoid activation function.

The loss function of this discriminative network, $\mathcal{D}$, is:

$$L_{\mathcal{D}} = \mathbb{E}[\mathcal{D}(\boldsymbol{X}), 1] + \mathbb{E}[\mathcal{D}(\mathcal{G}(\boldsymbol{G})), 0], \tag{2}$$

where $\mathbb{E}$ denotes the binary cross-entropy. Specifically, $\mathbb{E}[p, q] = q\log(p) + (1-q)\log(1-p)$ and $\mathbb{E}[\mathcal{D}(\boldsymbol{X}), 1]$ is the binary cross-entropy between the discriminator's prediction on the GT image and the desired label 1. $\mathbb{E}[\mathcal{D}(\mathcal{G}(\boldsymbol{G})), 0]$ is the binary cross-entropy between the discriminator's decision on a restored image estimated from the generative network and the desired label 0.

Finally, by adding the content loss and adversarial loss, we formulate the generative loss function of our GARNN as

$$L_{\mathcal{G}} = L_{\text{Content}} + \beta\mathbb{E}[\mathcal{D}(\mathcal{G}(\boldsymbol{G})), 1], \tag{3}$$

where $\mathbb{E}[\mathcal{D}(\mathcal{G}(\boldsymbol{G})), 1]$ is the adversarial loss term. We calculate it as the binary cross-entropy between the discriminator's prediction on the restored image and label 1 so that our generative network restored image output is encouraged to be like te GT data. β is a tunable weighting parameter that depends on the

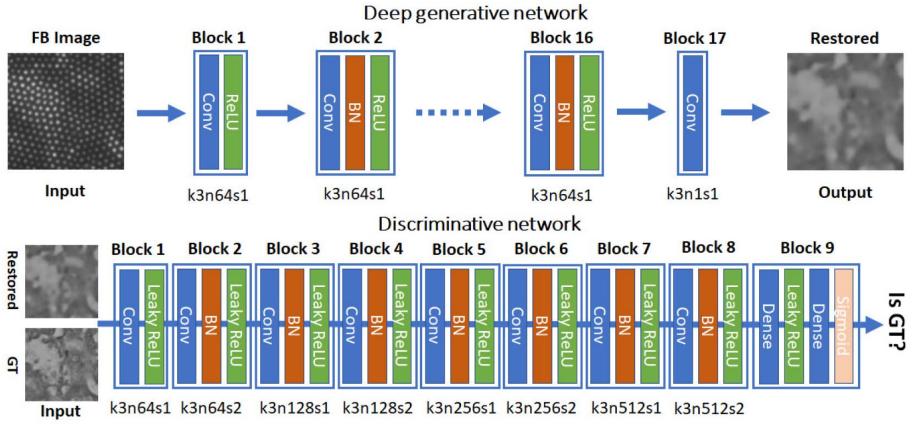

Fig. 2. Architecture of the GARNN. k is the kernel size, n represents the number of filters, and s is the stride size for each Conv layer.

content loss data range, which varies for different kinds of samples. During training, we first initialize the network by only minimizing the content loss by updating the trainable parameters in network $\mathcal{G}$. Then we alternately minimize $L_{\mathcal{D}}$ and $L_{\mathcal{G}}$, with trainable parameters in both $\mathcal{G}$ and $\mathcal{D}$ being updated. In testing, we use the trained network $\mathcal{G}$.

To be able to experiment with different illumination conditions, we captured four image pairs for each physical region of the sample by adjusting the power of an external white LED light source. In other words, each sample region has image pairs under four different brightness levels. Initial experiments demonstrated that normalization is more effective for handling images of different brightness levels compared to training with images of different brightness levels. Hence, in all reported experiments, we use the brightness normalization method described next. Such normalization is also called image whitening and has improved performance in some neural network tasks [12].

