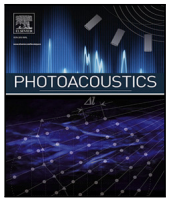
Photoacoustics 38 (2024) 100608

Contents lists available at ScienceDirect

Photoacoustics

journal homepage: www.elsevier.com/locate/pacs



Research article

UPAMNet: A unified network with deep knowledge priors for photoacoustic microscopy☆

Yuxuan Liu [a], Jiasheng Zhou [a], Yating Luo [a], Jinkai Li [a], Sung-Liang Chen [b], Yao Guo [a,*], Guang-Zhong Yang [a,*]

[a] Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[b] University of Michigan–Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai, 200240, China

ARTICLE INFO

Keywords:
Photoacoustic microscopy
Deep neural network
Image super-resolution
Image denoising

ABSTRACT

Photoacoustic microscopy (PAM) has gained increasing popularity in biomedical imaging, providing new opportunities for tissue monitoring and characterization. With the development of deep learning techniques, convolutional neural networks have been used for PAM image resolution enhancement and denoising. However, there exist several inherent challenges for this approach. This work presents a **U**nified **P**hoto**A**coustic **M**icroscopy image reconstruction **Net**work (UPAMNet) for both PAM image super-resolution and denoising. The proposed method takes advantage of deep image priors by incorporating three effective attention-based modules and a mixed training constraint at both pixel and perception levels. The generalization ability of the model is evaluated in details and experimental results on different PAM datasets demonstrate the superior performance of the method. Experimental results show improvements of 0.59 dB and 1.37 dB, respectively, for 1/4 and 1/16 sparse image reconstruction, and 3.9 dB for image denoising in peak signal-to-noise ratio.

1. Introduction

Photoacoustic microscopy (PAM) is now an established technique in biomedical imaging that acoustically detects the optical absorption contrast via the photoacoustic effect [1,2]. Using a pulsed laser beam, biological tissues undergo thermoelastic expansion, which is proportional to the optical absorption, and release acoustic signals that can be detected by an ultrasound transducer [3,4]. Due to the weaker scattering of ultrasound in tissue than that of optical scattering, PAM can achieve non-invasive high-resolution images at greater depths without radiation compared to traditional optical microscopy technologies [5].

Thus far point-by-point scanning has been widely used for biomedical PAM imaging, which scans a point of the tissue at each step [6–8]. In practice, when attempting to achieve higher acquisition speed, scanning the tissues with a larger step size is feasible. However, the use of larger step sizes may result in poorer image quality and lower resolution, which hinders the effectiveness of the obtained information. Super-resolution makes it possible to sparsely scan the points of the target and reconstruct high-resolution images from downsampled low-resolution images via advanced practical algorithms, which provides the opportunities to accelerate the imaging speed and efficiency of PAM system as well as improving the imaging quality of PAM images. During the PAM data acquisition, noise is unexpectedly generated from different aspects [9,10], making the obtained PA signal suffer from a low signal-to-noise ratio. Consequently, high signal-to-noise ratio PAM images are required and noise is required to be suppressed. Denoising has been an alternative and promising technique to remove the generated background noise and enhance the signal-to-noise ratio of PAM images for further animal experiments and clinical studies. Generally speaking, there is always a need for low-cost and advanced post-processing algorithms for both PAM images super-resolution and denoising which can serve as an addition to the advanced hardware.

Recent advances in deep learning for biomedical imaging reconstruction and restoration [11–13] provide new opportunities for fast and high-quality PAM imaging. Deep Convolutional Neural Networks (CNN) have been proposed for PAM image super-resolution [6,14–16] with different CNN architectures. DiSpirito et al. [6] proposed a fully dense U-net deep learning model to reconstruct images at different sampling ratios. Zhou et al. [14] proposed a CNN model with residual blocks and channel-wise attention to contribute to the performance of sparser image reconstruction. Vu et al. [15] proposed a self-supervised network for downsampled image reconstruction via deep image prior. In parallel, PAM image denoising based on deep learning has also

☆ This work was supported by Shanghai Municipal Science and Technology Major Project 2021SHZDZX, and in part supported by the Science and Technology Commission of Shanghai Municipality, China under Grant 20DZ2220400 and China Postdoctoral Science Foundation, China 2023M732246.

* Corresponding authors.
E-mail addresses: 20000905lyx@sjtu.edu.cn (Y. Liu), yao.guo@sjtu.edu.cn (Y. Guo), gzyang@sjtu.edu.cn (G.-Z. Yang).

https://doi.org/10.1016/j.pacs.2024.100608
Received 27 January 2024; Received in revised form 10 March 2024; Accepted 16 April 2024
Available online 25 April 2024

been explored [17–19]. He et al. [19] proposed an image noise model for photoacoustic microscopy and proposed a generative adversarial network for PAM image denoising.

However, there are several inherent challenges for PAM image reconstruction and denoising compared to conventional natural image processing. First, there is a lack of available training datasets as point-by-point scanning for PAM is time-consuming and not widely used clinically, making it difficult to collect high-quality ground truth data [20]. Existing large models for natural image processing cannot be readily used for PAM due to the potential overfitting problem associated with large model parameters. Second, compared to natural scenes, PAM images are not rich in textures and distinctive features. Traditional pixel-level constraints used in natural image reconstruction, like Mean Square Error (MSE), can easily lead to over-smoothing [14,19]. Directly using a popular super-resolution network in natural image reconstruction for PAM super-resolution makes the generated images with bad quality. Third, significant domain shifts exist between different PAM datasets due to the heterogeneity of the imaging environment and hardware setup [20]. Commonly used and pre-trained models on source datasets may result in poor performance when directly applied to an unseen dataset under different experimental settings. Transfer learning is required for better evaluation of cross-dataset performance and the generalization ability of existing methods. More detailed descriptions of these challenges can be found in supplementary materials.

To address the aforementioned challenges, we propose in this paper a **U**nified **P**hoto**A**coustic **M**icroscopy image reconstruction **N**etwork (UPAMNet) for two essential PAM image reconstruction tasks: super-resolution and denoising. The proposed UPAMNet is built on a CNN backbone with three additional attention blocks, which take full advantage of PAM image features from spatial attention, oriented attention, and positional attention. Spatial attention attempts to extract global features among different regions and channels, oriented attention aims to extract more discriminative local features from image edges and textures, and positional attention purposes to refine the element-wise value of reconstructed images with more fine details. To ensure the reconstruction quality, deep image priors are introduced in our network architecture to impose training constraints. More specifically, we propose a mixed training constraint by exploiting semantic features of the PAM images. The mixed training constraint includes both pixel-level and perception-level terms. Subsequently, to further improve the generalization ability of our method, we implement transfer learning to perform detailed cross-dataset evaluations and find that adaptation plays an important role in different datasets with significant domain gaps. During the inference stage, only low-quality PAM images are required and our method can generate super-resolved high-quality images. Experiments on our collected *in vivo* images and public datasets (i.e., leaf vein data, mouse cerebrovascular data, mouse ear vessels data) are performed, demonstrating the advantages of the method compared to the current state-of-the-art methods.

The main contributions of this paper include:

- An end-to-end deep network termed UPAMNet with built-in attention blocks is proposed by fully leveraging the PAM image characteristics for both image super-resolution and denoising.
- Both pixel-level and perception-level priors are incorporated by combining semantic image segmentation and deep image features to ensure high-fidelity image reconstruction.
- Few-shot and zero-shot transfer learning algorithms are implemented to ensure the generalization ability of the method for cross-dataset processing such that smaller training images are needed on the target domain, maximizing the potential clinical adoption of the method.

2. Methodology

2.1. Problem statement

Before introducing our detailed network architecture, we first define the problems and related notations. Let $\mathbf{I}_{\mathrm{i}} \in \mathcal{R}^{1\times H\times W}$, $\mathbf{I}_{\mathrm{o}} \in \mathcal{R}^{1\times H\times W}$, and $\mathbf{I}_{\mathrm{gt}} \in \mathcal{R}^{1\times H\times W}$ represent the input image, the output image, and the ground truth image, where H and W refer to, respectively, the height and width of the image. Although the original PAM signal is stored in 3D space, in this work we process it under the 2D Cartesian coordinate system via Maximum Amplitude Projection (MAP). The proposed UPAMNet $\mathcal{H}(\cdot)$ with parameters θ is to transform the images from low quality (low-resolution and noisy images) to high quality (high-resolution and clean images). The tasks can be defined as:

$$\mathcal{H}_{\theta^*} = \arg\min_{\theta} \mathcal{L}(\mathcal{H}(\mathbf{I}_{\mathrm{i}}, \mathcal{P}(\mathbf{I}_{\mathrm{i}})), \mathbf{I}_{\mathrm{gt}}) \tag{1}$$

where $\mathcal{L}(\cdot)$ represents the combined loss function and $\mathcal{P}(\cdot)$ represents the information priors.

PAM super-resolution refers to the recovery of fullsampled high-quality images from downsampled low-quality images. Given the ground truth high-quality images $\mathbf{I}_{\mathrm{gt}}$, we obtain the downsampled images $\mathbf{I}_{\mathrm{i}}^{\mathrm{LR}}$ as the input of the proposed model: $\mathbf{I}_{\mathrm{i}}^{\mathrm{LR}} = \downarrow_s \mathbf{I}_{\mathrm{gt}}$, where s represents the scale factor. The purpose of denoising is to learn the map function from the noisy images $\mathbf{I}_{\mathrm{i}}^{\mathrm{N}}$ to clean images $\mathbf{I}_{\mathrm{gt}}$. To form the image pairs for training the model, we manually add different types of noise to the original high-quality images: $\mathbf{I}_{\mathrm{i}}^{\mathrm{N}} = \mathbf{I}_{\mathrm{gt}} + \mathcal{N}$. Following [19], we model in this work the PAM image noise $\mathcal{N}$ as a combination of Gaussian noise, Poisson noise and Rayleigh noise [21,22].

