Zero-Shot Learning for Real-Time Ultrasound Image Enhancement

Yuxuan Li
Department of Automation
Tsinghua University
Beijing, China
liyx21@mails.tsinghua.edu.cn

Wenkai Lu
Department of Automation
Tsinghua University
Beijing, China
lwkmf@mail.tsinghua.edu.cn

Patrice Monkam
Department of Automation
Tsinghua University
Beijing, China
patrice123china1@gmail.com

2022 IEEE International Ultrasonics Symposium (IUS) | 978-1-6654-6657-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/IUS54386.2022.9958334

***Abstract*—Ultrasound(US) imaging has been widely used for clinical diagnosis. However, ultrasound images inherently suffer from speckle noise and low contrast. Although deep learning-based approaches have been proven to outperform traditional filtering algorithms in image enhancement tasks, they usually require high-quality images for training, which are unavailable in practice. In this paper, we develop a zero-shot learning framework for real-time ultrasound image enhancement. In our framework, relatively low-quality(LQ) and high-quality(HQ) images are beamformed using raw Synthetic Aperture Ultrasound (SAU) data with different channel numbers. The generated LQ-HQ image pairs are used to train two designed U-Net variants whereby one is for inference on GPU and the other with depth-wise separable convolution for inference on CPU and mobile devices. The trained models are directly used for enhancing noisy HQ US images without requiring speckle-free and high contrast targets for training. Experiment results indicate that our method performs more favorably against state-of-the-art image enhancement approaches in terms of both image quality and inference time.**

***Index Terms*—zero-shot learning, ultrasound image enhancement, synthetic aperture ultrasound**

I. Introduction

Ultrasound(US) imaging is widely used in clinical practice because of its non-invasive, real-time, and low-cost characteristics. Due to the physical properties of ultrasound imaging, the imaging results inevitably contain speckle noise which reduces contrast and resolution. Numerous methods have been developed for ultrasound image enhancement which can be broadly divided into two categories: traditional image processing-based methods and learning-based methods. Traditional image processing approaches including anisotropic diffusion [1], image and transform domain filtering [2], [3], and low-rank approximation [4] have been proven to be effective in speckle reduction task. With the fast development of deep learning technology, many data-driven methods [5]–[8] for ultrasound image enhancement have been proposed which outperform image processing approaches. Taking advantage of complex structural design and trained on large datasets, deep learning models have continually improved the performance of image enhancement tasks. However, the training process of these methods relies on high-quality images as ground truth targets which are generally unavailable in practice.

In recent years, target-free algorithms for image denoising [9]–[12] and super-resolution [13], [14] have been proposed in the computer vision community. Self-supervised learning is adopted in these studies to perform image enhancement which exploits the statistical properties of noise models and the internal structure of the input images. However, self-supervised learning methods are not very practical for ultrasound image enhancement since they typically assume that the image noise is Gaussian, and require time-consuming and complex training steps.

In this paper, inspired by the zero-shot super-resolution algorithm [13], we develop a zero-shot learning framework for ultrasound image enhancement. Considering that the useful structural features of two different US images taken at the same location should be identical while the noise should be different, Synthetic Aperture Ultrasound (SAU) channel data is used to generate the training set. That is, beamforming of the same raw US signal is performed using different number of channels, resulting in pairs of US images with different levels of speckle noises but the same important structural features. Deep learning models trained with the generated image pairs can successfully extract useful signals from corrupted images and produce high-quality ultrasound images. Two U-Net variants whereby one for inference on GPU and the other with depth-wise separable convolution for inference on CPU and mobile devices are designed and trained. Experiment results suggest that our method performs more favorably against traditional image processing methods [1], [2] and state-of-the-art self-supervised approach [12] in terms of both image quality and inference time.

II. Methods

A. General Framework

Our zero-shot learning framework(Fig. 1) is divided into two phases: the training phase and the inference phase.

