D-STAR: Diffusion-Based Sparse Tomographic Angular Recovery for Isotropic-Resolution Photoacoustic Imaging

Di Kong, Haoyu Yang, Yan Luo, Zhiqiang Chen, Bo Lei, Yi Zhong, Mingyuan Liu, Yuwen Chen, and Cheng Ma, *Member, IEEE*

***Abstract*— Anisotropy in imaging systems often results in directional degradation, impairing image quality and complicating subsequent analyses. While multiangle imaging has proven effective in mitigating these effects, it introduces challenges such as extended imaging times and increased excitation doses. To address these limitations in Photoacoustic Tomography (PAT), we propose a novel approach—Diffusion-based Sparse Tomographic Angular Recovery (D-STAR). D-STAR significantly reduces the number of required angles for high-resolution PAT while maintaining image quality comparable to full tomographic angular imaging. By training a diffusion model on a custom 3D PAT dataset, we optimize the balance between spatial and temporal resolutions, signal-to-noise ratio (SNR), and laser exposure. Our experiments with excised brain and vessel phantoms demonstrate that D-STAR produces high-fidelity images suitable for both structural and molecular imaging. This method outperforms existing approaches in static structural recovery and quantitative data extraction, offering substantial improvements in imaging quality, particularly in resolution and contrast. Furthermore, D-STAR enhances flexibility in imaging system design, reducing the need for hardware upgrades while improving temporal resolution and minimizing laser exposure.**

***Index Terms*— Photoacoustic tomography, diffusion probabilistic models (DPMs), molecular photoacoustic imaging, isotropic-resolution imaging.**

Received 10 April 2025; revised 22 May 2025; accepted 26 May 2025. Date of publication 30 May 2025; date of current version 30 October 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62475129 and Grant 82000429, in part by China Postdoctoral Science Foundation under Grant 2024M761726, in part by the Zhongguancun Academy under Grant 20240311, and in part by Fuzhou Institute for Data Technology. Recommended by Associate Editor C. Kim. *(Corresponding authors: Cheng Ma; Yuwen Chen.)*

This work involved animal subjects in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Animal Care and Use Committee (IACUC) of Tsinghua University under Protocol No. 21-ZY1, and performed in line with the Guide for the Care and Use of Laboratory Animals of Tsinghua University.

Di Kong is with the Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China, and also with the Zhongguancun Academy, Beijing 100094, China (e-mail: kd24@mails.tsinghua.edu.cn).

Haoyu Yang and Yi Zhong are with the School of Life Sciences and Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing 100084, China, and also with the IDG/McGovern Institute of Brain Research, Beijing 100084, China (e-mail: yhy21@mails.tsinghua.edu.cn; zhongyithu@tsinghua.edu.cn).

Yan Luo and Yuwen Chen are with the Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: lythu1999@163.com; chen-yw97@mail.tsinghua.edu.cn).

Zhiqiang Chen and Bo Lei are with Beijing Academy of Artificial Intelligence, Beijing 100084, China (e-mail: chenzhiqiang14@mails.ucas.ac.cn; b.lei.2022@hotmail.com).

Mingyuan Liu is with the Department of Vascular Surgery, Beijing Friendship Hospital, Capital Medical University, Beijing 100084, China (e-mail: dr.mingyuanliu@pku.edu.cn).

Cheng Ma is with the Department of Electronic Engineering, Beijing National Research Center for Information Science and Technology, the Institute for Precision Healthcare, and the Institute for Intelligent Healthcare, Tsinghua University, Beijing 100084, China, also with the IDG/McGovern Institute of Brain Research, Beijing 100084, China, and also with the Zhongguancun Academy, Beijing 100094, China (e-mail: cheng_ma@tsinghua.edu.cn).

Digital Object Identifier 10.1109/TMI.2025.3574946

I. Introduction

IN IMAGING systems, resolution describes the ability to distinguish fine details—think of it as how sharply a camera or microscope can capture the edges of an object. Ideally, resolution is isotropic, meaning it's consistent in all directions. For instance, a tiny circle would appear equally clear whether viewed horizontally, vertically, or diagonally. However, many real-world systems exhibit anisotropic resolution, where the sharpness of details depends on their orientation. Such limitation often arises from directionally biased sampling schemes or system geometry. As a result, tasks such as image registration and quantitative analysis become more challenging. To mitigate this issue, multiangle imaging and reconstruction have emerged as effective strategies (Fig. 1). This approach has been adopted by a variety of imaging techniques such as Optical Diffraction Tomography (ODT) [1], Light-Sheet Microscopy [2], Structured Illumination Microscopy (SIM) for super-resolution [3], Ultrafast Doppler Tomography (UFD-T) [4], and Photoacoustic Tomography (PAT) [5], [6], [7]. These methods utilize optical or mechanical scanning to alter the direction of degradation and integrate data from multiple angles to enhance image quality. They have demonstrated efficacy in their respective applications.

However, this multiangle approach introduces notable drawbacks. The scanning process prolongs imaging time, which is problematic for applications requiring high temporal resolution. Additionally, the increased excitation dose—such as laser exposure or ultrasound emission—can lead to sample bleaching [8] or exceed safety limits [9]. Consequently, a pertinent question arises: What is the minimum number of angles required? Existing methods employing full tomographic

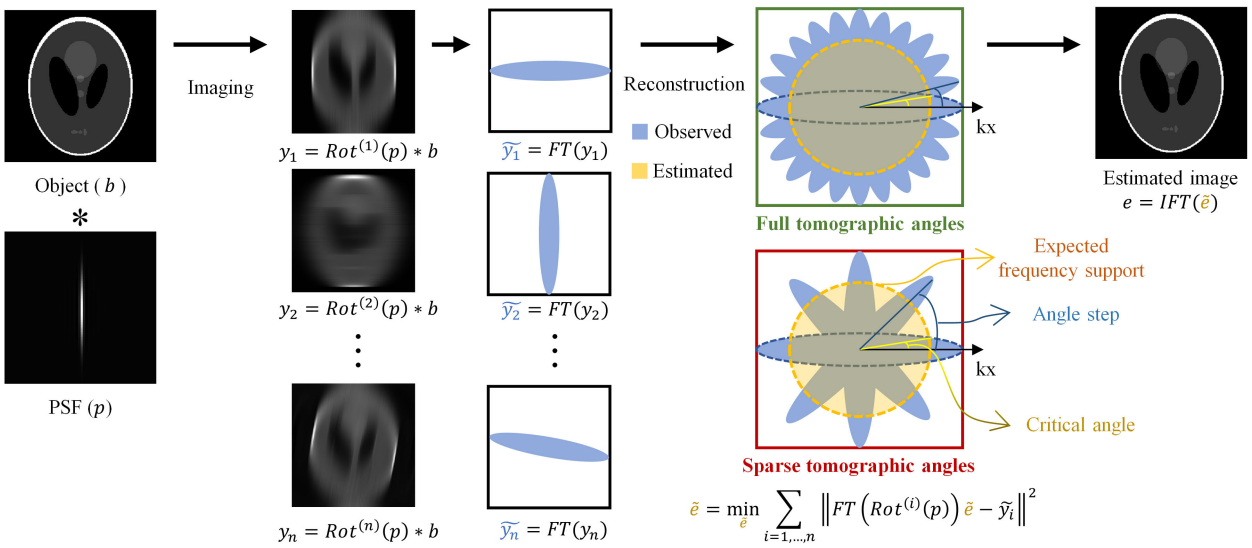

Fig. 1. Multiangle fusion process and issues of STAR. The object b is convolved with the PSF p, producing images $y_i = \text{Rot}(\theta_i)p * b$, which are Fourier-transformed into $\widetilde{y_i} = \text{FT}(y_i)$. The right panel compares full (green) and sparse (red) angle sampling in Fourier space, showing expected frequency support (orange), angle step (blue), and critical angle (yellow). Sparse angles cause reconstruction errors in the estimated image.

angular imaging typically involve rotation steps of less than twice the critical angle (the half-angle of the arc on the expected boundary of the frequency support occupied by a single-angle image, Fig. 1), ensuring comprehensive coverage of the frequency spectrum through the union of observed spectra from different angles [10]. Fortunately, sparse sampling theory [11] allows for recovery of the original image using a limited number of angles (termed Sparse Tomographic Angular Recovery, STAR) by incorporating prior information. In this context, the reconstruction problem transitions from being overdetermined to underdetermined, with its success heavily dependent on the quality of the prior information [12].

