Unsupervised Super-Resolution Framework for Medical Ultrasound Images Using Dilated Convolutional Neural Networks

Jingfeng Lu, Wanyu Liu
Department of Automatic Measurement and Control
Harbin Institute of Technology
Harbin, China
e-mail: lujingfeng@hit.edu.cn

Abstract—**Ultrasound Imaging is one of the most widely used imaging modalities for clinic diagnosis, but suffers from a low resolution due to the intrinsic physical flaws. In this paper, we present a novel unsupervised super-resolution (USSR) framework to solve the single image super-resolution (SR) problem in ultrasound images which lack of training examples. Our method utilizes the powerful nonlinear mapping ability of convolutional neural networks (CNNs), without relying on prior training or any external data. We exploit the multi-scale contextual information extracted from the test image itself to train an image-specific network at test time. We utilize several techniques to improve the convergence and accuracy, including dilated convolution and residual learning. To capture valuable internal information, dilated convolution is employed to increase the receptive field without increasing the network parameters. To speed up the convergence of the training, residual learning is used to directly learn the difference between the high-resolution and low-resolution images. Quantitative and qualitative evaluations on real ultrasound images demonstrate that the proposed method outperforms the state-of-the-art unsupervised method.**

Keywords-super-resolution; ultrasound; medical imaging; convolution neural networks; dilated convolution

I. INTRODUCTION

With progress in ultrasound imaging technology, medical ultrasound imaging has become one of the most widely used imaging modalities for clinical diagnosis, including soft-tissues and organ delineation, measurement of flow in large blood vessels, estimation of cardiac function, etc. [1]. Ultrasound imaging has many advantages over other medical imaging modalities such as X-ray, computed tomography (CT) and magnetic resonance imaging (MRI) due to its noninvasive, harmless, cost efficient, portable and real time properties. However, there are intrinsic flaws in ultrasound imaging due to the probe bandwidth, diffraction effects and trade-off between frequency and penetration (the resolution of images is better when higher frequencies are used, but, at the same time, limits the penetration). As a result, ultrasound images suffer from a low resolution, limiting their clinical applications. Therefore, accurate resolution enhancement methods for ultrasound images have become the focus of much recent research and have important practical and clinical applications [2] [3].

As far as ultrasound images resolution enhancement is concerned, traditional methods have been investigated by optimizing the imaging devices or sensors, such as back-projection image recovery method [4] and adaptive beamforming (ABF) method [5]. Although these methods have improved the resolution of ultrasound images, the dependence on the devices and frequency poses tremendous instrumentation constraints on the application. With the physical limitations of device-based approaches and the close combination of image processing techniques and medical image analysis, employing image-based methods to achieve resolution enhancement of ultrasound images is a promising alternative besides the device-based methods.

Super resolution (SR) is an effective method to reconstruct a high resolution (HR) image from one or several low resolution (LR) images [6], which is widely used for medical imaging [7], microwave imaging [8], remote sensing imaging [9], and so on. So far, various super-resolution methods have been developed for ultrasound images, including interpolation-based methods, reconstruction-based methods and learning-based methods. Interpolation-based methods [10] are widely used due to their simplicity and easy implementation. However, these methods share the same problem of over-smoothing. Reconstruction-based methods, most of which employed in ultrasound images are based on deconvolution techniques, solve the SR problem as a reconstruction process, incorporating prior information or introducing regularizations into the process. Although reconstruction-based methods are effective to preserve geometric structure, they generally fail to restore sufficient high frequency details, which is detrimental for visual quality [3].

Learning-based methods try to learn the mappings between LR and HR images from a large number of LR-HR examples. Then the learned mappings are used to reconstruct the desired HR images [11-15]. Learning-based methods have achieved convincing performance in SR problems in recent years. In particular, deep learning methods and especially convolutional neural networks (CNNs) have led to a dramatic progress in performance of SR, exceeding previous traditional SR methods. This progress in performance is obtained with very deep and well-designed CNNs. These CNNs are trained on external databases containing sufficient LR-HR examples, which are, however, extremely scarce in ultrasound images. These methods are

978-1-5386-4991-6/18/$31.00 ©2018 IEEE

able to obtain outstanding performance when the test images are restricted to the conditions the networks are trained on. But the performance will significantly deteriorate once these conditions are not satisfied, limiting their application in ultrasound images. In addition, these methods are generally trained for a single scale. Although some methods are able to solve the multi-scale problem, the scale factors are fixed rather than arbitrary. This limits the further practical application of these methods.

