Hindawi
IET Signal Processing
Volume 2023, Article ID 6615953, 19 pages
https://doi.org/10.1049/2023/6615953

Research Article

IRR-Net: A Joint Learning Framework for Image Reconstruction and Recognition of Photoacoustic Tomography

Zheng Sun [1,2] **Bing Ai** [1] **Meichen Sun** [1] **and Yingsa Hou** [1]

[1]*Department of Electronic and Communication Engineering, North China Electric Power University, Baoding 071003, Hebei, China*
[2]*Hebei Key Laboratory of Power Internet of Things Technology, North China Electric Power University, Baoding 071003, Hebei, China*

Correspondence should be addressed to Zheng Sun; sunzheng_tju@163.com

Received 16 August 2023; Revised 30 November 2023; Accepted 7 December 2023; Published 22 December 2023

Academic Editor: Wanli Wen

In photoacoustic tomography (PAT), object identification and classification are usually performed as postprocessing processes after image reconstruction. Since useful information about the target implied in the raw signal can be lost during image reconstruction, this two-step scheme can reduce the accuracy of tissue characterization. For learning-based methods, it is time consuming to train the network of each subtask separately. In this paper, we report on an end-to-end joint learning framework for simultaneous image reconstruction and object recognition, named IRR-Net. It establishes direct mapping of raw photoacoustic signals to high-quality images with recognized targets. The network consists of an image reconstruction module, an optimization module, and a recognition module, which achieved signal-to-image, image-to-image, and image-to-class conversion, respectively. We built simulation, phantom and *in vivo* data sets to train and test IRR-Net. The results show that the proposed method successfully yields concurrent improvements in both the quality of the reconstructed images and the accuracy of target recognition at a lower time cost compared to the separately trained networks.

1. Introduction

Photoacoustic tomography (PAT) is a newly developed hybrid biological imaging modality that distinguishes tissue compositions based on high specificity of optical absorption [1]. The use of machine learning technology to automatically identify objects of interest in images and characterize tissue types is of great significance for improving the clinical value of PAT and improving the efficiency of computer-aided diagnosis [2].

Most existing approaches employ a two-step strategy that considers image reconstruction and image segmentation or target recognition independently. First, the image representing the distribution of the initial pressure or optical absorption in the imaging plane is reconstructed from the pressure signal collected by ultrasonic transducers. Then, the image segmentation or classification is performed as a postprocessing procedure. Image reconstruction aims to recover the structural and functional information of the imaging target from the photoacoustic signal through acoustic inversion or optical inversion. Traditional methods include back projection (BP) [3], time reversal (TR) [4], delay and sum (DAS), iterative reconstruction technique (IRT) [5], etc. These methods usually make idealized assumptions about the imaging scenario and have limitations in noise suppression and artifact elimination. In recent years, learning-based image reconstruction has received extensive attention [6]. Based on the working domains, current methods can be divided into two categories: image-to-image conversion and signal-to-image conversion. The former uses a deep neural network (DNN) to optimize images reconstructed by the traditional methods. The initial image reconstructed from the pressure signal using a traditional method is fed into the network, which outputs the optimized image with high quality [7–9]. The process of this method is simple and fast. However, the image quality is limited by the initial reconstruction. Some information lost during initial reconstruction is difficult to recover through learning. The latter uses DNN to map signals to images. The photoacoustic signal is directly fed into the network, which is trained to learn the prior knowledge of the target. This process is complex and computationally intensive, but most of the image features can be recovered [10–13].

In the early stage, traditional image processing techniques were typically used to achieve PAT image segmentation and target recognition. For example, Zhang et al. [14] used

nonparametric smoothing and Gaussian low-pass spatial filtering to search for the skin surface. Mandal et al. [15] used an active contour model to segment target contours to identify regions with different speed of sound. Meiburger et al. [16] used Frangi vessel filtering [17] and the skeleton extraction algorithm to extract blood vessels from 3D PAT images. Raumonen and Tarvainen [18] adopted reliability assessment based on a probabilistic framework to extract vessels. Sun et al. [19] designed a hybrid method that combines Otsu thresholding with a 3D Hessian matrix to extract tumor vessels. Lutzweiler et al. [20] developed a signal domain algorithm to segment images. In PAT images, the contrast between the background and anatomical structures is generally lower than in the digital photography. The presence of image artifacts such as reflection artifacts, blurring artifacts, and limited view artifacts increases the difficulty of using traditional nonlearning methods to segment images. In addition, these methods generally design image processing steps and the corresponding parameters (such as filtering parameters and segmentation threshold) for specific applications and imaging objects, with poor generalization and a low degree of automation. In recent years, the application of deep learning in photoacoustic image processing has been extensively studied. For example, Chlis et al. [21] built sparse U-Net (S-UNet) to automatically extract vessel contours. Lafci et al. [22] employed U-Net to segment dual-modal optoacoustic ultrasound (OPUS) images. Rajanna et al. [23] used a DNN to realize the three-level classification (malignant, benign, and normal) of prostate tumors. Zhang et al. [24] employed AlexNet and GoogLeNet, respectively, to characterize tissues from PAT images of breast. In the two-step scheme, the quality of the reconstructed image affects the accuracy of the image processing. The errors that occur in image reconstruction will accumulate in subsequent image processing, thereby reducing segmentation and recognition accuracy.

Joint image reconstruction and segmentation can achieve both tasks simultaneously, utilizing all of the correlation and mutual information between the two tasks [25]. Taking into account the heterogeneity of the excitation field and the characteristics of the imaging region, image segmentation can improve image reconstruction [26]. Reconstruction can benefit from the target information obtained by segmentation and vice versa, thus improving the precision and consistency of reconstruction and segmentation [27]. In the PAT study, Boink et al. [28] used the partially learned primal-dual (L-PD) algorithm for the first time to achieve simultaneous image reconstruction and segmentation, obtaining binary images of the segmented blood vessels. However, they did not specify whether the method could be used to segment targets other than the blood vessels.

In this study, a joint learning framework named Image Reconstruction and Recognition Network (IRR-Net) is proposed for the concurrent PAT image reconstruction and target recognition. Our contributions are summarized as follows:

(1) We propose a joint learning framework that incorporates the image reconstruction subnetwork, the image optimization subnetwork, and the object recognition subnetwork. The reconstruction subnetwork achieves signal-to-image conversion by using a deep gradient descent (DGD) architecture, where the forward imaging operator and its adjoint operator are incorporated into the gradient calculation and separated from the network training. Gradient information is used to reduce the impact of the acoustic heterogeneity of the medium on the quality of the reconstruction. The image optimization subnetwork achieves image-to-image conversion, where the initially reconstructed image is optimized in the image domain to improve quality. The recognition subnetwork achieves the conversion from images to feature classification, with the aim of extracting features from optimized images, identifying and classifying objects of interest. To the best of our knowledge, this is the first work to achieve end-to-end mapping from photoacoustic pressure signals to high-quality images, where the target of interest has been identified.

(2) Compared to separately trained networks for subtasks, our end-to-end framework strikes a balance between computational efficiency and accuracy, making it more suitable for practical applications.

(3) We provide a general joint learning framework for simultaneous image reconstruction and object recognition. Although we designed, trained, and validated the network for PAT, the framework can be used for other tomographic imaging modalities by changing the forward imaging operator. It has clinical importance in facilitating tissue characterization and lesion recognition.

(4) We validated IRR-Net on simulation, phantom, and *in vivo* data sets to show its feasibility and generalizability. We achieved consistent improvements in terms of accuracy, sensitivity, specificity, and F1 score for both image reconstruction and object recognition tasks on each data set compared to the separately learned networks.

2. Materials and Methods

2.1. Overview of the Problem of PAT Image Reconstruction. PAT image reconstruction is essentially the inversion of the forward imaging model. We consider the following imaging scenario. A fixed-position light source emits short laser pulses that produce a uniform coverage of light on the surface of the imaging object. Pressure waves are collected by an array of ultrasonic transducers in which each transducer element is idealized as a point detector. A complete set of pressure signals is collected in the tomographic slice through full-view scanning. The absorbed optical energy density (AOED) is determined by the local light fluence and optical absorption coefficient, as shown below:

$$A(\boldsymbol{r}) = \mu_{\mathrm{a}}(\boldsymbol{r})\Phi(\boldsymbol{r}), \tag{1}$$

where $\boldsymbol{r}$ is a location in a 2D bounded imaging domain Ω with a boundary $\partial\Omega$, A is AOED in J/cm^3, and μ_{a} and Φ represent optical absorption coefficient and light fluence, respectively. The initial pressure photoacoustically induced by optical absorbers is proportional to AOED:

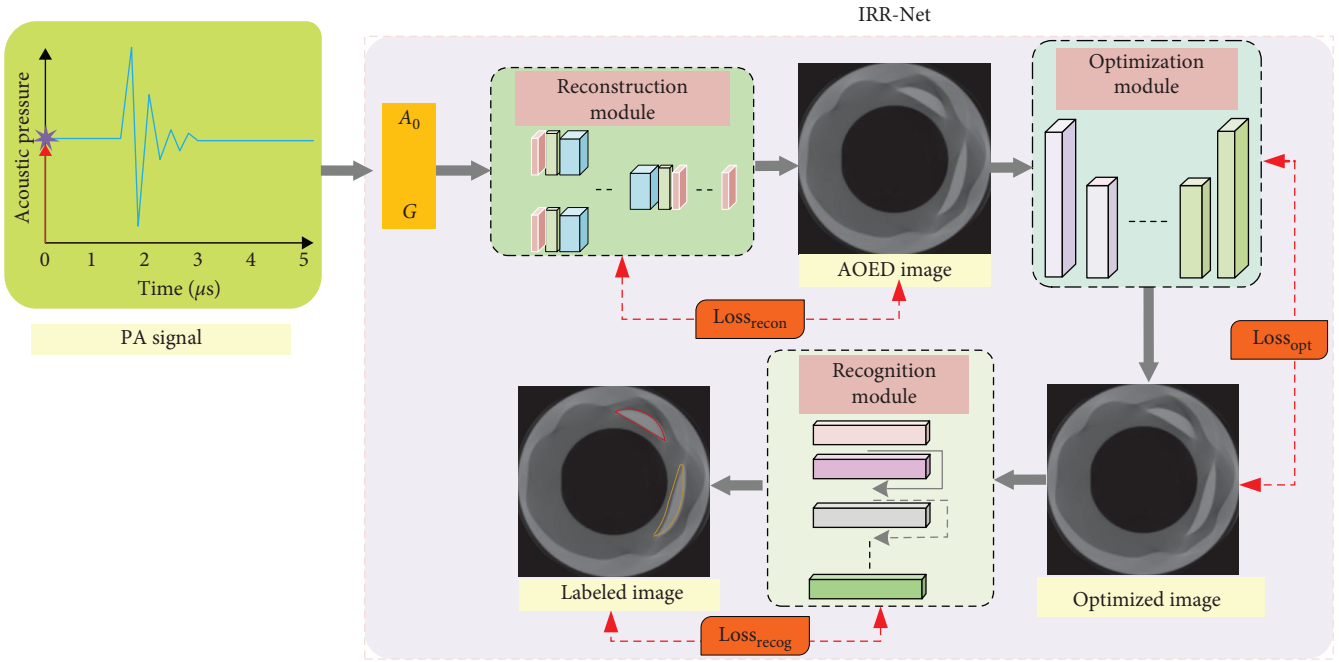


FIGURE 1: Flowchart of the proposed method.

$$p_0(\boldsymbol{r}) = \Gamma A(\boldsymbol{r}), \tag{2}$$

where $p_0(\boldsymbol{r})$ is the initial pressure generated by the optical absorber at $\boldsymbol{r}$, and Γ is the Grüneisen coefficient describing the thermoelasticity of the tissue. In this work, Γ is assumed to be constant in space and equal to 1.