Deep learning, especially generative models [13], [14], [15], has emerged as a promising solution to ill-posed inverse problems, where traditional methods may falter. This capability has led to deep learning-based methods being extensively studied and widely adopted for tackling problems similar to STAR, such as image denoising [16], [17] and super-resolution [18], [19], particularly in the realm of light microscopy. Research has consistently demonstrated the effectiveness of deep learning techniques in overcoming the challenges associated with STAR, highlighting their potential to enhance image quality and analytical accuracy [20].

In this study, we focus on the photoacoustic imaging modality, specifically Photoacoustic Tomography (PAT), due to its unique combination of optical contrast and acoustic penetration depth [21]. This integration provides fine structural and molecular information within deep tissues. Although isotropic photoacoustic imaging systems exist [22], [23], [24], their systems are complex, expensive and bulky. Anisotropic systems, by contrast, offer greater flexibility and are more commonly employed in clinical and preclinical research [25], [26], [27]. But the drawbacks associated with full tomographic angular imaging are particularly pronounced in these contexts; for instance, a 32-angle scan may be required to achieve a resolution of 150 μm within a 24 × 24 × 24 mm^3 FOV, taking up to 3 minutes for single-wavelength imaging. This duration is often unacceptably long for clinical applications and poses a risk of tissue damage when using contrast agents [5]. Therefore, it is essential to minimize the number of angles required for effective imaging.

However, STAR poses additional challenges in photoacoustic imaging, where bipolar photoacoustic signals can result in the loss of directional features at certain angles, reducing the information available for recovery and complicating the problem [28], [29]. Furthermore, the spatially varying point spread function (PSF), influenced by non-uniform sound speed distributions in biological samples, limits the effectiveness of traditional deconvolution methods. Additionally, the scarcity of physically acquired 3D PAT datasets means that most existing studies rely on simulated data for training, constraining the performance of their networks and leaving the STAR problem in PAT largely unexplored [30], [31], [32].

To address these challenges, we propose a diffusion-based sparse tomographic angular recovery (D-STAR) method to reduce the number of angles required for high-resolution PAT. This approach involves constructing a 3D PAT dataset using a custom-built rotatable PAT system and training a diffusion-based network specifically designed for PAT. This methodology aims to optimize the balance between imaging time, resolution, signal-to-noise ratio (SNR), and laser exposure dose. Our results, demonstrated on both excised brains and vessel phantoms, indicate that the trained D-STAR method can approximate the quality of full tomographic angular imaging. Furthermore, we show that the high-fidelity outputs from D-STAR can be directly employed in molecular photoacoustic imaging, a capability that represents a significant improvement over the limitations faced by previous network.

II. Preliminaries

A. Multiangle Image Fusion

Assuming an imaging system characterized by anisotropic resolution (Fig. 1), with a PSF p elongated in the y-direction,

the imaging or observation process can be described as a convolution operation: $y = p*b$, where b denotes the imaging object and $*$ represents the convolution operator. It is evident that the resulting observed image y will exhibit blur along the y-direction. To mitigate this directional blur, the imaging system is rotated, equivalently rotating the PSF to alter the direction of the blur. Consequently, a series of observed images are obtained:

$$y_i = \mathrm{Rot}^{(i)}(p) * b, \ \forall i \in \{1, \ldots, n\}, \tag{1}$$

where $\mathrm{Rot}^{(i)}(\cdot)$ denotes the i-th rotation of the PSF. In this context, the objective of multiangle imaging fusion is to reconstruct the original image b with isotropic resolution from the blurred observations $\{y_i\}_{i=1}^{n}$. The problem is formulated as:

$$\min_{e} \sum_{i=1}^{n} \left\| \mathrm{Rot}^{(i)}(p) * e - y_i \right\|^2, \tag{2}$$

where e represents the estimated image. If the PSF is known from simulation or experimental methods, this problem can be addressed using traditional deconvolution techniques. In the absence of prior PSF knowledge, deconvolution methods without prior knowledge must be employed, which entails substantial computational complexity. In PAT, the PSF lacks both spatial and rotational invariance, further exacerbating the computational challenges.

As an alternative approach, the image can be transformed from the spatial domain to the spatial frequency domain (k-space). For each spatial frequency, values are selected from the angle with the largest magnitude and incorporated into the fused data e. This method inherently satisfies Equation 2 without requiring prior knowledge of the PSF. However, this approach necessitates full tomographic angular scan. In cases of STAR, missing spatial frequencies can introduce significant artifacts.

B. Diffusion Probabilistic Model for Content Restoration

Before the advent of Diffusion Probabilistic Models (DPMs) [33], the most widely used generative models were Generative Adversarial Networks (GANs) [13]. While GANs are known for generating high-quality images, they present challenges such as training instability and mode collapse. Variational Autoencoders (VAEs) [14] provide a more stable training process but tend to produce blurrier reconstructions due to factors like simplified reconstruction loss and the regularization imposed by the KL divergence term. DPMs, by framing image generation and restoration as a gradual diffusion process, offer a strong balance between image quality and stability. This multi-step approach addresses some of the key challenges faced by GANs and VAEs, making DPMs particularly suitable for tasks requiring precise image restoration.

1) Noise Addition: In the forward process, a sequence of images $\boldsymbol{x}^0, \ldots, \boldsymbol{x}^T$ is generated by gradually adding noise to the original image. The amount of noise increases as timestep t progresses, where $\boldsymbol{x}^0$ is the original clean image and $\boldsymbol{x}^T$ is essentially pure Gaussian noise.

2) Denoising: A denoising neural network, often parameterized as $\epsilon_\theta(\boldsymbol{x}^t, t)$, learns to predict the added noise at each timestep t. The reverse process then uses these predictions to iteratively refine the noisy image back towards its clean form.

3) Content Restoration in STAR: In STAR for PAT, incomplete imaging data can be viewed as noisy or missing information. The diffusion model addresses this by filling in the gaps using prior knowledge learned during training, effectively mitigating anisotropic degradation caused by sparse angular acquisitions, which results in directional information loss. Compared to GANs and VAE, the diffusion model's gradual reconstruction process aligns closely with the fundamental nature of content restoration, allowing it to progressively remove noise or add details. Its single training objective makes the training process more stable, without needing complex adjustments, and it avoids problems like mode collapse. The model also provides better probabilistic interpretation, which helps prevent blurring issues that are common with other methods. Additionally, the multi-stage reconstruction process allows for more control, as conditions can be added at each step, unlike in one-stage methods.

III. Methods

A. Imaging System Setup

The homemade imaging system includes an excitation laser, scanning stages, and an ultrasound detection module (Fig. 2). The excitation of PA signals was carried out using an optical parameter oscillator (LP604, Solar Laser) together with its pump source (LQ929B, Solar Laser). The scanning strategy was accomplished by two motors: a direct-drive motor (ADRS-200-M-A-NS, Aerotech Inc.) for rotational movement and a linear motor (ANT 25L, Aerotech Inc.) for translational motion. We integrated a half-ring array transducer from ULSO TECH Co., Ltd., featuring a 55 mm radius. This transducer operates at a central frequency of 5.5 MHz and has a detection bandwidth of -6 dB at 60 %. It consists of 128 elements, each with an arc-rectangular aperture of 1.32 mm pitch and an elevational length of 20 mm, which collectively provide a cylindrical focus with a numerical aperture of 0.2. The ultrasound signals were recorded using a low-noise data acquisition system with 128 channels (MarsonicsDAQ128, Tianjin Langyuan Inc.).