2.2. UPAMNet for PAM image super-resolution and denoising

An overview of the proposed UPAMNet is shown in Fig. 1, which takes low-quality image $\mathbf{I}_{\mathrm{i}}$ as the input and obtains the reconstructed image $\mathbf{I}_{\mathrm{o}}$ as output. The backbone of UPAMNet includes three modules, i.e., a feature contraction module, a feature connection module, and a feature expansion module. These modules share the same basic residual convolution blocks (ResConv) as shown in the left bottom of Fig. 1. Each residual convolution block contains a combination of two group norm layers $(\cdot)^n$, two sigmoid layers $\sigma(\cdot)$, and two convolution layers $Conv(\cdot)$ with kernel size 3×3. With the input feature map noted as X_t, the output feature map X_{t+1} through each residual convolution block can be represented as:

$$Y_t = Conv(X_t^n \cdot \sigma(X_t^n)), X_{t+1} = X_t + Conv(Y_t^n \cdot \sigma(Y_t^n)) \tag{2}$$

The feature contraction module contains four sub-modules composed of two basic ResConv blocks, two attention blocks, and one downsample layer. After each downsample layer, the size of the feature map is reduced by half. This module extracts the shallow feature maps from the input image which preserves the low-level components. The feature connection module is followed by the feature contraction module. It is composed of two basic ResConv blocks and the size of the output feature map in this stage is the same as the input feature map. It extracts the deep feature from the latent feature map, which refers to the high-level components of the photoacoustic images. The feature expansion module is implemented to reconstruct high-quality images based on the feature maps from different feature extraction stages. It also consists of four sub-modules with a combination of basic ResConv blocks, attention blocks, and upsample layers. The feature maps from the shallow feature extraction are skipped and concatenated with the feature maps from the feature expansion stage, avoiding information loss from the early feature extraction stage. Finally, a convolution layer with kernel size 3×3 is used to recover the output feature map with the same shape as the input image.

In our proposed UPAMNet, different attention blocks are designed to improve the reconstruction performance. More specifically, three different types of self-attention mechanisms are used, i.e., spatial attention

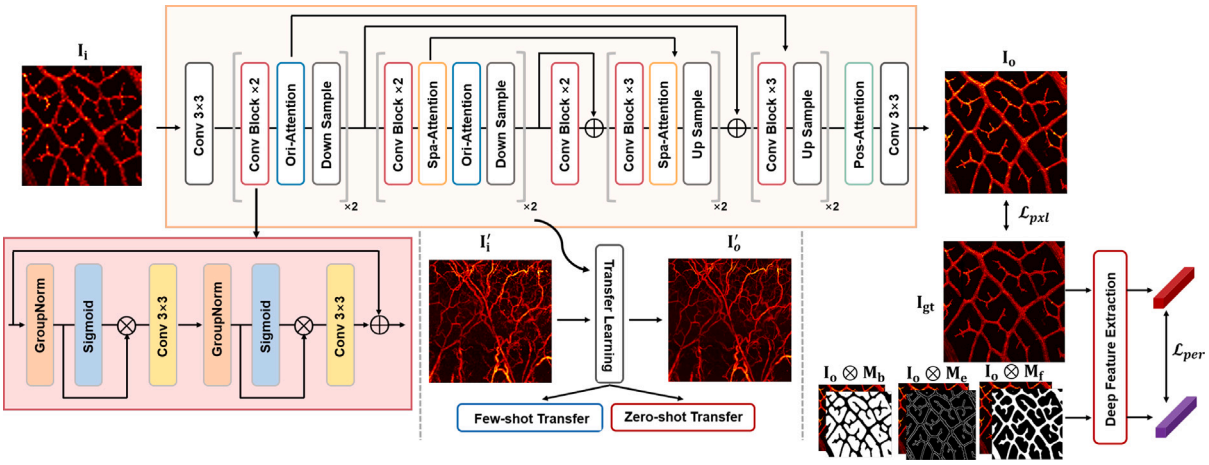


Fig. 1. Illustration of the network architecture of the proposed method. The upper part shows the key components of our UPAMNet. UPAMNet consists of three modules, i.e., the feature contraction module, the feature connection module, and the feature expansion module. Based on deep image priors, we design three attention blocks to improve the performance and exploit the semantic segmentation to propose a combined training constraint at both pixel and perception levels. To better evaluate the generalization ability of pre-trained models, we implement few-shot and zero-shot transfer learning to conduct experiments on unseen datasets. The detailed architecture of the ResConv block is shown at the left bottom.

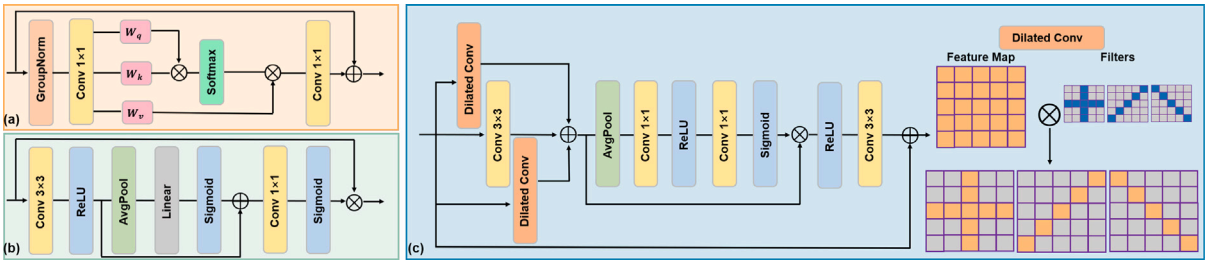


Fig. 2. Visualization of the detailed architecture of three proposed attention blocks. (a) The orange part represents the spatial attention block. (b) The green part represents the positional attention block. (c) The blue part represents the oriented attention block.

block (SPA), oriented attention block (ORA), and positional attention block (POA). The details of these attention blocks are shown in Fig. 2. The spatial attention block is designed to model the relationships among different regions and channels globally, which is a commonly adopted strategy in natural image reconstruction [23,24]. With this approach, it computes the compacted contextual attention maps from the input features and emphasizes the important regions with high self-similarity measures. Given the input feature map X, SPA projects it to three components $\{Q, K, V\} = Split(Conv(X^n))$ [25] and computes the self-attention to obtain the output feature Y:

$$Y = X + Conv(Softmax(Q \cdot K^T / \sqrt{d}) \cdot V) \tag{3}$$

The oriented attention block is proposed herein to extract more discriminative feature representations. Since PAM images have less texture than natural images, features extracted from different convolution perceptual fields can be used. Traditional convolution layers focus on local regions while dilated convolution layers are talented at modeling global correspondence. Based on [26], we design orientation-aware dilated convolution $dConv(\cdot)$ to extract oriented feature maps. Two types of orientation-aware dilated convolution kernels are defined on the right of the blue part in Fig. 2. One type is to extract vertical and horizontal features denoted as Ω_1 and the other is to extract diagonal features denoted as Ω_2. The weights not in the convolution region are set to zero and fixed during the network training. The oriented attention block computes feature maps from different orientations and concatenates them to perform channel attention:

$$F = Concat(Conv(X), dConv(X|\Omega_1), dConv(X|\Omega_2)) \tag{4}$$

In this work, the channel attention layer is composed of two convolution layers with kernel size 1×1, an adaptive average pooling layer $(\cdot)^p$, and an activation function $R(\cdot)$. The weight map is computed through a sigmoid function, and a convolution layer with kernel size 3×3 is introduced to transform the output feature map with the same shape as the input:

$$Y = X + Conv(R(F \cdot \sigma(Conv(R(Conv(F^p)))))) \tag{5}$$

The positional attention block is implemented as the last step of the feature expansion module to improve the performance of image reconstruction. The input feature map goes through a 3×3 convolution layer and an activation function, followed by a combination of an adaptive average pooling layer, a fully connected layer with weights W, and a sigmoid layer. The 1×1 convolution layer is introduced to refine the pixel-level feature map, which facilitates the restoration of more fine details. The weight map is calculated from the obtained pixel-level feature map and multiples the input feature map to get the final refined construction:

$$Y = X \cdot \sigma(Conv(F \cdot \sigma(W \cdot F^p))), F = R(Conv(X)) \tag{6}$$

2.3. Combined training constraint for improved performance

Traditional pixel-level constraints easily lead to over-smoothing for PAM image reconstruction since there are fewer textures and distinctive features for PAM images. Consequently, perceptional-level constraints are required to improve the visual quality of generated images. The proposed training constraint consists of two terms, one term is the pixel-level constraint $\mathcal{L}_{pxl}$, and the other is the perception-level constraint $\mathcal{L}_{per}$.

The pixel-level constraint refers to the well-known MSE error, which is one of the most commonly used training constraints for image

reconstruction. It is achieved by measuring the L2 distance between the ground truth images and the reconstructed images:

$$\mathcal{L}_{pxl} = \frac{1}{N}\sum_{k=1}^{N}\sum_{j=1}^{W}\sum_{i=1}^{H}\|I_{gt}^{k}(i,j) - I_{o}^{k}(i,j)\|_2^2 \tag{7}$$

However, models trained only on pixel-level constraint tend to be over-smoothing and lack of detailed textures, especially for PAM image reconstruction.

To improve the perceptual similarity between the ground truth image and the generated image, we propose an additional perception-level constraint by leveraging the semantic priors of PAM images. Although existing works [14,19] have tried to use the perception-level loss to train the model based on the pre-trained VGG model, they extract the perception feature maps from the same layer and distribute the same weights for the entire image. It can be found that the generated images are more likely to have undesirable noise and textures in the background area. As we discuss before, PAM images have several distinct characteristics compared to natural images, making it easy to determine the foreground and the background in PAM images, while it is hard to perform that for natural images since they are more diverse and complex. Consequently, it is feasible for PAM images to be segmented into the foreground and background, which provides opportunities to distribute different weights to different semantic parts of images.

Given the ground truth image $\mathbf{I_{gt}}$, we implement segmentation algorithms to obtain several binary images, i.e., image foreground $\mathbf{M_f}$, image background $\mathbf{M_b}$ and image edges $\mathbf{M_e}$, where 1 represents the pixels that fall into the segmentation and 0 represents not. We compute the perception-level constraint for each region with different layers from the pre-trained VGG model. Previous studies [27–29] have found that the early stage of a CNN model retains the low-level features, such as the high-frequency component like edges and boundaries, while the complex and compacted semantic features are restored by the deep layers of the CNN model. Consequently, low-level feature extracted by the early stage of VGG model is utilized for edges, mid-level feature extracted by the intermediate stage is used for image background, and high-level semantic object segmentation is reconstructed by deep layers.

