PAPER

Hybrid-supervised deep learning for domain transfer 3D protoacoustic image reconstruction

IOP Publishing Phys. Med. Biol. 69 (2024) 085007 https://doi.org/10.1088/1361-6560/ad3327

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
21 July 2023

REVISED
26 February 2024

ACCEPTED FOR PUBLICATION
12 March 2024

PUBLISHED
3 April 2024

PAPER

Hybrid-supervised deep learning for domain transfer 3D protoacoustic image reconstruction

Yankun Lang[1], **Zhuoran Jiang**[2], **Leshan Sun**[3], **Liangzhong Xiang**[3] **and Lei Ren**[1]

[1] Department of Radiation Oncology Physics, University of Maryland, Baltimore, Baltimore, MD 21201, United States of America
[2] Department of Radiation Oncology, Duke University, Durham, NC 27710, United States of America
[3] Department of Biomedical Engineering and Radiology, University of California, Irvine, Irnive, CA, 92617, United States of America

E-mail: lren@som.umaryland.edu

Keywords: proton therapy, dose verification, protoacoustic reconstruction, self-supervised deep learning

Abstract

Objective. Protoacoustic imaging showed great promise in providing real-time 3D dose verification of proton therapy. However, the limited acquisition angle in protoacoustic imaging induces severe artifacts, which impairs its accuracy for dose verification. In this study, we developed a hybrid-supervised deep learning method for protoacoustic imaging to address the limited view issue. *Approach*. We proposed a Recon-Enhance two-stage deep learning method. In the Recon-stage, a transformer-based network was developed to reconstruct initial pressure maps from raw acoustic signals. The network is trained in a hybrid-supervised approach, where it is first trained using supervision by the iteratively reconstructed pressure map and then fine-tuned using transfer learning and self-supervision based on the data fidelity constraint. In the enhance-stage, a 3D U-net is applied to further enhance the image quality with supervision from the ground truth pressure map. The final protoacoustic images are then converted to dose for proton verification. *Main results*. The results evaluated on a dataset of 126 prostate cancer patients achieved an average root mean squared errors (RMSE) of 0.0292, and an average structural similarity index measure (SSIM) of 0.9618, out-performing related start-of-the-art methods. Qualitative results also demonstrated that our approach addressed the limit-view issue with more details reconstructed. Dose verification achieved an average RMSE of 0.018, and an average SSIM of 0.9891. Gamma index evaluation demonstrated a high agreement (94.7% and 95.7% for 1%/3 mm and 1%/5 mm) between the predicted and the ground truth dose maps. Notably, the processing time was reduced to 6 s, demonstrating its feasibility for online 3D dose verification for prostate proton therapy. *Significance*. Our study achieved start-of-the-art performance in the challenging task of direct reconstruction from radiofrequency signals, demonstrating the great promise of PA imaging as a highly efficient and accurate tool for *in vivo* 3D proton dose verification to minimize the range uncertainties of proton therapy to improve its precision and outcomes.

1. Introduction

Proton therapy is a radiation treatment where proton beams are delivered to the target to disrupt and destroy tumor cells. After the protons enter the patient's body, the absorbed dose increases gradually at the beginning and then substantially at the end of the proton travel path, reaching a peak called Bragg peak (BP), before dropping off sharply. This finite range and sharp dose falloff at the distal end of the BP increase our ability to conform radiation therapy treatment dose to the tumor and minimize collateral damage to neighboring critical organs. However, the precision of proton therapy is highly affected by the variations of patient positioning, anatomic structures, and dose calculation errors due to the sharp dose falloff of the BP. A small delivery error could cause a significant underdose to the target and an overdose to the healthy tissues. Therefore, online 3D dose verification during treatment is highly desirable in proton therapy to verify and minimize dose delivery errors to maximize its efficacy.

Over the years, many *in vivo* dose verification methods have been developed to address this clinical need. For example, methods were developed to verify the proton dose range by measuring the dose or fluence with wireless implantable dosimeters (Lu *et al* 2010, Bentefour *et al* 2012, Telsemeyer *et al* 2012). However, these methods are not capable of fully verifying the tumor and organ at risk (OAR) dose since they don't provide the 3D volumetric information. Proton Radiology (Schneider *et al* 2004, Penfold *et al* 2009, Schneider *et al* 2012) technique is designed for the direct measurement of the range of proton beams by using dedicated proton beams for delivery and imaging, distinct from the beams used for treatment. However, these methods come with limited image resolution and they do not provide verification for the range of the actual treatment beams, making them lack the capability to confirm the precise delivered dose during the treatment process. Proton dose deposition can also be verified by measuring the surrogate data generated by proton irradiation. For example, Positron emission tomography (PET) (Fiedler *et al* 2008, 2010, Miyatake *et al* 2010, Nishio *et al* 2010) and prompt gamma (PG) imaging (Polf *et al* 2009, Kormoll and Compton *et al* 2011, Min *et al* 2012, Draeger *et al* 2018, Pietsch *et al* 2023) detects the gamma rays generated by irradiation along the proton beam path. Yuan *et al* (2013) used magnetic resonance imaging (MRI) to detect the radiobiological change of liver tissue after radiation. Specifically, MRI images were registered to the planning computed tomography (CT) images. Then MR signal intensity (SI) was correlated to the radiation dose. Finally, dose-SI correlation was employed on registered MR images to estimate the proton end-of-range. In summary, methods utilized in PET and MRI lack real-time dose verification capabilities during treatment, while prompt gamma imaging methods still contend with the challenges posed by limited accuracy arising from low signal intensity and the absence of 3D volumetric information. Although recent studies employed deep learning to obtain volumetric information in PG imaging, its efficacy and robustness in real patient applications remain to be validated.

In recent years, protoacoustic (PA) imaging has been developed to detect proton-induced raw acoustic (RA) signals for dose verification (Ahmad *et al* 2015, Carlier *et al* 2020, Yu *et al* 2021). Specifically, the proton beam creates heat during the dose deposition, causing tissue expansion and contraction to generate acoustic waves, which can be detected by ultrasound transducers. Positioned strategically, these transducers detect the acoustic waves and convert them into digitized acoustic signals. Subsequently, these raw acoustic signals are utilized in the reconstruction of a pressure map and derive the corresponding dose deposition. Many researchers have conducted simulations on 2D CT images to verify dose range with protoacoustic signals (Yu *et al* 2019b, Freijo *et al* 2021, Yao *et al* 2021). More recently, matrix array transducers (Yu *et al* 2019a, Wang *et al* 2020) have been utilized for 3D ultrasound imaging, which showed a potential to provide real 3D online dose verification. The initial pressure map is reconstructed from the RA signals, and then related to the dose deposition. Traditional algorithms of reconstruction from the signal domain has been proposed. For example, universal back projection (UBP) (Xu and Wang 2005) projects the quantity calculated from the transducer measurements backward on a spherical surface within a solid angle, which is integrated to obtain the pressure with respect to position. This method suffers from distortion due to that tissue heterogeneity was not considered. Time reveral (TR) (Hristova *et al* 2008, Treeby *et al* 2010) is a method that iteratively updates the current pressure by adding the residual errors calculated with time reversed back-projection. Despite the progress, the reconstructed PA pressure map still suffers from severe distortion and artifacts due to the limited-angle view of the matrix array detector, limting its accuracy for dose verification.