To normalize the brightness of captured images, we subtract off their mean intensity, μ, followed by dividing by their standard deviation (STD) σ of intensity. In training, we also fit the two mapping relationships ($\mu_{\mathbf{FB}} \rightarrow \mu_{\mathbf{GT}}$ and $\sigma_{\mathbf{FB}} \rightarrow \sigma_{\mathbf{GT}}$). Figure 3 shows the mapping for the lens tissue experiment. The red lines show the fitted regression model for μ and σ. In testing, we normalize the test image, recording its μ and σ, and then feed the normalized image into the network, to get an estimated normalized restored image. We then adjust the brightness of that output using the input image's μ and σ, and the mapping fit during training.

We first evaluated our method on lens tissue data. We captured 800 × 800 image pairs for 79 (training) and 13 (testing) distinct regions at four illumination levels. We normalized these images and then extracted patches with a size 40 × 40. Each patch had a 50% overlap with neighboring ones giving 39 × 39 windows per region and 480636 total training windows. For each epoch, we broke the set of windows into disjoint subsets of size 128 (batch size). Each batch was the input for an iteration of minimizing the loss function. We initialize our network by only training the generative network with content loss for the first 50 epochs; then we alternately train both discriminative and generative networks for another 50 epochs. β was set to 0.15 for the lens tissue study.

We compared the GARNN with two state-of-the-art restoration neural networks: (1) RED30 [13], which is a very deep residual encoder-decoder network; and (2) a generative restoration neural network (GRNN) [7]. GRNN only minimizes the content loss in Eq. (3), and it can be considered as a special case for the GARNN where $\beta = 0$. We also provide results from our MAP method [5]. Despite being designed for multiple FB frames, this method can efficiently remove fixed-pattern noises and recover limited details for a single FB image due to its Laplacian smoothing prior.

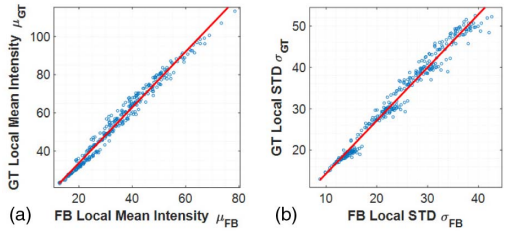

Fig. 3. Plots for μ and σ mapping relationships from FB to GT images in a lens tissue training dataset: (a) mean intensity μ mapping and (b) STD σ mapping. Each blue dot is from one captured image. The red lines mark the predicted linear models of μ and σ.

Table 1. Average PSNR, SSIM, and MOS Results for the Lens Tissue Test Dataset

	FB	MAP	RED30	GRNN	GARNN
PSNR	19.4 dB	17.5 dB	**31.8** dB	31.7 dB	31.4 dB
SSIM	0.54	0.80	**0.89**	**0.89**	0.87
MOS	N/A	1.37	3.38	3.32	**3.80**

In Table 1, we use a peak signal-to-noise ratio (PSNR) and a structural similarity index (SSIM) [14] to objectively measure differences. Specifically, $\mathrm{PSNR} = 20\log_{10}(I_{\max}/\sqrt{\mathrm{MSE}})$, where $I_{\max}$ is the possible largest pixel value, and MSE is the mean squared error between a noise-free image (GT image) and its noisy measurement (FB or restored FB image). RED30 and GRNN both achieve large PSNR and SSIM gains, and the GARNN has a slightly smaller PSNR and SSIM.

As noted by others (e.g., [10]), the PSNR or SSIM is not necessarily a good proxy for perceptual image quality. Hence, we conducted mean opinion score (MOS) studies. We asked 10 raters to score all the restored images with an integral grade from 1 (worst) to 5 (best) based on sharpnesses and visual image fidelity compared to GT data. The last row of Table 1 provides average MOS results. Our GARNN achieves the best MOS among all three methods, since the GARNN is benefiting from adversarial learning to output images more visually similar to GT data [10].