In this paper, we present a novel unsupervised super-resolution (USSR) framework to solve the single image SR problem in ultrasound images. The contributions of this paper can be summarized as follows:

1) Unsupervised property: Our method exploits the internal recurrence of information contained in a single image itself, without any external data or prior training. This unsupervised property can solve the SR problem in ultrasound images that lack of LR-HR examples.

2) Arbitrary scale factor: In our USSR framework, an image-specific CNN is trained at test time on examples extracted from the single image, with an arbitrary scale factor.

3) Residual learing: We utilize residual learning to speed up the convergency of the training. As the LR and HR images are highly correlated and share the same information to a large extent, we use residual architecture to directly learn the residual image, which is the difference between HR and LR images.

4) Receptive Field size: We demonstrate that the receptive field size is a significant factor in the SR task. To increase the receptive field size, we propose dilated convolution to replace the regular convolution. With the help of dilated convolution, our method achieves better performance with larger receptive field, without increasing computational complexity.

The remainder of the paper is organized as follows: Section II describes the details of our USSR framework, including data processing, CNN design and training implementation. Section III demonstrates the experiment results that validate the effectiveness of the proposed method. The comparison with a state-of-the-art unsupervised SR method is also presented in this section. Lastly, we conclude this work and its research outlook in the future in section IV.

II. Proposed Method

In an ultrasound system, the observed LR images can be generally regarded as a deteriorated product of corresponding HR images. The process of deterioration can be formulated as

$$x = DHy + n \tag{1}$$

where x and y is the observed LR images and the original HR images respectively (as the goal of this work is to reconstruct HR images from LR images, we take x as the LR images and y as the original HR images), D denotes the down-sampling operator, H denotes the decimation and blurring operators, and n represents the independent and identically distributed additive noise [16]. Based on the formula above, the SR can be regarded as the process of estimating HR images y from LR images x, which can be solved by CNNs to learn the LR-HR mappings.

Our USSR framework is shown in Figure 1. The framework exploits the multi-scale internal recurrence of information contained in a single image, rather than extra information from external data. Given a test image, an image-specific CNN is trained to infer complex image-specific HR-LR relations with examples generated from the image itself. We then apply those learned relations on the test image to produce the HR output. To achieve clear description, we describe the crucial contributions to this framework in three steps: data processing (generating and augmentation), CNN design, and USSR model training and testing.

