SwinIR for Photoacoustic Computed Tomography Artifact Reduction

Varun Shijo
Department of Computer Science and Engineering
University at Buffalo
Buffalo, NY, United States
varunshi@buffalo.edu

Tri Vu
Department of Biomedical Engineering
Duke University
Durham, NC, United States
tri.vu@duke.edu

Junjie Yao
Department of Biomedical Engineering
Duke University
Durham, NC, United States
junjie.yao@duke.edu

Wenyao Xu
Department of Computer Science and Engineering
University at Buffalo
Buffalo, NY, United States
wenyaoxu@buffalo.edu

Jun Xia
Department of Biomedical Engineering
University at Buffalo
Buffalo, NY, United States
junxia@buffalo.edu

2023 IEEE International Ultrasonics Symposium (IUS) | 979-8-3503-4645-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/IUS51837.2023.10307937

***Abstract*— Photoacoustic (PA) imaging is an emerging hybrid medical imaging modality involving the optical excitation of chromophores - light-sensitive molecules like hemoglobin and lipids, to infer underlying vascular structure. Supplying them energy in the form of pulsed laser results in rapid successive thermoelastic expansion and contraction, resulting in the generation of ultrasound, which can then be measured using transducer arrays. Raw sensor data is represented in k-Space from which the Cartesian equivalent is reconstructed using rule-based algorithms. These reconstructions tend to be noisy and have artifacts, but the recent widespread adoption of deep learning has facilitated the post-processing of reconstructions to significantly improve them. UNet, in particular, has had a far-reaching impact on the medical imaging domain, and PA imaging has been no exception, seeing a myriad of solutions based on it. In this paper, we investigate the efficacy of replacing convolution-based feature generation for post-processing PA reconstructions with a Vision Transformer-based (ViT) approach owing to its recent success in computer vision. Specifically, we examine the ability of Shifted Window (Swin) ViTs to restore an artifact-free vascular image from an artifact-heavy image reconstructed using the time-reversal algorithm.**

Keywords—deep learning, photoacoustic computed tomography, artifact removal, SwinIR, swin transformer

I. Introduction

PA signals, as measured by ultrasonic transducer arrays, are represented in k-space, which encodes the time taken by sound to reach the transducer in the spatial-frequency domain. This representation needs to be transformed into Cartesian space to be interpretable for medical diagnostic use. Deterministic reconstruction algorithms like back-projection and time-reversal, which have been the standard, yield low-resolution, noisy results that amplify the artifacts introduced by the hardware setup. Reconstructed results from scans using linear array transducers are especially susceptible to limited-view artifacts, which manifest as curved stripe features around the object being imaged[1].

Convolution-based architectures like UNet for post-processing reconstructions from deterministic methods have proven effective at reducing artifacts. However, compared to convolution-based methods, Vision-Transformer (ViT) [2] based methods that embed patches of a given image have proven superior in the ability to capture long-range dependencies that exist beyond a convolutional receptive field[3], [4]. This is achieved by modeling image patches like text and generating embeddings that map each patch into a global semantic space.

We propose to use a ViT-based architecture – SwinIR[5], which is an image-to-image mapping technique primarily targeted at image superresolution and denoising that uses a type of ViT called a Shifted Window (Swin) Transformer[6]. The motivation for this approach is the potential for the model to learn when vasculature is implicitly continuous, unlike in convolution-based methods, which often yield results with gaps in vasculature. Our results show that SwinIR outperforms existing state-of-the-art methods like WGAN-GP.

One of the major reasons behind the ubiquity of UNet has been the accuracy it affords while also being memory efficient, and most attempts to improve upon accuracy have been at the cost of this efficiency[7]. Furthermore, there is a lack of data owing to the nascence of the PA field, and due to the nature of the problem, it is prohibitive to visualize the ground truth vascular structure being imaged except through the use of alternative imaging modalities, which is why the domain sees extensive use of simulated data to train models.

Swin Transformers addressed the computational complexity and scalability challenges intrinsic to ViTs, as a result of which, the application of transformer-based architectures became viable while also being comparable to UNet with regard to inference speed and memory efficiency.

This paper investigates the viability of using ViT-based architectures as an alternative to conventional convolution-based architectures for artifact removal from reconstructed photoacoustic images. Specifically, we propose to use SwinIR to post-process reconstructed photoacoustic images of Two-Photon Microscopy (TPM) brain scans to denoise them and reduce limited-view, limited bandwidth, and sparse sampling artifacts.

This work was supported by grants from the National Institutes of Health (Nos. R01EB029596 and R01EB028978).