Deep learning-based methods have been developed in recent years to improve image reconstruction (Chen *et al* 2018, Lan *et al* 2020, Luo *et al* 2021, Chen *et al* 2022). Zhu *et al* (2018) proposed a network that performs image reconstruction from RA signals directly by mapping the dual domain (signal-to-image) correlations with fully connected (FC) layers. Then they used a series of convolutional layers to denoise the output. However, this method is limited to memory capacity when dealing with high-resolution protoacoustic images. Häggström *et al* proposed encoder-decoder architecture called DeepPET for direct PET image reconstruction (Häggström *et al* 2019). To reduce memory consumption, DeepPET used convolutional layers rather than FC layers to learn a latent space representing the dual domain correlations. The latent space was then upsampled in the decoder to restore the image. However, this method ignored the consistency in the signal domain without accounting for the data fidelity constraint. Zhang *et al* (2021) proposed a self-supervised learning method for ultrasound image reconstruction. The model is trained based on the data fidelity constraint, which minimizes the difference between the sinogram projected from the reconstructed image and the initially measured sinogram. Although this method has demonstrated improved reconstruction accuracy in ultrasound images, it still suffers from severe distortion artifacts when applied for protoacoustic images due to the limited angle view issue. In response to constraints encountered in image reconstruction, the utilization of deep learning has been advanced for the purpose of enhancing images post-reconstruction, as evidenced by the work of Jiang *et al* (2019). In mitigating challenges associated with limited view PA reconstruction, Jiang *et al* (2022) utilized a 3D U-net that enhances an initial pressure map reconstructed by TR to reduce the distortion artifact, then derived a 3D dose map for dose verification. Despite the improvements, the efficacy of deep learning enhancement is limited by the quality of the initial reconstruction. The initial pressure map reconstructed by the TR method suffers from severe distortion

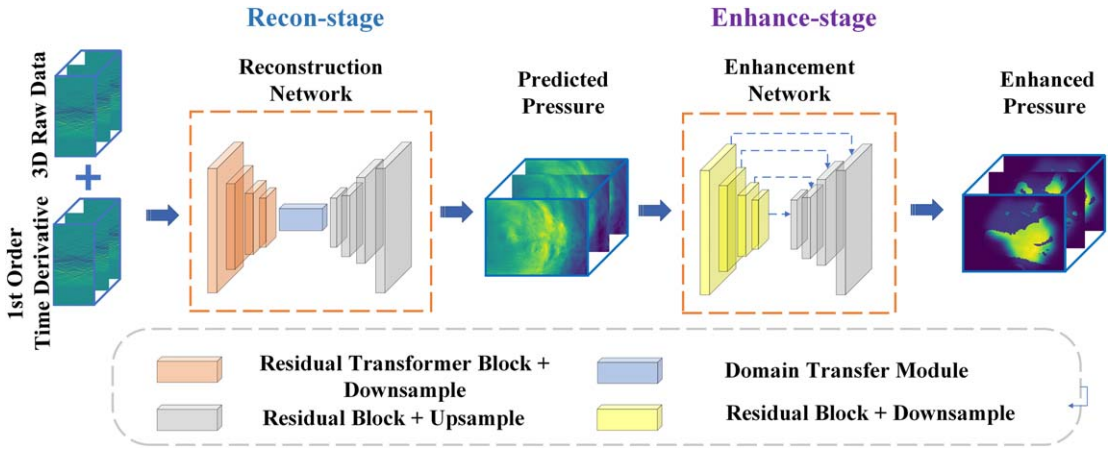

Figure 1. The workflow of our approach for protoacoustic image reconstruction with a Recon-enhance strategy.

with many detailed anatomical structures lost, which consequently impairs the accuracy of the image enhancement afterward. Meanwhile, this method suffers from time consuming in testing stage since TR method needs numerous time (120 s) for reconstruction. The low efficiency makes this method impractical for online dose verification. Recently, transformer network has been applied in various medical imaging research (Parmar *et al* 2018, Matsoukas *et al* 2021) due to its long-range dependency and adaptive self-attention characteristics. Swin Transform (Liu *et al* 2021) was proposed with moving receptive field windows of reduced size to greatly reduce the computational complexity. Huang *et al* (2022) utilized a Swin transformer-based generator to enhance the quality of k-space downsampled MRI images. A discriminator was used to distinguish the enhanced result from ground truth to improve the accuracy further. This work demonstrated that the transformer-based models showed great performance in enhancing MRI image quality after reconstruction.

To address the limited angle view problem in PA imaging and further improving the reconstruction quality, in this study, we proposed a deep learning-based protoacoustic image reconstruction method, where a Recon-enhance two-stage strategy is applied as shown in figure 1 to harness the power of deep learning for both image reconstruction and post-reconstruction enhancement. Specifically, in the Recon-stage, the proposed network directly reconstructs the image from RA signals with hybrid supervision and transfer learning. In the enhance-stage, a 3D U-net is applied to further improve the image quality. Compared with the method in Jiang *et al* (2022), where the reconstruction was implemented by Time Reversal, our approach directly reconstructs the initial pressure map from raw RA signals, which can reduce the processing time and improve the accuracy since more essential structural information can be preserved. The main contributions of our article are multi-fold: (1) an end-to-end image reconstruction and enhancement strategy using deep learning is developed for PA imaging to improve its quality; (2) we apply convolutional layers rather than fully connected layers to construct a domain-transfer module to address the memory consumption problem, while maintaining a higher inference speed; (3) we replace the general convolutional layers with transformers to build our network for its long-range dependency, and proposed a novel hybrid supervison method to keep the data fidelity consistency; (4) the proposed method is evaluated on protoacoustic data generated from the CT images and clinical treatment plans of prostate cancer patients, demonstrating the feasibility of high precision 3D dose verification in proton therapy.

2. Methods

2.1. Problem formulation

During protoacoustic process, proton deposits energy when traveling through the patient's body, causing tissue temperature to rise and generating acoustic signals, which can be formulated as:

$$\left(\nabla^2 - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\right)p(\boldsymbol{r},\, t) = -\frac{\Gamma}{c^2}H(\boldsymbol{r})\frac{\partial \delta(t)}{\partial t}, \tag{1}$$

where $p(\boldsymbol{r}, t)$ denotes the measured pressure at location $\boldsymbol{r}$ at time t. $H(\boldsymbol{r})$ denotes the initial pressure. c is speed of sound in the medium. Γ is the dimension less Grüneisen parameter, and $\delta(t)$ denotes the delta function. The objective of our study is to reconstruct the initial map $H(\boldsymbol{r})$ from the measurements $p(\boldsymbol{r}, t)$.

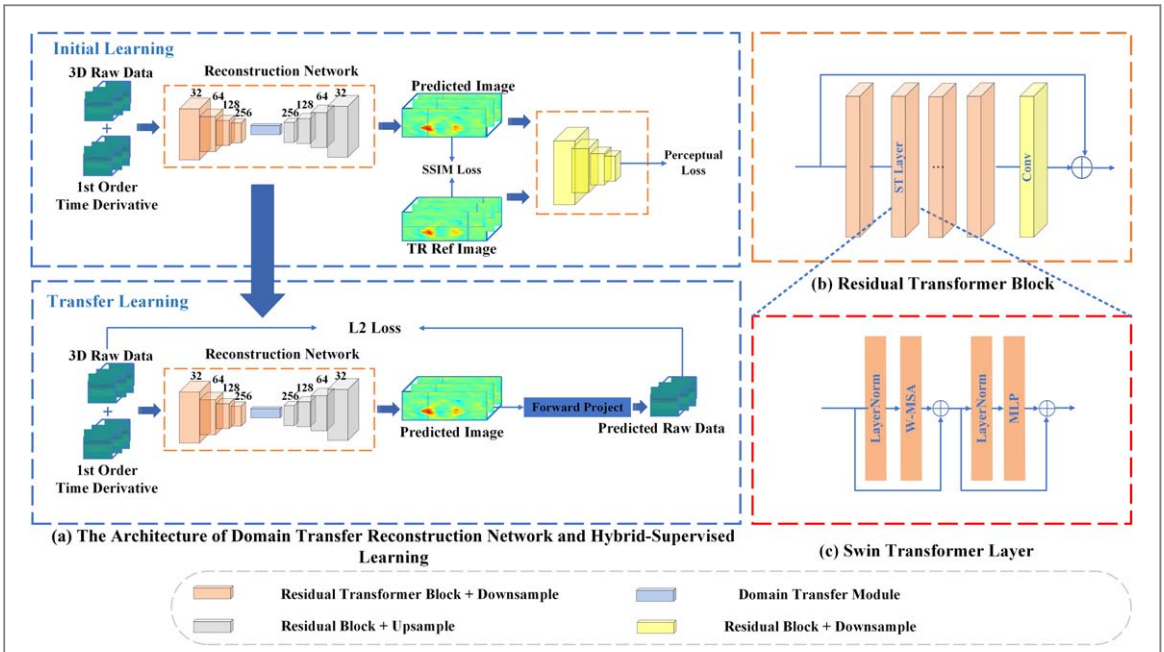

Figure 2. The architecture of the proposed DTR network. (a) Two-step hybrid-supervised training where the network learning weights are shared across the initial learning and transfer learning steps. (b) Details of the residual transformer block. (c) Details of the swin transformer layer.