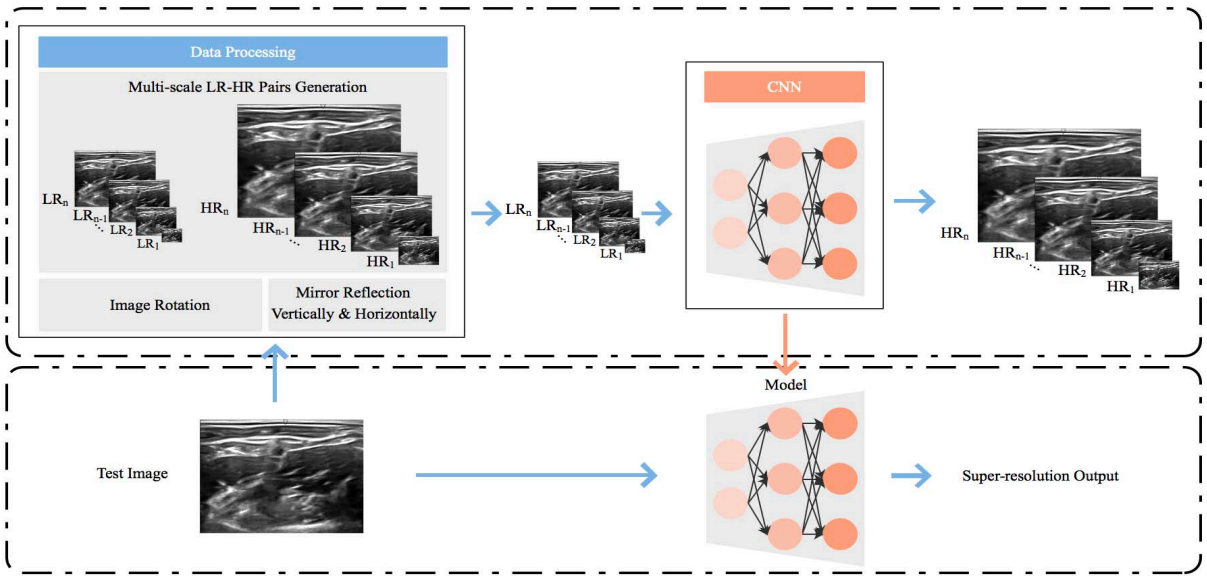

Figure 1. The proposed USSR framework. Given one single ultrasound image, a CNN is trained on LR-HR examples generated from the image itself. The CNN learns the mappings from LR images to HR images. The resulting image-specific CNN is then employed to the test image to reconstruct the HR output.

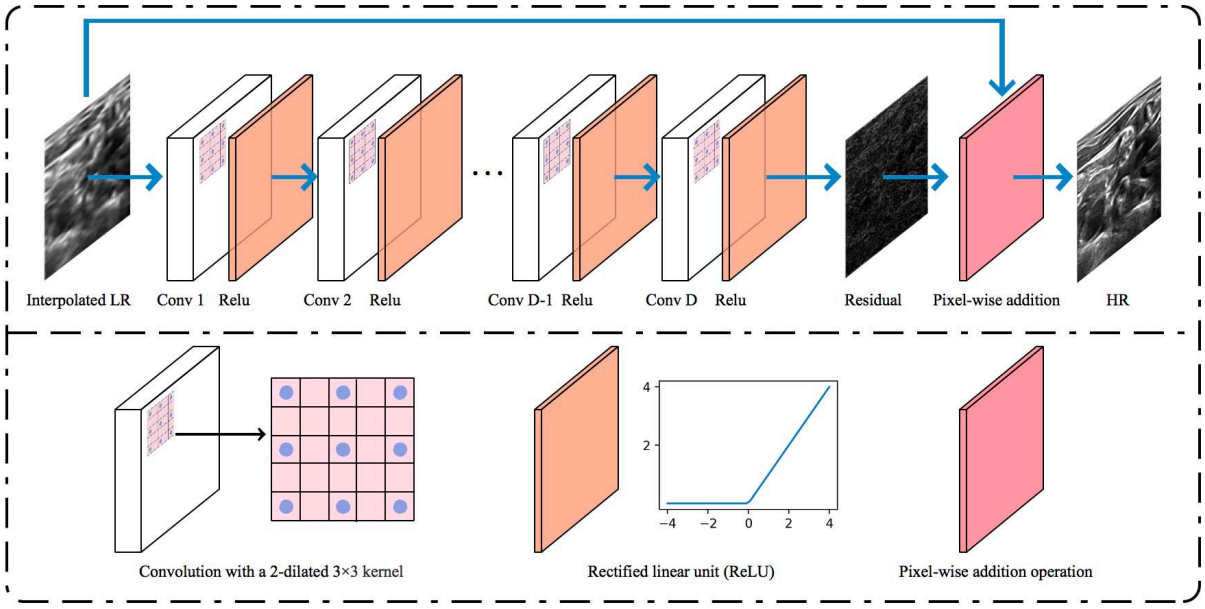

Figure 2. The structure of proposed network (full model). We stack D (in our full model, D is set to 8) dilated convolution layers and ReLU. The LR image is interpolated to the same size of the desired HR image, as the input of the network. The residual image is produced and added with the input image via the pixel-wise addition layer, to produce the HR output.

A. Data Processing

The data processing approach is shown on the top of Figure 1. Since we train the network with examples extracted from the test image itself, the image needs to be processed into LR image as the input and corresponding HR image as the training target. Given the SR scale factor s, the LR image is generated by downscaling the test image into the low-resolution version of itself by the SR scale factor s. The test image is directly taken as the HR target.