B. Multiangle PAT Imaging

The multiangle PAT employs a mechanical translation-rotation scanning strategy for the ultrasound transducer. Initially, the transducer is linearly scanned along the elevational direction with a step size of 0.6 mm, approximately half of the elevational resolution. Following this, the transducer is rotated to perform the subsequent linear scan. The angular step size for rotation is determined based on experimental requirements, specifically 5.625 ° for acquiring 32 angles and 45 ° for acquiring 4 angles.

The subsequent image reconstruction process is a simple solution to the problem raised in Section II-A, with the following steps:

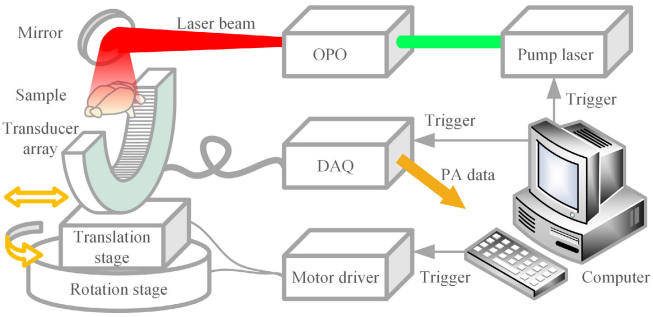

Fig. 2. Diagram of multiangle PAT system. The ultrasound transducer array is linearly scanned along the elevational direction (straight arrow) and rotated (curved arrow) to vary the linear scan orientation.

1. Reconstructing the raw PA signals into 2D images using the delay and sum algorithm.

2. Stacking these images along the elevational direction with respect to their scanning angle.

3. Filtering the stacked 3D images to flatten the uneven sampling density in k-space [5]. The filter $H(kx, kz)$ in the k-space of the plane perpendicular to y-axis can be described as:

$$H(kx, kz) = S(kx, kz; 0) \cdot \frac{1}{\sum_\theta S(kx, kz; \theta)}, \tag{3}$$

where $S(kx, kz; \theta)$ represents a transfer function corresponding to a translationally-scanned tomogram at angle θ:

$$S(kx, kz; \theta) = \text{step}(kx\cos(\theta) - kz\sin(\theta) + W_e) - \text{step}(kx\cos(\theta) - kz\sin(\theta) - W_e). \tag{4}$$

W_e is the cutoff spatial frequency in the elevational direction and step$(\cdot)$ denotes the step function.

4. Rotating and registering 3D images of different angles in the same coordinates and then summarizing them. The resulting near-isotropic 3D volume is ready for dataset constructing.

C. STAR Using Conditional Diffusion Model

Given an input 4-angle image, our model takes it as a condition $\boldsymbol{c}$, then gradually refines the reconstruction from noisy initialization to high-quality outputs using a guided reverse process (Fig. 3).

Following [15] and [33], the forward diffusion process incrementally adds Gaussian noise to the ground truth data, progressively corrupting the high-quality full-angle image $\boldsymbol{x}^0$. This process can be modeled as a Markovian chain, denoted by q, where each noisy image $\boldsymbol{x}^t$ at timestep t is obtained by conditioning on the previous step:

$$q(\boldsymbol{x}^t|\boldsymbol{x}^{t-1}) = \mathcal{N}\left(\boldsymbol{x}^t;\ \sqrt{1-\beta_t}\,\boldsymbol{x}^{t-1},\ \beta_t \boldsymbol{I}\right), \tag{5}$$

Here, β_t represents the noise schedule, which controls the variance of the added Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ at each timestep t, where I denotes the identity matrix. From Equation 5, $\boldsymbol{x}^t = \sqrt{1-\beta_t}\,\boldsymbol{x}^{t-1} + \sqrt{\beta_t}\,\epsilon$. The forward diffusion process continues until the final noisy image becomes indistinguishable from pure noise $\boldsymbol{x}^T \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

In the conventional, generation-oriented DDPM framework, the reverse sampling process trains the denoising network to predict the noise ϵ added at each step. This noise is then used to estimate the clean image $\boldsymbol{x}^0$. Specifically, at timestep t, given the noisy image $\boldsymbol{x}^t$, the denoiser estimates the added noise via a neural network $\epsilon_\theta(\boldsymbol{x}^t, t, \boldsymbol{c})$:

$$\hat{\boldsymbol{x}}^0 = \frac{\boldsymbol{x}^t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(\boldsymbol{x}^t, t, \boldsymbol{c})}{\sqrt{\bar{\alpha}_t}}, \tag{6}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ is the cumulative noise schedule, and $\alpha_t = 1 - \beta_t$. The reverse diffusion step then samples:

$$\boldsymbol{x}^{t-1} = \sqrt{\bar{\alpha}_{t-1}}\,\hat{\boldsymbol{x}}^0 + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_t,\ \ \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{7}$$

In contrast, we directly train the denoiser to predict the clean image $\boldsymbol{x}^0$ instead of the noise. This change better aligns with our reconstruction-oriented task, where structural fidelity

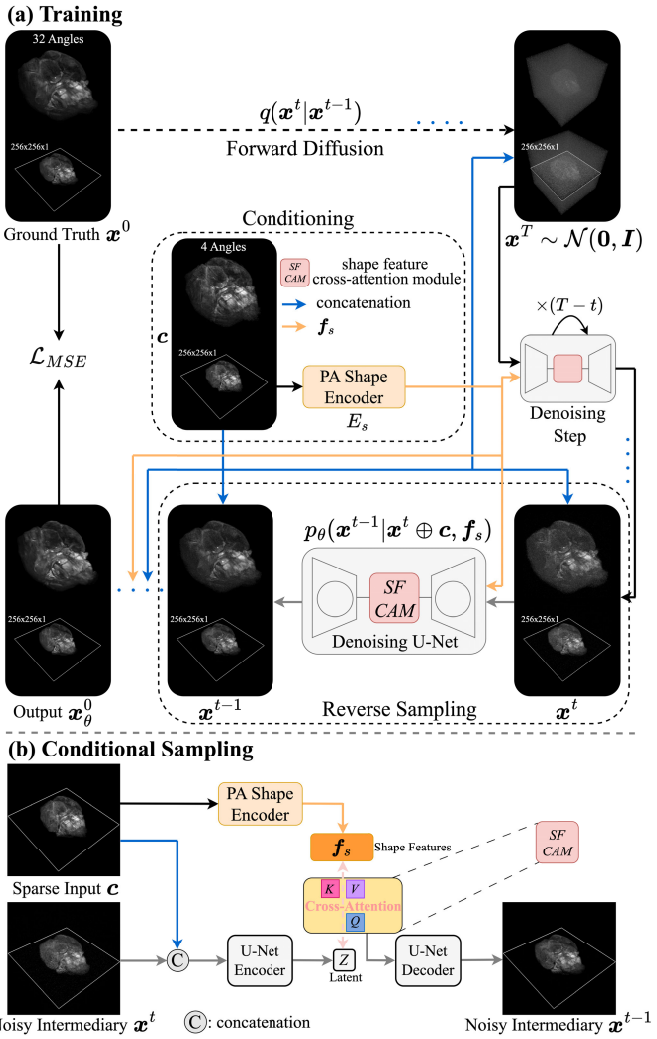

Fig. 3. Architecture along with training pipeline of D-STAR. a) The forward diffusion process progressively corrupts the 32-angle ground truth image x^0 to generate a noisy representation x^T. The reverse sampling process denoises x^T to recover the full-angle image, guided by both low-level (direct image concatenation) and high-level (features via cross-attention) conditioning from the sparse 4-angle input c. During training, sliced images from the GT voxel x^0 and the reconstructed voxel x_θ^0 are used for supervision. b) For low-level conditioning, at each denoising step, c is concatenated with x^t and fed into the denoising U-Net. Then a PA Shape Encoder E_S is trained to extract shape features f_S from c, which are integrated into the denoising process via cross-attention.