The propagation of a pressure wave prompted by an initial source in an acoustically inhomogeneous medium with spatially varying density and speed of sound is governed by the following coupling equations [29],

$$\begin{cases} \dfrac{\partial}{\partial t}\boldsymbol{u}(\boldsymbol{r},t) = -\dfrac{1}{\rho_0}\nabla p(\boldsymbol{r},t) \\ \dfrac{\partial}{\partial t}\rho(\boldsymbol{r},t) = -\rho_0 \nabla \cdot \boldsymbol{u}(\boldsymbol{r},t) \\ p(\boldsymbol{r},t) = [c(\boldsymbol{r})]^2\rho(\boldsymbol{r},t)\left\{1 - 2\mu[c(\boldsymbol{r})]^{a-1}\dfrac{\partial}{\partial t}(-\nabla^2)^{\frac{a}{2}-1} + 2\mu[c(\boldsymbol{r})]^a \tan(\pi a/2)(-\nabla^2)^{\frac{a+1}{2}-1}\right\}, \end{cases} \tag{3}$$

where $t \in [0, T]$ is the observation time for a final time T, $\boldsymbol{u}$ denotes the velocity of the acoustic particle, p is the acoustic pressure, ρ_0 denotes the ambient density, ρ represents the fluctuation of the acoustic density in the heterogeneous medium, c is the speed of sound, ∇ is the Hamiltonian operator, a is the power law exponent typically ranging from 1 to 1.5, and μ is the power law prefactor with a typical value of $10^{-7}/2\pi\,\text{cm}^{-1}\,\text{rad}^{-1}\text{s}$ [30]. The time-dependent pressure satisfies the initial conditions of $p(\boldsymbol{r}, 0) = p_0(r)$ and $\partial p(\boldsymbol{r}, t)/\partial t|_{t=0} = 0$ for $\boldsymbol{r} \in \Omega$, as well as the Neumann boundary condition of $\partial p/\partial n_0 = 0$ on $\partial\Omega \times [0, T]$, where n_0 denotes the outer unit normal to $\partial\Omega$. The velocity of the particles satisfies the initial condition of $u(r, 0) = 0$. The k-space pseudospectral method (PSM) [31] is used to discretize Equation (3).

The forward process of generating the pressure signal from AOED can be expressed by the operator $\mathscr{H}(\cdot)$ as follows:

$$\boldsymbol{p} = \mathscr{H}(\boldsymbol{A}), \tag{4}$$

where $\boldsymbol{A}$ is the spatial distribution of the AOED and p is the pressure matrix reaching the detector surface. Reconstruction of the image representing the AOED distribution within a tomographic slice from the pressure time series measured by the detector is essentially the inverse problem of Equation (4),

$$\boldsymbol{A} = \mathscr{H}^*(\boldsymbol{p}), \tag{5}$$

where $\mathscr{H}^*(\cdot)$ is the adjoint operator of H.

2.2. IRR-Net Architecture. An overview of our approach is illustrated in Figure 1. The acquired pressure signal in a slice is recorded in a 2D matrix and then fed into the network,

which outputs a reconstructed image representing the AOED distribution with identified targets of interest.

As shown in Figure 2, the overall architecture of IRR-Net consists of three modules: reconstruction module, optimization module, and recognition module. In the reconstruction module, the DGD architecture [10, 32] is used to solve the acoustic inversion, and the AOED distribution is robustly recovered from the pressure signal to realize the conversion from the signal domain to the image domain. Its calculation is expressed as follows:

$$\boldsymbol{A}_1 = \mathscr{D}_\theta(\boldsymbol{A}_0, \boldsymbol{G}), \tag{6}$$

where $\mathscr{D}_\theta(\cdot)$ represents the DGD unit with the learning parameter θ, and $\boldsymbol{A}_1$ represents the AOED distribution map output from the DGD unit. $\mathscr{D}_\theta(\cdot)$ has two inputs, one of which is the initial AOED distribution $\boldsymbol{A}_0$:

$$\boldsymbol{A}_0 = \mathscr{H}^*(\boldsymbol{P}_\mathrm{m}), \tag{7}$$

where $\boldsymbol{P}_\mathrm{m}$ is the matrix for recording the measured acoustic pressure. Another input is the likelihood gradient that measures the data fitting between the measured pressure signal and the theoretical pressure calculated by the forward model as follows:

$$\boldsymbol{G} = \nabla d(\boldsymbol{P}_\mathrm{m}, \mathscr{H}(\boldsymbol{A}_0)) = \mathscr{H}^*(\mathscr{H}(\boldsymbol{A}_0) - \boldsymbol{P}_\mathrm{m}), \tag{8}$$

where $\nabla d(\cdot)$ denotes the calculation of the likelihood gradient and $\mathscr{H}(\boldsymbol{A}_0)$ is the theoretical pressure calculated from the forward model described in Section 2.1.

As shown in Equation (8), the calculation of the likelihood gradient involves the forward operator and its adjoint operator, which contain the information that the estimates need to be improved. The likelihood gradient reflects the error distribution between reconstruction and measurement. Therefore, it is used as data fidelity to enhance data consistency and produce efficient estimate updates. Inputting the concatenated AOED distribution map and its likelihood gradient into the network helps to obtain more reliable estimates than just inputting the image itself, while eliminating the need for the network to learn the entire physical prior from training data.

The DGD unit consists of nine convolutional layers that form the encoding–decoding architecture. The encoder extracts features from the input through convolution, while increasing the number of channels between layers to achieve structural refinement of the feature map from the previous layer. The decoder combines similar features of layers by reducing the number of filtering kernels between layers. The network uses the same convolution with a kernel size of $3 \times 3 \times M$ and a stride of 1 is used, where M is the number of feature maps input into the layer. The initial number of feature channels is 16, meaning channels = 16 for both inputs. The linear rectification unit (ReLU) is used as the activation function. The output of the DGD module is obtained by adding the initial input AOED map and the output AOED update via a skip connection.

The optimization module uses the forward propagation U-Net as the backbone to optimize the initial reconstruction output from the DGD module with the goal of converting low-quality images to high-quality images. The module consists of a total of 26 convolutional layers with a kernel size of 3×3 and a stride of 1 used throughout, except for the last layer, which is a 1×1 convolution. The number of feature channels is set to 64. ReLU and max pooling with a kernel size of 2×2 and a stride of 2 are used. Upsampling is performed using a kernel size of 2×2 and a stride of 2.

The recognition module uses the ResNet-50 architecture [33] as a feature classifier. It consists of a 7×7 convolution layer, 16 residual blocks (ResBlocks), a global average pooling layer and a fully connected layer. Each ResBlock consists of three convolutional layers, three batch normalization (BN) layers, and ReLU activation functions. The image output from the optimization module is fed into the recognition module. First, a 7×7 convolution with a stride of 2 and 64 feature channels is used for preprocessing, in which the ReLU activation function and maximum pooling with a kernel size of 2×2 and a stride of 2 are adopted. The preprocessed image is then fed into ResBlocks, where the input tensor is downsampled along the eigenmapping dimension by 1×1 convolution and the compressed tensor is filtered by 3×3 convolution. Finally, another 1×1 convolution is used to upsample the tensor to the size of the original feature map. The final output is an image with the labeled targets of interest.

2.3. Data Preparation and Experimental Setup. To build the simulation data set for training IRR-Net and testing its binary and multiclass classification performance, we generate numerical tubular phantoms containing different types of tissues to mimic endoscopic scenarios. We use Adobe Illustrator (v.2022, Adobe Systems Incorporated, San Jose, California) to draw the cross-sectional geometry of the numerical phantoms. We set optical and acoustic parameters for each tissue type based on histological references. In order to show the diversity of tissue components, the absorption coefficient, scattering coefficient, speed of sound, and mass density of each tissue type follow a Gaussian distribution, with the values listed in Table 1 as their average and a variance of 0.5. After generating the numerical phantoms, we use the Matlab MCXLAB software package [34] to simulate light transport in tissues, in order to obtain the light fluence in the imaging domain. In forward optical simulation, we refer to our experimental setup in the phantom study and set the irradiation source as a pulse laser with a wavelength of 880 nm and a pulse width of 7 ns. The photon emission source is located at the center of the imaging plane, with a total of 10^6 incident photons in one pulse. Then, based on the proportional relationship between AOED and the product of absorption coefficient and light fluence, the simulated AOED distribution is obtained. For forward acoustic simulation, we use the Matlab K-wave toolkit [35] to simulate acoustic propagation in heterogeneous media and obtain sound