It is obviously not sufficient to train CNNs with dataset containing only one pair of LR and HR images. Therefore, data augmentation is conducted to artificially enrich the data before training. Firstly, the test image is downscaled into numerous smaller versions in random sizes. These downscaled images are taken as the HR set and the training targets. Then every image in the HR set is downscaled by the SR scale factor to generate the LR set, as input of the network. In this way, the training data consisting of various of LR-HR pairs is formed. Note that the probability of generating a HR example is non-uniformly distributed and proportionate to the size of the desired HR image, which means the probability of the and HR example to be generated is higher if the size of the HR example is closer to the size of the desired HR image. This reflects the higher reliability of non-synthesized HR examples over synthesize ones. Secondly, each LR-HR pair is vertically and horizontally flipped, and rotated by 45°, 90°, 135°, 180°, 225°, 270°, and 315°, respectively. As a result, the dataset is enriched by forming nine additional sets of LR-HR pairs.

B. CNN Design

Supervised CNNs [14], which are exhaustively trained on large external collections, tend to contain diverse weights to capture all possibilities of LR-HR mappings by minimizing the loss function

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N}\left\|F(x_i;\theta)-y_i\right\|^2 \qquad (2)$$

where x_i and y_i represent LR image and HR image examples respectively, $\theta=\{W,B\}$ represents the weights and bias of convolutional layers, N represents the number of training examples which tend to be extremely large. Therefore, these networks tend to be extremely deep and complex. However, in this work, our network is trained on examples generated from the test image itself. We assume that the diversity of LR-HR mappings within one single image is relatively sparse, which are able to be extracted by thinner and simpler CNNs.

To achieve the best performance, different network structures are designed and tested on the same data. We take a simple network containing 3 convolution layers as the baseline network. Each layer consists of 64 channels. Rectified linear unit (ReLU) is used as the activation function in each layer. The LR image is interpolated to the same size of the HR image, as the input of the network. Different factors (network depth, receptive field size and residual learning) that may affect the SR performance are considered in the CNN design. Figure 2 shows the structure and components of the proposed network.

1) Network depth: As reported in [17], the depth of networks is crucial for most of deep learning tasks. In this work, however, the network is trained on examples generated from the single image. The diversity of LR-HR mappings within one single image are relatively sparse, which can be extracted by thinner and simpler CNNs. Therefore, the influence of network depth is firstly

evaluated. On the basis of baseline network, more layers are stacked to construct deeper networks. In this way, we build networks with different depths respectively (D=3,5,8,10).

2) Residual learning: Previous work reported the notorious gradients vanishing or exploding problem that makes training difficult to converge and causes decreased accuracy with deeper networks[18,19]. As the LR image and HR image share the similar information to a large extent, we argue that it is more effective to explicitly learn the residual image, which is the difference between HR and LR images. Therefore, we use the residual learning to exploit the similarity between the input LR image and the output HR image. We define the risidual image as

$$r_i = y_i - x_i \tag{3}$$

The residual image is learned by minimizing the loss function

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left\|F(x_i;\theta) - r_i\right\|^2 \tag{4}$$

3) Receptive field size: Since our method directly learns the mappings from the internal recurrence of information contained in the test image alone, it is crucial to exploit sufficient contextual information, which can be captured by larger receptive field. To further improve the performance of SR, we propose networks with larger receptive field. Instead of simply increasing the kernel size, we use dilated convolution operation which is capable of increasing the receptive field without increasing the parameters of the network. As shown in Figure 3, for instance, the receptive field of a 2-dilated 3×3 convolution is the same with a 5×5 general convolution, while the number of its parameters is the same with a 3×3 non-dilated convolution.

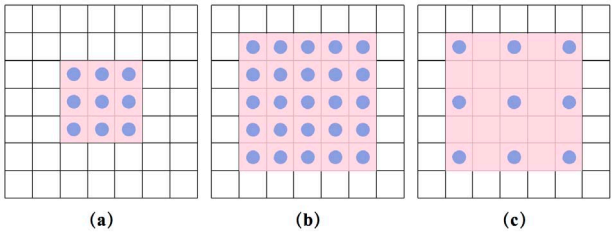

Figure 3. Example of the receptive field and the number of parameters of different types of kernels. The receptive field of a 2-dilated 3×3 convolution (c) is the same with a 5×5 non-dilated convolution (b), while the number of its parameters is the same with a 3×3 non-dilated convolution (a).

C. Implementation and Training Details

In the training stage, no weight is initialized by any pre-trained model. We use L1 loss as the loss function and Adam as the optimization algorithm with a fixed learning rate of 0.001. Pytorch package is used to train our network on a NVidia 1080ti GPU. To reduce the training time and ensure the training time does not depend on the image size, at each iteration, the LR-HR pair is randomly cropped in a fixed size of 128×128 (unless the sampled image-pair is smaller).