In the training phase, considering that high-quality target images do not exist, we leverage the US imaging process to generate different quality image pairs. Specifically, relatively low-quality(LQ) and high-quality(HQ) images are generated

This research is financially supported by the Tsinghua University Spring Breeze Fund under Grant No. 2021Z99CFZ009 and the Beijing Natural Science Foundation Project under Grant No. M21019.
Corresponding author: Wenkai Lu (lwkmf@mail.tsinghua.edu.cn).

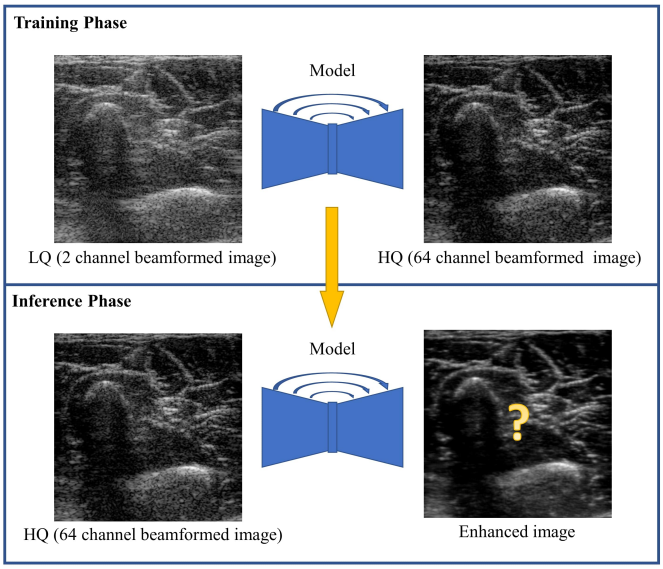

Fig. 1. Our zero-short learning framework for ultrasound image enhancement.

through beamforming with raw channel data of 2 transmitters and 64 transmitters, respectively. The generated LQ-HQ image pairs are used to train two designed U-Net variants.

In the inference phase, the trained models are directly used for enhancing the quality of noisy US images(HQ) without requiring ground truth targets for training.

B. Dataset preparation

A self-developed SAU imaging system was used for data collection. Our system's parameters are shown in Table I.

TABLE I
SYSTEM PARAMETERS

Array type	Linear array
Probe width	40mm
Transmitter number	64
Receiver number	64
Center frequency	7.5MHz

A total of 110 SAU imaging data were collected from 11 volunteers which are split into a training set and a test set. The portion of each part of the dataset is shown in Table II.

TABLE II
DATASET SPLITS

	Volunteers	**Images**
train	9	90
test	2	20

The raw channel data of the training set are used to form LQ-HQ image pairs using Fourier-based beamforming method. The test set data are kept unseen during the entire training phase and are used to evaluate the performance of our zero-shot learning framework.

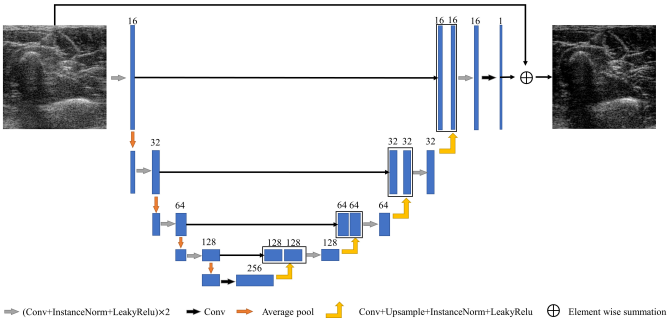

Fig. 2. Model architecture.

C. Model Architecture

Two deep learning networks are trained using the generated training image pairs.