UBP (Xu and Wang 2005) is a linear reconstruction method derived from equation (1), which can be formulated as:

$$H((\boldsymbol{r})) = \sum_{i=1}^{N} b(\boldsymbol{d_i}, t) = \frac{|\boldsymbol{r} - \boldsymbol{d_i}|}{c} \frac{\Delta\Omega_i}{\sum_{i=1}^{N}\Delta\Omega_i} \tag{2}$$

where $\boldsymbol{d_i}$ and $\Delta\Omega_i$ denote the position and solid angle, respectively. $b(\boldsymbol{d_i}, t)$ is the back projection term of the i-th transducer, which can be formulated as:

$$b(\boldsymbol{d_i}, t) = 2p(\boldsymbol{d_i}, t) - 2t\frac{\partial p(\boldsymbol{d_i}, t)}{\partial t} \tag{3}$$

Apparently, the reconstruction of the initial pressure map $H(\boldsymbol{r})$ from the measurements $p(\boldsymbol{r}, t)$ critically depends on the first-order partial derivative $\partial p(\boldsymbol{d_i}, t)/\partial t$, which can be used as prior knowledge for our model design.

Direct reconstruction of high-quality initial pressure map from RA signals is challenging since the network needs to balance the domain transfer for image reconstruction and the enhancement to correct the distortions caused by limited view in PA images. To address this problem, a recon-enhance strategy is proposed, as shown in figure 1, to first use a network for image reconstruction to generate an initial pressure map with reasonable quality. Then another network is applied afterward to further enhance the reconstructed images.

2.2. Domain transfer reconstruction network (DTR-Net)

The overview of the proposed DTR-Net is shown in figure 2(a). DTR-Net utilizes a contracting-expanding architecture, taking both the 3D RA image $\mathbf{S} \in R^{H^s \times W^s \times D^s}$ and the corresponding first order derivative image $\partial\,\mathbf{S}/\partial\,t \in R^{H^s \times W^s \times D^s}$ as input, where H^s, W^s and D^s represent the height, width and depth of the RA image, respectively. The contracting path consists of four residual transformer blocks (RTBs) followed by down-sampling layer to extract high level features as shown in figure 2(a). Each RTB shown in figure 2(b) is built by several 3D Swin transformer (ST) layers shown in figure 2(c) due to its characteristic of long-range dependency. Swin transformer (Liu *et al* 2021) was developed from the original transformer layer where window based multi-head self-attention (W-MSA) is implemented. Specifically, given a feature map, the ST layer first partitions the input into several non-overlapping windows. For each local window feature F, the query (Q), key (K) and value (V) matrices are calculated by:

$$Q = FP_Q,\ K = FP_K,\ V = FP_V, \tag{4}$$

where P_Q, P_K and P_V are the projection matrices. Then, a self-attention mechanism is applied to calculate the attention matrix by:

$$Att(Q,\ K,\ V) = \sigma(QK^T/+B)V, \tag{5}$$

where B is the relative positional encoding. σ denotes the softmax activation fuction. The final output of the ST layer is computed as:

$$F_{\text{att}} = W - \text{MSA}(\text{Norm}(F)) + F, \tag{6}$$

$$F_{\text{ST}} = \text{MLP}(\text{Norm}(F_{\text{att}})) + F_{\text{att}} \tag{7}$$

where Norm denotes layer normalization. MLP denotes multi-layer perceptron with two fully connected layers for further feature transformations. Residual connection is applied here for feature consistency as shown in figure 2(c). We applied 2, 4, 8 and 16 STs in each RTB respectively to extract hierarchy features. The final extracted feature map $F^S \in R^{\frac{H^s}{16}\times\frac{W^s}{16}\times\frac{D^s}{16}}$ is fed into a domain transfer module, which is simply built by a learnable convolution layer, resizing the feature map from $\frac{H^s}{16} \times \frac{W^s}{16} \times .\frac{D^s}{16}$ to $\frac{H^i}{16} \times \frac{W^i}{16} \times \frac{D^i}{16}$, where H^i, W^i and D^i represent the height, width and depth of the initial pressure map, respectively. The expanding path consists of 4 residual blocks, each of them is built by a up-sampling layer, and two consistent 3D convolution layers with $3 \times 3 \times 3$ kernel, followed by ReLU activation and group normalization layers. Finally, a convolution layer with a kernal size $1 \times 1 \times 1$ is applied to output the reconstructed initial pressure map $P \in R^{H^i\times W^i\times D^i}$. Notably, different to U-net, skip connection is not applied for the following reasons: (1), feature map size inconsistency. Since the size of the feature maps in the contracting and expanding paths are different, skip connection cannot be directly applied; (2) domain inconsistency. The features in the contracting and expanding paths are extracted from two different domains (signal domain and image domain), it is not reasonable to simply concatenate them by a skip connection. Moreover, adding domain transfer module to each skip connection could increase GPU memory consumption.

The reconstruction network is trained using hybrid supervision with transfer learning, as explained below:

2.2.1. Initial training

As shown in figure 2(a), in the initial training, the model is trained to reconstruct PA images by minimizing the difference between the reconstructed pressure map P by the model and the reference pressure map P^* reconstructed by the TR method. Since iterative TR can recover most of the reconstruction details, we utilize the TR results as the reference for initial training. Contrary to the $l2$ and $l1$ loss, the structural similarity index measure (SSIM) loss provides a measure of the similarity by comparing two images based on luminance similarity, contrast similarity and structural similarity information. As the main task in initial training procedure is to reconstruct the structural details, we apply the SSIM loss $L_{\text{ssim}}(P,\ P^*)$ to train the network. Besides, we also apply perceptual loss $L_{\text{perc}}(P, P^*)$ that calculates the difference between features yielded by a designed VGG network to further enhance the stability of the reconstruction. The training loss is defined as:

$$L_I = \alpha_1 L_{\text{ssim}}(P,\ P^*) + \beta_1 L_{\text{perc}}(P,\ P^*), \tag{8}$$

where $\alpha_1 = 1.0$ and $\beta_1 = 0.025$ are the training weights that have been set empirically. This step enables DTR-Net focus on discovering the most representative features for fast reconstruction.

2.2.2. Transfer learning

Considering that the inverse problem is ill-posed and the TR reconstruction is prone to artifacts itself, we applied self-supervised transfer learning to further improve the reconstruction network based on data fidelity constraint. Specifically, as shown in figure 2(a), the network is fine-tuned using transfer learning and self-supervision based on data fidelity constraint, which forces the projected RA data from the reconstructed images to match the measured raw data. The forward projection of RA data from the reconstructed images is carried out using Matlab k-wave toolbox (Treeby and Cox 2010). Empirically, we found that using a $l2$ loss is more efficient than SSIM loss or $l1$ loss for regression in signal domain. Thus, the loss function L_{TL} used for transfer learning (TL) is defined as a $l2$ loss to focus on eliminating the difference between the input RA signal S and the predicted RA data S^* in the data fidelity constraint, as shown below:

$$L_{\text{TL}} = \sum_{i=1}^{N}(S_i - S_i^*)^2 \tag{9}$$

where N denotes the entire number of image voxels. This step enables the network to further fine-tune the reconstruction solely based on the raw data, thus removing the impact of imperfect supervision by the TR reconstructed images in the initial training to improve the reconstruction quality.

2.3. Enhancement and dose conversion network

Due to the limited angle scan of PA imaging, the image generated by the reconstruction network can still have residual artifacts, such as image distortion. A 3D U-net will be applied to further enhance the reconstructed images to address the residual artifacts. The network has the same architecture and parameter setting as

Table 1. Tissue-specific parameter setting for RA signal simulation. v, ρ and Γ refer to the speed of sound, tissue density and the Grüneisen parameter, respectively. α denotes the attenuation coefficient.

Tissue	HU value	v (m s^{-1})	ρ (kg m^{-3})	Γ	$\rho \times \Gamma$ (kg m^{-3})	α (dB/cm/MHz)
Air	[–1000, –200)	—	—	—	—	—
Water	Air overwritten	1500	1000	0.11	110	0.0022
Fat	[–200, –50)	1480	920	0.80	736	0.5
Soft tissue	[–50, 100)	1540	1040	0.30	312	1
Bone	[100, max)	2000	1900	0.80	1520	10

proposed in Jiang *et al* (2022). Specifically, the network takes the reconstructed results as input, and output the residual difference between the input and ground truth. During this phase, the network is dedicated to the refinement of the reconstruction quality, the optimization of model weights is undertaken through the minimization of the mean squared error (MSE) loss, which quantifies the disparity between the enhanced images and their corresponding ground truth counterparts during the training process. The final result is obtained by adding the output of the enhancement network to the input.