III. Experiment and Results

In this section, we evaluate the performance of our method on three types of ultrasound images including brachial plexus, cardiac, and brain ultrasound images. Peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM) are used as the quantitative evaluation matrices. Contributions of different components of our network are analyzed. As our method aims at single image super-resolution problem in ultrasound images which are obtained with realistic acquisition process, and no external training data is used, we compare our method within the unsupervised SR field, with a state-of-the-art unsupervised SR method, SelfExSR [12].

A. Investigation of the Proposed Network

To analyze the property of the proposed network and the contributions of the different components of our network, we design a set of controlling experiments. The experimental results are as followed.

1) Network depth

To analyze the contributions of different network depths, we train the regular CNNs without dilation or residual learning, with different depths (D=3, 5, 8, 10). Figure 4(a) shows the convergence curves in terms of PSNR on the brain ultrasound images for 3× SR scale factor. We can see that deeper networks generally perform better than shallow ones. However, when D is larger than 8, the PSNR results stop increasing. This is corresponding to our assumption that the diversity of LR-HR mappings within one single image is relatively sparse. Simply increasing the network depth can not further improve the performance. Accordingly, we choose D=8 for our network to strike a balance between performance and efficiency.

2) Residual learning

To demonstrate the effect of residual learning, we add the residual connection to the 8-layer regular convolutional network and show the convergence curve in Figure 4 (b). We can see that with the help of residual learning, the residual network (orange curve) converges faster than the non-residual network (red curve). However, the residual network dose not outperform the non-residual network in terms of accuracy.

3) Receptive field size

To further improve the performance of our network, we replace the regular convolution with 2-dilated convolution to increase the receptive field. Figure 3 (b) shows that dilated network (blue curve) produces higher PSNR than the non-dilated networks (red and orange curves). This demonstrates that the receptive field size is crucial for exploiting sufficient contextual information. Finally, we train our full model with residual and dilated network. Our full model (green curve) achieves the best performance in both convergence and accuracy.

B. Comparisons with the State-of-the-Arts

In this section, we compare our full model with the state-of-the-art unsupervised SR methods, SelfExSR [12]. The implementation is from the publicly available codes provided by the authors. In Table I, we provide a summary of

TABLE I. AVERAGE PSNR/SSIM FOR SCALE FACTOR OF 2, 3, 4 ON BRACHIAL PLEXUS, CARDIAC, AND BRAIN ULTRASOUND IMAGES.

Data	Scale	Bicubic	SelfEx [12]	Dilated_8 (Ours)	USSR (Our full model)
		PSNR/SSIM	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Brachial Plexus	2	37.32/0.9778	38.17/0.9806	38.54/0.9808	**38.66/0.9809**
	3	29.30/0.9022	32.80/0.9440	34.09/0.9473	**34.11/0.9485**
	4	26.87/0.8496	29.98/0.9005	31.06/0.9019	**31.12/0.9023**
Cardiac	2	38.43/0.9638	39.47/0.9685	39.98/0.9701	**40.20/0.9718**
	3	30.91/0.8704	34.39/0.9195	36.60/0.9345	**36.68/0.9366**
	4	30.16/0.8507	32.71/0.8876	34.39/0.9024	**34.56/0.9030**
Brain	2	40.08/0.9496	40.36/0.9527	41.02/0.9562	**41.17/0.9575**
	3	33.21/0.8611	36.38/0.9060	38.25/0.9265	**38.47/0.9279**
	4	33.21/0.8613	35.50/0.8884	36.59/0.9009	**36.77/0.9016**

