Hybrid Neural Network for Photoacoustic Imaging Reconstruction

Hengrong Lan[1], *Student Member, IEEE,* Kang Zhou[1,2], Changchun Yang[1], Jiang Liu[2], Shenghua Gao[1], Fei Gao[1,*], *Member, IEEE*

***Abstract*— Photoacoustic imaging (PAI) is an emerging non-invasive imaging modality combining the advantages of ultrasound imaging and optical imaging. Image reconstruction is an essential topic in photoacoustic imaging, which is unfortunately an ill-posed problem due to the complex and unknown optical/acoustic parameters in tissue. Conventional algorithms used in photoacoustic imaging (e.g., delay-and-sum) provide a fast solution while many artifacts remain. Convolutional neural network (CNN) has shown state-of-the-art results in computer vision, and more and more work based on CNN has been studied in medical image processing recently. In this paper, we propose Y-Net: a CNN architecture to reconstruct the PA image by integrating both raw data and beamformed images as input. The network connected two encoders with one decoder path, which optimally utilizes more information from raw data and beamformed image. The results of the simulation showed a good performance compared with conventional deep-learning based algorithms and other model-based methods. The proposed Y-Net architecture has significant potential in medical image reconstruction beyond PAI.**

I. Introduction

Photoacoustic tomography (PAT) is a hybrid imaging modality that mixed both optical and ultrasonic advantages. PAT excites ultrasonic wave by pulsed laser, which has embodied both optical absorption contrast and ultrasonic deep-resolution [1, 2]. Many practical applications have been investigated to show its great potential in clinical and preclinical imaging, such as early-stage cancer and small animal whole body imaging [3-6]. To obtain the image from the PA signals, image reconstruction is a significant topic of concern. Conventional reconstruction algorithms, e.g., filtered back-projection, delay-and-sum, are very popular due to fast reconstructive time. However, the imperfection of conventional algorithms is severe artifact, which results in distorted images especially in limited view configuration.

Deep learning has been much developed in recent years, especially in computer vision. Recently, deep learning methods are beginning to attract intensive research interest in image reconstruction problems for photoacoustic imaging [7-9]. The two main methods are direct and post-processing [10], and the difference between them is input data: the former method feeds the raw PA data and convert into image at the output of the network (raw data-feature-image); the latter method feeds a poor quality PA image and convert the feature of image into the final image(image-feature-image).

In this paper, a CNN-based architecture, named Y-Net, is proposed to solve the image reconstruction problem for PAT, which simultaneously has two inputs and one output. The approach can be called hybrid processing: both the measured raw data and a beamformed image are used as inputs, which contain different types of information respectively: rich details and overall textures. In the numerical experiment, the training data is generated by MATLAB loading the factitious segmented vessels from a public database and verified by experiments. The trained model showed good performance in test dataset compared with conventional reconstruction algorithms and other deep-learning based methods, such as U-Net.

II. Method

A. Numerical vessels data generation for training

The deep-learning-based approach is data-driven method that requires a number of data for training to get the desired results. PAT does not have access to a large amount of clinical data to train the network. Especially for reconstruction problems, we often need raw data, which is only available in research lab. Therefore, we seek to train neural networks using simulation data and test the trained models in experiments.

The MATLAB toolbox k-Wave [11] is used to generate the training data. The simulation setup is shown in Figure 1, where a linear array transducer was placed at the top of the region of interest (ROI). The sample is placed in the 38.4×38.4 mm size of ROI, and the linear array probe with 128 elements can receive PA signals. We record the raw data of the sensor, generate beamformed images and ground truth for training and testing. All images are 128×128 pixels and acoustic speed is set as 1500 m/s.

The factitious segmented vessel from public fundus oculi CT imaging [12] can be deployed with initial pressure distribution. The vessels need to be segmented and pre-processed. After a series of operations (position change, rotation, etc.), the dataset will be loaded into k-Wave simulation tool as the initial pressure distribution.