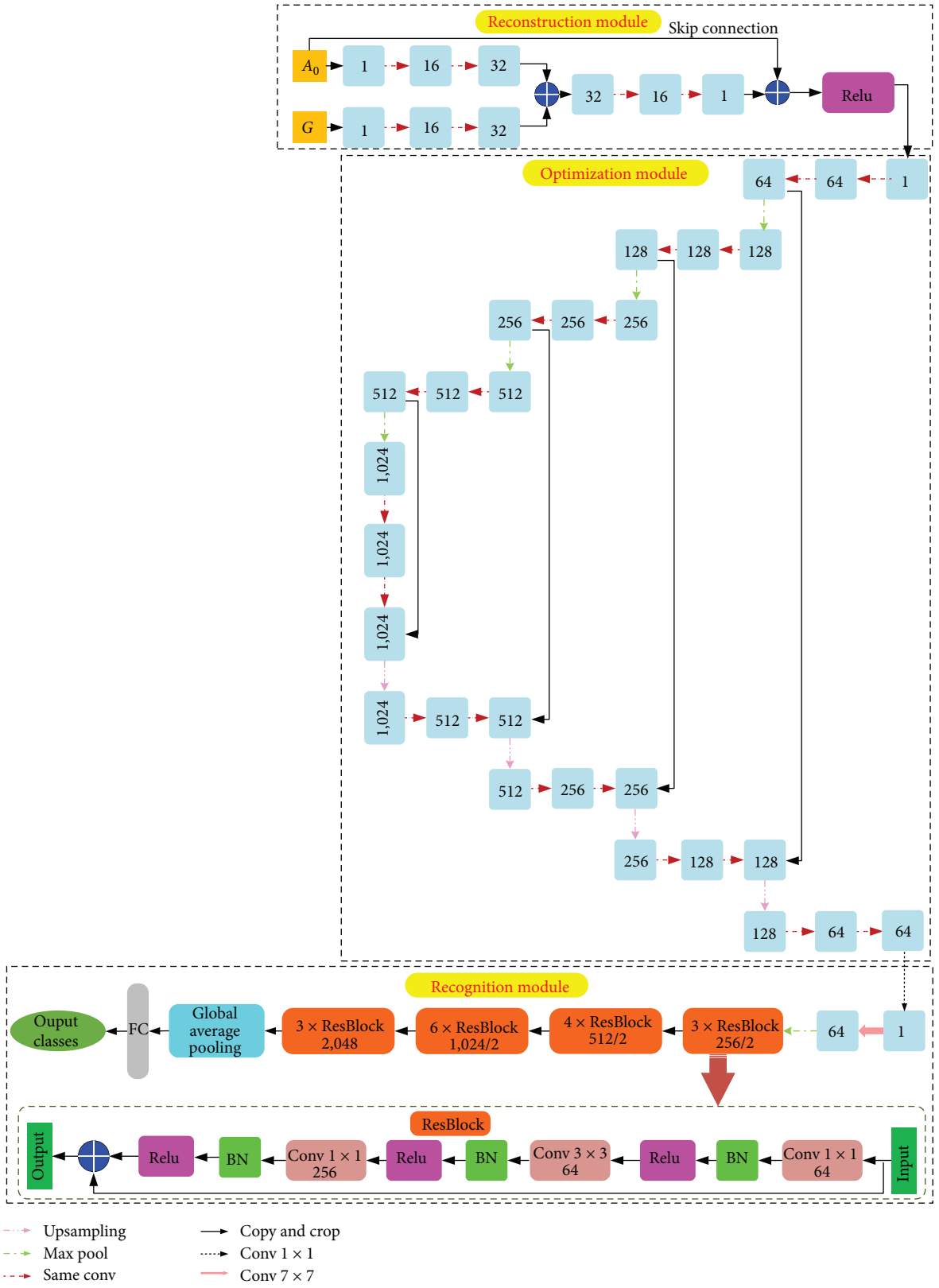


Figure 2: IRR-Net architecture with all labels.

Table 1: Characteristic parameters of different tissue types in numerical phantoms.

Tissue name	Refractive index	Reflectance	μ_a (cm^{-1})	μ_s (cm^{-1})	Anisotropy factor	Sound speed (m/s)	Mass density (kg/L)	Radial thickness (mm)
Outer wall (OW)	1.42	0.08	0.70	6	0.85	1,600	1.026	0–0.5
Inner wall (IW)	1.42	0.05	0.40	6	0.85	1,580	1.073	0.1–0.4
Fibrous cap (FC)	1.47	0.24	0.78	25	0.78	1,610	1.058	0.2–0.5
Calcification (Cal)	1.47	0.34	0.83	560	0.78	1,540	1.668	0.3–0.5
Lipid pool (LP)	1.47	0.42	0.90	520	0.82	1,650	0.947	0.1–0.3
Mixed calcification (MC)	1.47	0.23	0.96	550	0.80	1,650	0.947	0.01–0.3
Cavity	1.35	0.12	0	610	0.99	1.540	1.065	1.2–1.7

pressure signals reaching the detector surface. The time-dependent pressure is then collected by a finite-sized unfocused ultrasound transducer in the receiving mode, which is located at the center of the image plane and scans the surrounding tissue along a circular trajectory in a full view (360°). The size of the imaging plane is 67 mm × 67 mm, and the image size is 256 × 256 pixels.

To construct the phantom data set for training and testing the binary classification performance of IRR-Net, we fabricate 14 cylindrical phantoms of about 20 mm in diameter and 80 mm in height using materials such as low melting point agar, gelatin, castor oil, intralipid, etc., in which we design embeddings with different shapes, positions, and materials (such as India ink, pencil lead, iron wire, etc.) to mimic light absorbers of different compositions. We use a preclinical multispectral optoacoustic tomography (MSOT) system, inVision 256-TF (iThera Medical GmbH, Munich, Germany) to acquire the PA scanning data of the phantoms. The system uses an Nd:YAG pumped optical parametric oscillator as an excitation source to provide laser illumination at a 10 Hz repetition rate in the near-infrared spectral range of 680–900 nm. The pulse width is 7 ns and the maximum pulse energy is 120 mJ. The detector consists of 256 focused ultrasonic transducer arrays with a center frequency of 5 MHz and a bandwidth of 60% arranged in a ring array with a coverage angle of 270°. After data acquisition, we use the convolutional neural network (CNN) proposed in [12], which is composed of five cascaded DGD units, to reconstruct the image representing the AOED distribution in each slice from the measured pressure signal. The images are then manually labeled with binary labels, namely "embedding" and "other," which are used as the ground truth (GT) for network training and testing.

To construct the *in vivo* data set for training and testing the multiclass classification performance of the joint learning network, we obtain live mouse scan data from the Institute of Materia Medica of the Chinese Academy of Medical Sciences. All animal procedures during data collection are reviewed and approved by the Research Animal Care Subcommittee of the Institute of Materia Medica, Chinese Academy of Medical Sciences. The MSOT system inVision 128 (iThera Medical GmbH, Munich, Germany) is used for whole-body scanning in live mice. A tunable Nd:YAG laser provides laser irradiation at a 10-Hz repetition rate in the spectral range of 715, 730, 760, 800, and 850 nm. At 730 nm, the maximum incident pulse energy is 70 mJ and the pulse width is 8 ns. The detection device consists of 128 focused ultrasonic transducer arrays with a center frequency of 5 MHz and a bandwidth of 60%, arranged into a ring array with a curvature radius of 40 mm and a coverage angle of 270°. After data collection is completed, the images are reconstructed using the DGD-based method [12]. The images are then labeled manually by the medical experts.

2.4. Network Training. We used the error backpropagation algorithm to calculate the gradient of each layer in the network, and we used the gradient descent algorithm for adaptive moment estimation (Adam) [36] to optimize the parameters of each layer along the gradient direction to achieve end-to-end training. The learning rate, batch size, and epoch step were set to 0.001, 32, and 200, respectively.

The loss function for the joint learning framework consists of three parts as follows:

$$\mathscr{L} = \mathscr{L}_{\text{recon}} + \mathscr{L}_{\text{opt}} + \mathscr{L}_{\text{recog}}, \tag{9}$$

where $\mathscr{L}_{\text{recon}}$, $\mathscr{L}_{\text{opt}}$, and $\mathscr{L}_{\text{recog}}$ represent the reconstruction loss, optimization loss, and recognition loss, respectively.

$\mathscr{L}_{\text{recon}}$ indicates the loss of the initial reconstruction by a single DGD unit, which encourages the AOED image output from the reconstruction module to be as similar as possible to the GT image:

$$\mathscr{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{A}_{1,i} - \boldsymbol{A}_{\text{true},i}\right\|_2^2. \tag{10}$$

N is the number of samples in a batch of the training set, $\boldsymbol{A}_{\text{true},i}$ is the expected output of sample i, and $\boldsymbol{A}_{1,i}$ is the output image of sample i from the reconstruction module.

$\mathscr{L}_{\text{opt}}$ indicates the loss of optimization by U-Net, defined by:

$$\mathscr{L}_{\text{opt}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{A}_{i} - \boldsymbol{A}_{\text{true},i}\right\|_2^2, \tag{11}$$

where $\boldsymbol{A}_i$ is the output image of sample i from the optimization module.

$\mathscr{L}_{\text{recog}}$ is the multiclass cross-entropy loss [37], which is used to guarantee differentiation between feature classes, defined by:

$$\mathscr{L}_{\text{recog}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log p_{i,c}. \tag{12}$$

C denotes the number of feature classes, log represents the logarithm based on the natural constant e, $y_{i,c}$ represents the GT label of sample i, and $p_{i,c}$ denotes the prediction probability of classifying sample i as class c. $y_{i,c}$ is a symbolic function which takes 1 if the true class of sample i is c and 0 otherwise.

Tenfold cross-validation was performed to tune hyperparameters to make full use of all data and to avoid locally optimal hyperparameters due to data distribution deviation caused by improper data set partition. For binary classification, the data sets were evaluated using accuracy (Acc), sensitivity (Sens), specificity (Spec), and F1 score [38] as quantitative measures. For multiclass classifications, the data sets were evaluated by calculating the weighted accuracy of each class and the F1 score of all classes.

Table 2: Two-step methods for comparison.

Name of method for comparison		Image reconstruction method	Classifier
HH	5 DGD + GoogLeNet	Five cascaded DGD units	GoogLeNet
	DGD + U-Net+ResNet	A single DGD unit followed by a U-Net	ResNet50
LH	1 DGD + GoogLeNet	A single DGD unit	GoogLeNet
	TR + GoogLeNet	TR	GoogLeNet
	TR + AlexNet	TR	AlexNet
	TR + ResNet	TR	ResNet
HL (5 DGD + SVM)		Five cascaded DGD units	SVM
LL (TR + SVM)		TR	SVM

Note: HH, LH, HL and LL refer to "High-quality reconstruction + High-performance classifier", "Low-quality reconstruction + High-performance classifier", "High-quality reconstruction + Low-performance classifier", and "Low-quality reconstruction + Low-performance classifier," respectively.

3. Experiments and Results

3.1. Experimental Design

3.1.1. Validation Data Sets. We validated the ability of IRR-Net to directly map from raw scanning data to AOED images with recognized targets on simulated, phantom, and *in vivo* data sets, respectively. The construction details of these data sets can be found in Section 2.3

3.1.2. Baseline Methods. To test the performance of the joint learning framework and demonstrate its superiority, a two-step scheme is exploited as the baseline in which object recognition is performed in the image domain as a postprocessing process after image reconstruction. We compare IRR-Net with the eight two-step methods listed in Table 2. In these comparative methods, the network models used for image reconstruction and target recognition subtasks are trained separately.

3.1.3. Evaluation Metrics. We evaluate the algorithm performance in terms of commonly used quantitative metrics in image classification and recognition. Considering that the positive and negative classes in binary classification can be clearly determined, we use accuracy, sensitivity, specificity, and F1 score to measure the accuracy of binary classification. In a multiclass classification scenario, there are at least five categories, making it difficult to determine positive and negative categories. We use accuracy and F1 score to evaluate the accuracy of multiclass classification, as they can evaluate the overall performance of classification without considering specific categories.

3.1.4. Implementation Details. All simulations, image reconstruction by TR, and SVM classifiers are implemented through Matlab programing (R2016a, Math Works, Inc., Natick, Massachusetts) on an AMD Ryzen 7 4800H CPU, Radeon Graphics and Windows 11 64-bit operating system. Neural networks including our proposed IRR-Net and baseline network models are implemented, trained, and tested on an NVIDIA GeForce 3090Ti GPU using Python 3.7 with the deep learning framework of TensorFlow 2.6.0 and Keras 2.0. The data sets for training GoogLeNet, AlexNet, ResNet, and SVM classifiers are constructed as follows: the images reconstructed using the DGD-based method are used as sample inputs, and manually labeled images are used as expected outputs.