In this work, we implement patch-loss to compute the similarity between the ground truth and the reconstructed perception features since it is more reasonable to force the highly compacted deep features to be region-invariant instead of position-invariant. The patch similarity constraint function $\mathcal{P}(\cdot)$ over feature maps X and Y is defined as :

$$\mathcal{P}(X,Y) = 1 - \frac{1}{N}\sum_{k=1}^{N}\frac{\sum_{\Omega_k} X(i,j)\cdot Y(i,j)}{\sqrt{\sum_{\Omega_k} X^2(i,j)}\sqrt{\sum_{\Omega_k} Y^2(i,j)}} \tag{8}$$

where $(i,j) \in \Omega_k$ refers to the kth patch and N is the number of patches. Compared to the conventional L2 distance, the patch similarity function requires similar arrangements of pixels in fixed regions instead of single-pixel locations.

The perception-level constraint can be written as:

$$\mathcal{L}_{per}^{b} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{P}(\mathcal{V}^{b}(I_{o}^{k}\cdot M_{b}), \mathcal{V}^{b}(I_{gt}^{k}\cdot M_{b})) \tag{9}$$

$$\mathcal{L}_{per}^{f} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{P}(\mathcal{V}^{f}(I_{o}^{k}\cdot M_{f}), \mathcal{V}^{f}(I_{gt}^{k}\cdot M_{f})) \tag{10}$$

$$\mathcal{L}_{per}^{e} = \frac{1}{N}\sum_{k=1}^{N}\mathcal{P}(\mathcal{V}^{e}(I_{o}^{k}\cdot M_{e}), \mathcal{V}^{e}(I_{gt}^{k}\cdot M_{e})) \tag{11}$$

where $\mathcal{V}^{b}(\cdot)$, $\mathcal{V}^{f}(\cdot)$ and $\mathcal{V}^{e}(\cdot)$ represent different layers of pre-trained VGG network. Finally, the combined training constraint is the weighted sum of pixel-level $\mathcal{L}_{pxl}$ and perception-level $\mathcal{L}_{per}$ constraints with different weights ω:

$$\mathcal{L} = \omega_{per}(\omega_b\mathcal{L}_{per}^{b} + \omega_f\mathcal{L}_{per}^{f} + \omega_e\mathcal{L}_{per}^{e}) + \omega_{pxl}\mathcal{L}_{pxl} \tag{12}$$

2.4. Cross dataset transfer learning for better generalization

Another significant challenge in PAM image reconstruction is the heterogeneity of data due to different imaging hardware settings. To improve the generalization ability of our method, we implement transfer learning algorithms and evaluate the cross-dataset performance of different methods. More specifically, both few-shot transfer learning and zero-shot transfer learning are considered in this paper. In the following, the source domain or dataset is noted as D_s and the target domain or dataset is noted as D_t, and η refers to the transfer rate which represents the number of training images for the few-shot learning.

few-shot transfer learning refers to fine-tuning the pre-trained model with a small number of samples from the target domain D_d. In this work, to evaluate the generation ability of our pre-trained model, we perform few-shot transfer learning with only $\eta \in \{5\%, 10\%, 20\%\}$ target data. We freeze the feature contraction module and feature connection module and only finetune the parameters of the feature expansion module. For datasets with large domain shifts, few-shot transfer learning helps the pre-trained network rapidly get self-adapted on the target domain.

In practice, zero-shot transfer learning is much more challenging and commonly used for real-world scenarios since there is no available training data to assist model training in most cases for PAM image reconstruction. Only low-quality images for testing are provided for pre-trained models to perform transfer learning. In the following, we describe zero-shot transfer learning algorithms for PAM super-resolution and denoising tasks separately.

2.4.1. Zero-shot learning for super-resolution

In conventional supervised PAM image super-resolution, we have the paired images $\{\mathbf{I}_{\mathrm{i}}^{\mathrm{LR}}, \mathbf{I}_{\mathrm{gt}}\}$. However, for zero-shot transfer learning, we only have the downsampled or test images $\mathbf{I}_{\mathrm{i}}^{\mathrm{LR}}$. To form similar image pairs, the test images can be regarded as ground truth images from their coarser resolutions. It means that by downsampling the test images, we can regard the downsampled images as input and the test images with original resolution as ground truth: $\tilde{\mathbf{I}}_{\mathrm{i}}^{\mathrm{LR}} = \downarrow_s \mathbf{I}_{\mathrm{i}}^{\mathrm{LR}}$. With the paired images $\{\tilde{\mathbf{I}}_{\mathrm{i}}^{\mathrm{LR}}, \mathbf{I}_{\mathrm{i}}^{\mathrm{LR}}\}$, we optimize the pre-trained model:

$$\mathcal{H}_{\theta^*}^{D_t} = \arg\min_{\theta}\mathcal{L}(\mathcal{H}(\tilde{\mathbf{I}}_{\mathrm{i}}^{\mathrm{LR}}, \mathcal{P}(\tilde{\mathbf{I}}_{\mathrm{i}}^{\mathrm{LR}})), \mathbf{I}_{\mathrm{i}}^{\mathrm{LR}}) \tag{13}$$

Since we have frozen the early feature extraction stages, our model does not lose the ability to represent the texture and the edges of the new images mentioned above. Via transfer learning on target datasets, our model not only retains the reconstruction ability on large source datasets but also is able to adaptively achieve better performance on new images.

2.4.2. Zero-shot learning for denoising

Similar to zero-shot transfer learning for super-resolution, only noisy input $\mathbf{I}_{\mathrm{i}}^{\mathrm{N}}$ is accessible. Based on [30], the noisy image can be decomposed by the estimated clean image and the additional noise: $\mathbf{I}_{\mathrm{i}}^{\mathrm{N}} = \tilde{\mathbf{I}}_{\mathrm{c}} + \mathcal{N} = \mathcal{H}(\mathbf{I}_{i}^{N}, \mathcal{P}(\mathbf{I}_{i}^{N})) + \mathcal{N}$. During the transfer learning stage, we force the network to be consistent for pure noise and clean images. It means that when the estimated clean image is fed to the network, the output is supposed to be the same clean image, and when the pure noise is fed to the network, the output is consistent to be the zero image:

$$\mathcal{L}_{zs}^{1} = \mathcal{L}(\mathcal{H}(\tilde{\mathbf{I}}_{\mathrm{c}}, \mathcal{P}(\tilde{\mathbf{I}}_{\mathrm{c}})), \tilde{\mathbf{I}}_{\mathrm{c}}), \mathcal{L}_{zs}^{2} = \mathcal{L}(\mathcal{H}(\mathcal{N}, \mathcal{P}(\mathcal{N})), \mathbf{0}) \tag{14}$$

The final zero-shot transfer learning constraint is the combination of these two terms:

$$\mathcal{H}_{\theta^*}^{D_t} = \arg\min_{\theta}(\mathcal{L}_{zs}^{1} + \mathcal{L}_{zs}^{2}) \tag{15}$$

3. Experiments

3.1. Datasets

To evaluate the performance of our proposed method, we conduct experiments on three different datasets and on our collected *in vivo* images. The images from different datasets are distinctive both in structure and texture, which provides more challenges for supervised learning and transfer learning tasks. The visualization of images from these datasets is shown in supplementary materials.

3.1.1. Leaf vein data (D-I)

The leaf vein data is described in [14]. The original size of the 3D data is $256 \times 256 \times 180$ and we apply MAP to acquire 2D images. There are 187 images in the training dataset, 54 images in the validation dataset, and 27 images in the test dataset. This dataset is used for supervised PAM super-resolution and denoising and serves as the source dataset for transfer learning as well.

3.1.2. Mouse cerebrovascular data (D-II)

The mouse cerebrovascular vessel data is described in [6]. There are 337 images in the training dataset and 38 images in the test dataset. Since the original size of the images is too large (more than 800×800), we cut the images into patches for training and testing. This dataset is used for supervised PAM super-resolution and denoising, as well as for being served as the target dataset for transfer learning.

3.1.3. Mouse ear vessels data (D-III)

Part of the mouse ear vessels data is described in our previous work [19] and we also collect new PAM images for better evaluation. There are 28 images in total with image size 512×512. Since the number of this dataset is relatively small, we regard this dataset only as the target dataset for transfer learning.

3.1.4. In vivo mouse vessels data (D-IV)

The PAM system used for collecting new images is described in our previous work [19] and we collect mouse brain cerebrovascular images and mouse ear vessels images for better evaluation of the generalization ability of our proposed method.[1]

3.2. Implementation details

3.2.1. Data preparation

For PAM image super-resolution, we downsample the original image to form the image pairs. In this work, we exploit the sampling ratios $\times 2$ and $\times 4$, which refer to 1/4 and 1/16 pixels for the low-sampling data, respectively. For PAM image denoising, we treat the noise as a combination of Gaussian noise, Poisson noise and Rayleigh noise. The Gaussian noise is sampled from the zero-mean distribution, and the standard deviation is randomly selected from 0.01 to 0.04 for 3D data. To obtain the semantic segmentations of the images, we perform edge detection via the morphological operations and compute dilation to close the hole and generate the foreground segmentation as well as the complementary background segmentation. Note that the segmentation maps are only required during the training stages and are not required in the inference time. We also generate patches with size 128×128 for the training stage and perform image augmentation including vertical or horizontal flips and 90 degrees rotation to make the training stage robust and avoid overfitting. Images are normalized to $[0, 1]$ by dividing the maximum value.

[1] All experimental animal procedures are carried out in conformity with the laboratory animal protocol approved on March 17, 2023 by Institutional Animal Care and Use Committee (IACUC) at the Shanghai Jiao Tong University, Shanghai, China. The protocol number is A2023032.