Finally, the enhanced pressure map is converted to dose map for proton dose verification. Specifically, an initial dose map was calculated by dividing the reconstructed pressure map by the dose conversion coefficient map derived from patient CT images. A 3D U-net was developed with the same architecture and training settings as in Jiang *et al* (2022) to predict the residual errors compared with the ground truth to generate the final dose map.

2.4. Training implementation and inference

The models in both Recon-enhance stages were trained by an ADAM optimizer with an initial learning rate of 0.001, reduced by a factor of 5 after every 500 000 epochs. In the Recon-stage, we set $\alpha_1 = 1.0$ and $\beta_1 = 0.025$ for initial training with the loss defined by equation (8). After 3000 000 epochs, we started the transfer learning with the loss defined in equation (9) for another 1000 000 epochs. Finally, in the Enhance-stage, we train the enhancement network for another 1000 000 epochs. The entire training process takes about 3 days to be finished.

During the inference, the trained DTR-Net uses RA data measured by limited angle PA imaging to reconstruct the pressure map, which is then enhanced by the enhancement network to generate the final PA images. This recon-enhance approach takes less than 6 s to process a 3D RA signal image with the size of $32 \times 32 \times 112$ to reconstruct a PA pressure map with the size of $48 \times 48 \times 112$. The network was implemented based on Pytorch with a 40 GB Nvidia server GPU and a 64 GB RAM.

2.5. Data collection

In this study, a dataset consisting of 126 anonymized patients with prostate cancers was collected under an IRB approved protocol. Data of each patient contains the planning CT scan and the corresponding clinical treatment plan. Dose map of the plan was provided by a commercial software named RayStation (RaySearch Laboratories, Stockholm/Sweden), and then normalized to the maximum dose. Each CT scan was firstly segmented into four categories: air, fat, soft tissue and bone according to the predefined HU value thresholding. All the tissue-specific parameters including the density, speed of sound, and the Grüneisen parameter are predefined in table 1.

The acoustic simulation for generating the RA signal started with the calculation of the initial pressure ($P0$) by multiplying the dose map with the tissue density and the Grüneisen parameter:

$$P0 = \text{dose\_map} \times \rho \times \Gamma, \tag{10}$$

Then, the simulation was performed using the open-source k-wave toolbox on Matlab. Specifically, a planar detector of 8 cm×8 cm with a 64×64 ultrasound transducer array was simulated below the prostate and near the perineum area with a $\frac{\pi}{6}$ tilt angle to cover the prostate area and avoid the pelvic bones. The central frequency of each transducer element was set to 500 kHz with 100% bandwidth and a sampling rate of 5 MHz. Tissue-specific heterogeneity and attenuation were considered during the acoustic signal propagation. Finally, a Gaussian white noise with 10 dB signal-to-noise ratio (SNR) was added to the acquired RA signals, which is used as input of our network. TR method was applied for 10 iterations to reconstruct the initial pressure maps from the simulated RA signals, which are used as ground truth for the initial training of DTR-Net in the Recon-stage in figure 2(a). The initial pressure map $P0$ and the dose map were used as the ground truth for training the pressure map enhancement network and the dose conversion network. Both the pressure map and dose map were resampled to the resolution of $2.50 \times 2.50 \times 1.25$ mm^3 with the size of $48 \times 48 \times 112$, and the simulated RA signal was resampled to the size of $32 \times 32 \times 112$ to reduce the memory consumption.

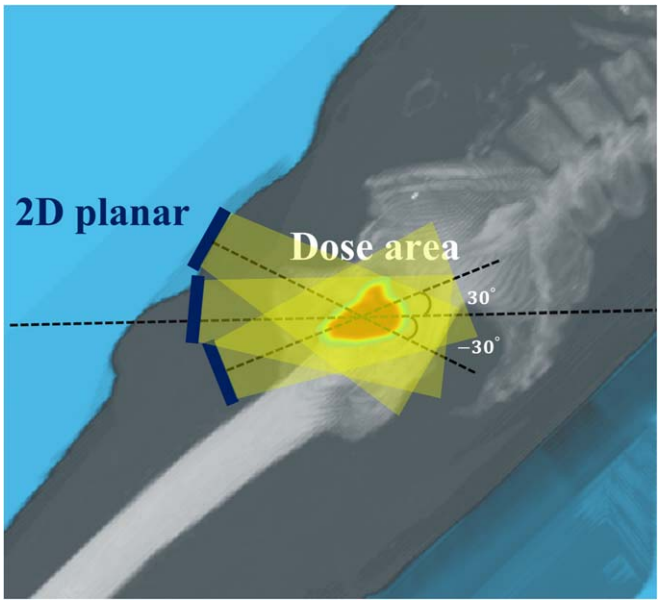

Figure 3. Augmentation setup. To simplify the visualization, we fix the patient body and rotate the 2D plannar matrix along the lateral direction to cover the entire prostate that is illustrated by the dose area. Three examples of acquisition are given in this figure.

3. Experiments and results

3.1. Data augmentation

We perform data augmentation to improve the model generalization while avoiding over-fitting. In this study, the PA detector was simulated at different positions in the perineum area to generate more raw acoustic-initial pressure (RA-*P*0) pairs to enlarge the training set. Specifically, the detector was located below the prostate and near the perineum area with an initial $\frac{\pi}{6}$ degree tilt angle. Then the detector was rotated along the lateral axis by different angles that are equally sampled within a range of $[-\frac{\pi}{6}, \frac{\pi}{6}]$ that covers the whole prostate area, as shown in figure 3. For each sampled angle, protoacoustic simulation procedures was performed to generate the corresponding RA signals from the initial pressure map *P*0. The augmentation was repeated for 20 times with equally spaced angles for each patient. The augmented dataset are used for training the proposed network. Using 5-fold cross-validation, we randomly selected 66 patients for training, 20 patients for validation, and the rest 40 patients for testing. No augmentation was performed for validation/testing sets.

3.2. Competing methods

We quantitatively and qualitatively compared our method with twobaseline methods:

- **Time reversal**: An iterative method for image reconstruction. In each iteration, a pressure map is reconstructed based on forward projection, then the time parameter is reversed and a RA signal is calculated from the reconstructed pressure map and compared with the acquired RA signals. The current pressure is updated by adding a residual pressure obtained by back-projecting the RA signal differences. In this experiment, TR method was repeated for 10 iterations empirically considering the balance between reconstruction quality and time consumption to reconstruct the initial map.
- **Method in** Jiang *et al* (2022): A state-of-the-art deep-learning method that jointly performs initial pressure reconstruction and dose verification. The first network takes the pressure reconstructed by TR method as input, and outputs a result with enhanced quality. An initial dose map is generated by multiplying the reconstructed results with the dose coefficients derived from the CT scans, and then further refined by the second network. We trained the network for pressure reconstruction using the same architecture and paremeter setting as described in Jiang *et al* (2022) with the input size of $48 \times 48 \times 112$. Same training/validation set and augmentation method were applied for the training.