Figure 4 shows the test results from one lens tissue sample region. Images in (a) and (b) represent matched FB and GT image pairs which are under four different illumination brightnesses. Images in (c)–(f) show results from the MAP, RED30, GRNN, and GARNN methods, respectively. All four methods can efficiently remove fixed-pattern noises in FB images, but images from the MAP have the poorest visual quality. Furthermore, the images suggest that the GARNN is able to reduce blur and reveal finer details than RED30 and GRNN.

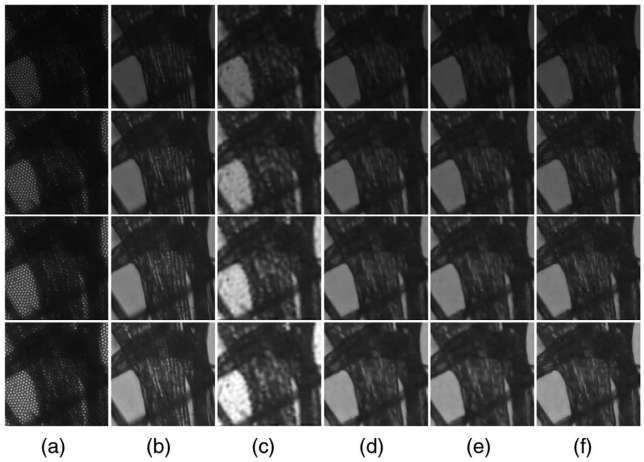

Fig. 4. Experimental results with a lens tissue under four illuminations conditions: (a) raw FB images, (b) GT images, (c) results from MAP, (d) results from RED30, (e) results from GRNN, and (f) results from the GARNN. The brightness increases from top to bottom.

Table 2. Cross-Validation Experiments for a Lens Tissue Sample

	Level 1	Level 2	Level 3	Level 4	Mean
TAF	35.1 dB	32.4 dB	30.1 dB	28.1 dB	31.4 dB
LOOF	35.1 dB	32.1 dB	30.0 dB	27.7 dB	31.2 dB

Table 3. PSNR, SSIM, and MOS Results for Human Specimens Test Dataset, Trained Using a Combined Liver and Kidney Data Set

		FB	MAP	RED30	GRNN	GARNN
Kidney	PSNR	15.9 dB	18.7 dB	**30.1** dB	30.0 dB	29.8 dB
	SSIM	0.33	0.77	**0.83**	**0.83**	0.82
	MOS	N/A	1.31	2.93	2.95	**3.81**
Liver	PSNR	16.6 dB	18.6 dB	**30.3** dB	**30.3** dB	30.2 dB
	SSIM	0.34	0.77	**0.84**	**0.84**	0.83
	MOS	N/A	1.23	2.83	2.80	**3.86**
Tonsil	PSNR	17.1 dB	24.6 dB	**25.7** dB	**25.7** dB	25.5 dB
	SSIM	0.42	0.69	**0.72**	**0.72**	**0.72**
	MOS	N/A	1.40	2.55	2.59	**3.07**

Next, we evaluate our GARNN under different illumination brightness conditions by conducting leave-one-out-of-four (LOOF) cross-validation experiments. Each time we train only on images from three brightness levels and test on the images with a left-out level. We compare this performance to training on all four (TAF) illumination conditions. The PSNR results for all levels (1 for darkest and 4 for brightest) are shown in Table 2. We see that our model is insensitive to different brightness conditions.

Finally, we trained our framework on human histological specimens. Similar to the lens tissue experiment, we captured FB and GT images from 73 (training) and 12 (testing) regions for both stained human kidney and liver slides, again at four different illumination conditions. We combined liver and kidney training data to get 584 training images to study training on multiple types. We also captured images for 12 regions of a tonsil tissue only for testing. We set β to 0.6. All other training parameters were the same as in the lens tissue experiments.