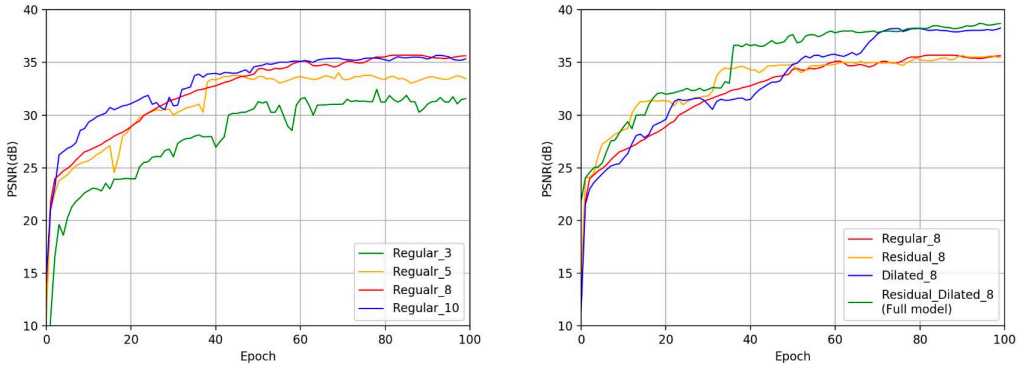

Figure 4. Convergence analysis on the network depth, residual learning and dilated convolution. The results are obtained on the brain ultrasound images with the 3× SR scale factor. (a). Convergence curves of regular networks with different depths (D=3,5,8,10). Deeper networks generally perform better than shallow ones. When D is larger than 8, the PSNR results stop increasing. (b) Convergence curves of 8-layer networks with different components (residual learning and dilated convolution). Incorporating with residual learning makes the network converge faster. Utilizing dilated convolution produces higher PSNR than the non-dilated networks. Our full model achieves the best performance in both convergence and accuracy.

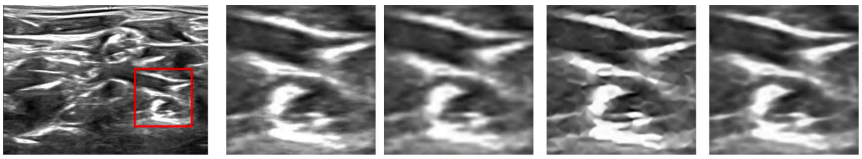

Figure 5. Visual comparison of reconstructed results on a brachial plexus ultrasound image.

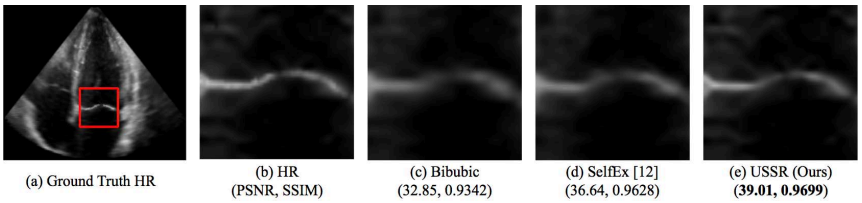

Figure 6. Visual comparison of reconstructed results on a cardiac ultrasound image.

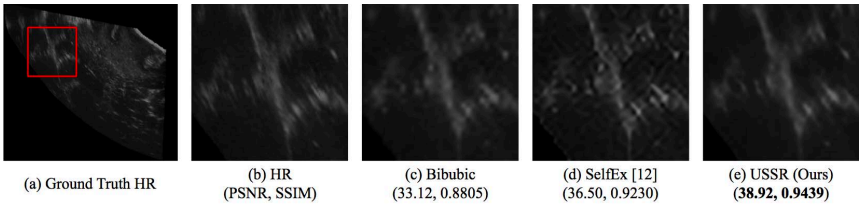

Figure 7. Visual comparison of reconstructed results on a brain ultrasound image.

quantitative evaluations on all three types of ultrasound images. We can see that our dilated-8 model (8 dilated convolution layers without the residual connection) is able to outperform the compared method. Furthermore, our full USSR model (8 dilated convolution layers with the residual connection) achieves the highest average PSNR and SSIM in all the three types of ultrasound images. Typical examples of reconstructed ultrasound images of USSR and the compared methods are shown in Figure 5, 6 and 7. In Figure 5, the reconstructed image of USSR is much sharper and clearer than the other results. Similarly, in Figure 6 and Figure7, the mitral valve and brain tissue are clearer in our method but blurred or distorted in other methods.

IV. CONCLUSION

In this work, we propose an unsupervised super-resolution (USSR) framework for medical ultrasound images, which lack of LR-HR examples. The proposed method employs the powerful nonlinear mapping ability of CNNs, without relying on external training data or pre-train models. By replacing the regular convolution with dilated convolution and incorporating the residual learning, our network further improves the performance in both convergence and accuracy. Evaluations on realistic ultrasound images demonstrate that our method outperforms the state-of-the-art unsupervised method in terms of accuracy and visual quality. This end-to-end framework can be utilized as a straightforward and powerful tool for a wide range of SR tasks.