3.2.2. Training details

The proposed method is implemented by PyTorch and is end-to-end trained on an NVIDIA TITAN RTX 3090 GPU with 24 GB memory. All the networks are optimized by the Adam optimizer. During the supervised training stage, we set the learning rate as 5e−4 and set $\omega_b = 0.1$, $\omega_f = 1.0$, $\omega_e = 2.0$, $\omega_{per} = 2.0$ and $\omega_{pxl} = 1.0$. During the few-shot training stage, we set the learning rate as 1e−4 and use 10% images for finetuning. During the zero-shot transfer learning stage, we optimize the network with a small learning rate 1e−5 and set $\omega_b = 0.01$, $\omega_f = 0.5$, $\omega_e = 1.0$, $\omega_{per} = 2.0$ and $\omega_{pxl} = 1.0$. The weights mentioned above are derived from parameter ablation studies and we choose the weights for the best performance. The codes of this work will be available at https://github.com/Lrnyux/UPAMNet.

3.3. Evaluation metrics

We evaluate the generated images from both pixel level and perception level. Four evaluation metrics are implemented. Root Mean Square Error (RMSE↓) calculates the square root of the second sample moment between the predicted image and the ground truth. Peak Signal to Noise Ratio (PSNR↑) calculates the log ratio between the maximum value and the mean distance error between two images. Structural Similarity Index Measure (SSIM↑) calculates the image similarity between two given images based on the statistical distribution of the images. These three evaluation metrics are used as quantitative measurements for evaluating the difference between the reconstruction and ground truth in pixel space. In addition, we implement Learned Perceptual Image Patch Similarity (LPIPS↓) to evaluate the perceptual similarity [31]. LPIPS uses off-the-shelf deep classification networks to calculate the distance of deep features.

3.4. Comparison methods

For detailed performance assessment, we compare existing methods developed for each task separately. To ensure a fair comparison, we retrain the existing methods with these three datasets and optimize the parameters based on experimental results.

For the task of super-resolution, we compare our method with three existing deep learning methods for PAM super-resolution [6,14,15]. We also compare our method with the method proposed for natural image super-resolution [32]. The comparison experiments are conducted on datasets D-I and D-II for supervised learning. Meanwhile, the transfer learning strategies are implemented on datasets D-II and D-III while the source dataset is dataset D-I. For the task of denoising, we compare our method with four methods, including one deep learning method designed for PAM image denoising [19] and two methods for natural image denoising [33,34]. Similar to the task of super-resolution, we conduct experiments via supervised learning on datasets D-I and D-II, and via the transfer learning from D-I to D-II and D-III.

3.5. Ablation study

We also conduct ablation studies of the proposed method to demonstrate the effectiveness of different network architectures and training strategies. We remove the attention blocks to evaluate the effectiveness of implemented attention mechanisms noted as w/o attn. Similarly, we remove the semantic image priors and constraint the whole image in perceptual space noted as w/o prior. For this ablated model, we set ω_b and ω_e to zero and the mask $\mathbf{M}_b$ to an identity matrix. We evaluate the balance between pixel-level and perception-level constraints by changing the weight coefficients between $\mathcal{L}_{pxl}$ and $\mathcal{L}_{per}$. We also evaluate the influence of the size of the training dataset η for few-shot transfer learning.

Table 1
Results of PAM super-resolution via supervised learning.

Dataset	Approaches	×2				×4			
		RMSE↓	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
D-I	SEResnet [14]	10.2456	27.7188	0.8769	0.0250	15.4909	24.1724	0.7632	0.0813
	FDUnet [6]	9.5322	27.2466	0.8814	0.0513	**12.7731**	23.5129	0.7689	0.1611
	DIP [15]	13.0587	25.8240	0.7808	0.1702	16.9233	22.8117	0.6866	0.3356
	CAT [32]	9.7763	27.8657	0.9058	0.0259	15.1487	23.9712	0.7544	0.0917
	Ours w/o attn	9.6244	28.0994	0.9082	0.0270	14.5231	24.0737	0.7660	0.0801
	Ours w/o prior	10.1597	27.5898	0.8896	0.0331	13.8994	24.2969	0.7637	0.0800
	Ours	**9.3946**	**28.3056**	**0.9115**	**0.0249**	13.9266	**24.6112**	**0.7921**	**0.0720**
D-II	SEResnet [14]	2.9200	37.0301	0.9731	0.0166	6.1051	31.1535	0.8766	0.0721
	FDUnet [6]	2.5719	37.5262	**0.9801**	0.0219	**6.0725**	31.2851	0.8378	0.0904
	DIP [15]	6.0387	31.1351	0.8990	0.0974	9.6668	26.9683	0.8318	0.1516
	CAT [32]	2.7415	37.5920	0.9767	0.0144	6.3434	30.2522	0.8996	0.0771
	Ours w/o attn	2.7594	37.2456	0.9748	0.0146	6.2939	30.8722	0.8991	0.0691
	Ours w/o prior	3.1323	35.8150	0.9684	0.0138	6.5032	30.4274	0.8964	0.0536
	Ours	**2.5568**	**38.5378**	0.9796	**0.0115**	6.2082	**31.3645**	**0.9048**	**0.0526**

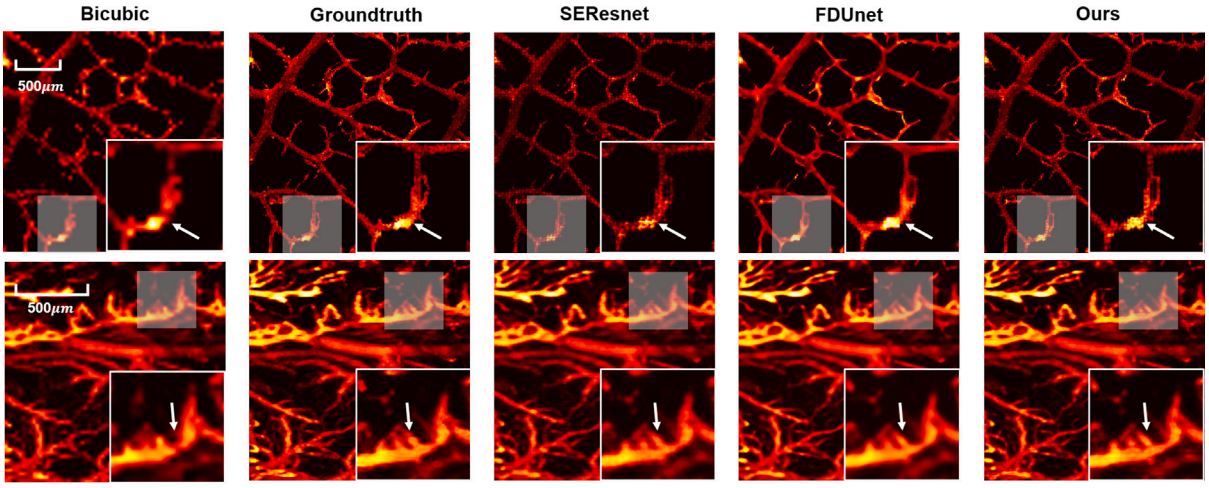


Fig. 3. Visualization results of super-resolution (×4) via supervised learning. The first row refers to the representative results in dataset D-I and the second row refers to that in dataset D-II. We show the zoomed images of the white shaded area in the lower right part.

4. Results and analysis

In subsequent sections, the **bold** and underline values in the table indicate the best and the second-best performance in each task, respectively.

4.1. PAM super-resolution via supervised learning

The results of different methods on PAM image super-resolution via supervised learning are represented in Table 1. We train the models in a supervised manner on D-I and D-II. The upper part of Table 1 shows the results on dataset D-I. For ×2 image reconstruction, our method achieves the best on four metrics, and for ×4 case our method achieves the best on three metrics. The lower part of Table 1 shows the results on dataset D-II. For the ×2 case, our method achieves the best on three metrics and the second best on the other one metric, and for the ×4 case, our method achieves the best on three metrics. It can be seen that our method significantly outperforms the existing methods for PAM super-resolution with different sampling ratios on different datasets. In Fig. 3, we also provide the corresponding visualization results (×4) of different methods for super-resolution. From the zoomed images in the lower right part, it can be found that our proposed method reconstructs the detailed structure best compared with existing methods and avoid over-smoothing for super-resolution.

4.2. PAM super-resolution via transfer learning

The results of transfer learning algorithms on PAM super-resolution are shown in Table 2. We report the performance of few-shot transfer learning and zero-shot transfer learning compared with existing methods [6,14]. The first part of Table 2 shows the results via transfer learning from dataset D-I to dataset D-II. Compared with the two existing methods, our method with few-shot transfer learning achieves the best on three metrics for ×2 case. With zero-shot transfer learning our method achieves the best performance on three metrics for ×2 case and on all metrics for ×4 case, which significantly outperforms the other two methods. The second part of Table 2 shows the results via zero-shot transfer learning from dataset D-I to dataset D-III. The zero-shot transfer learning improves the performance of three different methods and among them, our method achieves the best performance on three metrics for ×2 case and three metrics for ×4 case. The qualitative results of different types of transfer learning are shown in Fig. 4. It can be found that both few-shot and zero-shot transfer learning algorithms improve the visual quality of reconstructed images a lot compared with low-resolution images. We also provide the results without performing transfer learning in supplementary materials.