The first network is designed for inference on the GPU. It is a typical U-Net structure network with encoder-decoder architecture and skip connection paths(Fig. 2). Basic blocks in the model consist of 3×3 convolution layers, instance normalization layers, and rectified linear unit (ReLU) activation functions. Average pooling and bi-linear up-sampling layers are used to scale the feature maps to construct the up-sampling and down-sampling paths. Following standard practice in image denoising, Global residual learning [17] is also employed in our network for training stability and better performance.

The second network is a lightweight version of the first network intended for inference on CPU and mobile devices. By replacing all convolution layers with depth-wise separable convolution [16] and reducing channel numbers, we construct a lightweight U-Net which can run in real-time on the CPU.

D. Training Strategy

The image pairs of size 256×620 are randomly cropped to 224×448 and flipped horizontally to generate diverse training samples. L1 loss is employed in our study as the optimization objective which can be expressed as follows:

$$min[L1(I_{HQ}, I^*_{HQ})] \tag{1}$$

where I_{HQ} and I^*_{HQ} are the target image and the predicted output respectively.

All models are trained using the Adam optimizer with a learning rate of 0.001 for 20 epochs. Hyper-parameters β_1 and β_2 are set to 0.9 and 0.999 respectively for the Adam optimizer. The batch size is set to 4 for all training processes.

E. Evaluation

Considering that no ground truth is available in our image enhancement task, the contrast-to-noise ratio(CNR) is employed to quantitatively evaluate our method. Specifically, foreground and background regions are manually annotated using segmentation masks for each test image. An example

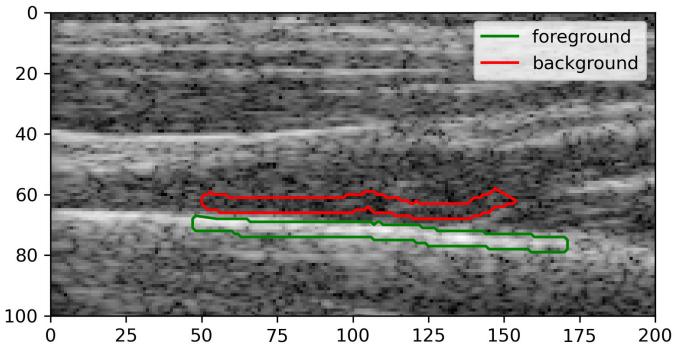

Fig. 3. An example of annotated regions for CNR calculation.

of annotated regions is shown in Fig. 3. CNR is calculated as follows:

$$CNR = \frac{1}{N}\sum_{i=1}^{N}\frac{\mu_F^i - \mu_B^i}{\sqrt{\sigma_F^i + \sigma_B^i}} \quad (2)$$

where N represents the number of foreground and background region pairs. μ_F^i, μ_B^i, σ_F^i, and σ_B^i represent the average value of the foreground region, the average value of the background region, the standard deviation of the foreground region, and the standard deviation of the background region respectively.

III. Results

Our models are compared with traditional image processing methods SRAD [1] and OBNLM [2], as well as state-of-the-art self-supervise denoising method Self2Self [12]. The visualized enhanced results are shown in Fig. 4 and the quantitative evaluation is shown in Table III.

TABLE III
AVERAGE CNR CALCULATED ON IN-VIVO US IMAGES

	Average CNR
Original Image	2.76
SRAD	3.98
OBNLM	3.34
Self2Self	3.37
Ours-Light	3.16
Ours	3.57

As shown in Fig. 4, our algorithm has the capability of removing more speckle noises without damaging fine textures. Compared with our proposed framework, filtering methods (SRAD and OBNLM) tend to introduce over-smoothness, and the self-supervised approach (Self2Self) retains more noise. The quantitative evaluation shows that our method can noticeably increase the average CNR of the input images (from 2.76 to 3.57). It is worth noting that the CNR of blurry images tends to be higher. Therefore, although the CNR of SRAD is higher than that of our method, our method produces more appealing results without blurriness. Our lightweight model is also capable of removing most speckle noises. Compared with our normal-size model, the contrast of the images generated by our lightweight model can still be improved in future works.