II. Background and Preliminaries

A. PACT Theory and Methods

A typical photoacoustic imaging setup consists of a pulsed energy source - particularly, 1064nm laser pulsed at 10HZ - for optical excitation, and ultrasonic transducer arrays that measure the pressure distribution resulting from rapid thermoelastic expansion and contraction of lipids and hemoglobin. Multiple form factors of ultrasound transducer arrays are documented in contemporary literature - linear arrays, 2D arrays, and ring arrays [8] being a few of the most commonly discussed. Linear arrays are the most economical and are used in consumer ultrasound owing to their cost-effectiveness, convenience of scanning, and the relative ease with which k-space signals can be reconstructed.

B. Deep Learning-based methods

Deep learning-based approaches have enabled rapid advancements in PA reconstruction quality by doing away with the necessity for hand-crafted deterministic algorithms while matching or improving upon computational performance. Waibel et al. [9] broadly categorize deep learning-based approaches into post-processing and direct reconstruction, where the former consists of models that take the result of a deterministic approach as the input and further process them, whereas the latter operates directly on sensor data in k-space.

C. Dataset

To train and evaluate SwinIR, we use Vu et al.'s adapted Two-Photon Microscopy (TPM) brain vasculature dataset from [1] since the vasculature structures are similar to PACT target vascular structures in their complexity and distribution. In addition to the TPM images themselves, the dataset contains TPM structures processed in the k-Wave toolbox in MATLAB to generate simulated sensor readings. These are then subjected to random addition of noise, reconstructed using a deterministic method - the results of which constitute what we refer to as bipolar domain inputs and contain limited-view and limited-bandwidth artifacts. These bipolar domain inputs are converted to the unipolar domain using the Hilbert transform, subjected to random addition of Gaussian noise, and normalized between 0 and 1. Models trained and evaluated in [1] use the Hilbert-space/unipolar images as inputs and TPM images as targets. We do the same for SwinIR and P-SwinIR to compare performance.

TABLE I. Simulation Parameters

Parameter	Value
Transducer elements	128
Central frequency	5MHz
60% detection bandwidth	3MHz
Speed of sound	1540ms^{-1}

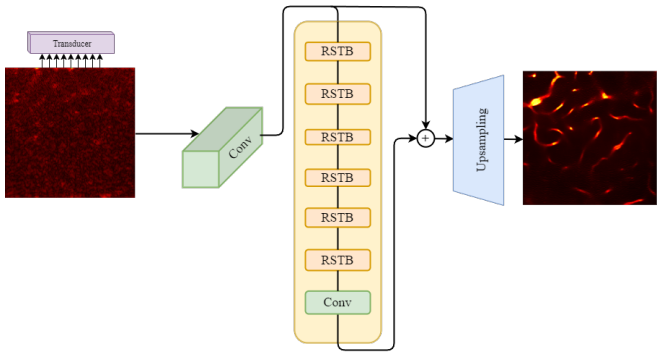

Fig. 1. The SwinIR architecture consists of a shallow feature extraction convolution layer followed by the deep feature extraction block and a final upsampling convolution layer. The output of the shallow feature extraction is concatenated with the output of the deep feature extraction block.

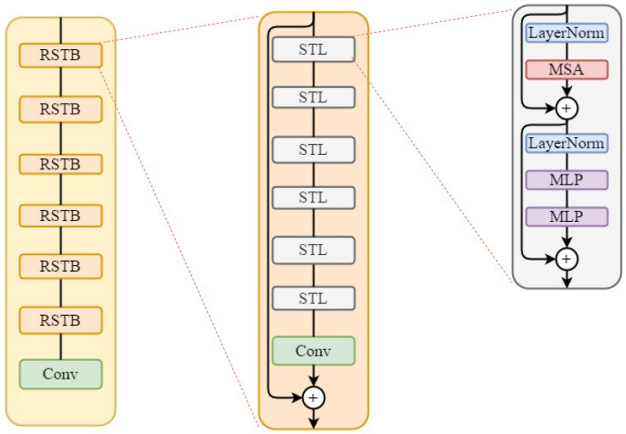

Fig. 2. The architecture of the deep feature extraction block is comprised of Residual Swin Transformer Blocks (RSTBs), which in turn are built by stacking Swin Transformer Layers (STLs). Each STL uses Multihead Self-Attention (MSA) followed by MultiLayer Perceptrons (MLPs) with Layer Normalization.

The simulation parameters in k-Wave are configured to approximate the L7-4 linear transducer array specification, as outlined in Table 1.