4.3. PAM denoising via supervised learning

The results of different methods on PAM denoising via supervised learning are listed in Table 3. We conduct experiments on the dataset D-I and D-II. It can be found that for these two datasets, our proposed method achieves superior performance on three reported metrics,

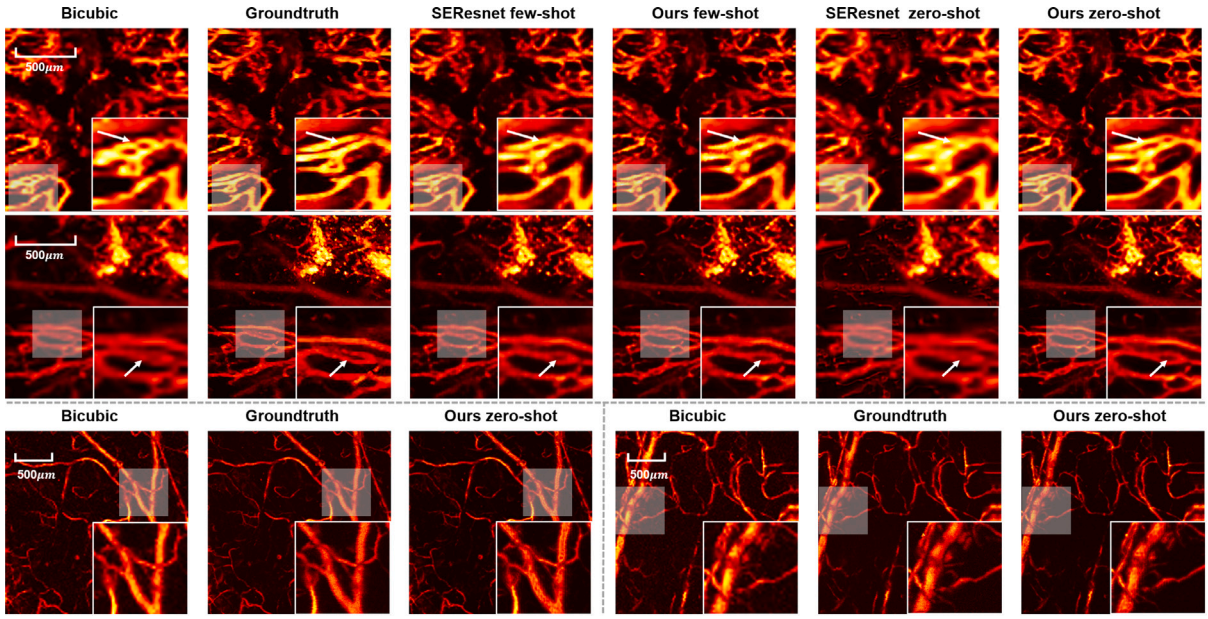


Fig. 4. Visualization results of super-resolution (×4) via transfer learning. The first and the second rows refer to the representative results in dataset D-II and we show the results for the cases of few-shot transfer learning and zero-shot transfer learning of two methods. The third row refers to the results in dataset D-III and we show the case of zero-shot transfer learning of our method. We show the zoomed images of the white shaded area in the lower right part.

Table 2
Results of PAM super-resolution via transfer learning.

Dataset	Type	Approaches	×2				×4			
			RMSE↓	PSNR↑	SSIM↑	LPIPS↓	RMSE↓	PSNR↑	SSIM↑	LPIPS↓
D-I ↓ D-II	Few-shot ($\eta = 10\%$)	SEResnet [14]	3.1241	36.3303	0.9682	**0.0170**	6.5455	29.6784	**0.8914**	**0.0785**
		FDUnet [6]	4.2557	34.4619	0.7651	0.0250	6.5270	**30.2390**	0.8360	0.0959
		Ours	**2.9756**	**36.3317**	**0.9719**	0.0241	**6.4024**	30.1091	0.8687	0.0817
	Zero-shot	SEResnet [14]	3.1767	36.1877	0.9223	0.0317	9.3548	27.0769	0.7397	0.2132
		FDUnet [6]	3.9332	33.1554	0.9504	**0.0278**	8.6538	26.9877	0.7105	0.0933
		Ours	**3.0040**	**36.3106**	**0.9710**	0.0283	**7.0240**	**29.3718**	**0.8757**	**0.0891**
D-I ↓ D-III	w/o transfer	SEResnet [14]	4.4523	32.4405	0.8540	**0.0687**	**5.5903**	**30.0697**	0.7657	0.1612
		FDUnet [6]	5.1118	30.4644	0.8408	0.0828	9.6649	25.2142	0.5757	0.3051
		Ours	**4.3473**	**32.6127**	**0.8683**	0.0798	5.9410	29.4831	**0.7714**	**0.1254**
	Zero-shot	SEResnet [14]	3.7470	31.6709	0.8726	**0.0940**	6.0690	29.3001	**0.7950**	0.1766
		FDUnet [6]	4.3773	30.9451	0.8368	0.1028	8.7595	27.2327	0.6273	0.1888
		Ours	**3.7011**	**32.7212**	**0.8883**	0.1023	**5.3650**	**29.7472**	0.7905	**0.1299**

which outperforms the existing methods proposed for both PAM and natural image denoising. The qualitative results of PAM denoising are shown in Fig. 5. It can also be found that our method attempts to recover more detailed image textures compared to existing methods under different noise levels. For the area severely degraded by noise, our method attempts to denoise the local region based on the global texture features of surrounding areas.

4.4. Results on PAM denoising via transfer learning

The results of transfer learning on PAM denoising are reported in Table 4 with dataset D-I as the source dataset. Since the domain gap between the dataset D-I and D-II is too large, we find that zero-shot transfer fails to work in this case and therefore we only implement few-shot transfer learning with 10% new training images to evaluate the performance. The upper part of Table 4 shows the results on the target domain of D-II via few-shot transfer learning. It can be found that there is a marked improvement for transfer learning with only a small number of new training images. The lower part of Table 4 shows the results of the target dataset D-III via zero-shot transfer learning. It can be found that our method with zero-shot learning achieves the best performance on three metrics, and the improvements before and after zero-shot learning for our method are the largest which demonstrates the effectiveness of the zero-shot denoising transfer algorithms. Qualitative results of the transfer learning are shown in Fig. 6.

Table 3
Results of PAM denoising via supervised learning.

Dataset	Approaches	PSNR↑	SSIM↑	LPIPS↓
D-I	Noisy input	18.0911	0.3012	0.2101
	DnGAN [19]	28.8881	0.7778	0.0449
	DnCNN [33]	31.3200	0.7771	0.0535
	CBDNet [34]	31.6526	0.7869	0.0458
	Ours w/o attn	31.5690	0.7900	0.0382
	Ours w/o prior	31.4929	0.8027	**0.0367**
	Ours	**31.8866**	**0.8167**	0.0377
D-II	Noisy input	18.0814	0.3221	0.6375
	DnGAN [19]	24.9031	0.7225	0.0991
	DnCNN [33]	29.2022	0.7768	0.0827
	CBDNet [34]	29.7377	0.7419	0.0641
	Ours w/o attn	28.1147	0.6352	0.0724
	Ours w/o prior	29.6356	0.7211	**0.0497**
	Ours	**30.1451**	**0.7941**	0.0654

We also conduct statistical analysis for our reported results in Tables 1–4 and visualize the statistical results in Fig. 7, which gives

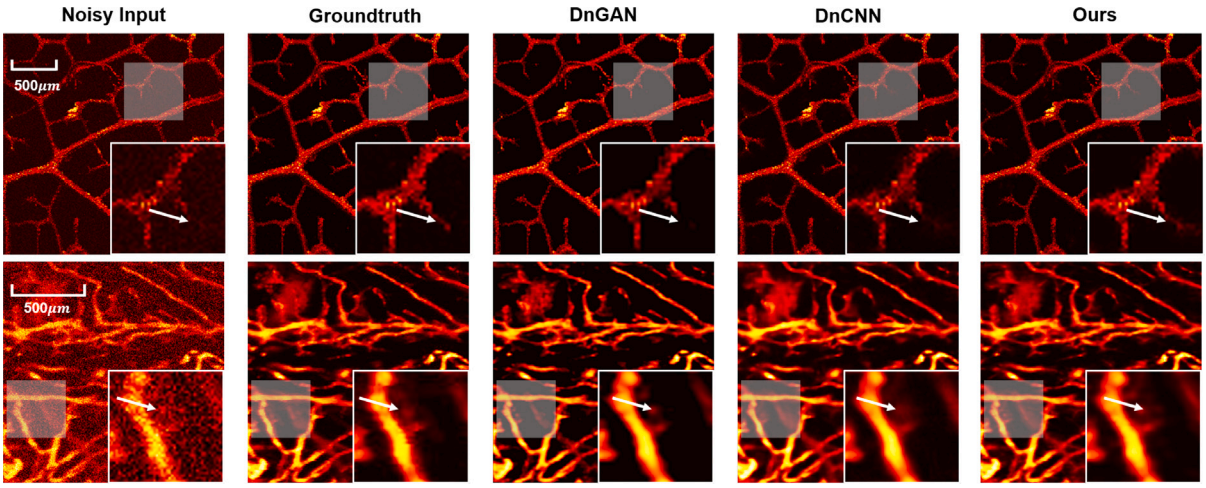


Fig. 5. Visualization results of denoising via supervised learning. The first row refers to the representative results in dataset D-I and the second row refers to that in dataset D-II. We show the zoomed images of the white shaded area in the lower right part.

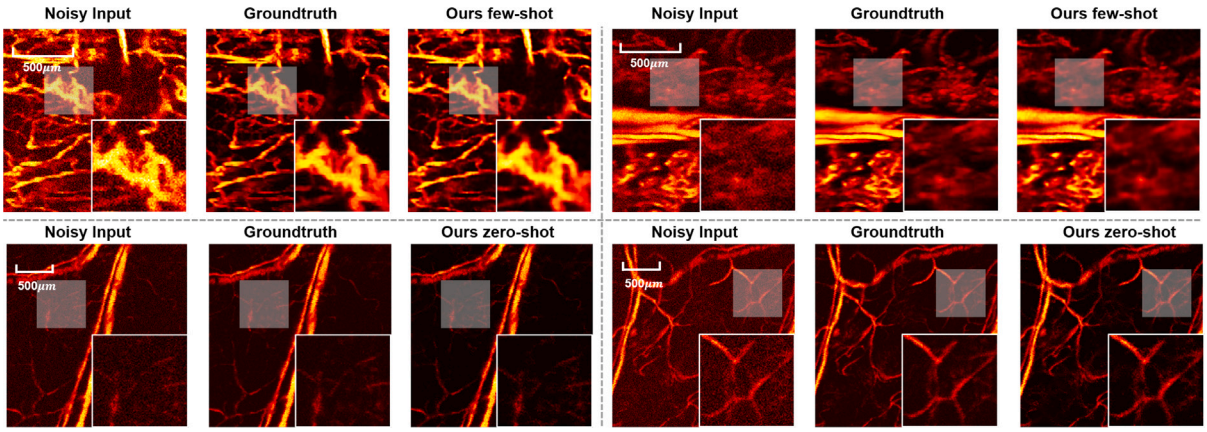


Fig. 6. Visualization results of denoising via transfer learning. The first row refers to the representative results in dataset D-II for few-shot transfer learning by our method and the second row refers to the results in dataset D-III for zero-shot transfer learning by our method. We show the zoomed images of the white shaded area in the lower right part.

Table 4
Results of PAM denoising via transfer learning.