3.2. Simulation Data Sets and Test Result. We take the simulated pressure signal matrix in the imaging plane as the input of the sample, and the corresponding simulated AOED image with the manually labeled targets as the expected output. In order to train the performance of the network in classifying different tissue types, two simulation data sets are built: binary classification data set and multiclass classification data set. In the binary classification data set, binary labels, i.e., "lesion" and "other," are used, each accounting for about 50%. In the multiclass classification data set, a multiclass label containing five categories is used, that is, "lesions" are further divided into "calcifications," "lipid pools," "fibrous caps," and "mixed calcifications," accounting for approximately 15%, 5%, 10%, 20%, and 50% respectively. By changing the shape, number, location, and type of tissues, we generate 2,000 pairs of simulated samples. They are shuffled and randomly partitioned into training and testing in a ratio of 8 : 2. To avoid overfitting, we augment the original training set by random rotation (−180° to 180°) and random shift (less than 10% of the image width or height), resulting in 8,000 pairs of samples.

Figure 3 shows the results of image reconstruction and target recognition on a simulation test set with GT labels. Tables 3–5 present the statistical results of the evaluation metrics for target identification using IRR-Net, LL, LH, HL, and HH methods, respectively. The runtime provided in Tables 3 and 4 refers to the time taken from the input of the pressure signal matrix collected in a tomographic slice into the trained network to the output of the labeled grayscale image that represents the AOED distribution. The time spent in training the classifiers and image reconstruction networks is not included. For IRR-Net, this time consists of two parts: the time to calculate the likelihood gradient in Matlab and the time to test the network. For the two-step approach, this time refers to the total time taken to reconstruct and label the image, without taking into account the time taken by any intermediate steps involved.

From Table 3, we find that the overall accuracy, specificity and sensitivity of binary classification by IRR-Net and the HH methods are significantly improved compared with the

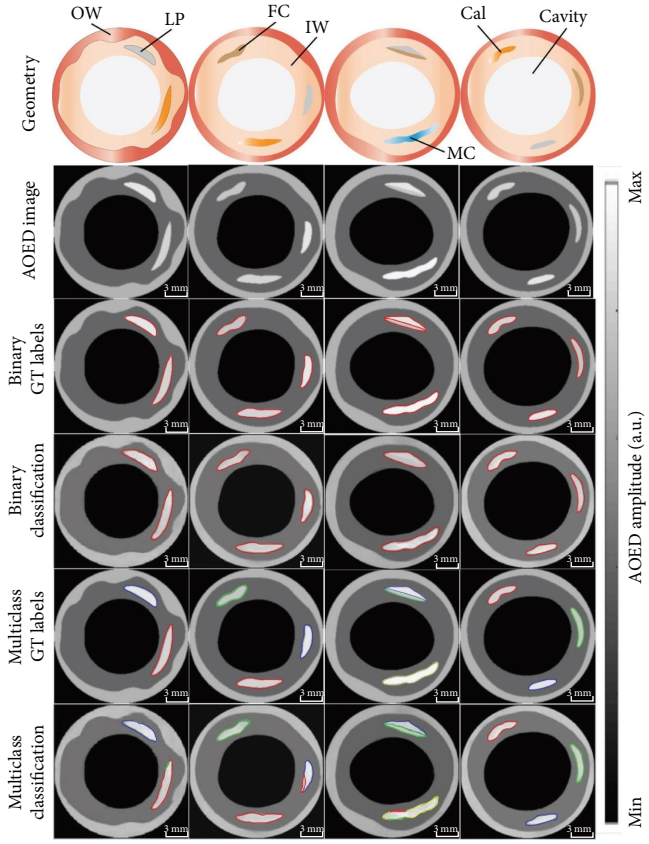


FIGURE 3: Example of image reconstruction and object recognition from the simulated data. In the binary labels, the "lesions" are the regions enclosed by red lines, and the unmarked ones are the "others". In the multiclass labels, "lipid pool," "calcification," "fibrous cap," and "mixed calcification" are the regions enclosed by the blue, red, green, and yellow lines, respectively. The unmarked parts are "other".

LH, LL, and HL methods. The HH (5 DGD + GoogLeNet) reaches the performance upper limit in all methods. In multiclass classification, the classification of the four tissue types of calcification, lipid pool, fibrous cap, and mixed calcification can influence each other, resulting in one tissue type being misclassified into the other three. For example, fibrosis is often mistakenly classified as lipids, and vice versa. From Table 4, we can find that the classification accuracy of calcification is the highest among the four types of tissues, while the accuracy of mixed calcification is relatively low, mainly due to its complex composition and the small number of samples in the training set. Similar to binary classification, HH (5 DGD + GoogleLeNet) performs the best among all methods in terms of accuracy and F1 score in multiclass classification. The improvement is attributed to the high quality of the reconstructed images using five cascaded DGD units, as well as the use of GoogLeNet for high-performance classification. However, high-quality reconstruction is achieved at the cost of extending training and computational time. Due to the fact that the reconstruction network and the classification network are trained separately, the training time of the DGD + U-Net + ResNet and 5 DGD + GoogLeNet methods is much longer than that of IRR-Net, as shown in Table 5. Based on the running time listed in Tables 3 and 4, we can find that the LL method is very fast, less than 2 s. The high-time cost of the IRR-Net and HH methods is attributed to the calculation of the likelihood gradient involved in DGD-based image reconstruction. The time cost of 5 DGD + GoogLeNet is significantly higher than the IRR-Net because it uses five cascaded DGD units to guarantee the high quality of the reconstructed images. In summary, IRR-Net achieves a tradeoff between computational efficiency and accuracy.

By comparing the results in Tables 3 and 4, we find that the accuracy and F1 score of the multiclass classification are significantly lower than those of binary classification, indicating the uncertainty of these methods in the classification of specific types of lesions.

In addition, to verify the necessity of data augmentation, we train IRR-Net using the original simulation data set and the augmented data set, respectively, while other conditions remained unchanged. Table 6 presents the metrics for evaluation binary and multiclass classification. We can find that the performance of IRR-Net in binary and multiclass classification is significantly improved after training on the augmented data set.

3.3. Phantom Data Set and Test Result. We obtain a total of 4,379 slices by scanning the cylindrical phantoms along their long axes. We take the pressure signal matrix collected by the detector as the input of the sample and the manually labeled image as the expected output to construct the original phantom data set containing 4,379 pairs of samples. The samples are shuffled, and 1,305 pairs of samples are randomly selected for testing. The remaining 3,074 pairs are augmented to 6,596 pairs for training.

We validate the performance of the trained IRR-Net in reconstructing images and recognizing embedded targets on the experimental phantom test set. Figure 4 presents the results for four phantoms. Nine different algorithms are compared regarding the accuracy, sensitivity, specificity, F1 score, and computational time for the binary classification. Table 7 provides the result of evaluation metrics for binary classification. From the table, we find that deep learning-based classifiers have significantly higher accuracy and F1 scores than SVM. Our IRR-Net outperforms LL, LH, and HL on the phantom data set in terms of all evaluation metrics, which can improve F1 scores by 8.5%. In addition, the performance of IRR-Net is very close to that of HH (5DGD + GoogLeNet), which is the upper performance limit of the comparative methods, while the computational time is reduced by half. The results show that IRR-Net strikes a balance between accuracy and time cost.

3.4. In Vivo Data Set and Test Result. We collect a total of 1,970 cephalic, thoracic and abdominal slices from whole-body scans of anesthetized male nude mice (6–8 weeks of age, supine, and prone). A scanning signal slice and its corresponding manually labeled image with multiclass labels are taken as the input and expected output of an *in vivo* sample,

Table 3: Evaluation metrics for binary classification in the simulation study.

Method		Acc	Sens	Spec	F1 score	Runtime (s)
HH (5 DGD + GoogLeNet)		0.9574	0.9535	0.9691	0.9711	21.4
IRR-Net		**0.9474**	**0.9324**	**0.9524**	**0.9572**	**10.6**
HH (DGD + U-Net+ResNet)		0.9474	0.9324	0.9524	0.9572	28.6
LH	1 DGD + GoogLeNet	0.9277	0.9189	0.9486	0.9503	9.3
	TR + GoogLeNet	0.9184	0.9098	0.9440	0.9436	3.2
	TR + AlexNet	0.9012	0.8910	0.9316	0.9311	2.7
	TR + ResNet	0.8987	0.8892	0.9303	0.9275	2.7
HL (5 DGD + SVM)		0.8926	0.8874	0.9271	0.9208	20.1
LL (TR + SVM)		0.8739	0.8665	0.8961	0.9117	1.4

Table 4: Evaluation metrics for multiclass classification in the simulation study.

Method		Acc				F1 score	Runtime (s)
		Cal	FC	LP	MC		
HH (5 DGD + GoogLeNet)		0.9243	0.9146	0.9322	0.8801	0.8980	29.1
IRR-Net		**0.9191**	**0.9089**	**0.9161**	**0.8702**	**0.8878**	**10.7**
HH (DGD + U-Net+ResNet)		0.9191	0.9089	0.9161	0.8702	0.8878	33.6
LH	1 DGD + GoogLeNet	0.9002	0.8913	0.8867	0.8561	0.8705	9.8
	TR + GoogLeNet	0.8976	0.8891	0.8555	0.8405	0.8627	3.1
	TR + AlexNet	0.8776	0.8702	0.8415	0.8245	0.8229	2.7
	TR + ResNet	0.8714	0.8649	0.8358	0.8191	0.8203	2.7
HL (5 DGD + SVM)		0.8617	0.8587	0.8302	0.8238	0.8189	28.6
LL (TR + SVM)		0.8575	0.8461	0.8266	0.8212	0.7968	1.4

CAL, FC, LP, and MC stand for calcification, lipid pool, fibrous cap, and mixed calcification, respectively.

Table 5: Training time in hours.