Hengrong Lan and Kang Zhou contributed equally to this work.

Hengrong Lan, Kang Zhou, Changchun Yang, Shenghua Gao and Fei Gao are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (*corresponding author: gaofei@shanghaitech.edu.cn).

Kang Zhou and Jiang Liu are with Ningbo Institute of Material Technology and Engineering, Chinese Academy of Sciences, Ningbo 315201, China.

978-1-5386-1311-5/19/$31.00 ©2019 IEEE

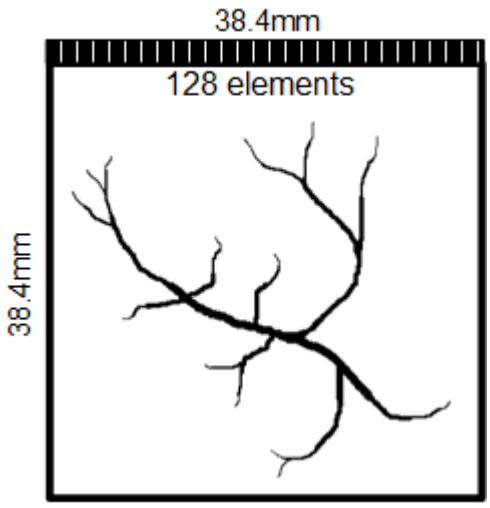

Figure 1. The illustration of the simulation setup.

B. *Architecture of Y-Net*

Most CNN architecture only establishes a single input-output stream for reconstruction (signals or image). For the image reconstruction from original raw PA signals, it is an ill-posed problem considering relationship between raw PA signals and ground truth, although the raw data contains more information about the object. For the post-processing operation, it inputs the conventional results (e.g. delay-and-sum beamforming), and loses some information about the object. Figure 2 indicates the information loss after beamforming, which loses backbone information from the difference about 80%. It is difficult to distinguish some branches and artifacts of matched difference from Figure 2(c). The elimination result of the artifact is not satisfactory for non-trained data, but it provides rich texture information of the target. Therefore, we assume that it may be a good solution to combine the raw PA signals and beamformed images as input data. It deserves noting that the raw PA signals and beamformed image have different size and features, which inspired us to build the neural network with two inputs.

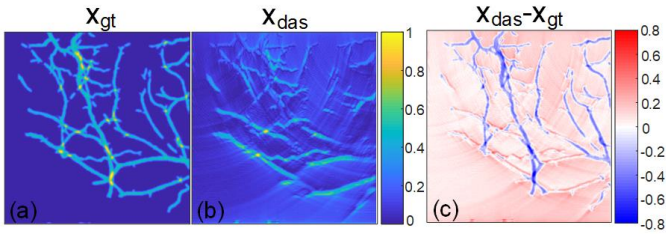

Figure 2. Comparison of information loss in traditional beamforming method. (a) The ground truth; (b) The delay-and-sum beamforming result of (a); (c) The difference between (a) and (b).

The proposed Y-Net integrates both of features with two inputs by two different encoders. The global architecture of Y-Net is shown in Figure 3, which inputs the raw PA signals to an encoder (Encoder II), and processes the raw data to obtain an imperfect beamformed image as the input of another encoder (Encoder I). Being different from U-Net [13], the proposed Y-Net enables two inputs for different types of training data that is optimized for hybrid reconstruction. The Y-Net consists of two contracting paths and a symmetric expanding path. Encoder I and Encoder II encode the texture features and physical features respectively, and the final decoder concatenates the features of both encoder outputs and generates the final result.