is crucial. The network $x_\theta^0(\boldsymbol{x}^t, t, \boldsymbol{c})$ directly estimates the underlying clean image from $\boldsymbol{x}^t$, and the reverse step becomes:

$$\boldsymbol{x}^{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, x_\theta^0(\boldsymbol{x}^t, t, \boldsymbol{c}) + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_t,\ \ \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{8}$$

Given the sparse tomographic angular acquisition, we condition the reverse sampling process using the input 4-angle images $\boldsymbol{c}$. For this conditional control, we use two approaches to implement it. In the first approach, inspired by the super-resolution works [34], we use a simple and effective method to modify the intermediate noisy image $\boldsymbol{x}^t$ by channel-wise concatenating the 4-angle images. Our model uses $\hat{\boldsymbol{x}}^t_{(2,w,h)} := \boldsymbol{x}^t_{(1,w,h)} \oplus \boldsymbol{c}_{(1,w,h)}$ as input to the denoising process at each timestep, where $\oplus$ denotes channel-wise concatenation. This formulation incorporates an additional 4-angle image channel to guide the generation process during denoising. The second approach, in order to reduce the interference of redundant noise information as well as to enhance the fusion of information between the condition and the network output, we provide conditional control of the feature dimension by introducing the cross-attention mechanism [35], implemented in the middle block of the denoising U-Net:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V, \tag{9}$$

with $Q = W_Q \cdot \varphi(\boldsymbol{x}^t \oplus \boldsymbol{c})$, $K = W_K \cdot \boldsymbol{f}_s$, $V = W_V \cdot \boldsymbol{f}_s$. Here, d is the dimensionality of the query and key vectors, while W_Q, W_K and W_V are learnable matrices that project the input features into the query, key, and value spaces. And $\varphi(\boldsymbol{x}^t \oplus \boldsymbol{c})$ represents a flattened intermediate feature map from the denoising U-Net, parameterized by θ. Additionally, $\boldsymbol{f}_s = E_s(\boldsymbol{c})$, where E_s is a pretrained ResNet-50 [36], referred to as the PA Shape Encoder. This encoder captures the morphological information of the objects imaged from the 4-angle input, enhancing the model's ability to faithfully restore the structural details in the decoding phase. By leveraging these two modes of conditional control, our model is able to reconstruct a high-resolution image along the slice direction. The 4-angle input data provides crucial structural information that guides the denoising process, allowing the restoration of fine features that are otherwise lost. The reverse sampling process is defined as:

$$p_\theta(\boldsymbol{x}^{t-1}|\boldsymbol{x}^t,\ \boldsymbol{c},\ \boldsymbol{f}_s) = \mathcal{N}\big(\boldsymbol{x}^{t-1};\ \mu_\theta(\boldsymbol{x}^t,\ \boldsymbol{c},\ \boldsymbol{f}_s,\ t),\ \sigma_t^2 \boldsymbol{I}\big), \tag{10}$$

where μ_θ is the predicted mean from the denoising network, and θ is the learnable parameter. The variance σ_t^2, with some algebraic manipulation, is derived from β_t. The reverse process iterates from $t = T$ (final timestep) down to $t = 0$, gradually removing the noise introduced in the forward process.

By concatenating the input image $\boldsymbol{c}$ and intermediate image $\boldsymbol{x}^t$, we create a 2-channel input for the denoising network, which is structured as a convolutional U-Net. A key strength of our U-shaped network is its ability to maintain a broad bandwidth, capturing and preserving detailed structural information essential for high-fidelity 2D image recovery and richly detailed 3D volume generation. The network's convolutional layers fully leverage the geometric priors in the input data, effectively managing complex spatial relationships among different slices, resulting in consistent and geometrically accurate 3D representations, evident in the spatial alignment of the output images along both the x-axis and z-axis.

To further enhance the quality and robustness of our model, we introduce key improvements in training the conditional diffusion models: **(1) Zero-SNR Training** [37]. We apply the zero-SNR technique, which helps address the discrepancy between the initial Gaussian noise in the sampling process and the noisiest training samples. **(2) Sample Prediction.** Instead of predicting the added noise ϵ, our model is trained to directly predict the full-angle image $\boldsymbol{x}^0$. This approach shifts the focus from noise estimation to image reconstruction, making the model more adept at restoring the high-resolution image from each noisy step. The training is structured to minimize the difference between the predicted image $\boldsymbol{x}_\theta^0(\boldsymbol{x}^t, \boldsymbol{c}, t)$ and the full-angle image $\boldsymbol{x}^0$. The loss function is defined as the mean squared error (MSE) between these two images:

$$\mathcal{L}_{\text{MSE}} = \mathbb{E}_{\boldsymbol{x}^0,\,\boldsymbol{c},\,t}\left[\left\|\boldsymbol{x}^0 - \boldsymbol{x}_\theta^0(\boldsymbol{x}^t,\ \boldsymbol{c},\ E_s(\boldsymbol{c}),\ t)\right\|^2\right]. \tag{11}$$

D. Temporal Unmixing For Fluorescent Proteins

For most fluorescent proteins, the excitation light causes photobleaching during the imaging process, leading to an exponential decay in PA signals [38]. In our temporal unmixing algorithm, we model this PA amplitude decay using the equation:

$$A(t) = a \cdot \exp(-bt) + c, \tag{12}$$

where b represents the bleaching rate, a denotes the PA signal strength contributed by the fluorescent proteins, and c accounts for the PA signal strength from any unbleached background chromophores. In this work, the 4-angle imaging cycle was performed 24 times to generate 24 frames. To achieve a more precise analysis, the multiangle fusion procedure was applied to every set of 4 adjacent angles throughout the scanning process, resulting in a temporal interpolation that produced 93 virtual frames for exponential fitting.

E. Sample Preparation

All experimental procedures of rodents were approved by the Institutional Animal Care and Use Committee (IACUC) of Tsinghua University and were performed using the principles outlined in the Guide for the Care and Use of Laboratory Animals of Tsinghua University.

The brains used to construct the dataset were prepared through standard transcardiac perfusion with phosphate buffered saline (PBS) followed by 4% paraformaldehyde (PFA). After this processing, the samples were ready for PAT imaging. The imaging data presented in Fig. 11 were obtained from a brain that had been injected with AAV-hsyn-iRFP-EGFP.

The vessel dataset was constructed by imaging vessel-like seaweed samples submerged in a 3.6% v/v 30%-intralipid emulsion solution.

The in vivo dataset was established by imaging subcutaneous tumors in tumor-bearing mice. For this experiment, NU/NU mice were employed as the animal model, and 4T1 cells were orthotopically implanted into their dorsal region. Tumors were grown for approximately 7 days before being used for experiments.

IV. Experiments and Results

A. Datasets and 3D Image Preprocessing

We prepared the brain, tumor and seaweed samples as described in Section III-E for the dataset, followed by multiangle imaging procedure described in Section III-B using the homemade PAT imaging system (see Section III-A for details). In our proposed dataset, we used 3D images fused from 32 different angles as the Ground Truth and 4 angles as the input image of the model. In total, we collected 86 mouse brains, 32 mouse tumors and made 75 vessel phantoms to make our dataset, the STAR-PAT. The data in STAR-PAT are divided into training, validation and testing datasets, containing [76, 5, 5] mouse brains, [24, 4, 4] mouse tumors, and [65, 5, 5] vessel phantoms. The brain and vessel 3D volumetric images each have a data size of $256 \times 256 \times 256$, while the tumor data has a size of $288 \times 288 \times 288$.