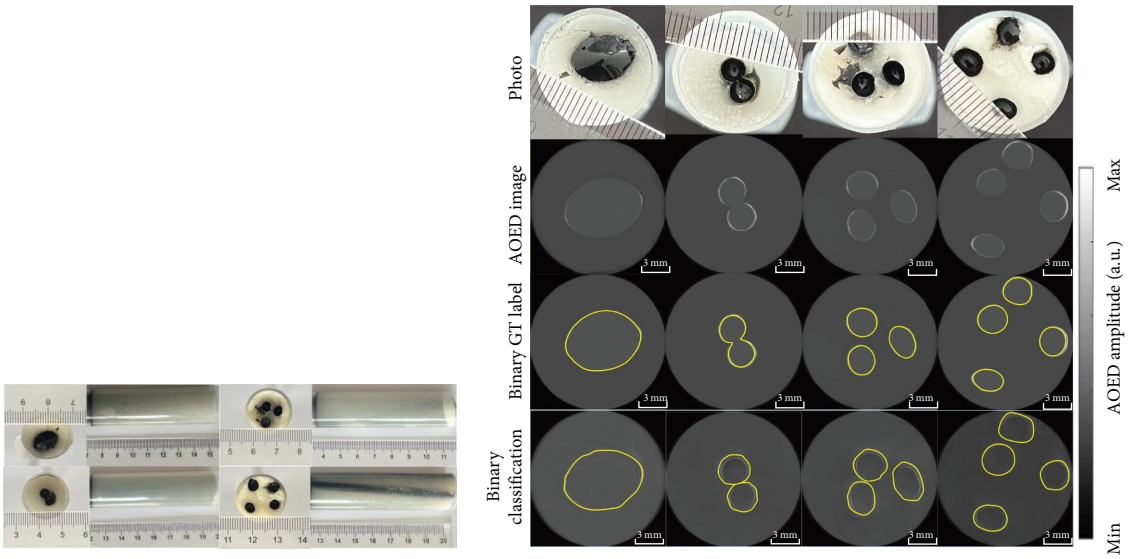
	5 DGD + GoogLeNet	DGD + U-Net + ResNet	1 DGD + GoogLeNet	IRR-Net
Total training time	27.6	11	6.8	5.3
Training time of reconstruction network	26	6.5	5.2	—
Training time of optimization network	—	2.4	—	—
Training time of classifier	1.6	2.1	1.6	—

Table 6: Evaluation metrics for binary classification and multiclass classification before and after data augmentation in the simulation study.

Data set	Binary classification				Multiclass classification				
	Acc	Sens	Spec	F1 score	Acc				F1 score
					Cal	FC	LP	MC	
Original data set	0.9330	0.9012	0.9467	0.9310	0.8274	0.8173	0.8168	0.8108	0.8569
Augmented data set	**0.9474**	**0.9324**	**0.9524**	**0.9572**	**0.9191**	**0.9089**	**0.9161**	**0.8702**	**0.8878**

respectively. We randomly select 750 samples for testing, and augment the remaining 1,220 samples to 4,025 samples for training.

To further analyze the accuracy and generalizability of our proposed IRR-Net, we test it on the *in vivo* data set. Figure 5 presents the results for image reconstruction and tissue identification of cephalic, thoracic, and abdominal slices, in which a multiclass classification was performed, including mandible, tongue, spine, kidney, liver, abdominal aorta, and ribs. Table 8 presents the accuracy, F1 score, and runtime for nine methods. Compared with the simulation and phantom experiments, the overall performance of IRR-Net in multiclass classification of *in vivo* images containing complex anatomical structures has decreased, but it still outperforms LL, LH, and HL. IRR-Net has the highest classification precision for the spine (81.09%) and the lowest for the



(a) (b)

FIGURE 4: Example of image reconstruction and object recognition from the scanning data of the experimental phantoms. (a) Photos of four cylindrical phantoms and (b) images reconstructed using the DGD method [12] and the object recognition result. The GT of the binary labels was obtained manually on the images. In the binary labels, "embedding" refers to the region enclosed by the yellow line. Unmarked parts are labeled as "other".

TABLE 7: Evaluation metrics for binary classification of experimental phantom images.

Method		Acc	Sens	Spec	F1 score	Runtime (s)
HH (5 DGD + GoogLeNet)		0.9211	0.9338	0.8879	0.9502	23.6
IRR-Net		**0.9108**	**0.9237**	**0.8721**	**0.9395**	**11.8**
HH (DGD + U-Net+ResNet)		0.9108	0.9237	0.8721	0.9395	30.2
LH	1 DGD + GoogLeNet	0.8907	0.9123	0.8516	0.9206	10.3
	TR + GoogLeNet	0.8743	0.8873	0.8354	0.9137	3.1
	TR + AlexNet	0.8562	0.8694	0.8166	0.9007	2.6
	TR + ResNet	0.8511	0.8647	0.8132	0.8916	2.6
HL (5 DGD + SVM)		0.8461	0.8573	0.8094	0.8812	22.7
LL (TR + SVM)		0.8088	0.8221	0.7689	0.8658	1.5

ribs (75.75%) among all types of tissues. This is because the number of spinal samples in the training set is greater than that of other structures. Therefore, the network can learn more features of the spine. Furthermore, the GT labels in the data set were manually marked, which can result in unclear boundary representation and a decrease in classification accuracy. In addition, compared to the simulation experiment, the evaluation indicators of both phantom and *in vivo* images have decreased, which is related to the decrease in image quality in practical application scenarios.

4. Discussion

4.1. Influence of the Number of DGD Units. In this section, we analyze the impact of the number of DGD units in IRR-Net on the quality of reconstructed images through simulation, phantom and *in vivo* experiments. Figure 6 shows the images reconstructed using our joint framework and the metrics for evaluating image quality, where the image reconstruction modules consists of 1, 2, 3, and 4 DGD units, respectively. From the figure, we can find that increasing the number of DGD units does not significantly improve the quality of the reconstructed image due to the architecture of initial reconstruction followed by image optimization.

In the simulation experiment, compared with the architecture of "1 DGD unit + U-Net," the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) of the images reconstructed using "4 DGD units + U-Net" are only increased by 1.847% and 1.185%, respectively. In both phantom and *in vivo* experiments, as the number of DGD units increases, there is no significant difference in the visual effect of the images, and the metrics only slightly improved. Compared with "1 DGD unit + U-Net," the contrast (CR) and noise ratio (CNR) of the phantom images reconstructed by "4 DGD units + U-Net" are increased by approximately 1.7635% and 0.8942%, respectively, and the CR and CNR of the *in vivo*

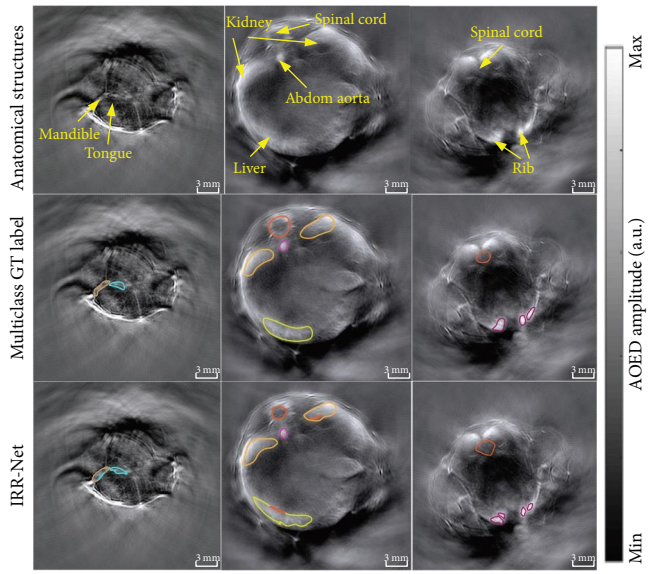


Figure 5: Example of image reconstruction and tissue characterization from the whole-body scanning pressure data of living mice. The GT of the multiclass labels was obtained manually by medical experts on the reconstructed images using the DGD-based approach [12]. In the multiclass GT labels, the anatomical structures such as mandible, tongue, liver, kidney, spinal cord, abdom aorta, and rib are the regions enclosed by lines in different colors.

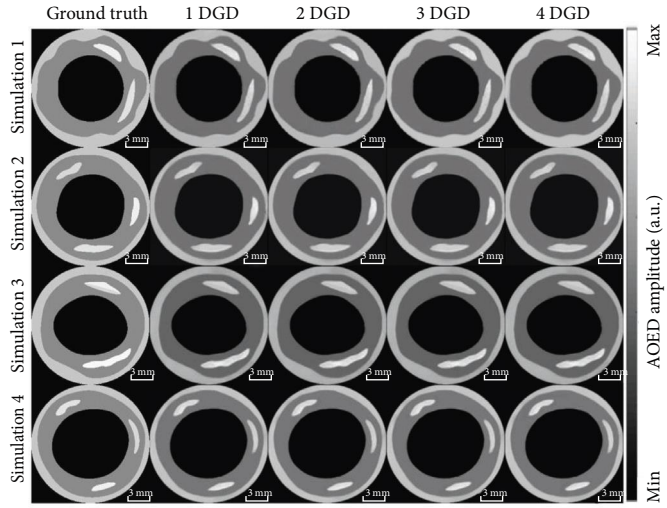
Table 8: Evaluation metrics for multiclass classification of *in vivo* images.

Method		Acc							F1 score	Runtime (s)
		Mandible	Tongue	Liver	Kidney	Spinal cord	Abdom aorta	Rib		
HH (5 DGD + GoogLeNet)		0.8134	0.8067	0.8203	0.7946	**0.8286**	0.7798	0.7802	0.6251	24.1
IRR-Net		0.7941	0.7855	0.7995	0.7667	**0.8109**	0.7657	0.7575	0.6135	12.0
HH (DGD + U-Net+ResNet)		0.7941	0.7855	0.7995	0.7667	**0.8109**	0.7657	0.7575	0.6135	33.7
LH	1 DGD + GoogLeNet	0.7884	0.7811	0.7805	0.7602	**0.7968**	0.7613	0.7536	0.5904	10.9
	TR + GoogLeNet	0.7733	0.7766	0.7741	0.7545	**0.7839**	0.7574	0.7504	0.5745	3.1
	TR + AlexNet	0.7686	0.7705	0.7687	0.7478	**0.7743**	0.7536	0.7488	0.5544	2.7
	TR + ResNet	0.7677	0.7700	0.7583	0.7443	**0.7725**	0.7524	0.7465	0.5537	2.7
HL (5 DGD + SVM)		0.7670	0.7698	0.7452	0.7401	**0.7708**	0.7515	0.7440	0.5531	21.5
LL (TR + SVM)		0.7560	0.7681	0.7227	0.7354	**0.7686**	0.7503	0.7380	0.5433	1.5

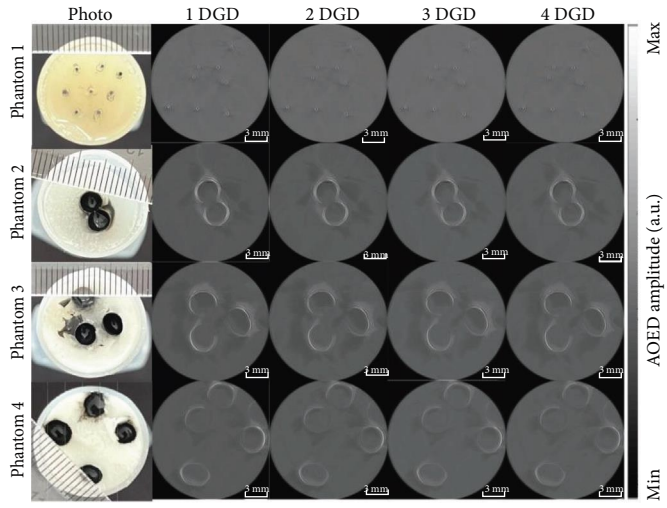
images are increased by 1.4880% and 0.7974%, respectively. However, adding additional DGD units in the joint framework greatly extends the training time and image formation time, as shown in Figure 6(g). Therefore, in order to find a balance between time cost and image quality, we use a single DGD unit for the initial reconstruction.