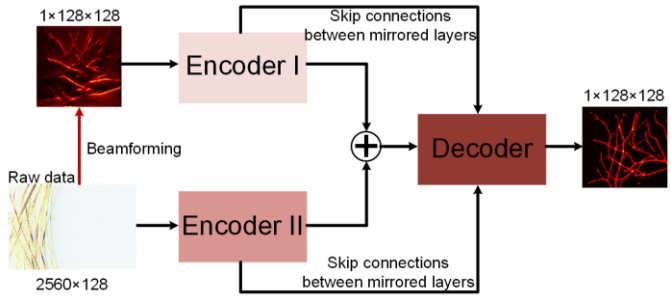

Figure 3. The global architecture of Y-Net. Two encoders extract different input feature, which concatenates into the decoder. Both encoders have skip connections with the decoder. More detail of the encoders and decoder will be illustrated in next section.

Encoder I: The Encoder I module takes the image reconstructed from raw PA data by conventional beamforming (we used the delay-and-sum). Figure 4 shows the Encoder I in detail, which is similar to the contracting path of U-Net. Every layer unit is composed of two 3×3 convolution, batch normalization and leaky rectified linear unit (ReLU), and a maximizing pooling to downsample the features. The image is passed through a series of layers that gradually downsample, and every layer acquired different information respectively. Meanwhile, every layer shared their information with the Decoder mirrored layers by skip connection. It is desirable to concatenate many low-level information such that the location of texture will be passed to the Decoder.

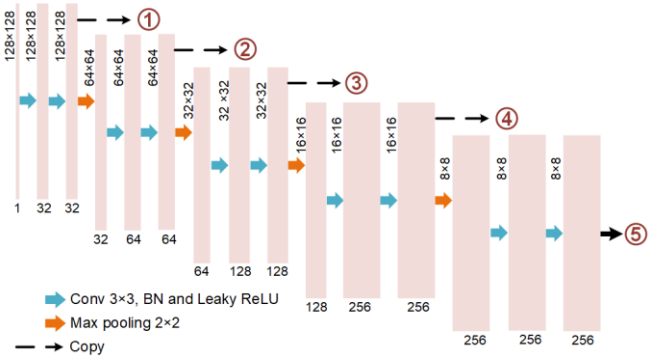

Figure 4. The Encoder I module of Y-Net. The copied features concatenate into Decoder in sequence number.

Encoder II: The Encoder II module is shown in Figure 5, which takes the raw PA signals as input. The structure of every layer is the same as Encoder I except the bottom layer. An extra 20×3 convolution is put on the middle of the bottom layer, which translates the 160×8 features map to 8×8. Meanwhile, the signals have a longer size in time-dimension, and a larger receptive field is desirable to focus more information in this dimension. Every layer also shared their information with the Decoder mirrored layers by resizing and skipping connection. The raw data contains complicated feature, and Encoder II filtrates the feature as a supplement for the information loss of input of Encoder I during the beamforming process.

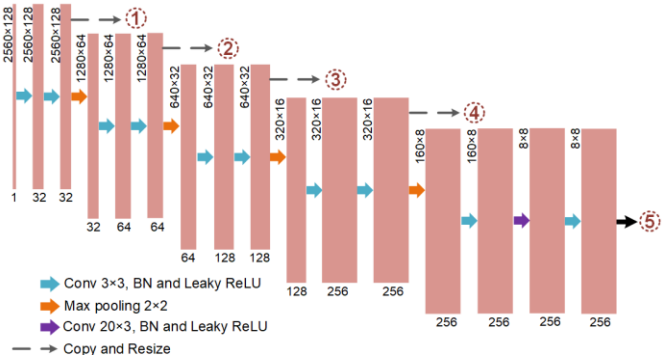

Figure 5. The Encoder II module of Y-Net. The copied features concatenate into Decoder in sequence number.

Decoder: The outputs of the two encoders are taken to the decoder after concatenation. The detail of the Decoder is shown in Figure 6, which are reversed layers compared with Encoder I. Every layer unit is composed of two 3×3 convolution, and an up-convolution to upsample the features. On the other hand, every layer receives low-level information from two encoders' mirrored layers and concatenating with the feature from before layer of the decoder. The final layer will generate a 128×128 image.