For 3D image preprocessing, following [32], with efficiency concerns in 3D network to be compared in the experiments, the 3D voxels were cropped into 8 cubes with 128 pixels along each dimension as the input of the network. And after data augmentation, $4864, 320, 320$ brain matrices, in addition of $4160, 320, 320$ vessel matrices, both with dimensions of $128 \times 128 \times 128$, make up the 3D STAR-PAT dataset, which is the largest real 3D dataset in photoacoustic imaging neural network training. Next, we processed the original $256 \times 256 \times 256$ / $288 \times 288 \times 288$ 3D image data into 256×256 / 288×288 2D slices along the y-axis to fit our proposed D-STAR model. We only took out the slices containing object information, and finally obtained $[7243, 454, 457]$ mouse brain slices, $[4847, 614, 672]$ mouse tumor slices and $[14728, 1175, 1170]$ vessel phantom slices. The index of each 2D slice was recorded so that they could be stacked into the 3D voxel again after the network processing, and the same went for 3D cubes.

B. Network Architecture

The conditional diffusion model is based on a U-Net architecture and contains around $605M$ parameters. The denoising network follows an encoder-decoder structure with feature channels of $[128, 256, 512, 512, 1024, 1024]$. Each downsampling level consists of three residual blocks followed by a downsampling operation. Self-attention blocks are applied at resolution $[32, 16, 8]$ to enhance feature representation. The middle block contains two residual blocks and one cross-attention block, which facilitates interaction between sparse input features and the latent representation for improved recovery. The upsampling phase mirrors the downsampling phase, with each level containing three residual blocks and self-attention blocks at the same resolutions.

C. Implementation Details

The conditional diffusion model was trained on 4 NVIDIA A100 40GB GPU cards for 3 days with 150 epochs (that is $130k$ iterations) for ex vivo Brain data and for 4 days with 180 epochs (that is $320k$ iterations) for Vessel phantom data. The model was trained with batch size 8. We used the AdamW optimizer with initial learning rate $5e-5$ and adjusted the learning rate to $1e-5$ after epochs 100. We set the number of timesteps $T = 1000$. During sampling, the conditional diffusion model was sampled with 50 steps using DDIM [40].

D. Quantitative and Qualitative Results

We validated the performance of our method in two categories of data, namely, the ex-vivo data of mouse brain and the phantom data of vessels (shown in Fig. 4 and Fig. 5).

1) Qualitative Results: To validate the effectiveness of our method, we qualitatively compare our results with previous works, including 3DFD U-Net [32], Deep-E [30], SRGAN [41] and PAT-Diffusion [42]. Since 3DFD U-Net is not open-sourced, we use Open3DFD U-Net [43], an open-source implementation of 3DFD U-Net, for comparison. For the other baselines, we use their official codes and retrain the checkpoints. Clearly, our method reconstructs 3D images with higher fidelity and reduced background noise (Fig. 4 (a) and Fig. 5 (a)). The reconstructed 3D images from other baselines looks blurry, as evidenced by the enlarged images in the (Fig. 5 (b)). This is because our model fully utilizes the pixel spatial alignment of every input 2D slice along the longitudinal space of the y-axis, and distinguishes the object information and interference information such as reflection artifacts and noise in each slice, then completes the missing information of sparse angular scan, and finally superimposes them to output a 3D image expressed in the form of voxel.

Additionally, the images generated by our method have more natural and smooth edge transitions, better structural details, and the most accurate restoration that is closest to the 32-angle images (Fig. 4 (b) and Fig. 5 (c)). It is worth mentioning that our method is the only one succeeded in restoring the hippocampus structure, whereas other compared baselines failed (Fig. 4 (c)).

Moreover, the spatial frequency spectrums indicate significant improvement of our method compared to the 4-angle images (Fig. 4 (d) and Fig. 5 (d)). Compared to other baseline methods, our spatial frequency spectrum is the closest to that of the 32-angle images in the meaningful spectrum distribution area where we focus on, and we have achieved a great expansion on the basis of the original sparse spectrum distribution.

2) Quantitative Results: We quantitatively evaluated the quality of outputs from our method and baselines by calculating the root mean square error (RMSE), the peak signal-to-noise ratio (PSNR) and the structural similarity index measure (SSIM), following [32], which measure the resemblance in appearance between the generated 3D image and the original ground truth (documented in Table I). Our method consistently surpasses the baselines across all metrics, demonstrating its effectiveness in producing accurate reconstructions, reducing noise, and achieving greater similarity to the ground

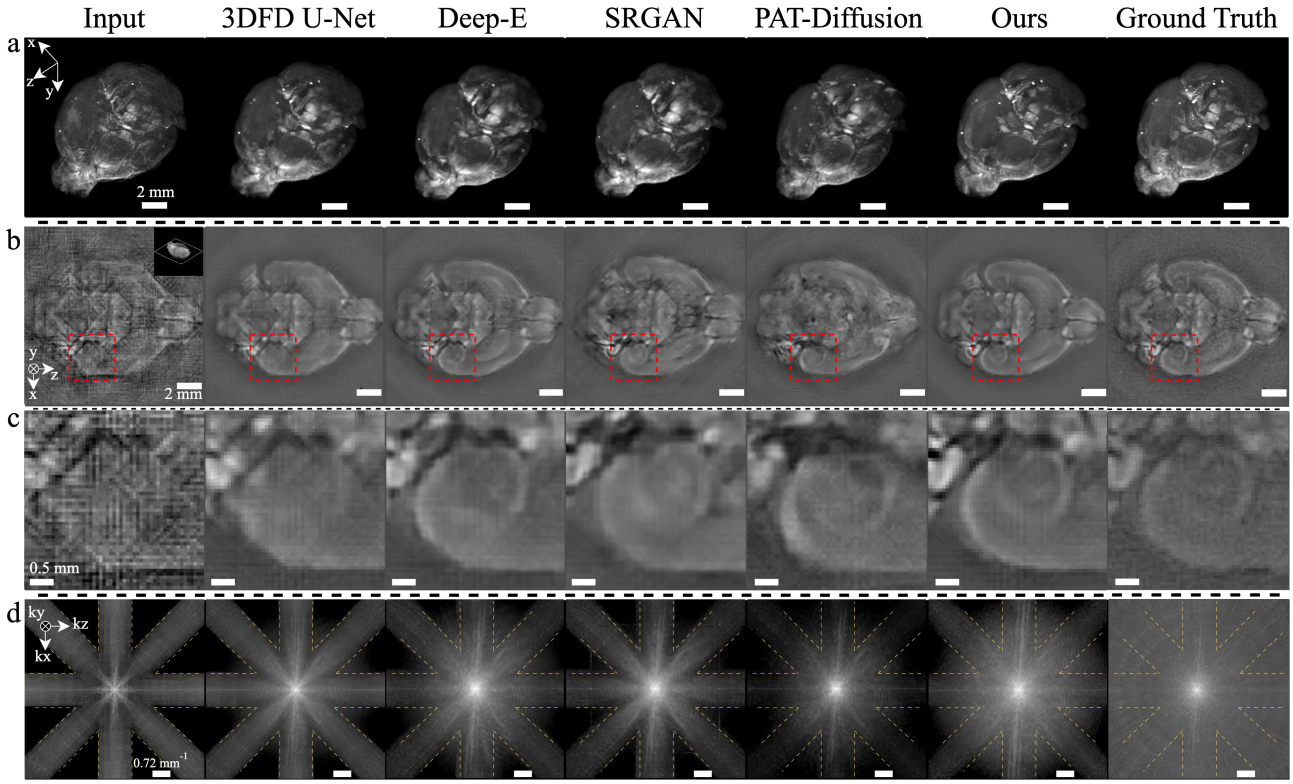

Fig. 4. Qualitative comparison with baselines using the mouse brain dataset. a) Maximum intensity projection (MIP) of the reconstructed brain images. b) 2D slices extracted from the 3D images in (a), as shown in the inserted pannel. c) Zoom in on the red boxes in (b), respectively. d) MIP of spatial frequency spectrum of each image in (a).