4.2. Necessity of Optimization Module. In this section, we discuss the necessity of an optimization module in reconstructing high-quality images in the proposed framework. We compare the images reconstructed using a DGD unit followed by an optimization module (i.e., 1 DGD + U-Net), a single DGD unit, and traditional BP, respectively. Figure 7 presents the results of image reconstruction. From the figure, we can observe the poor quality of BP reconstructions, with obvious blurring, distortion, and overall low brightness. In the images reconstructed using a single DGD unit, the regions around the vessel wall and the lesions appear blurred, and the visualization of areas containing multiple tissue components is poor. In contrast, the images reconstructed using 1 DGD + U-Net have significantly improved quality, with clear tissue boundaries and high contrast between different tissue regions. This conclusion can also be drawn from the metrics for evaluating the image quality shown in Figure 7(d)–7(f), indicating that the use of optimization module improves the quality of reconstruction.

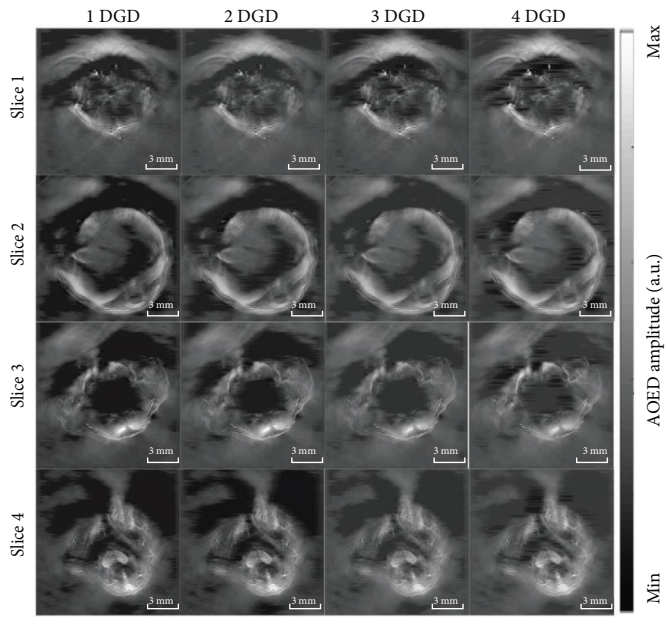
4.3. Advantages and Limitations. Based on the analysis of all the experimental results, we can find that our joint learning framework improves the efficiency of image reconstruction and object recognition compared with the separately learned subtask networks. For example, in the binary classification experiment of the phantom study, IRR-Net achieves a 12.6% improvement in classification accuracy over the traditional TR + SVM approach, despite approximately eight times longer computation time. Compared with the HH (5 DGD



(a)

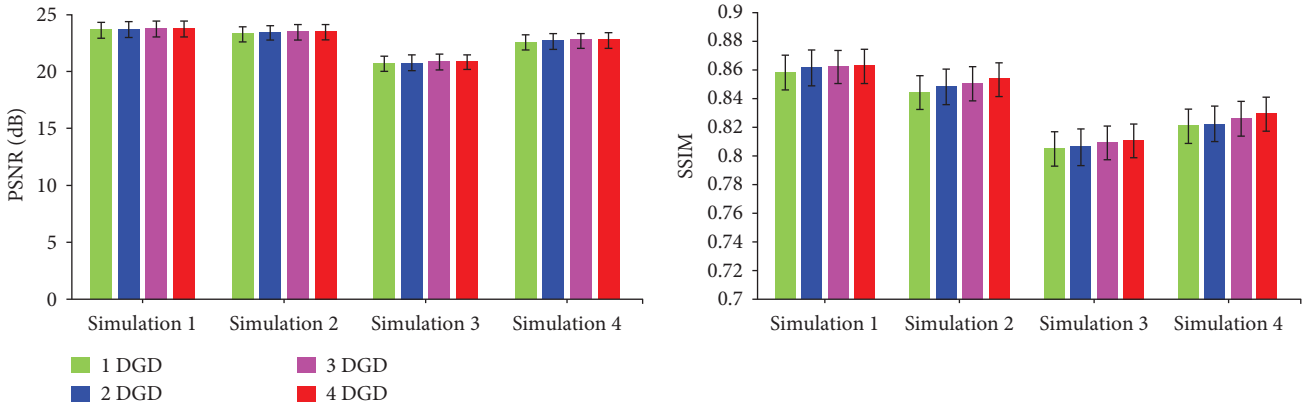


(b)

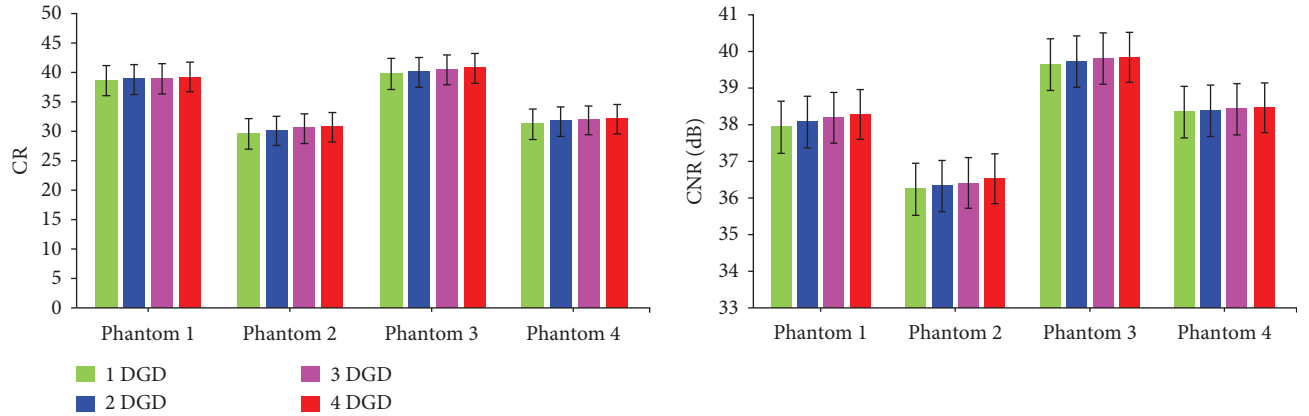


(c)

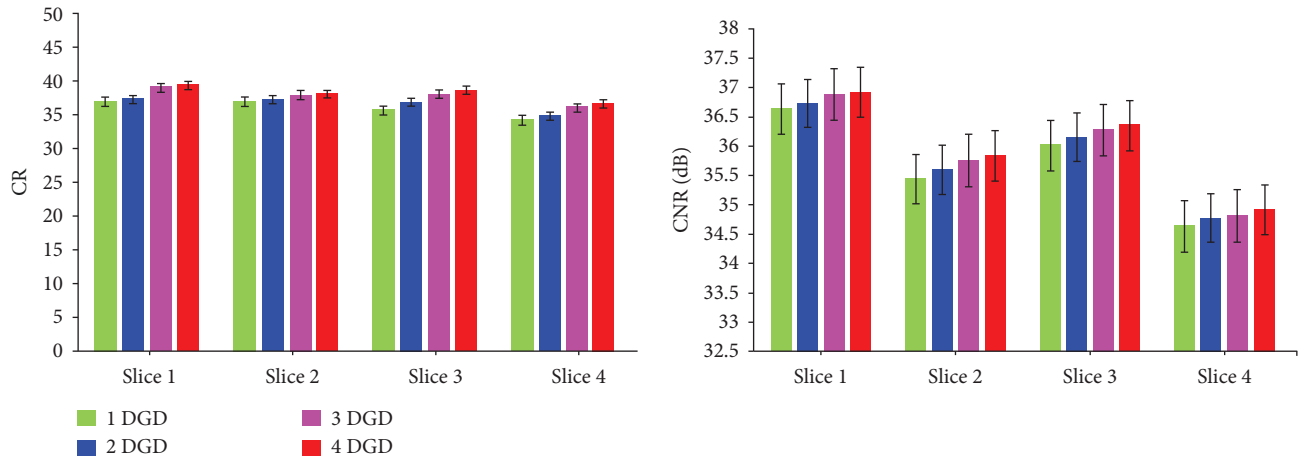
Figure 6: Continued.



(d)

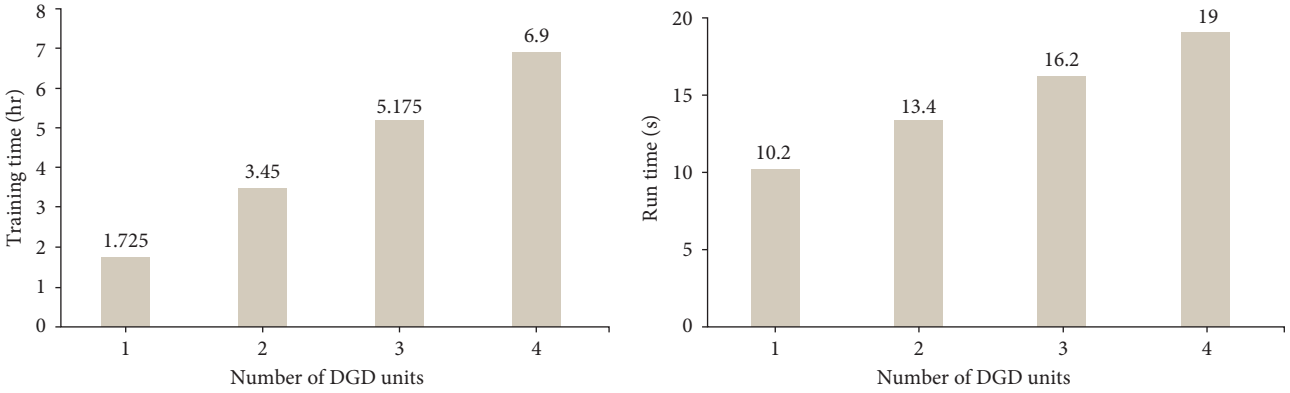


(e)



(f)

Figure 6: Continued.