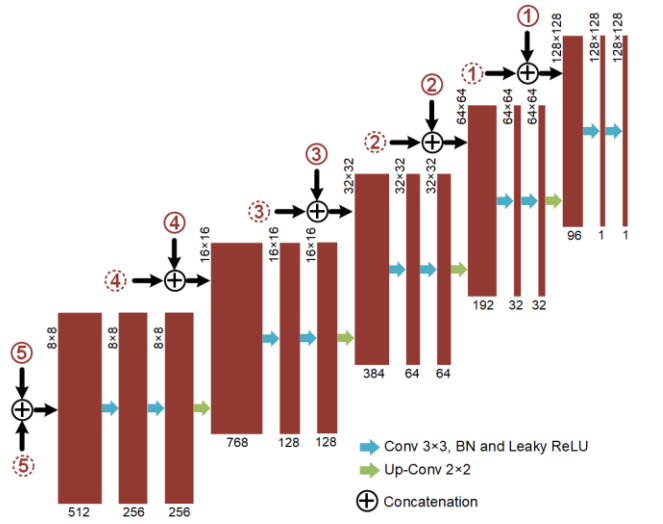

Figure 6. The Decoder module of Y-Net. The sequence number concatenation come from two encoders in order.

III. Training and Results

A. Training the network

The dataset consists 4700 training sets and 400 test sets, which are generated by MATLAB k-Wave toolbox for PA simulation introduced before. We use the mean squared error (MSE) loss function to evaluate the training error, and the Adam optimization algorithm [14] is used to optimize the network iteratively. The MSE loss is shown as:

$$L_{MSE}(x) = \|x - gt\|_F^2 \qquad (1)$$

where x is the reconstruction image, and gt is the ground truth.

The deep learning framework Pytorch is used to implement the proposed Y-Net. The hardware platform we used is a high-speed graphics computing workstation consisting of two Intel Xeon E5-2690 (2.6GHz) CPUs and four NVIDIA GTX 1080Ti graphics cards. The batch size is set as 64, and the running time is 0.453 second. The iteration is set as 1000 epochs, and initial learning rate is 0.005.

B. Results

Three indexes of quantitative evaluation are used as the metric to evaluate the performance of different methods, which are structural similarity index (SSIM), peak signal-to-noise ratio (PSNR) and signal-to-noise ratio (SNR) respectively. We compared two different conventional algorithms and three different models with our proposed approach. Time reversal (TR) and delay-and-sum (DAS) are selected as conventional algorithms for evaluating performance. We also compare two variant Y-Net with our approach, which removes the connection of raw data (Encoder II) and the connection of the beamformed image (Encoder I) with the Decoder respectively. Meanwhile, the post-processing method based U-Net that only input an image after beamforming is also demonstrated for evaluation.

The performance comparison is shown in TABLE I. The conventional algorithms show obvious artifacts, and the information's synthesis for the position far away from the detector is insufficient. On the other hand, the deep-learning-based methods has significant advantages, and the results are encouraging with regard to the proposed network's performance in comparison with the other networks.

TABLE I. Quantitative Evaluation of Different Methods for Test Sets

Algorithms	SSIM	PSNR	SNR
delay-and-sum (DAS)	0.2032	17.3626	1.7493
time reversal (TR)	0.5587	17.8482	2.2350
Y-Net	**0.9119**	**25.5434**	**9.9291**
Y-Net (concatenate BF)	0.8988	25.2708	9.6577
Y-Net (concatenate signals)	0.8622	23.9152	8.105
U-Net	0.9002	25.0032	9.3233

To further visually compare the performance of different methods, four image examples of results are shown in Figure 7. We compare four rows and every method in the same column. From left to right, the method is DAS, TR, Y-Net only concatenates BF into the Decoder, Y-Net only concatenates raw signals into the Decoder, U-Net and the proposed complete Y-Net.