Dataset	Type	Approaches	PSNR↑	SSIM↑	LPIPS↓
D-I ↓ D-II	Few-shot (η = 10%)	DnGAN [19]	22.5934	0.6666	0.1070
		DnCNN [33]	25.2376	0.6337	0.2347
		CBDNet [34]	27.1534	0.6741	0.1329
		Ours	**28.3624**	**0.7317**	**0.0950**
D-I ↓ D-III	w/o transfer	DnCNN [33]	24.5111	0.3703	0.3972
		CBDNet [34]	24.2294	0.3470	0.4132
		Ours	**26.8752**	**0.5181**	**0.2157**
	Zero-shot	DnCNN [33]	23.5357	0.3615	0.4044
		CBDNet [34]	24.5313	0.3593	0.4037
		Ours	**28.4220**	**0.5162**	**0.1413**

a more intuitive representation to demonstrate the best performance of our method. With the statistical tests, it can be found that our proposed method has significant improvements over existing methods.

4.5. Results on ablation study

The results of ablation studies are reported in Table 5. We remove part of the proposed modules and change the hyperparameters in our experiments to show the different performances on two tasks. We perform the super-resolution on dataset D-I on ×4 case and the denoising on dataset D-II and the few-shot transfer learning from D-I to D-II.

Do the different attention blocks work? We conduct four different experiments to prove the effectiveness of the proposed attention blocks. We remove all three attention blocks and report the results in Tables 1 and 3, noted as w/o attn, and give more detailed results in Table 5-(A) by removing one by one noted as w/o SPA, w/o ORA and w/o POA. It can be found that the model without three blocks achieves the worst performance on four metrics. The spatial attention block plays the most important role in deep feature extraction since this block is designed to model the global relationships among different regions and channels. After removing the SPA, the performance decreases the most compared to the other two blocks. The oriented attention block attempts to extract more discriminative feature representations since the PAM images have less texture than natural images. The dilated convolution layers are talented at modeling global correspondence. Two different dilated convolution kernels can be used to extract features from different directions in feature space. After removing the ORA, the performance also decreases significantly but is better than only removing SPA, which demonstrates the ORA plays the second important role in our method. The positional attention block is designed as the last step of the feature expansion module to refine the pixel-level feature map, which facilitates more image details and contributes to improving the visual quality. It can be found that with POA block, LPIPS decreases a lot which demonstrates that POA mainly plays the role of improving the visual quality. The attention modules enable the

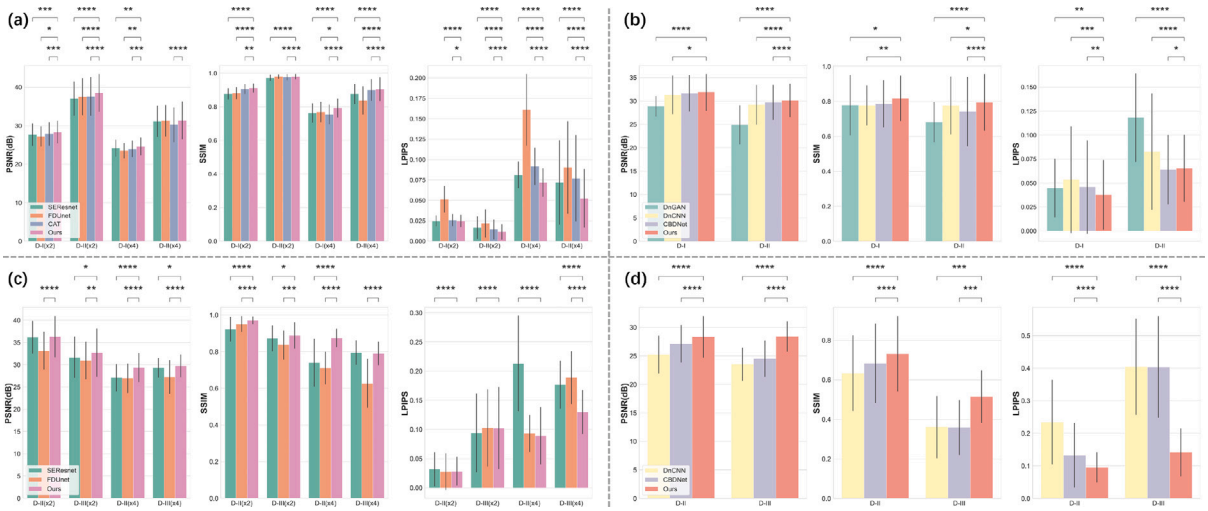


Fig. 7. Statistical analysis of quantitative results of our proposed method compared with existing methods. The left part shows the results for image super-resolution and the right part shows the corresponding results for image denoising. The first row refers to supervised learning reported in (a)-Table 1 and (b)-Table 3. The second row refers to transfer learning reported in (c)-Table 2 and (d)-Table 4. P-values are calculated based on the comparison with our results. The symbols ****, ***, ** and * represent p-values ≤ 0.0001, ≤ 0.001, ≤ 0.01, ≤ 0.05, respectively.

network to capture more efficient image features with a limited number of model parameters, which significantly improves the performance of PAM image reconstruction.

Do the image priors of the semantic segmentations work? In Tables 1 and 3, we remove the semantic segmentation masks and report the results noted as w/o prior. It can be found that our method outperforms the ablated model in all experiments on two datasets and on two tasks. Compared with constraining the entire image in latent space, we distribute different weights and layers for regions with different semantic meanings and it forces the model to be aware of more semantic features in latent space. Meanwhile, considering the computation efficiency, we choose to compute the segmentation masks in latent space instead of in image space since the latent feature map in $\mathcal{L}_{per}$ is much smaller than the reconstructed image in $\mathcal{L}_{pxl}$, and the feature map in latent space represents more detailed texture and edges, which can improve the perception-level quality.

What is the influence of pixel-level and perception-level constraints? We have mentioned that only pixel-level constraints will lead to over-smoothing and we conduct several ablated studies to examine the balance of these two constraints by changing the value of ω_{per} from 0.0 to 5.0 and the corresponding results are shown in Table 5-(B). It can be found that RMSE decreases as the coefficient ω_{per} decreases, which means that the pixel-level constraints play a more important role, while LPIPS increases which indicates that the perception-level quality is decreased and over-smoothing is severe. Although our method with $\omega_{per} = 2.0$ achieves the worst RMSE in image space, the other three metrics are the best demonstrating that our proposed combination loss function achieves the best balance between the pixel level and the perception level.

What is the influence of the number of images used for few-shot transfer learning? In Table 5-(C), we report the results on different numbers of images used for few-shot transfer learning on two tasks. It can be found that with a small number of training images on the target domain, our method can achieve good performance compared with a full-supervised training case, which demonstrates the generalization ability of our proposed network.

What is the influence of different noise levels for denoising? To evaluate the performance of our proposed method on the image denoising task, we conduct experiments to obtain real noisy PAM images and corresponding clean images to demonstrate the influence of different noise levels. We use low excitation light energy to collect noisy images and high excitation light energy to collect clean images, respectively. The visualization results are shown in Fig. 8. We also calculate the non-reference metrics of signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) and report values in the lower left part of the images. The upper part refers to the images collected from the leaf vein samples. We use 120 nJ excitation light energy to collect the ground truth clean images and use 10 nJ and 20 nJ low excitation light energy to collect the noisy images. It can be found that our method achieves good performance compared to the collected ground truth images under high excitation light energy. Especially for the 10 nJ case, the collected image is severely degraded by noise and our method attempts to recover most of the original image. The lower part refers to the images collected from the *in vivo* mouse ear blood vessels. Similarly, we use 400 nJ excitation light energy to collect the clean images and 80 nJ and 100 nJ to collect noisy images. It can also be found that our method generalizes well to denoising the real-world PAM images.

4.6. Results of in vivo images for super-resolution and denoising

Several experiments have been conducted on three datasets to demonstrate that our method can achieve the best performance on PAM image super-resolution and denoising separately compared with existing methods. To prove that our proposed method can generalize well to real-world scenarios, we conduct additional experiments on D-III and D-IV to evaluate our method on the combination of super-resolution(×4) and denoising which means that we collect low-resolution and noisy images as input to our network and attempt to reconstruct the high-resolution and clean images. Different from evaluating the performance on super-resolution and denoising separately, recovering the high-resolution and clean images from both low-resolution and noisy images simultaneously is much more challenging since the process of down-sampling generates more undesirable noise and artifacts compared with traditional noise like Gaussian noise. We collect *in vivo* PAM images of blood vessels of the mouse ear and mouse brain for qualitative evaluation. Similarly, we use low excitation energy to collect noisy images and high excitation energy to collect ground truth clean images.

The results are shown in Fig. 9 which includes four examples. The first and second examples (a,b) are images of mouse ear blood vessels. We use 80 nJ excitation light energy to collect the noisy and low-resolution images shown in (a,b)-1 and use 320 nJ excitation light energy to collect the ground truth clean images shown in (a,b)-3. The reconstruction results of our method are shown in (a,b)-2. The other two examples (c,d) refer to mouse cerebrovascular images.

Table 5
Ablation results of PAM super-resolution and denoising.