An Nvidia 2080Ti GPU and an Intel(R) Core (TM) i7-9800X CPU are utilized in our study to assess the inference time of different methods. With the input image size of 256×620, the inference time is shown in Table IV. It can be observed from Table IV that our lightweight model can achieve real-time inference(25fps) on CPU, making it very suitable for portable ultrasound devices and our normal-sized model can run at a very high frame rate (250fps) on GPU. Compared with our method, image processing approaches fail to perform real-time processing and the Self2Self algorithm requires self-supervised training at test time which increases inference time.

In conclusion, although trained without ground-truth target images, our method can achieve real-time processing and produce more visually appealing enhancement results.

TABLE IV
INFERENCE TIME COMPARISON

	Model Size(MB)	**CPU(s)**	**GPU(s)**
SRAD	/	0.3	/
OBNLM	/	8	/
Ours-Light	0.275	0.04	/
Self2Self	4.36	/	180
Ours	8.24	/	0.004

IV. Discussion

In this work, a zero-shot learning framework is proposed which achieves real-time and high-quality ultrasound image enhancement results.

There are a few limitations of our work. First, the collected dataset is relatively small. A large scale dataset will be collected to improve the performance of deep learning models and better evaluate the performance of the image enhancement result. Second, the current model architectures are simple U-Net variants. More advance model designs will be investigated in future works for better enhancement results and faster inference speed. Third, our algorithm is not deployed in clinical practice yet. This will be topic for future studies.

Acknowledgment

The authors would like to thank Mr. Weiguang Zhang for hardware support and all volunteers who participated in the data collection process.

References

[1] Yongjian Yu and S. T. Acton, "Speckle reducing anisotropic diffusion," IEEE Trans. Image Process., vol. 11, no. 11, pp. 1260-1270, Nov. 2002.
[2] P. Coupe, P. Hellier, C.Kervrann, and C. Barillot, "Nonlocal means-based speckle filtering for ultrasound images," IEEE Trans. Image Process., vol. 18, no. 10, pp. 2221–2229, 2009.
[3] A. Garg, J. Goal, S. Malik, and K. Choudhary, "De-speckling of medical ultrasound images using Wiener filter and wavelet transform," Int. J. Electron. Commun. Technol, vol. 2, no. 3-1, 2011.
[4] L. Zhu, C.-W. Fu, M. S. Brown, and P.-A. Heng, "A non-local low-rank framework for ultrasound speckle reduction," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5650-5658.
[5] D. Mishra, S. Chaudhury, M. Sarkar, and A. S. Soin, "Ultrasound image enhancement using structure oriented adversarial network," IEEE Signal Process. Lett., vol. 25, no. 9, pp. 1349-1353, 2018.