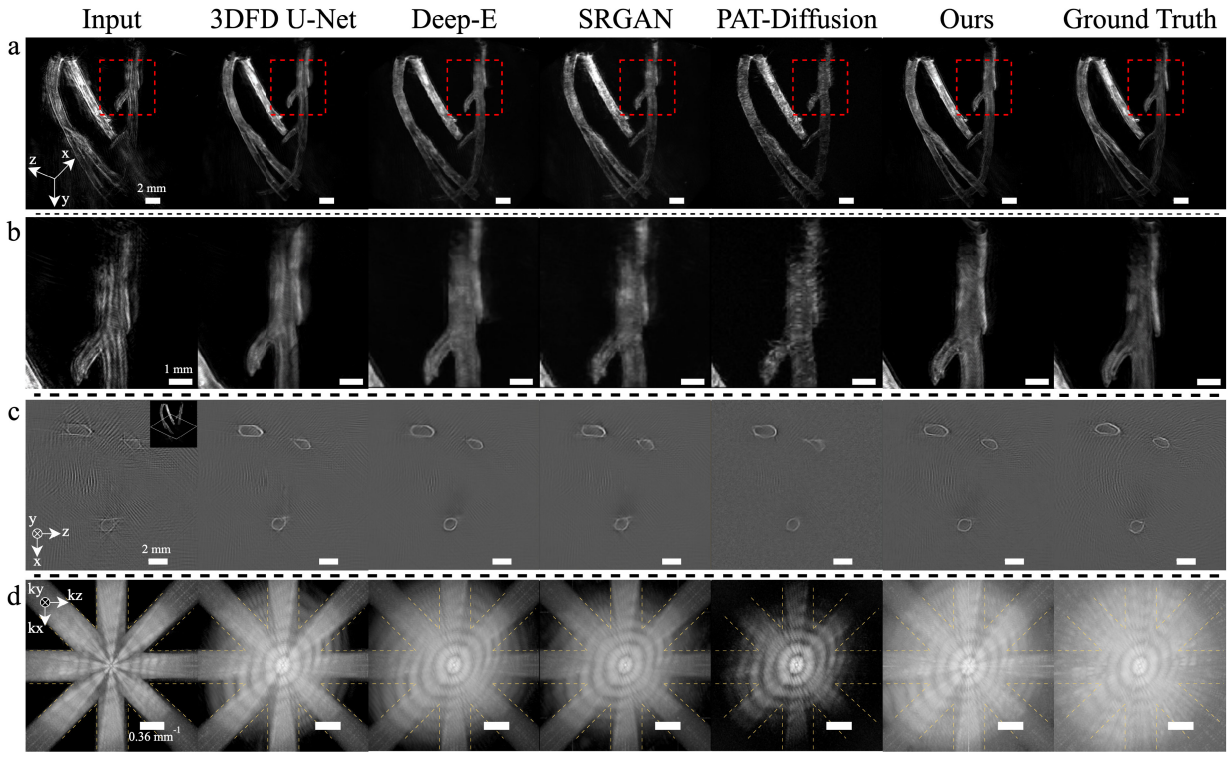

Fig. 5. Model performance evaluation using the vessel phantom dataset. a) MIP of the reconstructed vessel phantom images. b) Zoom in on the red boxes in (a), respectively. c) 2D slices from 3D images in (a), indicated by the inserted pannel. d) MIP of spatial frequency spectrum of each image in (a).

truth. For voxel geometry evaluation, we report Volume IoU, which measures the geometric similarity between the outputs and ground truths. The 3D images generated by our model also outperform those of all baselines.

TABLE I

QUANTITATIVE COMPARISON WITH BASELINES. PERFORMANCE COMPARISON OF DIFFERENT METHODS ON THE EX VIVO MICE BRAIN DATASET AND THE VESSEL PHANTOM DATASET. ALL METRICS ARE COMPUTED ON THE TEST SET, WITH HIGHER VALUES INDICATING BETTER PERFORMANCE FOR PSNR, SSIM, AND VOL. IOU, AND LOWER VALUES INDICATING BETTER PERFORMANCE FOR RMSE

Methods	Brain				Vessel			
	RMSE↓	PSNR↑	SSIM↑	Vol. IoU↑	RMSE↓	PSNR↑	SSIM↑	Vol. IoU↑
3DFD U-Net [32]	0.0569***	29.057***	0.909***	0.756***	0.0540***	26.597***	0.872***	0.778***
Deep-E [30]	0.0358***	30.758***	0.913***	0.763***	0.0419***	29.565***	0.911***	0.808***
SRGAN [41]	0.0357***	30.254***	0.911***	0.781***	0.0528***	28.155***	0.902***	0.792***
PAT-Diffusion [42]	0.0403***	29.930***	0.895***	0.753***	0.0493***	28.173***	0.829***	0.758***
Ours	**0.0249**	**33.063**	**0.947**	**0.824**	**0.0318**	**32.736**	**0.927**	**0.859**

*p<0.05, **p<0.01, ***p<0.001 (based on paired t-tests conducted across all metrics, comparing our method with each baseline.)

TABLE II

EVALUATION OF D-STAR'S GENERALIZATION CAPABILITY AND ADAPTABILITY ON IN VIVO MICE TUMOR DATA

Metric	Input	Zero-Shot	Fine-Tuned
RMSE ↓	0.0879	0.0484 (-0.0395)	**0.0257** (-0.0622)
PSNR ↑	22.009	28.439 (+6.430)	**34.160** (+12.151)
SSIM ↑	0.579	0.929 (+0.350)	**0.961** (+0.382)
Vol. IoU ↑	0.614	0.855 (+0.241)	**0.893** (+0.279)

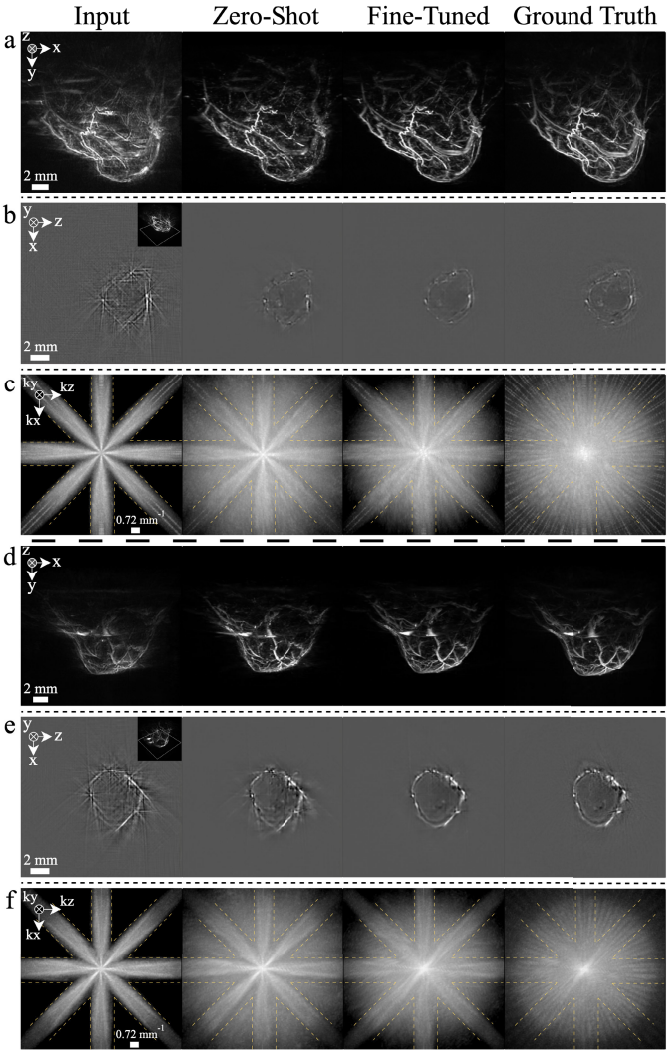

Fig. 6. Performance validation on in vivo mouse tumor data. a) and d) MIP of the reconstructed in vivo tumor images. b) and e) 2D slices from 3D images in (a) and (d), indicated by the inserted pannel. c) and f) MIP of spatial frequency spectrum of each image in (a) and (d).