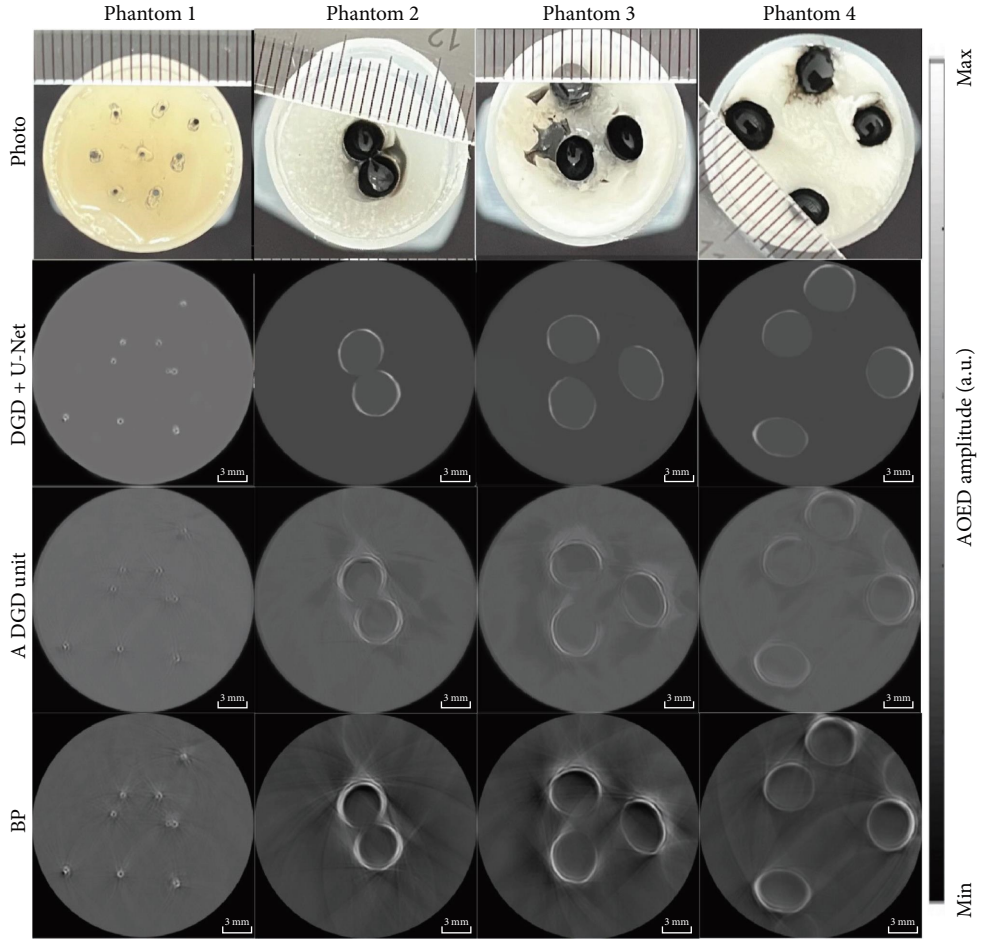


(g)

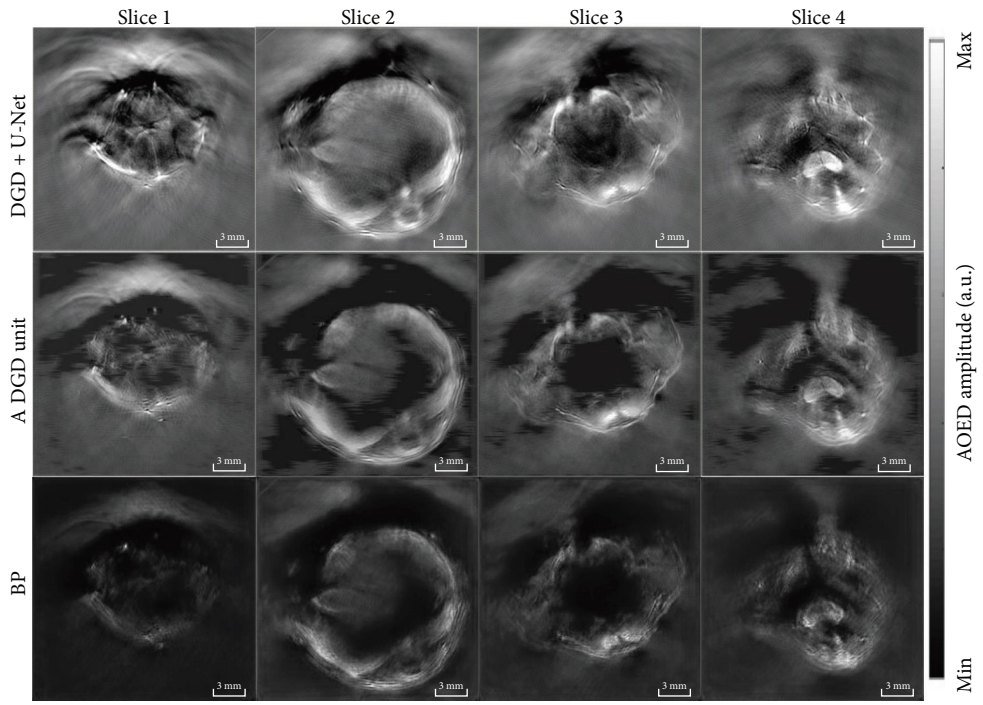
Figure 6: Reconstructed images using our joint learning framework, where the image reconstruction modules consists of 1, 2, 3, and 4 DGD units, respectively, and the metrics for evaluating the image quality. (a) Images reconstructed from the simulation data; (b) images reconstructed from the scanning data of the experimental phantoms; (c) images reconstructed from the whole-body scanning data of living mice; (d) PSNR and SSIM of the reconstructed images from the simulation data; (e) CR and CNR of the phantom images; (f) CR and CNR of the *in vivo* images; and (g) training time and running time.

(a)

Figure 7: Continued.

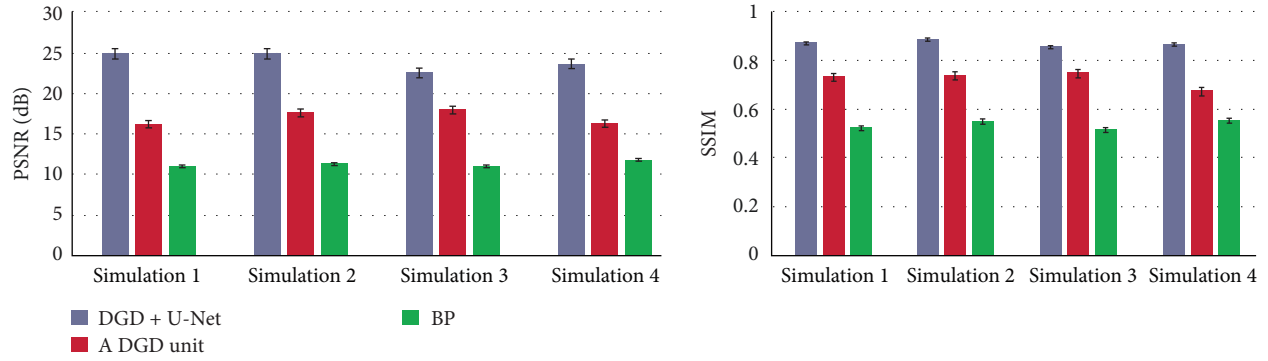


(b)

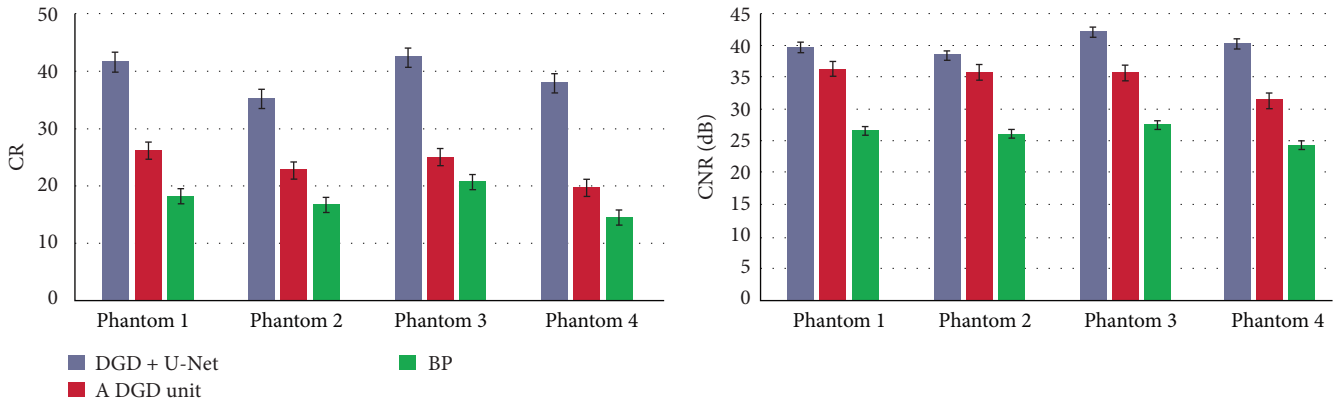


(c)

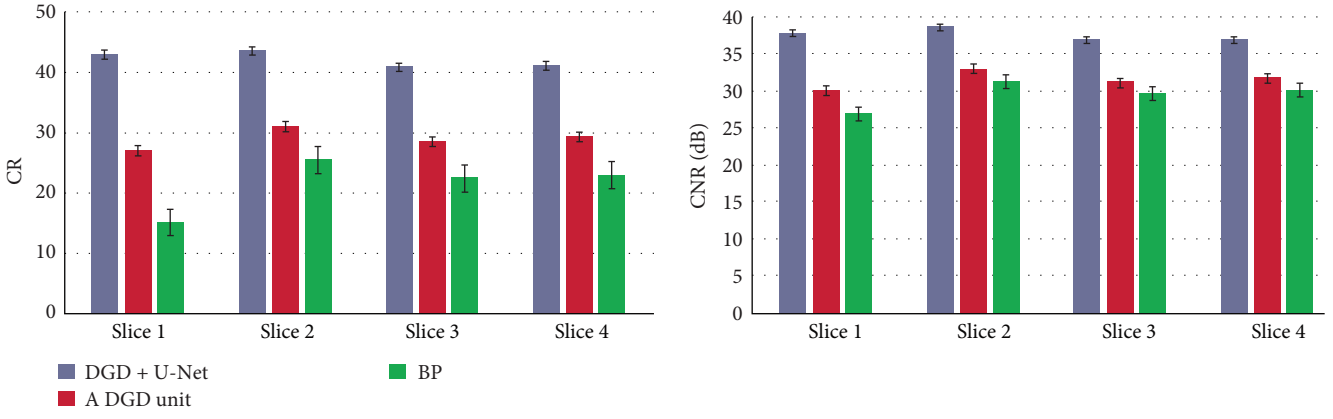
Figure 7: Continued.



(d)



(e)



(f)

FIGURE 7: Images reconstructed using a single DGD unit and 1 DGD + U-Net, respectively, and the metrics for evaluating the image quality. (a) Images reconstructed from the simulation data; (b) images reconstructed from the scanning data of the experimental phantoms; (c) images reconstructed from the whole-body scanning data of living mice; (d) PSNR and SSIM of the reconstructed images from the simulation data; (e) CR and CNR of the phantom images; and (f) CR and CNR of the *in vivo* images.

+ GoogLeNet) method with the highest accuracy in our experiment, the accuracy of IRR-Net is slightly lower by about 1.13%, while the computation time on the GPU is reduced from 23.6 to 11.8 s, and the training time is significantly reduced from 27.6 to 5.3 hr. Our method utilizes a CNN to implement joint learning for image reconstruction and recognition, seeking a tradeoff between computational efficiency and accuracy.