In future research, we plan to apply our method to images of various domains. Conducting in-depth analysis of USSR framework to gain a better comprehension of this framework is another research direction we plan to explore.

REFERENCES

[1] A. M. Tanter and M. Fink, "Ultrafast imaging in biomedical ultrasound," IEEE Trans. Ultrason. Ferroelectr. Freq. Control, vol. 61, no. 1, pp. 102–119, 2014.

[2] J. M. Hudson, R. Williams, C. Tremblay-Darveau, P. S. Sheeran, L. Milot, and G. A. Bjarnason, "Dynamic contrast enhanced ultrasound for therapy monitoring," European Journal of Radiology,vol. 84, no. 9, pp. 1650–1657, 2015.

[3] R. Morin, S. Bidon, A. Basarab, and D. Kouame, "Semi-blind deconvolution for resolution enhancement in ultrasound imaging," in IEEE International Conference on Image Processing, pp.1413–1417, 2013.

[4] G. T. Clement, J. Huttunen, and K. Hynynen, "Superresolution ultrasound imaging using back-projected reconstruction," Journal of the Acoustical Society of America, vol. 118, no. 6, pp. 3953–3960, 2005.

[5] M. A. Ellis, F. Viola, and W. F. Walker, "Super-resolution image reconstruction using diffuse source models," Ultrasound in Med. & Biol., vol. 36, no. 6, pp. 967–977, 2010.

[6] S. Farsiu, D. Robinson, M.Elad, and P.Milanfar, "Advances and challenges in superresolution," International Journal of Imaging Systems and Technology, vol. 14, no. 2, pp.47–57. 2004.

[7] R. Morin, A. Basarab, and D. Kouame, "Alternating direction method of multipliers framework for super-resolution in ultrasound imaging," in IEEE International Symposium on Biomedical Imaging, pp. 1595–1598, Barcelona, Spain, 2012.

[8] I. Yanovsky, B. H. Lambrigtsen, A. B. Tanner, and L. A. Vese, "Efficient deconvolution and super-resolution methods in microwave imagery," IEEE J. Sel. Topics Appl. Earth Observations and Remote Sens., vol. 8, no. 9, pp. 4273–4283, 2015.

[9] G. Martin and J. M. Bioucas-Dias, "Hyperspectral compressive acquisition in the spatial domain via blind factorization", in IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, pp. 1–4, Tokyo, Japan, 2015.

[10] S. Mallat and G. Yu, "Super-resolution with sparse mixing estimators," IEEE Trans. Image Process., vol. 19, no. 11, pp. 2889–2900, 2010.

[11] H. He, W. C. Siu, "Single image super-resolution using Gaussian process regression," in IEEE Conf. Comput. Vis. Pattern Recognit., pp. 449–456, 2011.

[12] J. B. Huang, A. Singh, and N. Ahuj, "Single image superresolution from transformed self-exemplars," in IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5191–5206, 2015.

[13] W. Gong, Y. Tang, X. Chen, Q. Yi, and W. Li, "Combining Edge Difference with Nonlocal Self-similarity Constraints for Single Image Super-Resolution," Neurocomputing, vol. 249, pp. 157–170, 2017.

[14] C. Dong, C. Loy, K. He, and X. Tang, "Image super-resolution using deep convolutional networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 2, pp. 295–307, 2016.

[15] J. Kim, J. K. Lee, and K. M. Lee, "Accurate Image Super-Resolution Using Very Deep Convolutional Networks," in IEEE Conf. Comput. Vis. Pattern Recognit., pp. 1646–1654, 2016.

[16] N. Zhao, Q. Wei, A. Basarab, D. Kouamé, and J. Y. Tourneret, "Single image super-resolution of medical ultrasound images using a fast algorithm," in International Symposium on Biomedical Imaging. IEEE, pp. 473–376, 2016.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in International Conference on Learning Representations, 2015.

[18] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

[19] X. Glorot, and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," Journal of Machine Learning Research vol. 9, pp. 249–256, 2010.

[20] D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.