E. Generalization Capability of D-STAR

We then conducted in vivo experiments using mice tumor dataset, where we first evaluated the performance of D-STAR in a zero-shot setting (using a model trained on the ex vivo brain dataset) and after fine-tuning on the in vivo tumor dataset for only 8 epochs. The quantitative results are summarized in Table II, and qualitative results are presented in Fig. 6. These results demonstrate that even without fine-tuning, the model trained on the brain dataset effectively denoises the sparse-angle input and partially recovers structural details of the tumor, as evident in both 3D and 2D reconstructions, suggesting that D-STAR has inherent generalization capability, even when applied to different anatomical regions without additional training. And with only 8 epochs of fine-tuning on the tumor dataset, D-STAR achieves a substantial improvement across all metrics, recovering fine-scale structures such as tumor morphology and vasculature. The 3D FFT visualization further supports these findings, showing that the fine-tuned model effectively completes missing frequency components, enhancing spatial resolution. These results confirm that D-STAR is not limited to a specific dataset or anatomical region but can adapt effectively to new imaging conditions with minimal fine-tuning (only 8 epochs, requiring less than 2 hours on a single NVIDIA A100 GPU), highlighting its potential for real-world clinical applications.

F. Resolution Enhancement Brought by D-STAR

We subsequently performed PSF measurements to demonstrate the substantial improvements in spatial and temporal resolution achieved by our method, addressing the inherent trade-offs typically encountered in imaging systems. To approximate real imaging conditions, we selected a microbubble near the brain from our test dataset for PSF evaluation (Fig. 7 (a)). Artifacts were clearly evident in the data acquired with 4 angles, whereas D-STAR produced images with spatial sharpness comparable to those obtained using the full 32 angles (Fig. 7 (b)). Quantitatively, each volume was fitted to a 3D-Gaussian distribution to determine the full width at half maximum (FWHM). D-STAR enhanced the spatial response bandwidth by up to 6.9 times, with no compromise in imaging speed. This advancement results in

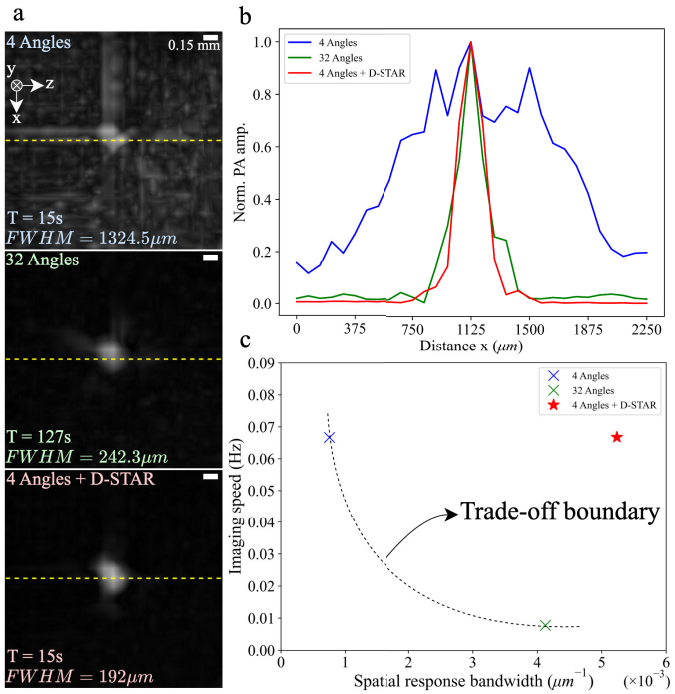

Fig. 7. Resolution Enhancement brought by D-STAR. a) MIP of PSFs of 4-angle image, 32-angle image and D-STAR output. b) PA amplitude profile along the dashed yellow lines in (a). Blue, green and red curves represent the input, ground truth and output, respectively. c) Comparison on imaging speed and spatial response bandwidth.

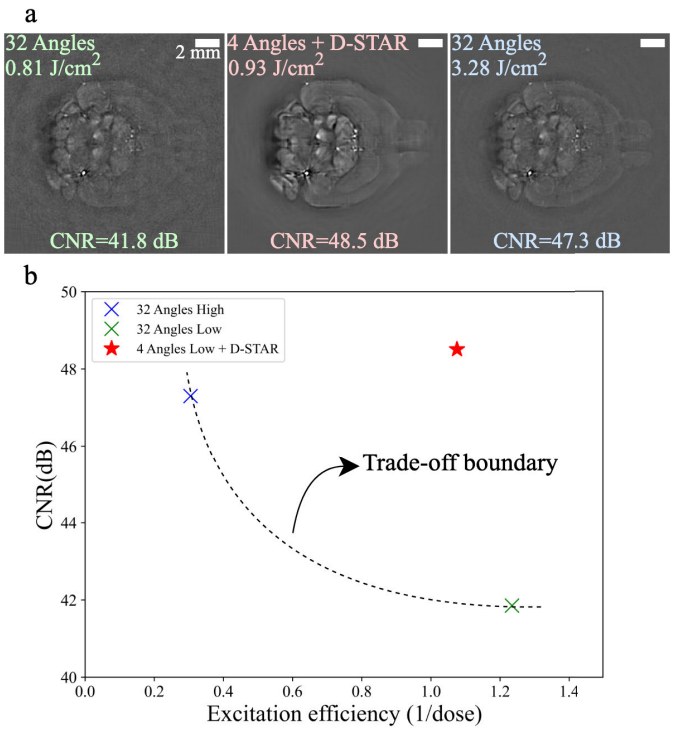

Fig. 8. Low-Dose, High-Contrast Imaging with D-STAR. a) Transectional slices of a brain imaged under conditions of 32 angles with low dose, D-STAR with low dose and 32 angles with high dose, respectively. b) Comparison on excitation efficiency and CNR.

an imaging approach that is 7 times faster than the full angular imaging method, effectively breaking the spatial and

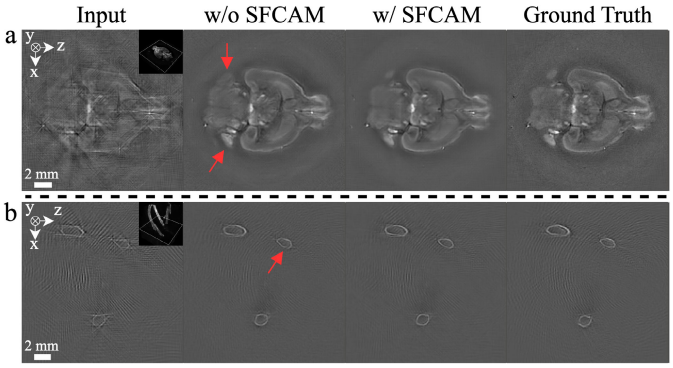

Fig. 9. The Shape Feature Cross-Attention Module (SFCAM) is beneficial for our model. a) Without SFCAM, the predicted brain misses key structures and fails to restore the overall outline. With SFCAM, the result is more accurate. b) For vessel reconstruction, the result without SFCAM shows twisted and unsealed shape, while SFCAM improves shape consistency and completeness.