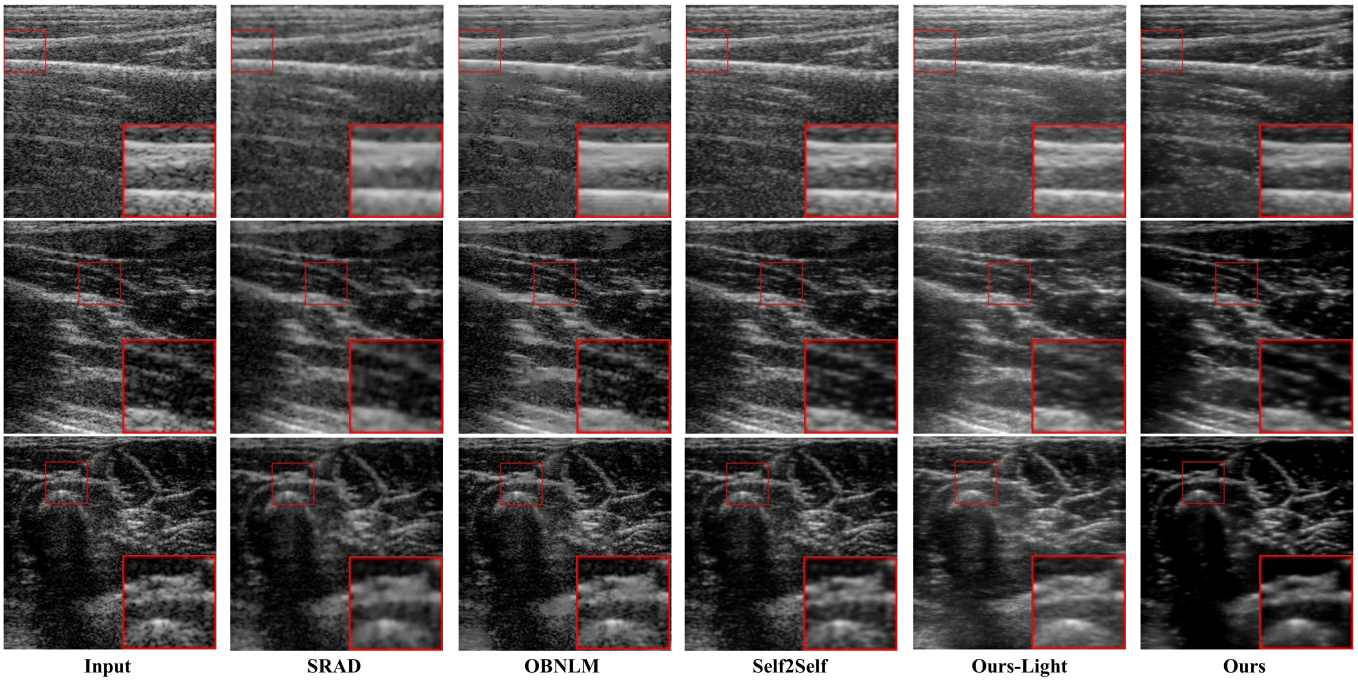

Fig. 4. Qualitative result comparison on in-vivo US images in the test set.

[6] Y. Lan and X. Zhang, "Real-time ultrasound image despeckling using mixed-attention mechanism based residual UNet," IEEE Access, vol. 8, pp. 195327-195340, 2020.

[7] P. Kokil and S. Sudharson, "Despeckling of clinical ultrasound images using deep residual learning," Comput. Methods Programs Biomed., vol. 194, p. 105477, 2020.

[8] A. Sadeghi, I. Apostolakis, C. Meral, F. Vignon, J. S. Shin, and J.-L. Robert, "Improving Contrast of Fundamental Ultrasound Imaging Using a Deep Neural Network," in 2021 IEEE International Ultrasonics Symposium (IUS), 2021, pp. 1-3: IEEE.

[9] J. Lehtinen et al., "Noise2Noise: Learning image restoration without clean data," arXiv preprint arXiv:1803.04189, 2018.

[10] A. Krull, T.-O. Buchholz, and F. Jug, "Noise2void-learning denoising from single noisy images," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2129-2137.

[11] J. Batson and L. Royer, "Noise2self: Blind denoising by self-supervision," in International Conference on Machine Learning, 2019, pp. 524-533: PMLR.

[12] Y. Quan, M. Chen, T. Pang, and H. Ji, "Self2self with dropout: Learning self-supervised denoising from single image," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1890-1898.

[13] A. Shocher, N. Cohen, and M. Irani, ""zero-shot" super-resolution using deep internal learning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3118-3126.

[14] Q. Sun, Y. Liu, T.-S. Chua, and B. Schiele, "Meta-transfer learning for few-shot learning," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 403-412.

[15] O. Ronneberger, P. Fischer and T.Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. MICCAI, 2015,pp. 234–241.

[16] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251-1258.

[17] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, "Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising," IEEE Trans. Image Process., vol. 26, no. 7, pp. 3142-3155, 2017.