III. Materials and Methods

In this paper, we propose the use of a SwinIR – a transformer-based deep neural architecture designed for image restoration, denoising, and mapping; to post-process PA images reconstructed by time-reversal.

A. Swin Transformer

Swin (Shifted Window) Transformers[6] were introduced as a solution to challenges unique to adapting Transformers – initially introduced to solve natural language processing problems by modeling a latent language space and embedding tokens within it. Vision Transformers (ViTs) were the first solution to adapt Transformers to the image domain without using any form of convolution. ViTs split an image into patches of a constant dimensionality and embed them along with their position using a standard transformer

encoder. While its results were promising, the high variance in the scale of objects and the generally higher resolution of information to be modeled compared to language remained to be addressed. Swin Transformers, by dividing the image into patches of multiple scales, can model long-range dependencies within the input image while simultaneously being capable of handling large images owing to local attention.

B. SwinIR

SwinIR was proposed as a baseline method for image restoration using attention instead of convolution as the "spatial token mixer" - following the nomenclature from [3]; similar to how UNet has become the de-facto baseline for most medical imaging tasks as reported in [7]. It is a three-stage architecture that performs shallow and deep feature extraction followed by image reconstruction. The deep feature extraction block consists of Residual Swin Transformer Blocks (RSTBs), which, we hypothesize, are more effective at modeling vasculature than convolution-based methods owing to a larger effective receptive field (ERF)[4]. SwinIR was chosen as the foundation for investigating the effectiveness of transformer-based image reconstruction for PA images on the basis of its performance on standard benchmarks in the computer vision domain. Since it is built upon Swin Transformers, it solves the problem of modeling long-range dependencies within an image and is content-aware, unlike convolution-based methods like UNet.

$$L = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{W}\sqrt{\left(I_{\text{pred}}(i,j) - I_{\text{gt}}(i,j)\right)^2 + \epsilon^2} \qquad (1)$$

We minimize the Charbonnier loss function (1) using the AdamW[10] optimizer to guide the training process. The use of Charbonnier loss allows for the benefits of L1 loss, yielding sharper results than L2 while avoiding division by 0 through the addition of a small constant ϵ term set to 1e-9 following the original SwinIR implementation. We use 6 RSTBs with 6 Swin Transformer Layers (STLs) each, where each STL has a Multilayer Perceptron (MLP) ratio of 2. We set the embedding dimension to 180 to allow for effective modeling of the variance of the distribution of vascular structures. Patches of size 128x128 were used to train the model, with a window size of 8x8, resulting in a larger ERF than in conventional convolution-based models. The optimizer's learning rate was set to 5e-5 with a weight decay factor of 0.01 and a learning rate scheduler that halves the learning rate every 20 epochs. Training was performed on a single Nvidia RTX 2080Ti using the PyTorch framework. The total dataset consisted of 9531 TPM images with their corresponding TR reconstructions of simulated PACT. This dataset was then subjected to a train-test split of 80:20, resulting in a training set of 7625 image pairs and a test set of 1906 image pairs. The model was trained for 85 epochs over a span of 64 hours.

IV. Results

Structural Similarity (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Multiscale Structural Similarity (MS-SSIM) were calculated using the predictions of each method for each sample and averaged over the test set. SwinIR achieves significantly higher quality than UNet, which is reflected in the 1dB improvement in PSNR, and the positive changes in SSIM and MS-SSIM. SwinIR also outperforms WGAN-GP on all three metrics.

SwinIR is able to implicitly learn the continuity intrinsic to vessel structures by modeling it in a manner similar to natural language using the attention mechanism, allowing it to outperform purely convolution-based feature extraction methods.