The acoustic inversion of PAT is ill-posed due to the influence of complex factors involving incident irradiation, acoustic propagation, and ultrasonic detection, such as pulsed laser energy fluctuation, nonuniform light coverage, incorrect sound speed assumption, heterogeneously distributed acoustic properties, imaging system calibration error, limited-view scanning geometry, limited detector bandwidth, and incomplete measurement data. In this work, to improve

the image quality, we adopt the scheme of initial reconstruction followed by optimization. For image optimization or enhancement tasks, it can be difficult to accurately identify the causes of image quality degradation and implement targeted solutions. In the initial reconstruction, the forward imaging model can fully consider the nonideal scenarios to avoid the factors that may lead to the degradation of image quality. However, the computational cost of a physical model is proportional to its completeness. The more comprehensive and realistic the imaging process described by the model, the higher the computational complexity required to solve its inverse problem. In the forward imaging model established in this work, the inhomogeneous acoustic medium with spatially varying sound speed and density is taken into account. Other nonideal factors mentioned above are not included. The experimental results show that IRR-Net improves the accuracy of multiclass classification compared to HL, LL, and LH methods. However, the classification accuracy of IRR-Net is slightly lower than that of HH (5 DGD + GoogLeNet), although the latter has a much higher time cost. This may be due to the fact that artifacts are not completely eliminated during image formation, resulting in artifacts being misclassified as lesions. One of our future work is to build a forward physics model that can adequately describe real-world imaging scenarios, further improve the quality of reconstructed images, and find a tradeoff between accuracy and computational cost. In addition, we plan to conduct experiments to test which model architecture (such as U-Net, ResNet, and GAN) or U-Net parameters (such as the number of convolutional filters and kernel size) can be used as optimization modules to yield an optimal balance between image quality and inference time.

5. Conclusions

This work presents a joint learning framework for simultaneous image reconstruction and target recognition. The framework consists of the DGD module, the U-Net module, and the ResNet-50 module. The DGD module maps the raw photoacoustic signal to the optical absorption distribution image. The forward imaging operator and its adjoint operator define the mapping relationship between the optical absorption and the measured acoustic pressure, which are included in the likelihood gradient calculation, but are separated from the network training, thus reducing the complexity of the network architecture and training. The U-Net module optimizes the initial reconstruction and outputs high-quality images. The ResNet-50 module is used for image feature extraction and target recognition and outputs the labeled image. The feasibility of this method has been verified by the simulations, phantom, and *in vivo* studies. In addition, experiments have been conducted to compare the performance of the proposed method with state-of-the-art nonlearning and learning-based methods. The results show that IRR-Net achieves a balance between computational efficiency and accuracy compared to separately trained networks for subtasks. This method provides a universal deep learning scheme for simultaneous image reconstruction and object recognition. When the forward imaging operator is changed, it can be used for other tomographic imaging modalities.

Data Availability

The data sets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank the Beijing University of Chemical Technology and the Institute of Materia Medica of the Chinese Academy of Medical Sciences for their support and assistance in the acquisition of the imaging data. This work was financially supported by the National Natural Science Foundations of China (no. 62071181).

References

[1] W. Choi, D. Oh, and C. Kim, "Practical photoacoustic tomography: realistic limitations and technical solutions," *Journal of Applied Physics*, vol. 127, no. 23, Article ID 230903, 2020.

[2] R. Manwar, M. Zafar, and Q. Xu, "Signal and image processing in biomedical photoacoustic imaging: a review," *Optics*, vol. 2, no. 1, pp. 1–24, 2021.

[3] M. H. Xu and L. V. Wang, "Universal back-projection algorithm for photoacoustic computed tomography," in *Proceeding of SPIE International Conference on Photons Plus Ultrasound: Imaging and Sensing 2005*, pp. 251–254, SPIE, San Jose, USA, 2005.

[4] Z. Sun, D. Han, and Y. Yuan, "2-D image reconstruction of photoacoustic endoscopic imaging based on time-reversal," *Computers in Biology and Medicine*, vol. 76, pp. 60–68, 2016.

[5] J. Poudel, L. Yang, and M. A. Anastasio, "A survey of computational frameworks for solving the acoustic inverse problem in three-dimensional photoacoustic computed tomography," *Physics in Medicine & Biology*, vol. 64, Article ID 14TR01, 2019.

[6] P. Rajendran, A. Sharma, and M. Pramanik, "Photoacoustic imaging aided with deep learning: a review," *Biomedical Engineering Letters*, vol. 12, no. 2, pp. 155–173, 2022.

[7] D. Allman, A. Reiter, and M. A. L. Bell, "Photoacoustic source detection and reflection artifact removal enabled by deep learning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1464–1477, 2018.

[8] 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," *Experimental Biology and Medicine*, vol. 245, no. 7, pp. 597–605, 2020.

[9] G. Godefroy, B. Arnal, and E. Bossy, "Compensating for visibility artefacts in photoacoustic imaging with a deep learning approach providing prediction uncertainties," *Photoacoustics*, vol. 21, Article ID 100218, 2021.

[10] A. Hauptmann, F. Lucka, M. Betcke et al., "Model-based learning for accelerated, limited-view 3-D photoacoustic tomography," *IEEE Transactions on Medical Imaging*, vol. 37, no. 6, pp. 1382–1393, 2018.

[11] F. K. Joseph, A. Arora, P. Kancharla, M. K. A. Singh, W. Steenbergen, and S. S. Channappayya, "Generative adversarial network based photoacoustic image reconstruction from band limited and limited-view data," *Proceedings of SPIE International Conference on Photons Plus Ultrasound: Imaging and Sensing*, vol. 11642, Article ID 1164235, 2021.

[12] Z. Sun, X. Wang, and X. Yan, "An iterative gradient convolutional neural network and its application in endoscopic photoacoustic image formation from incomplete acoustic measurement," *Neural Computing and Applications*, vol. 33, no. 14, pp. 8555–8574, 2021.

[13] H. Lan, C. Yang, and F. Gao, "A jointed feature fusion framework for photoacoustic image reconstruction," *Photoacoustics*, vol. 29, Article ID 100442, 2023.

[14] H. F. Zhang, K. Maslov, and L. V. Wang, "Automatic algorithm for skin profile detection in photoacoustic microscopy," *Journal of Biomedical Optics*, vol. 14, no. 2, Article ID 024050, 2009.

[15] S. Mandal, X. L. Deán-Ben, and D. Razansky, "Visual quality enhancement in optoacoustic tomography using active contour segmentation priors," *IEEE Transactions on Medical Imaging*, vol. 35, no. 10, pp. 2209–2217, 2016.

[16] K. M. Meiburger, S. Y. Nam, E. Chung, L. J. Suggs, S. Y. Emelianov, and F. Molinari, "Skeletonization algorithm-based blood vessel quantification using *in vivo* 3D photoacoustic imaging," *Physics in Medicine and Biology*, vol. 61, no. 22, Article ID 7994, 2016.

[17] T. Oruganti, J. G. Laufer, and B. E. Treeby, "Vessel filtering of photoacoustic images," *Proceeding of SPIE International Conference on Photons Plus Ultrasound: Imaging and Sensing 2013*, vol. 8581, Article ID 85811W, 2013.

[18] P. Raumonen and T. Tarvainen, "Segmentation of vessel structures from photoacoustic images with reliability assessment," *Biomedical Optics Express*, vol. 9, no. 7, pp. 2887–2904, 2018.

[19] M. Sun, C. Li, N. Chen et al., "Full three-dimensional segmentation and quantification of tumor vessels for photoacoustic images," *Photoacoustics*, vol. 20, Article ID 100212, 2020.

[20] C. Lutzweiler, R. Meier, and D. Razansky, "Optoacoustic image segmentation based on signal domain analysis," *Photoacoustics*, vol. 3, no. 4, pp. 151–158, 2015.

[21] N.-K. Chlis, A. Karlas, N.-A. Fasoula et al., "A sparse deep learning approach for automatic segmentation of human vasculature in multispectral optoacoustic tomography," *Photoacoustics*, vol. 20, Article ID 100203, 2020.

[22] B. Lafci, E. Merčep, S. Morscher, X. L. Deán-Ben, and D. Razansky, "Deep learning for automatic segmentation of hybrid optoacoustic ultrasound (OPUS) images," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 3, pp. 688–696, 2021.

[23] A. R. Rajanna, R. Ptucha, S. Sinha, B. Chinni, V. Dogra, and N. A. Rao, "Prostate cancer detection using photoacoustic imaging and deep learning," *Electronic Imaging*, vol. 28, Article ID art00007, 2016.

[24] J. Zhang, B. Chen, M. Zhou, H. Lan, and F. Gao, "Photoacoustic image classification and segmentation of breast cancer: a feasibility study," *IEEE Access*, vol. 7, pp. 5457–5466, 2019.

[25] Z. Wei, B. Liu, B. Dong, and L. Wei, "A joint reconstruction and segmentation method for limited-angle X-ray tomography," *IEEE Access*, vol. 6, pp. 7780–7791, 2018.

[26] V. Corona, M. Benning, M. J. Ehrhardt et al., "Enhancing joint reconstruction and segmentation with non-convex Bregman iteration," *Inverse Problems*, vol. 35, no. 5, Article ID 055001, 2019.

[27] L. Qiu and H. Ren, "RSegNet: a joint learning framework for deformable registration and segmentation," *IEEE Transactions on Automation Science and Engineering*, vol. 19, no. 3, pp. 2499–2513, 2022.

[28] Y. E. Boink, S. Manohar, and C. Brune, "A partially-learned algorithm for joint photo-acoustic reconstruction and segmentation," *IEEE Transactions on Medical Imaging*, vol. 39, no. 1, pp. 129–139, 2020.

[29] B. E. Treeby and B. T. Cox, "Modeling power law absorption and dispersion for acoustic propagation using the fractional Laplacian," *The Journal of the Acoustical Society of America*, vol. 127, no. 5, pp. 2741–2748, 2010.

[30] L. Mohammadi, H. Behnam, J. Tavakkoli, and M. Avanaki, "Skull's photoacoustic attenuation and dispersion modeling with deterministic ray-tracing: towards real-time aberration correction," *Sensors*, vol. 19, no. 2, Article ID 345, 2019.

[31] M. Tabei, T. D. Mast, and R. C. Waag, "A *k*-space method for coupled first-order acoustic propagation equations," *The Journal of the Acoustical Society of America*, vol. 111, no. 1, pp. 53–63, 2002.

[32] J. Adler and O. Öktem, "Solving ill-posed inverse problems using iterative deep neural networks," *Inverse Problems*, vol. 33, no. 12, Article ID 124007, 2017.

[33] K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, IEEE, Las Vegas, NV, USA, 2016.

[34] L. Wang, S. L. Jacques, and L. Zheng, "MCML—Monte Carlo modeling of light transport in multi-layered tissues," *Computer Methods and Programs in Biomedicine*, vol. 47, no. 2, pp. 131–146, 1995.

[35] 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, Article ID 021314, 2010.

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

[37] X. Chen, S. Kar, and D. A. Ralescu, "Cross-entropy measure of uncertain variables," *Information Sciences*, vol. 201, pp. 53–60, 2012.

[38] G.-H. Fu, F. Xu, B.-Y. Zhang, and L.-Z. Yi, "Stable variable selection of class-imbalanced data with precision–recall criterion," *Chemometrics and Intelligent Laboratory Systems*, vol. 171, pp. 241–250, 2017.