In Table 3, we show the average PSNR, SSIM, and MOS results for each method and sample type using the liver-kidney training data. All three neural network methods achieve significant PSNR and SSIM improvements for all three test samples, although, not surprisingly, the cross-type experiments on the tonsil show less gains than the two that have representation in the training data. The MAP method generally has less PSNR improvement than the neural network methods, but is comparable to that on the cross-type experiment. Since the MAP method is not based on training data, we expect less advantage for the neural networks on the tonsil data. For subjective MOS studies, our GARNN shows the best perceptual performance among all methods for all samples, including cross-type sample tonsil. The qualitative results of this human specimen study can be found in Fig. 5.

In conclusion, we propose a deep learning FB image restoration method. We develop a dual-sensor imaging system to obtain aligned GT data for FB images and a GARNN to remove honeycomb patterns and improve spatial resolution. The experimental results on the lens tissue and human histological samples show that our network can remove fixed patterns in FB images and recover hidden information for resolution enhancement. Based on subjective MOS studies, we further show that restored images are generally sharper and more realistic with adversarial learning. Our future plans for better restoration accuracy include collecting larger datasets from extensive sample types and modifying the network to exploit information from multiple input FB frames.

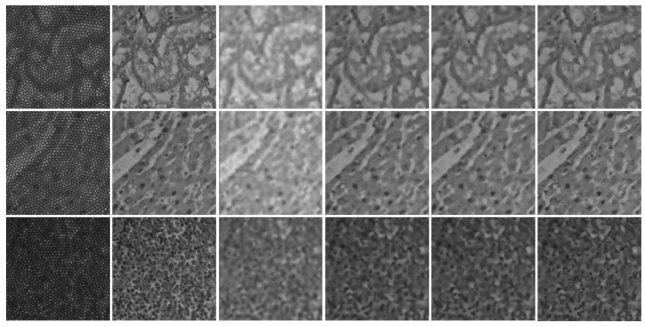

Fig. 5. Experimental results with human histological specimens (from top to bottom: kidney, liver, and tonsil). (a) Raw FB images, (b) GT images, (c) results from MAP, (d) results from RED30, (e) results from GRNN, and (f) results from the GARNN.

Funding. National Institute of Biomedical Imaging and Bioengineering (NIBIB) (R21EB022378).

REFERENCES

1. A. F. Gmitro and D. Aziz, Opt. Lett. **18**, 565 (1993).
2. A. Shinde and M. V. Matham, J. Med. Imaging Heal. Inform. **4**, 203 (2014).
3. J. H. Han and S. M. Yoon, Opt. Lett. **36**, 3212 (2011).
4. G. W. Cheon, J. Cha, and J. U. Kang, Opt. Lett. **39**, 4368 (2014).
5. J. Shao, W.-C. Liao, R. Liang, and K. Barnard, Opt. Lett. **43**, 1906 (2018).
6. D. Ravì, A. B. Szczotka, D. I. Shakir, S. P. Pereira, and T. Vercauteren, Int. J. Comput. Assist. Radiol. Surg. **13**, 917 (2018).
7. K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, IEEE Trans. Image Process. **26**, 3142 (2017).
8. V. Nair and G. E. Hinton, *Rectified Linear Units Improve Restricted Boltzmann Machines* (Omnipress, 2010), pp. 807–814.
9. S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," arXiv:1502.03167 (2015).
10. C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," arXiv:1609.04802 (2016).
11. A. L. Maas, A. Y. Hannun, and A. Y. Ng, *Rectifier Nonlinearities Improve Neural Network Acoustic Models* (2013).
12. A. Coates, A. Ng, and H. Lee, in *Proceedings of Machine Learning Research (PMLR)* (2011), Vol. **15**, pp. 215–223.
13. X.-J. Mao, C. Shen, and Y.-B. Yang, *Image Restoration Using Very Deep Convolutional Encoder-decoder Networks with Symmetric Skip Connections (NIPS)* (Curran Associates Inc., 2016), pp. 2810–2818.
14. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, IEEE Trans. Image Process. **13**, 600 (2004).