Approaches		Super-resolution (×4)				Denoising		
		RMSE↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
A	w/o SPA	14.6515	24.1933	0.7656	0.0795	28.8993	0.6897	0.0690
	w/o ORA	14.4265	24.2273	0.7658	0.0778	29.7473	0.7783	0.0742
	w/o POA	14.1362	24.2448	0.7805	0.0778	29.7321	0.7946	0.0678
	w/o attn	14.5231	24.0737	0.7660	0.0801	28.1147	0.6352	0.0724
B	$\omega_{per} = 5.0$	14.6237	24.0915	0.7818	0.0769	28.8327	0.7855	0.0611
	$\omega_{per} = 2.0$	13.9266	24.6112	0.7921	0.0720	30.1451	0.7941	0.0654
	$\omega_{per} = 1.0$	13.7859	24.1631	0.7898	0.0717	29.6377	0.7693	0.0703
	$\omega_{per} = 0.1$	12.5720	24.1334	0.8023	0.0819	28.4275	0.6784	0.0930
	$\omega_{per} = 0.0$	11.4327	23.8255	0.8158	0.2275	28.0750	0.6339	0.0994
C	$\eta = 20\%$	6.3868	29.9863	0.8723	0.0826	28.7635	0.7398	0.0846
	$\eta = 10\%$	6.4024	30.1091	0.8687	0.0817	28.3624	0.7317	0.0950
	$\eta = 5\%$	6.8128	28.8170	0.8882	0.0820	26.2563	0.7303	0.1115

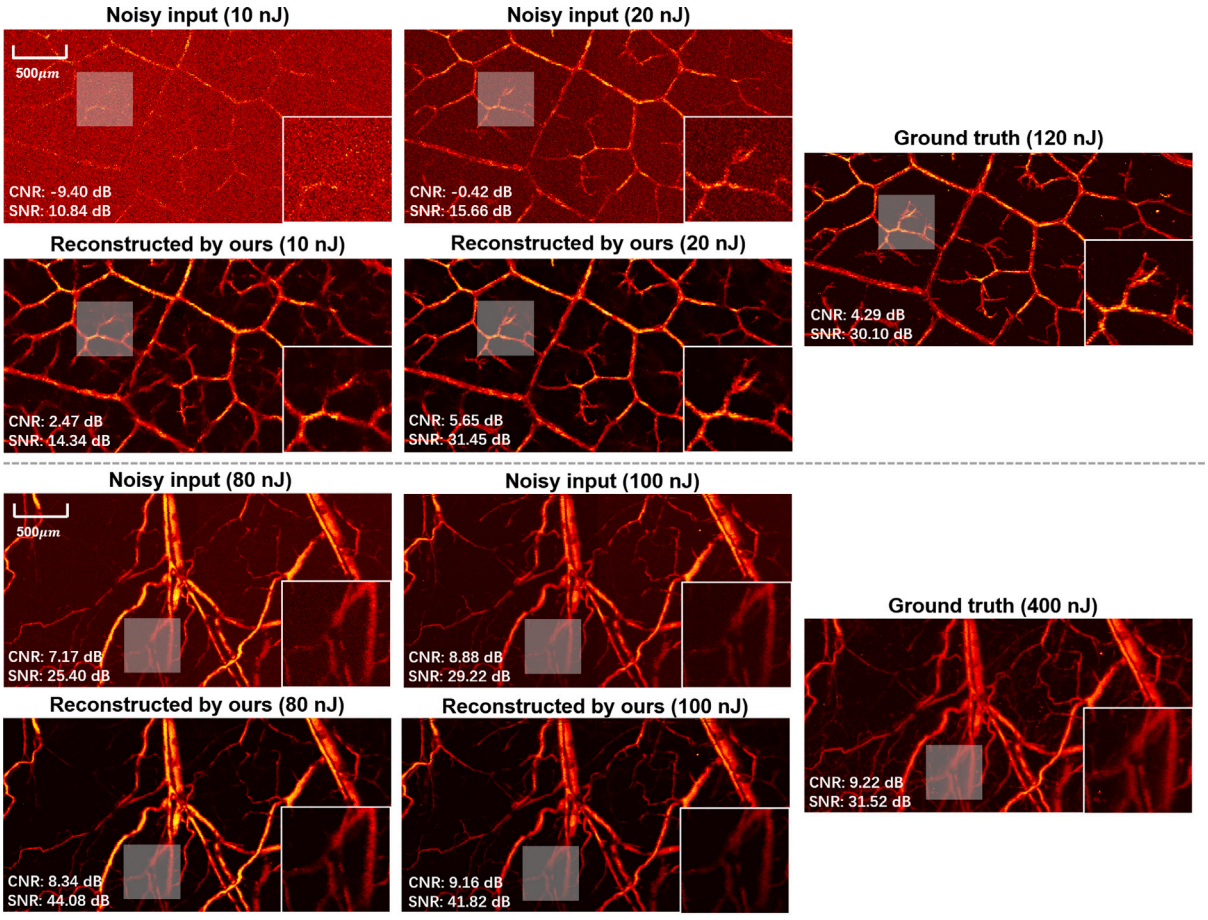


Fig. 8. Visualization results of real-world PAM denoising images. High excitation light energy can be used to obtain clean images and low excitation light energy can be used to collect noisy images. We show the zoomed images of the white shaded area in the lower right part.

Compared with the low-quality images as input, our method achieves to perform both resolution enhancement and denoising simultaneously which demonstrates the effectiveness and generalization ability to *in vivo* images.

5. Conclusions

This paper presents a deep neural network for PAM image reconstruction to perform the tasks of super-resolution and denoising. In this work, we implement three attention blocks to extract multi-level features and perform effective feature fusion. Based on the deep knowledge priors of PAM images, we design a mixed training constraint to leverage both the pixel-level and perception-level constraints. By introducing additional image segmentations, our method attempts to embed semantic features for improved quality of image reconstruction. To fully evaluate the performance of our method, we conduct detailed experiments on three datasets and implement transfer learning to improve the performance across different datasets. In the *in vivo* experiments performed in this study, the super-resolution process of our proposed method achieves an improvement of 0.59 dB for ×2 case and 1.37 dB for ×4 case in peak signal-to-noise ratio, as well as 0.02 and 0.07 in structural similarity. For image denoising, our method achieves an improvement of 3.9 dB in peak signal-to-noise ratio and 0.15 in structural similarity. These experimental results demonstrate that our work can achieve sufficient performance gain for PAM image reconstruction for both super-resolution and denoising.

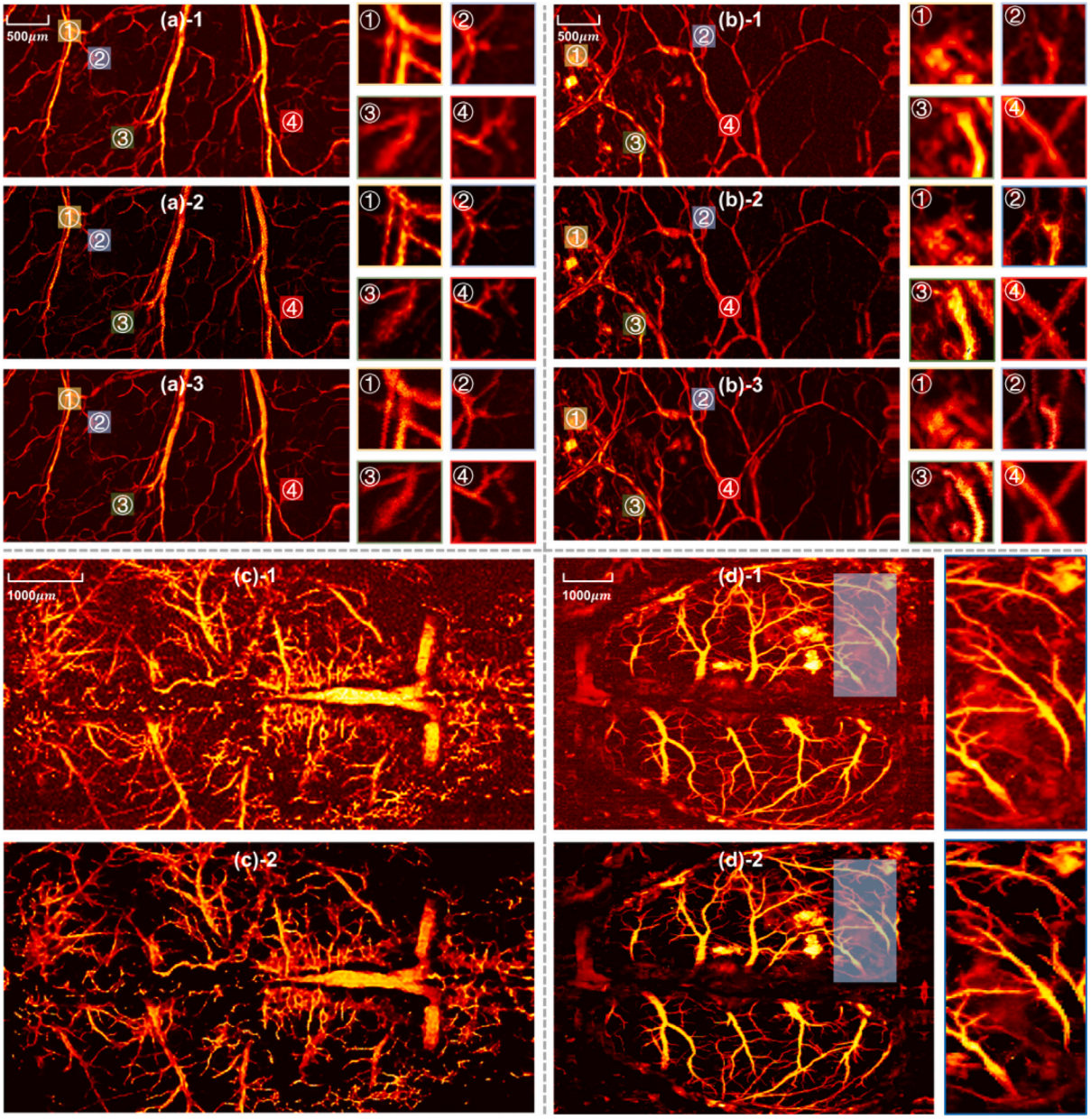


Fig. 9. Visualization results of *in vivo* PAM image super-resolution (×4) and denoising. (a,b)-1: Low-quality images with image size 256 × 128 collected by 80 nJ excitation light energy. (a,b)-2: Reconstructed image by our method with image size 1024 × 512. (a,b)-3: Ground truth images with image size 1024 × 512 collected by 320 nJ excitation light energy. (c)-1: Low-quality images with image size 256 × 136. (c)-2: Reconstructed image by our method with image size 1024 × 544. (d)-1: Low-quality images with image size 256 × 184. (d)-2: Reconstructed image by our method with image size 1024 × 736. We show the zoomed images of the colorful shaded areas on the right.

CRediT authorship contribution statement

Yuxuan Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization. **Jiasheng Zhou:** Writing – review & editing, Methodology, Data curation. **Yating Luo:** Writing – review & editing, Visualization, Methodology. **Jinkai Li:** Writing – review & editing, Software. **Sung-Liang Chen:** Writing – review & editing, Supervision, Data curation. **Yao Guo:** Writing – review & editing, Supervision, Methodology, Funding acquisition. **Guang-Zhong Yang:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.pacs.2024.100608.