The conventional algorithms are easily fooled by artifacts, and we can still see the appearance of the object roughly. The deep-leaning model based approach almost restores the rough outline of the object, and its performance differs for reconstructing of the details. It is interesting to note that all models connected to BF are susceptible to strong artifacts in BF, and occurred some errors in the details such as third row, fifth column. Y-Net (concatenate signals) can avoid the mentioned problems, but it is difficult to identify at small independent source. The proposed complete Y-Net provides a clearer texture in detail than the U-Net, which indicates that Y-Net is more anti-disturbing to artifacts in BF by integrating the information in raw data. So the performance of Y-Net may be further improved by utilizing more advanced BF algorithm,

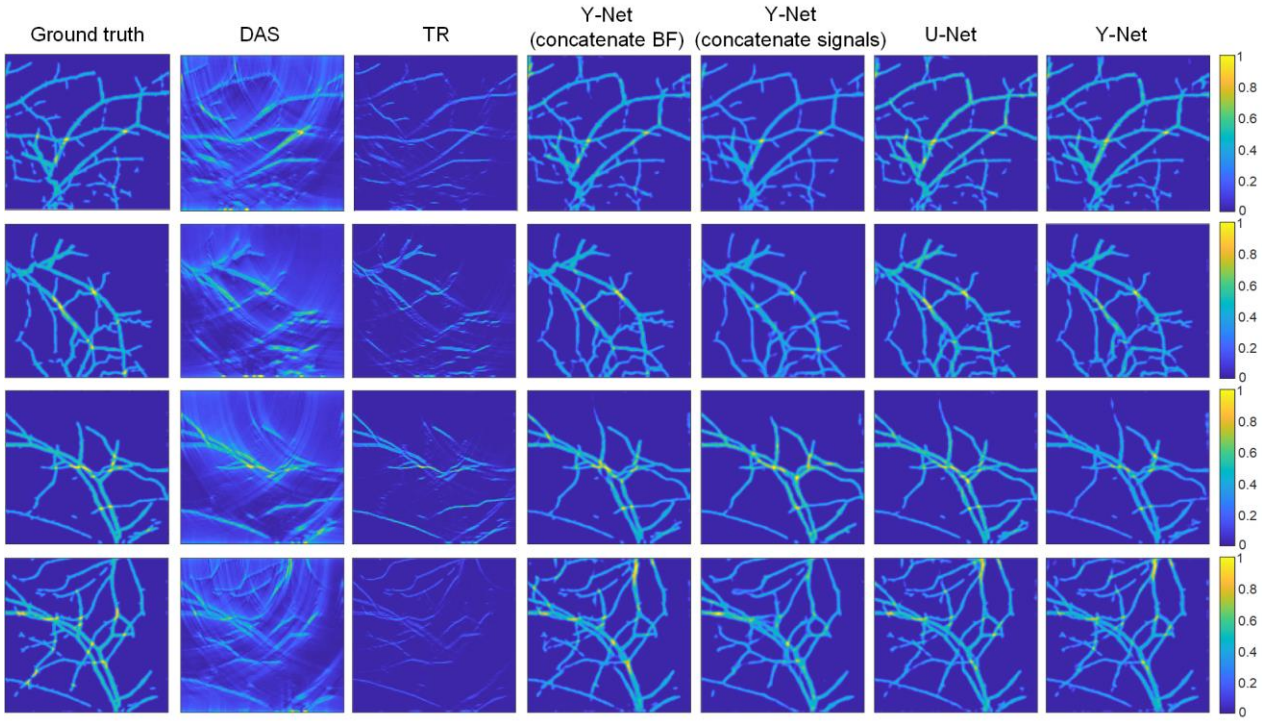

Figure 7. The example of performance comparison that different methods to reconstruct initial pressure. The four examples correspond to different four rows; every column correspond to different method, from left to right: ground truth, DAS, TR, Y-Net only concatenate BF into the Decoder, Y-Net only concatenate raw data into the Decoder, U-Net and Y-Net.

which seamlessly bridges the joint improvements of conventional reconstruction algorithms and deep learning.