3.3. Pressure map reconstruction results

The reconstruction quality was evaluated by comparing the predicted pressure with the ground truth using root mean squared errors (RMSE). Additionally, we also compared Peak signal-to-noise ratio (PSNR) and SSIM to further investigate the performance on details and basic structure reconstruction. The overall quantitative

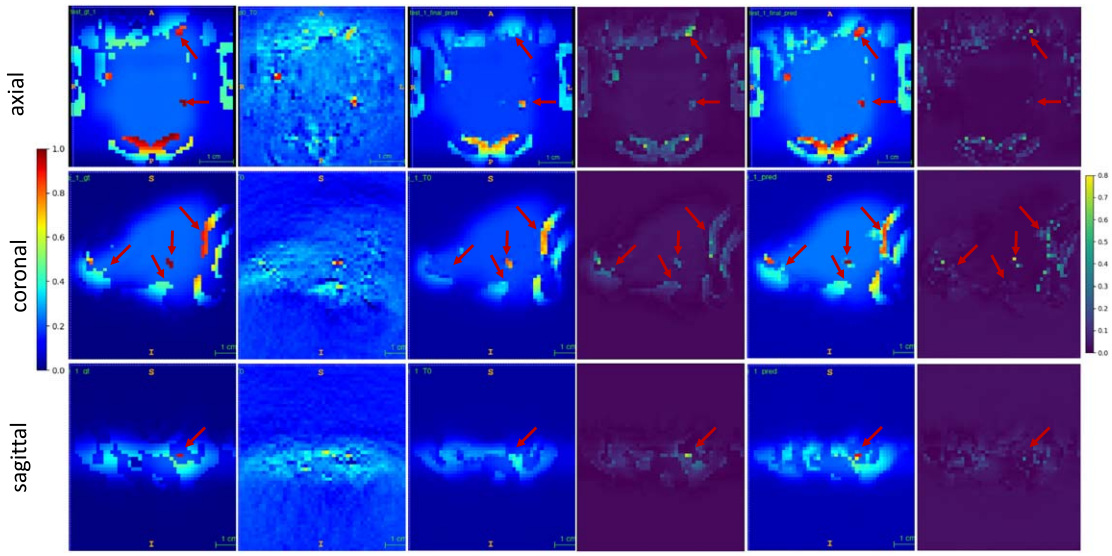

Figure 4. Example initial pressure reconstruction results (Normalized). From left to right: color bar of the initial maps, ground truth, results by using TR, results by using the method in (Jiang *et al* 2022), difference maps between ground truth and results from (Jiang *et al* 2022), results by our approach, and difference maps between ground truth and our results, color bar of the difference maps.

Table 2. Quantitative analysis of the reconstruction results of initial pressure maps and dose verification.

Modality	Method	RMSE	PSNR (dB)	SSIM	Speed (s)
PA image	Time reversal	0.145 ± 0.059	24.02 ± 0.58	0.854 ± 0.045	120
	Method in Jiang *et al* (2022)	0.033 ± 0.021	29.6 ± 0.34	0.939 ± 0.013	120
	DTR-A	0.042 ± 0.029	26.52 ± 0.41	0.892 ± 0.015	6
	DTR-B	0.030 ± 0.014	30.21 ± 0.37	0.959 ± 0.015	6
	Our approach	**0.029 ± 0.011**	**30.37 ± 0.26**	**0.962 ± 0.013**	**6**
Dose verification	Method in Jiang *et al* (2022)	0.026 ± 0.013	31.79 ± 0.34	0.973 ± 0.016	120
	Our approach	**0.018 ± 0.009**	**34.86 ± 0.27**	**0.989 ± 0.007**	**6**

results of pressure map reconstruction are summarized in table 2. The qualitative results are also shown in figure 4. Among the three compared methods, TR method results in the largest RMSE (0.145) and the lowest SSIM (0.854). The method in Jiang *et al* (2022) improved the reconstruction quality by reducing the RMSE to 0.033. Meanwhile, the SSIM was improved to 0.939, demonstrating the effectiveness of using 3D U-net for quality enhancement. However, details were still not reconstructed in some challenging locations, while the whole structure was blurred.

Our method is more accurate than all compared methods, with a RMSE error as low as 0.029. As shown in figure 4, most of the details were successfully reconstructed in the challenging areas while the blur effect was eliminated, suggesting the effectiveness of the explicit learning of correlation between the image and signal domains. Specifically, the SSIM was improved to 0.962, showing a high similarity of anatomic structure compared with the ground truth, confirming the effectiveness of using SSIM and perceptual losses for training. RMSE and SSIM results are boxplotted in figures 5(a) and (b). Notably, we also compared the runtime for testing using different methods. TR and method in Jiang *et al* (2022) both took about 2 minutes to process a single case due to iterations. Our approach achieved the fastest speed taking as low as 6 s, making the method much more applicable for online dose verification in proton therapy.

3.4. Dose verification results

We compared the dose maps that were predicted from the pressure maps recontructed by our approach and the method in Jiang *et al* (2022), in terms of RMSE, PSNR and SSIM. Table 2 also gives the quantitative results, where our approach gains significant improvements. Particularly, our approach reduces the RMSE from 0.026 to 0.018, and increases the SSIM from 0.973 to 0.989, showing a high similarity between the predicted and the ground truth 3D dose maps. Figure 6 shows qualitative results of several challenging cases, where the dose maps restored by using our method show more accuracy, due to the high quality of the input pressure maps. Finally, the pressure reconstruction and dose prediction with the proposed method only take about 6 s in total.

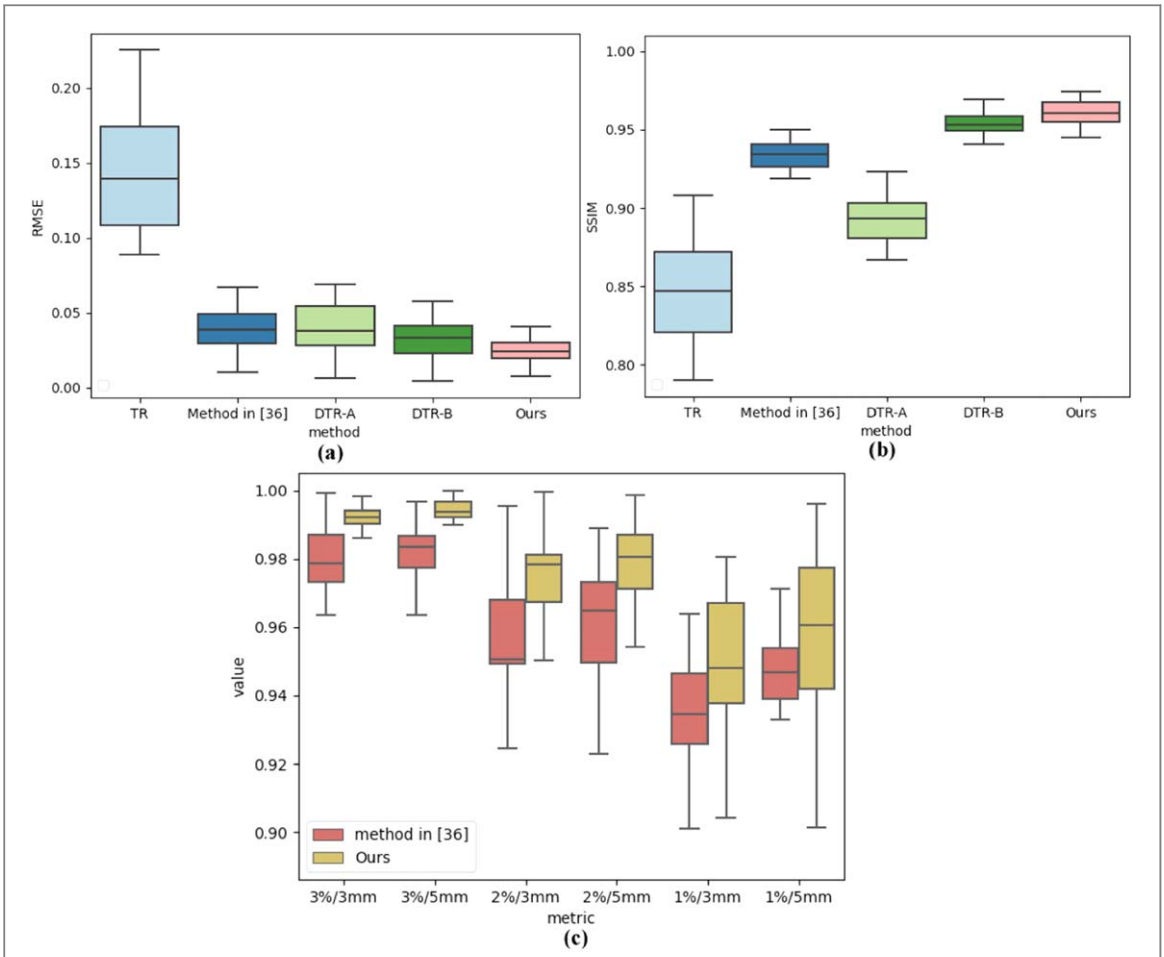

Figure 5. Boxplot of the (a) RMSE of the predicted pressure maps using different methods, (b) SSIM of the predicted presure maps using different methods, and (c) Gamma index of the predicted dose maps using different metrics.