References

[1] L.V. Wang, Multiscale photoacoustic microscopy and computed tomography, Nat. Photonics 3 (9) (2009) 503–509.

[2] L.V. Wang, S. Hu, Photoacoustic tomography: in vivo imaging from organelles to organs, Science 335 (6075) (2012) 1458–1462.

[3] H.F. Zhang, K. Maslov, G. Stoica, L.V. Wang, Functional photoacoustic microscopy for high-resolution and noninvasive in vivo imaging, Nature Biotechnol. 24 (7) (2006) 848–851.
[4] J. Yao, L.V. Wang, Photoacoustic microscopy, Laser Photonics Rev. 7 (5) (2013) 758–778.
[5] S. Jeon, J. Kim, D. Lee, J.W. Baik, C. Kim, Review on practical photoacoustic microscopy, Photoacoustics 15 (2019) 100141.
[6] A. DiSpirito, D. Li, T. Vu, et al., Reconstructing undersampled photoacoustic microscopy images using deep learning, IEEE Trans. Med. Imaging 40 (2) (2020) 562–570.
[7] J. Yao, L.V. Wang, Recent progress in photoacoustic molecular imaging, Curr. Opin. Chem. Biol. 45 (2018) 104–112.
[8] J. Yao, L. Wang, J.-M. Yang, K.I. Maslov, T.T. Wong, et al., High-speed label-free functional photoacoustic microscopy of mouse brain in action, 12 (5) (2015) 407–410.
[9] B. Stephanian, M.T. Graham, H. Hou, M.A.L. Bell, Additive noise models for photoacoustic spatial coherence theory, Biomed. Opt. Express 9 (11) (2018) 5566–5582.
[10] S. Telenkov, A. Mandelis, Signal-to-noise analysis of biomedical photoacoustic measurements in time and frequency domains, Rev. Sci. Instrum. 81 (12) (2010).
[11] C. Niu, M. Li, F. Fan, W. Wu, X. Guo, Q. Lyu, G. Wang, Noise suppression with similarity-based self-supervised deep learning, IEEE Trans. Med. Imaging (2022).
[12] Z. Chen, X. Guo, P.Y. Woo, Y. Yuan, Super-resolution enhanced medical image diagnosis with sample affinity interaction, IEEE Trans. Med. Imaging 40 (5) (2021) 1377–1389.
[13] H. Chung, E.S. Lee, J.C. Ye, MR image denoising and super-resolution using regularized reverse diffusion, IEEE Trans. Med. Imaging (2022).
[14] J. Zhou, D. He, X. Shang, Z. Guo, S.-L. Chen, J. Luo, Photoacoustic microscopy with sparse data by convolutional neural networks, Photoacoustics 22 (2021) 100242.
[15] T. Vu, A. DiSpirito III, D. Li, Z. Wang, X. Zhu, et al., Deep image prior for undersampling high-speed photoacoustic microscopy, Photoacoustics 22 (2021) 100266.
[16] D. Seong, E. Lee, Y. Kim, S. Han, J. Lee, M. Jeon, J. Kim, Three-dimensional reconstructing undersampled photoacoustic microscopy images using deep learning, Photoacoustics 29 (2023) 100429.
[17] A. Sharma, M. Pramanik, Convolutional neural network for resolution enhancement and noise reduction in acoustic resolution photoacoustic microscopy, Biomed. Opt. Express 11 (12) (2020) 6826–6839.
[18] H. Zhao, Z. Ke, F. Yang, K. Li, N. Chen, L. Song, et al., Deep learning enables superior photoacoustic imaging at ultralow laser dosages, Adv. Sci. 8 (3) (2021) 2003097.
[19] D. He, J. Zhou, X. Shang, X. Tang, J. Luo, S.-L. Chen, De-noising of photoacoustic microscopy images by attentive generative adversarial network, IEEE Trans. Med. Imaging 42 (5) (2023) 1349–1362.
[20] C. Yang, H. Lan, F. Gao, F. Gao, Review of deep learning for photoacoustic imaging, Photoacoustics 21 (2021) 100215.
[21] I.U. Haq, R. Nagaoka, S. Siregar, Y. Saijo, Sparse-representation-based denoising of photoacoustic images, Biomed. Phys. Eng. Express 3 (4) (2017) 045014.
[22] B. Cohen, et al., New maximum likelihood motion estimation schemes for noisy ultrasound images, Pattern Recognit. 35 (2) (2002) 455–463.
[23] Y. Li, Y. Fan, X. Xiang, D. Demandolx, R. Ranjan, R. Timofte, L. Van Gool, Efficient and explicit modelling of image hierarchies for image restoration, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023.
[24] J. Cao, Q. Wang, Y. Xian, Y. Li, B. Ni, et al., Ciaosr: Continuous implicit attention-in-attention network for arbitrary-scale image super-resolution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2023.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[26] R. Chen, H. Zhang, J. Liu, Multi-attention augmented network for single image super-resolution, Pattern Recognit. 122 (2022) 108349.
[27] A. Khan, A. Chefranov, H. Demirel, Image scene geometry recognition using low-level features fusion at multi-layer deep CNN, Neurocomputing 440 (2021) 111–126.
[28] M.S. Rad, B. Bozorgtabar, U.-V. Marti, M. Basler, H.K. Ekenel, J.-P. Thiran, Srobb: Targeted perceptual loss for single image super-resolution, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2710–2719.
[29] T. An, B. Mao, B. Xue, C. Huo, S. Xiang, C. Pan, Patch loss: A generic multi-scale perceptual loss for single image super-resolution, Pattern Recognit. 139 (2023) 109510.
[30] R. Neshatavar, M. Yavartanoo, et al., CVF-SID: Cyclic multi-variate function for self-supervised image denoising by disentangling noise from image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17583–17591.
[31] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[32] Z. Chen, Y. Zhang, J. Gu, L. Kong, X. Yuan, et al., Cross aggregation transformer for image restoration, Adv. Neural Inf. Process. Syst. 35 (2022) 25478–25490.
[33] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising, IEEE Trans. Image Process. 26 (7) (2017) 3142–3155.
[34] S. Guo, Z. Yan, K. Zhang, et al., Toward convolutional blind denoising of real photographs, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1712–1722.

Yuxuan Liu received his B.E. degree in Biomedical Engineering from Shanghai Jiao Tong University, Shanghai, China in 2022 and he is currently pursuing the Ph.D. degree in the Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include computer vision, biomedical image processing, biomedical signal processing and machine learning algorithms. Mr. Liu received National Scholarships and Shanghai Outstanding Graduates for his B.E. degree.

Jiasheng Zhou received his B.S. degree in Electronic Science and Technology from Shandong University and M.S. degree in Optics from East China Normal University. In 2022, he received his Ph.D. degree from the University of Michigan–Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai, China. Now, he works as a Postdoctoral Research Fellow in the Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include fast scanning photoacoustic imaging and non-contact photoacoustic imaging.

Yating Luo received her B.E. degree in ACM Honored Class from Shanghai Jiao Tong University, Shanghai, China in 2021 and she is currently pursuing the Ph.D. degree in the Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China. Her research interests include robot control, visual servoing, trajectory optimization and deep learning algorithms. She received Zhiyuan Honor Degrees of Bachelor of Engineering in Computer Science and Technology and Shanghai Outstanding Graduates for her B.E. degree.

Jinkai Li received his B.E. degree in Mechanical Engineering from Beihang University, Beijing, China in 2022 and he is currently pursuing the Master degree in the Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China. His research interests include computer vision, biomedical image processing, social robotics and machine learning algorithms.

Sung-Liang Chen received the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, MI, USA, and post-doctoral training at the University of Michigan Medical School, USA. He is currently an Associate Professor with the University of Michigan–Shanghai Jiao Tong University Joint Institute, Shanghai Jiao Tong University, Shanghai, China. His research interests include photoacoustic imaging technology and applications, artificial intelligence for optical microscopy and medical imaging, and optical neural networks. He was a recipient of the Shanghai Pujiang Talent Award.

Yao Guo received his B.S. degree in automation and M.S. degree in communication and information system from Sun Yat-sen University, Guangzhou, China in 2011 and 2014, respectively. He earned his Ph.D. degree in robotic vision from the City University of Hong Kong, Hong Kong in 2018. His postdoctoral training was at the Hamlyn Centre for Robotic Surgery, Imperial College London, London, UK from 2018 to 2020. Since 2020, he has been the tenure-track Assistant Professor with the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China. His main research interests include robotic vision, gait analysis, rehabilitation

and assistive robotics, human–machine interaction, and machine learning algorithms in healthcare applications. He received the Best Conference Paper Award at the IEEE International Conference on Mechatronics and Automation (ICMA) 2016.

Guang-Zhong Yang (Fellow, IEEE) was the Director and the Co-Founder of the Hamlyn Centre for Robotic Surgery and the Deputy Chairman of the Institute of Global Health Innovation, Imperial College London, London, U.K., where he also holds a number of key academic positions, such as the Director and the Founder of the Royal Society/Wolfson Medical Image Computing Laboratory, the Co-Founder of the Wolfson Surgical Technology Laboratory, and the Chairman of the Centre for Pervasive Sensing. He is currently the Founding Dean of the Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China. His main research interests are in medical imaging, sensing, and robotics. In imaging, he is credited for a number of novel MR phase contrast velocity imaging and computational modeling techniques that have transformed in vivo blood flow quantification and visualization. These include the development of locally focused imaging combined with real-time navigator echoes for resolving a respiratory motion for high-resolution coronary angiography, as well as the MR dynamic flow pressure mapping for which he received the ISMRM I. I Rabi Award. He pioneered the concept of perceptual docking for robotic control, which represents a paradigm shift of learning and knowledge acquisition of motor and perceptual/cognitive behavior for robotics, as well as the field of the body sensor network (BSN) for providing personalized wireless monitoring platforms that are pervasive, intelligent, and context-aware. Dr. Yang is a fellow of the Royal Academy of Engineering, the Institution of Engineering and Technology (IET), and the American Institute for Medical and Biological Engineering (AIMBE). He was a recipient of the Royal Society Research Merit Award. He is listed in The Times Eureka “Top 100” in British Science.