IV. Conclusion

In this paper, a new CNN architecture, named Y-Net, is proposed, which consists of two intersecting encoder paths. The Y-Net takes two types of inputs that represent the texture structure of the conventional algorithms and the high-dimensional features contained in the original raw signals respectively. We use k-Wave PA simulation tool to generate a large amount of training data to train the network, and evaluate our approach on the test set. In the experiment, we demonstrate the feasibility and robustness of our proposed method by comparing with other models and conventional methods. Y-Net still is affected by the artifacts of beamforming, which may be improved by using a better beamforming algorithm. In the future work, we will further validate Y-Net using *ex vivo* and *in vivo* data.

References

[1] L. V. Wang and J. Yao, "A practical guide to photoacoustic tomography in the life sciences," *Nat Methods,* vol. 13, no. 8, pp. 627-38, Jul 28 2016.

[2] H. Zhong, T. Duan, H. Lan, M. Zhou, and F. Gao, "Review of Low-Cost Photoacoustic Sensing and Imaging Based on Laser Diode and Light-Emitting Diode," *Sensors (Basel),* vol. 18, no. 7, Jul 13 2018.

[3] H. Lan, T. Duan, H. Zhong, M. Zhou, and F. Gao, "Photoacoustic Classification of Tumor Model Morphology Based on Support Vector Machine: A Simulation and Phantom Study," *IEEE Journal of Selected Topics in Quantum Electronics,* vol. 25, no. 1, pp. 1-9, 2019.

[4] F. Gao *et al.*, "Single laser pulse generates dual photoacoustic signals for differential contrast photoacoustic imaging," *Sci Rep,* vol. 7, no. 1, p. 626, Apr 04 2017.

[5] T. Duan, H. Lan, H. Zhong, M. Zhou, R. Zhang, and F. Gao, "Optical Spectroscopic Ultrasound Displacement Imaging: A Feasibility Study," *IEEE Journal of Selected Topics in Quantum Electronics,* vol. 25, no. 1, pp. 1-8, 2019.

[6] S. Mallidi, G. P. Luke, and S. Emelianov, "Photoacoustic imaging in cancer detection, diagnosis, and treatment guidance," *Trends Biotechnol,* vol. 29, no. 5, pp. 213-21, May 2011.

[7] C. Cai, K. Deng, C. Ma, and J. Luo, "End-to-end deep neural network for optical inversion in quantitative photoacoustic imaging," *Opt Lett,* vol. 43, no. 12, pp. 2752-2755, Jun 15 2018.

[8] A. Hauptmann *et al.*, "Model based learning for accelerated, limited-view 3D photoacoustic tomography," *IEEE Transactions on Medical Imaging,* pp. 1-1, 2018.

[9] D. Waibel, J. Gröhl, F. Isensee, T. Kirchner, K. Maier-Hein, and L. Maier-Hein, "Reconstruction of initial pressure from limited view photoacoustic images using deep learning," in *Photons Plus Ultrasound: Imaging and Sensing 2018*, 2018, vol. 10494, p. 104942S: International Society for Optics and Photonics.

[10] G. Wang, J. C. Ye, K. Mueller, and J. A. Fessler, "Image Reconstruction is a New Frontier of Machine Learning," *IEEE Trans Med Imaging,* vol. 37, no. 6, pp. 1289-1296, Jun 2018.

[11] B. E. Treeby and B. T. Cox, "k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields," *Journal of biomedical optics,* vol. 15, no. 2, pp. 021314-021314-12, 2010.

[12] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. van Ginneken, "Ridge-based vessel segmentation in color images of the retina," *IEEE Trans Med Imaging,* vol. 23, no. 4, pp. 501-9, Apr 2004.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234-241: Springer.

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