Additionally, we compared the predicted dose maps with the ground truth in terms of gamma index as shown in table 3 and figure 5(c). Our approach increased the gamma index from 97.9% to 99.3%, from 98.3% to 99.6%, from 95.7% to 97.1%, and from 96.5% to 97.8% for 3%/3 mm, 3%/5 mm, 2%/3 mm, and 2%/5 mm, respectively. Notably, our approach achieved high gamma index rates as 94.7% and 95.7% for 1%/3 mm and 1%/5 mm, showing a high agreement between the predicted and the ground truth dose maps, which further demonstrates the effectiveness of our approach.

3.5. Ablation study

We performed an ablation study by comparing our approach with two variants: (1) DTR-A, where transfer learning was not applied. We used the loss function defined in equation (8) to train the network in the Recon-stage, then performed enhancement in the enhance-stage; (2) DTR-B, where we kept the same network architecture and training losses that were used in our proposed method, except that we used the initial pressure rather than TR results as the ground truth in the Recon-stage. The reconstruction results were quantitatively evaluated with RMSE, PSNR and SSIM. All compared methods were trained using the same augmented dataset.

The results of the ablation study, denoted as DTR-A and DTR-B, are also summarized in table 2 and figures 5(a) and (b). Specfically, DTR-A had the highest RMSE (0.042) and the lowest SSIM (0.892). Compared with DTR-A, DTR-B further improves the reconstruction quality with a RMSE of 0.030 and SSIM of 0.959, confirming the effectiveness of transfer learning. Our approach achieved the lowest RMSE and the highest SSIM.

4. Discussion

4.1. Pressure and dose reconstruction for protoacoustic imaging

Our approach used transformer-based blocks to build the network, which is trained by hybrid-supervision for reconstructing the initial pressure map directly from the RA signals. Results showed that our approach has gained an improved accuracy and speed. For the compared TR method, due to the limited angle view of the 2D

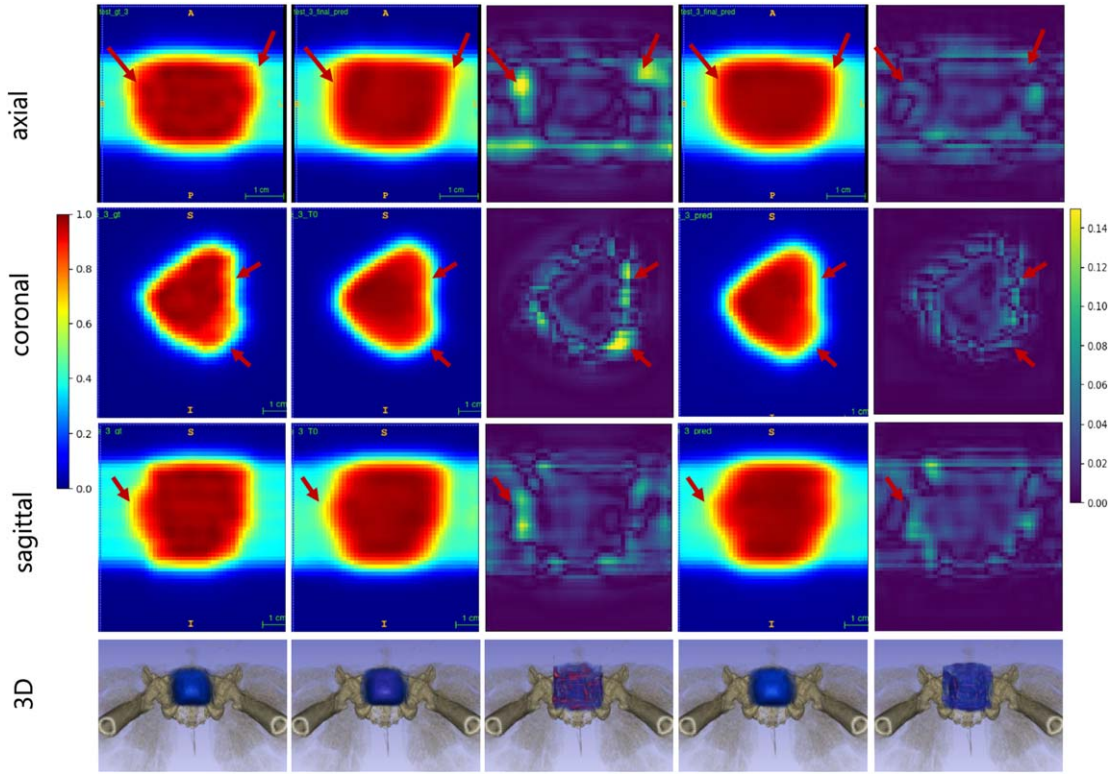

Figure 6. Example dose verification results (Normalized). From left to right: Color bar of the dose maps, ground truth, results by using the method in (Jiang *et al* 2022), difference maps between ground truth and results from Jiang *et al* (2022), results by our approach, difference maps between ground truth an our results, and color bar of the difference maps.

Table 3. Quantitative analysis of the reconstruction results of dose maps.

Modality	Metric	Method in (Jiang *et al* 2022)	Our approach
Dose	Gamma Index (3%/3mm)	97.9%± 1.1%	**99.3**%± **0.4**%
	Gamma Index (3%/5 mm)	98.3% ± 0.8%	**99.6**% ± **0.3**%
	Gamma Index (2%/3 mm)	95.7% ± 2.2%	**97.1**% ± **1.9**%
	Gamma Index (2%/5 mm)	96.5% ± 2.0%	**97.8**% ± **1.8**%
	Gamma Index (1%/3 mm)	92.7% ± 2.5%	**94.7**% ± **2.5**%
	Gamma Index (1%/5 mm)	93.7% ± 2.4%	**95.7**% ± **2.5**%

matrix array, the reconstructed pressure map suffers from severe distortions, where most of the structure details cannot be distinguished. The method in Jiang *et al* (2022) applied a 3D U-net to enhance the quality of the initial map reconstructed by TR method. However, the efficacy of the network enhancement is limited by the quality of the TR reconstruction. Specifically, in areas where the TR image is severely distorted with missing details, image enhancement will not be able to recover anatomical details that are completely lost in the input image as shown in figure 5 highlighted by red arrows. For DTR-A, without the transfer learning to tune the model based on data fidelity constraint, the reconstruction is highly affected by the limited quality of TR reconstruction used as the reference in the initial training, leading to suboptimal results. DTR-B further improved the accuracy. However, using initial pressure as the ground truth requires the network to perform both domain correlation learning for image reconstruction and correction of image distortion caused by the limited-angle acquisition, which is hard to balance during the training and leads to slightly lower quality compared with the proposed method. Increasing learning parameters could potentially solve this problem but will cause more memory consumption. Compared with the method in Jiang *et al* (2022), our approach directly reconstructs the initial pressure map from RA signals to preserve the essential structural information. Using TR results for initial training in the Recon-stage made the network focus on domain transfer mapping, thus improved the training efficiency and efficacy. Meanwhile, the transfer learning with self-supervision based on data fidelity constraint ensured consistency in both domains. Figures 4 and 5 showed that our approach can successfully reconstruct most structural details, leading to high quality 3d dose verification result. Quantitative results also demonstrated the