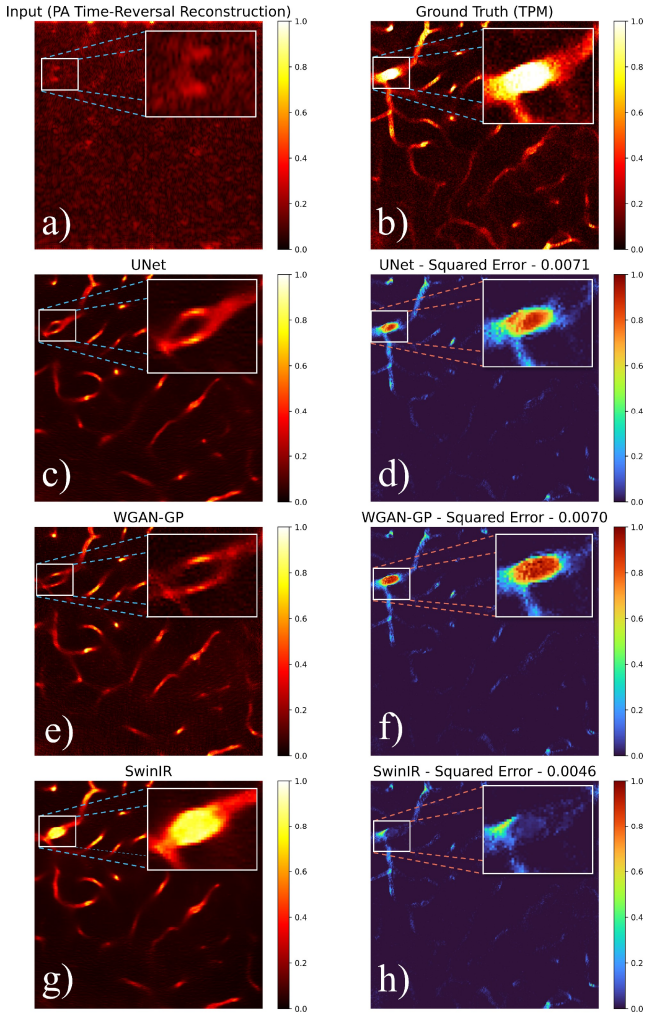

Fig. 3. Post-processing results from models. a) PACT image reconstructed using the deterministic Time Reversal algorithm. b) Two Photon Microscopy (TPM) vascular image of the brain used as ground truth for the models. c), e), and g) are predictions from UNet, WGAN-GP, and SwinIR, respectively, with their corresponding squared error maps d), f), and h). The insets highlight the ability of SwinIR to recover true vessel structure even when limited view artifacts – manifested as curved stripe features around the imaged object, are present.

TABLE II. QUANTITATIVE METRICS

Model	PSNR (dB)	SSIM	MS-SSIM
UNet	25.96 ± 2.51	0.64 ± 0.11	0.86 ± 0.07
WGAN-GP	26.14 ± 2.16	0.64 ± 0.08	0.87 ± 0.05
SwinIR	**26.93 ± 2.74**	**0.66 ± 0.11**	**0.89 ± 0.06**

V. CONCLUSION

In this work, we propose the use of SwinIR – an image-to-image mapping model originally developed to solve image denoising and superresolution problems in the computer vision domain – for post-processing PACT images that are reconstructed using conventional algorithms like Time Reversal. We demonstrate that SwinIR is effective at removing noise from the reconstructions and also recovering vessel structures that are otherwise obscured due to hardware constraints like limited view and limited bandwidth. Our approach outperforms existing methods like UNet, which is the de facto baseline for most medical imaging-related deep learning solutions, and WGAN-GP, which was introduced for the express purpose of reducing artifacts in PACT reconstructions.

VI. REFERENCES

[1] T. Vu, M. Li, H. Humayun, Y. Zhou, and J. Yao, "A generative adversarial network for artifact removal in photoacoustic computed tomography with a linear-array transducer," *Exp. Biol. Med.*, vol. 245, no. 7, pp. 597–605, Apr. 2020, doi: 10.1177/1535370220914285.

[2] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, Jun. 03, 2021. doi: 10.48550/arXiv.2010.11929.

[3] J. Dai *et al.*, "Demystify transformers & convolutions in modern image deep networks." arXiv, Nov. 10, 2022. Accessed: Nov. 14, 2022. [Online]. Available: http://arxiv.org/abs/2211.05781

[4] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "Understanding the Effective Receptive Field in Deep Convolutional Neural Networks." arXiv, Jan. 25, 2017. doi: 10.48550/arXiv.1701.04128.

[5] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "SwinIR: image restoration using swin transformer." arXiv, Aug. 23, 2021. doi: 10.48550/arXiv.2108.10257.

[6] Z. Liu *et al.*, "Swin transformer: hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[7] R. Azad *et al.*, "Medical image segmentation review: the success of U-net." arXiv, Nov. 27, 2022. Accessed: Dec. 08, 2022. [Online]. Available: http://arxiv.org/abs/2211.14830

[8] J. Xia, J. Yao, and L. V. Wang, "Photoacoustic tomography: principles and advances," *Electromagn. Waves Camb. Mass*, vol. 147, pp. 1–22, 2014.

[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*, SPIE, Feb. 2018, pp. 196–203. doi: 10.1117/12.2288353.

[10] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization." arXiv, Jan. 04, 2019. doi: 10.48550/arXiv.1711.05101.