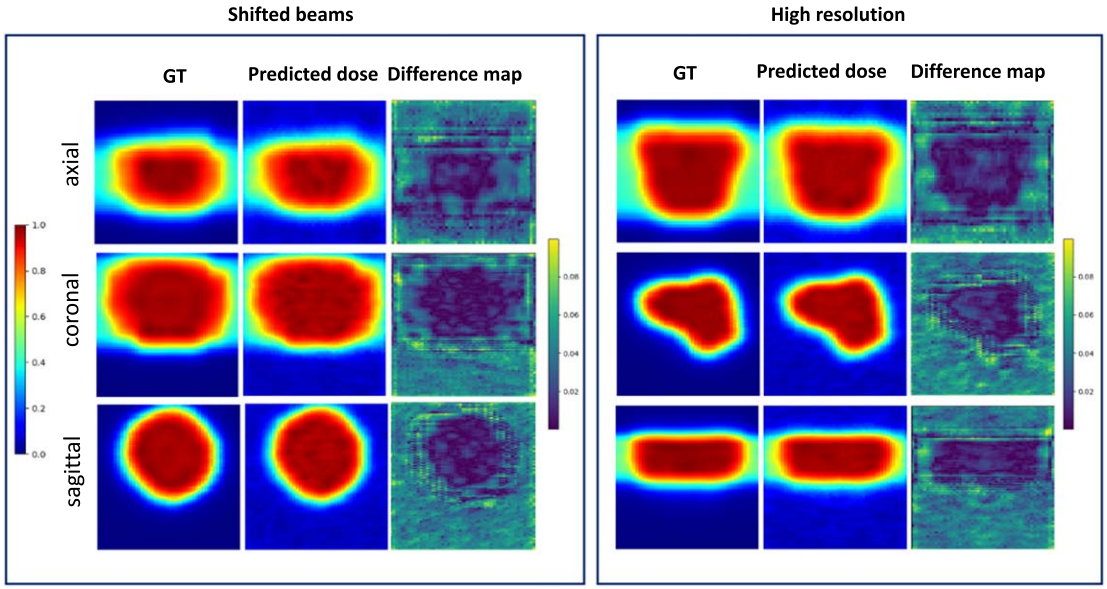

Figure 7. Example dose verification results (Normalized). The left part represents the results of using RA signals generated from shifted beams. The right part represents the results of high-resolution RA signals generated by using small grids. The ground truth (GT), predicted result and the difference map are given from left to right in both parts, respectively.

superiority of our method compared to other methods. Another major advantage of the deep learning reconstruction network is its high efficiency. The proposed network achieved an end-to-end processing time of 6 s, which is substantially shorter than the 2 min required by the TR method. This high efficiency is critical for the clinical adoption of the technique since time is of the essence when performing online dose verification during proton therapy.

To further verify the robustness of our approach, we conducted additional experiments wherein dose maps were subjected to random shifts of 1 cm along the axial, sagittal, and coronal directions. This simulation aimed to replicate scenarios resembling beam overshooting. The inputs are the RA signals generated from the PA simulation and the corresponding first order derivatives, and we calculate the RMSE, PSNR and SSIM for the predicted dose maps in comparison with the ground truth. The RMSE and SSIM reached 34.15 and 0.981, which is very close to the results without overshooting. This proximity indicates the robustness of our approach even under conditions where the beam overshoots. Qualitative results are also shown in figure 7.

4.2. Temporal characteristics of the proton pluse

It is worth to note that stress confinement is presumed during the PA simulation. In actual clinical or experimental settings, achieving perfect stress confinement is challenging, leading to the generation of degraded protoacoustic signal. In this study, the resolution of the generated PA signal is between 5 and 6 mm. The outcomes of the reconstruction and dose verification, conducted under such resolution, affirm the efficacy of our approach, signifying its applicability to 3D dose verification. We can improve the resolution in further research by reducing pulse duration and increasing frequency. For instance, employing a pulse duration of 0.5 ms with a frequency of 1 MHz can yield a generated PA signal resolution ranging from 1 to 2 mm.

4.3. Inverse crime

During the simulation, time reversal was performed to derive reconstruction results using as ground truth for the training of our network, where a uniform grid size (1.25 mm) was applied for both forward and backward projections, leading to a same resolution between the simulated signals and their corresponding reconstructions. In practice, this may give rise to an instance of inverse crime, yielding unrealistic good results, thereby compromising the model's generalizability. To address the inverse crime issue and mimic real situations, an additional study was conducted wherein a diminished grid size (1.00 mm) was employed during forward projection, resulting in a higher resolution of the generated signals. We performed such simulation to generate a testing set containing 80 cases. The dose verification results achieved to a PSNR of 33.69 and a SSIM of 0.978, closely approximating outcomes derived from low-resolution signals. Qualitative results, presented in figure 7, reveal accurate reconstruction of the majority of the dose distribution areas. The background is slightly noisy compared with the results shown in figure 6, which can be potentially refined by fine-tuning with an

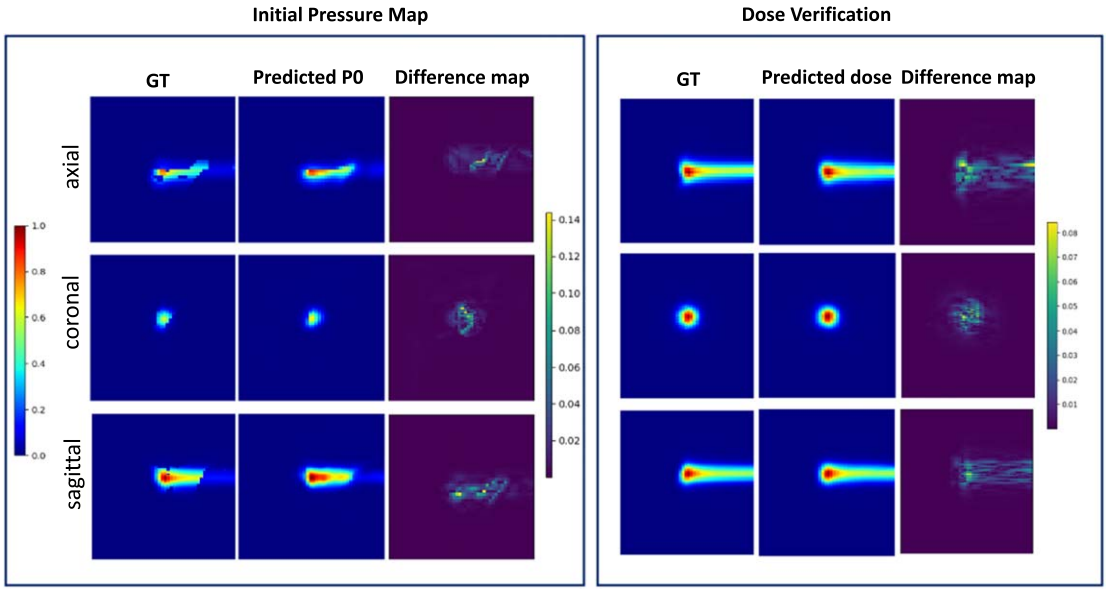

Figure 8. Example pensil beam dose verification results (Normalized). The left part represents the initial pressure maps reconstruction results for a pensil beam. The right part represents the corresponding dose verification results. The ground truth (GT), predicted result and the difference map are given from left to right in both parts, respectively.

additional high-resolution dataset. Nevertheless, the results affirm the generalization efficacy of our network in practical situations.

4.4. Clinical relevance

In practice, online dose verification necessitates the reconstruction of dose deposition from each individual pencil beam in real-time during its delivery. In this work, additional experiments were performed to validate our approach on initial pressure map reconstruction and dose verification for individual pencil beams. To achieve this, we extracted 120 individual pencil beams from diverse patient datasets and subjected them to the same simulation processes, thereby generating a comprehensive dataset instrumental in fine-tuning the pretrained network for pencil beam verification. The network then underwent testing on a seperate set of 30 pencil beams, yielding an average PSNR of 40.38 for the initial pressure reconstruction and an average PSNR of 41.98 for dose verification. Figure 8 presents the qualitative results, demonstrating the successful reconstruction of the initial pressure map for each pencil beam. This accomplishment contributes to a high-quality three-dimensional dose verification outcome, affirming the efficacy of our approach in the context of online dose verification.

4.5. Limitations and future work

There are some limitations of this study. First, since we applied k-wave toolBox to perform projection from the image domain back to the signal domain during the transfer learning, the training time has increased numerously. Second, the reconstruction quality is still expected to be improved, although our method has eliminated the distortion and artifacts caused by limited angle view.

In the future work, we will focus on developing a deep learning method to automatically learn the back-projection mapping to accelerate the training process. Besides, we will investigate RA signal pre-processing to improve the RA signal quality, which can further improve the performance of our proposed method. We will also apply our approach to other image modalities to verify the generalization of the proposed network.

Our approach was evaluated on simulated data due to the lack of patient experiments. Simulation data have the advantage of providing the ground truth of initial pressure and dose map for evaluation compared to real patient data where ground truth is often unavailable. The simulation parameters were set empirically to make the simulation results close to real data. Experimental and real patient studies are warranted in the future to further evaluate the clinical efficacy of the technique.

5. Conclusion

In this work, we have proposed a hybrid-supervised deep learning method to reconstruct PA images for proton therapy dose verification. DTR-Net using transformer blocks, transfer learning and hybrid supervision has been

developed for direct PA image reconstruction from the RA signals, and image enhancement has been applied to solve the limited angle view problem. The results show that our method outperforms competing state-of-the-art methods. Most importantly, our approach achieved superior performance on reconstructing 3D dose with a fast processing speed, making it very practical for online 3D dose verification in proton therapy.

Acknowledgments

This work was sponsored in part by National Institutes of Health Grants R01 EB032680, R01 EB028324 and R01 CA279013.

Data availability statement

Those clinical data was collected under an IRB approved protocol. The data that support the findings of this study are available upon reasonable request from the authors.

References

Ahmad M, Xiang L, Yousefi S and Xing L 2015 Theoretical detection threshold of the proton-acoustic range verification technique *Med. Phys.* **42** 5735–44

Bentefour E H, Shikui T, Prieels D and Lu H M 2012 Effect of tissue heterogeneity on an in vivo range verification technique for proton therapy *Phys. Med. Biol.* **57** 5473

Carlier B *et al* 2020 Proton range verification with ultrasound imaging using injectable radiation sensitive nanodroplets: a feasibility study *Phys. Med. Biol.* **65** 065013

Chen C, Xing Y, Gao H, Zhang L and Chen Z 2022 Sam's net: aself-augmented multi-stage deep-learning network for end-to-end reconstruction of limited angle CT *IEEE Trans. Med. Imaging* **41** 2912–24

Chen H *et al* 2018 LEARN: Learned experts' assessment-based reconstruction network for sparse-data CT *IEEE Trans. Med. Imaging* **37** 1333–47

Draeger E *et al* 2018 3D prompt gamma imaging for proton beam range verification *Phys. Med. Biol.* **63** 035019

Fiedler F *et al* 2008 In-beam PET measurements of biological half-lives of 12C irradiation induced $\beta\pm$ activity *Acta Oncol.* **47** 1077–86

Fiedler F *et al* 2010 On the effectiveness of ion range determination from in-beam PET data *Phys. Med. Biol.* **55** 1989

Freijo C, Herraiz J L, Sanchez-Parcerisa D and Udias J M 2021 Dictionary-based protoacoustic dose map imaging for proton range verification *Photoacoustics* **21** 100240

Häggström I, Schmidtlein C R, Campanella G and Fuchs T J 2019 DeepPET: a deep encoder-decoder network for directly solving the PET image reconstruction inverse problem *Med. Imag. Anal.* **54** 253–62

Hristova Y, Kuchment P and Nguyen L 2008 Reconstruction and time reversal in thermoacoustic tomography in acoustically homogeneous and inhomogeneous media *Inverse Prob.* **24** 055006

Huang J, Wu Y, Wu H and Yang G 2022 Fast MRI Reconstruction: how powerful transformers are? 2066–70, arXiv:2201.09400

Jiang Z *et al* 2019 Augmentation of CBCT reconstructed from under-sampled projections using deep learning *IEEE Trans. Med. Imaging* **38** 2705–15

Jiang Z *et al* 2022 3D in vivo dose verification in prostate proton therapy with deep learning-based proton-acoustic imaging *Phys. Med. Biol.* **67** 215012

Kormoll T *et al* 2011 A compton imager for *in vivo* dosimetry of proton beams design study", nuclear instruments and methods in physics research section a: accelerators, spectrometers *Detectors Associated Equipment* **626** 114–9

Lan H, Jiang D, Yang C, Gao F and Gao F 2020 Y-Net: hybrid deep learning image reconstruction for photoacoustic tomography in vivo *Photoacoustics* **20** 100197

Liu Z *et al* 2021 Swin transformer: Hierarchical vision transformer using shifted windows *Proc. of the IEEE/CVF Int. Conf. on Computer Vision* pp 10012–22

Lu H, Mann G and Cascio E 2010 Investigation of an implantable dosimeter for single-point water equivalent path length verification in proton therapy *Med. Phys.* **37** 5858–66

Luo Y *et al* 2021 3D transformer-GAN for high-quality PET reconstruction *Int. Conf. on Medical Image Computing and Computer-Assisted Intervention* pp 276–85

Matsoukas C, Haslum J F, Söderberg M and Smith K 2021 Is it time to replace cnns with transformers for medical images? arXiv:2108.09038

Min C H, Lee H R, Kim C H and Lee S B 2012 Development of array-type prompt gamma measurement system for *in vivo* range verification in proton therapy *Med. Phys.* **39** 2100–7

Miyatake A *et al* 2010 Measurement and verification of positron emitter nuclei generated at each treatment site by target nuclear fragment reactions in proton therapy *Med. Phys.* **37** 4445–55

Nishio T *et al* 2010 The development and clinical use of a beam ON-LINE PET system mounted on a rotating gantry port in proton therapy *Int. J. Radiat. Oncol. Biol. Phys.* **76** 277–86

Parmar N *et al* 2018 Image transformer *Int. Conf. on Machine Learning* pp 4055–64

Penfold S N, Rosenfeld A B, Schulte R W and Schubert K E 2009 A more accurate reconstruction system matrix for quantitative proton computed tomography *Med. Phys.* **36** 4511–8

Pietsch J *et al* 2023 Automatic detection and classification of treatment deviations in proton therapy from realistically simulated prompt gamma imaging data *Med. Phys.* **50** 506–17

Polf J C, Peterson S, Ciangaru G, Gillin M and Beddar S 2009 Prompt gamma-ray emission from biological tissues during proton irradiation: a preliminary study *Phys. Med. Biol.* **54** 731

Schneider U, Pedroni E, Hartmann M, Besserer J and Lomax T 2012 Spatial resolution of proton tomography: methods, initial phase space and object thickness *Z. Med. Phys.* **22** 100–8

Schneider U *et al* 2004 First proton radiography of an animal patient *Med. Phys.* **31** 1046–51

Telsemeyer J, Jäkel O and Martišíková M 2012 Quantitative carbon ion beam radiography and tomography with a flat-panel detector *Phys. Med. Biol.* **57** 7957
Treeby B E and Cox B T 2010 k-Wave: MATLAB toolbox for the simulation and reconstruction of photoacoustic wave fields *J. Biomed. Opt.* **15** 021314
Treeby B E, Zhang E Z and Cox B T 2010 Photoacoustic tomography in absorbing acoustic media using time reversal *Inverse Prob.* **26** 115003
Wang M *et al* 2020 Toward *in vivo* dosimetry for prostate radiotherapy with a transperineal ultrasound array: a simulation study *IEEE Trans. Radiat. Plasma Med. Sci.* **5** 373–82
Xu M and Wang L V 2005 Universal back-projection algorithm for photoacoustic computed tomography *Phys. Rev.* E **71** 016706
Yao S, Hu Z, Xie Q, Yang Y and Peng H 2021 Further investigation of 3D dose verification in proton therapy utilizing acoustic signal, wavelet decomposition and machine learning *Biomed. Phys. Eng. Express* **8** 015008
Yu J, Yoon H, Khalifa Y M and Emelianov S Y 2019a Design of a volumetric imaging sequence using a vantage-256 ultrasound research platform multiplexed with a 1024-element fully sampled matrix array *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* **67** 248–57
Yu Y, Li Z, Zhang D, Xing L and Peng H 2019b Simulation studies of time reversal-based protoacoustic reconstruction for range and dose verification in proton therapy *Med. Phys.* **46** 3649–62
Yu Y, Pengyuan Q and Hao P 2021 Feasibility study of 3D time-reversal reconstruction of proton-induced acoustic signals for dose verification in the head and the liver: a simulation study *Med. Phys.* **48** 4485–97
Yuan Y *et al* 2013 Feasibility study of in vivo MRI based dosimetric verification of proton end-of-range for liver cancer patients *Radiother. Oncol.* **106** 378–82
Zhang J *et al* 2021 Ultrasound image reconstruction from plane wave radio-frequency data by self-supervised deep neural network *Med. Image Anal.* **70** 102018
Zhu B, Liu J Z, Cauley S F, Rosen B R and Rosen M S 2018 Image reconstruction by domain-transform manifold learning *Nature* **555** 487–92