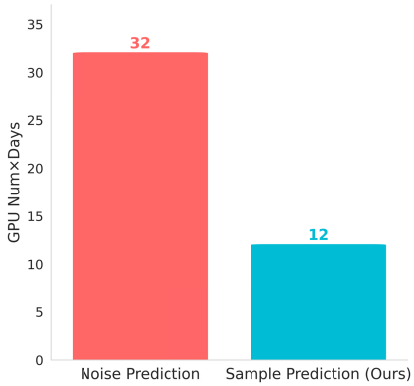

Fig. 10. Training cost comparison on brain dataset. Sample Prediction require much less computation cost than Noise Prediction.

temporal bandwidth product constraint of our system (a dashed reciprocal curve in Fig. 7 (c)).

G. Low-Dose, High-Contrast Imaging With D-STAR

Despite certain exceptions, PAT can generally be considered a linear system with respect to excitation laser energy [44]. Consequently, when the total exposure of laser energy is constrained, it is preferable to use fewer imaging acquisitions with higher pulse energy rather than more acquisitions with lower pulse energy. This preference arises because the noise level in imaging is inversely proportional to the square root of the number of imaging acquisitions. This observation led us to hypothesize that with fewer scanning angles, the SNR could be enhanced by a factor up to the square root of the compression ratio (32 angles / 4 angles = 8 in this study). To test this hypothesis, we compared the contrast-to-noise ratio (CNR) of the images and the corresponding total excitation dose acquired under different imaging parameters for the same brain: 32 angles with an excitation laser fluence of 0.79 mJ/cm^2, 4 angles with 7.3 mJ/cm^2, and 32 angles with 3.2 mJ/cm^2 (Fig. 8 (a)). For ease of comparison, we defined excitation efficiency as the inverse of the total excitation dose. Under conditions of high excitation efficiency, D-STAR yielded a significant increase in CNR—6.7 dB and 1.2 dB greater compared to the 32 angles with low and high excitation doses, respectively (Fig. 8 (b)). This improvement exceeded

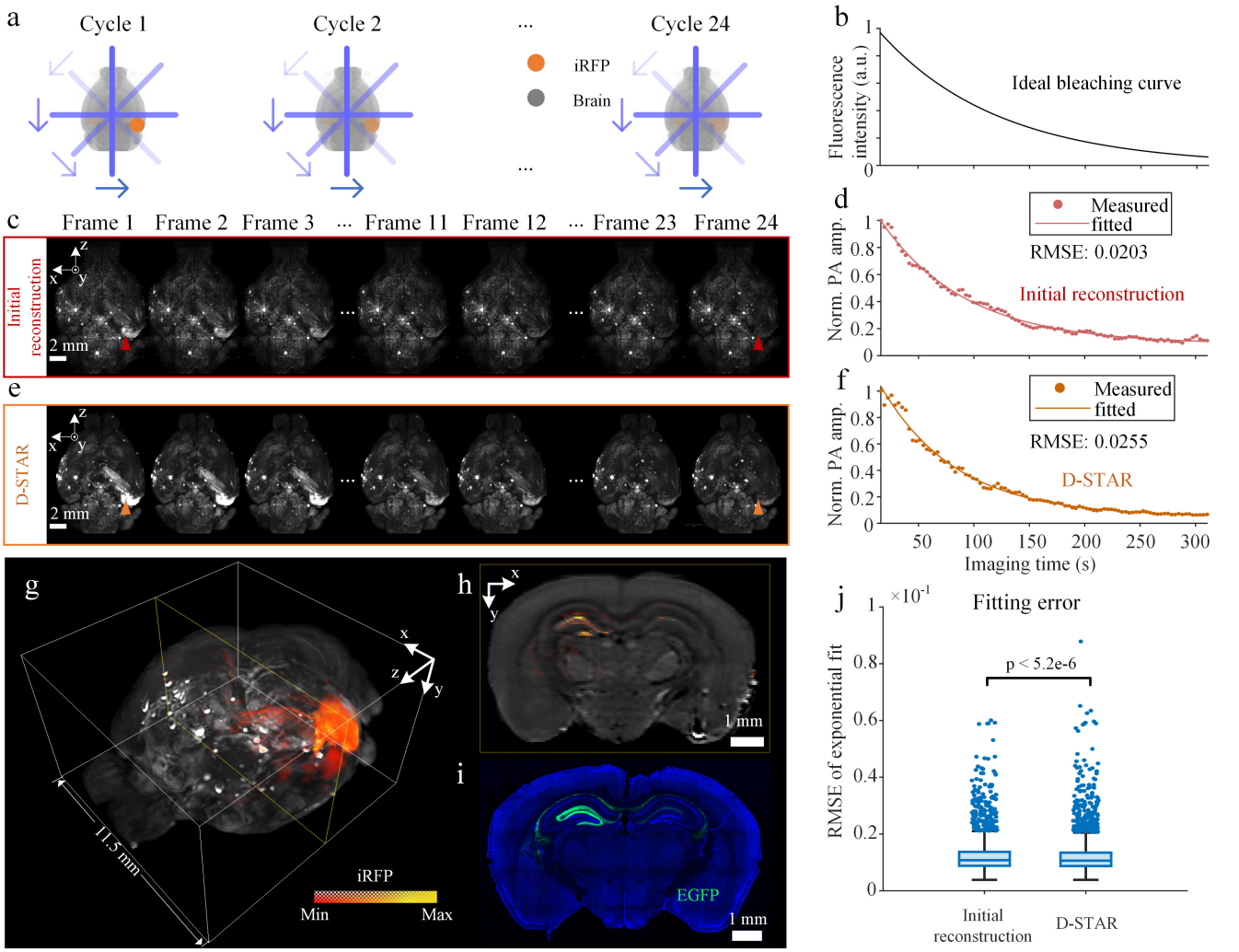

Fig. 11. Employment in molecular photoacoustic imaging using D-STAR. a) Imaging sequence for temporal unmixing. b) Expected bleaching curve. c) 3D-image stacks of initial reconstruction. d) Bleaching curve referred to the arrow in (c). e) 3D-image stacks of D-STAR output. f) Bleaching curve referred to the arrow in (e). g) 3D-rendered volume of unmixed iRFP (red-yellow) and tissue background (grayscale). h) Tomographic view related to the plane in (g). i) The same slice in (h) via confocal microscopy. j) Comparison of fitting errors between initial reconstruction and D-STAR.

the expected gain constrained by the square root of the compression ratio, which would predict increases of 5.1 dB and −0.96 dB after adjusting for total excitation dose. This evidence suggests that D-STAR effectively reduces image noise, thereby extending the trade-off boundary (a dashed exponentially decaying curve in Fig. 8 (b)) between image contrast and excitation efficiency.

TABLE III
ABLATION ANALYSIS ON BRAIN DATASET. THE TABLE ILLUSTRATES THE DIFFERENCES IN MODEL PERFORMANCE AFTER REMOVING THE DESIGN OF TWO KEY MODULES. HERE, SFCAM DENOTES THE SHAPE FEATURE CROSS-ATTENTION MODULE

Methods	RMSE↓	PSNR↑	SSIM↑	Vol. IoU↑
w/o SFCAM	0.0278	32.039	0.919	0.798
w/o Zero-SNR	0.0262	32.912	0.942	0.820
Ours	**0.0249**	**33.063**	**0.947**	**0.824**

H. Ablation Study and Analysis

1) Design of Conditional Diffusion: Here we examine the effectiveness of the design of conditional diffusion models. Table III provides a quantitative analysis of our Zero-SNR strategy and Shape Feature Cross-Attention Module (SFCAM). Removing SFCAM (denoted as w/o SFCAM) results in significant performance degradation, indicating that SFCAM plays a crucial role in preservation of structural details. Fig. 9 further provides qualitative evidence of the importance of SFCAM. Without this module, the reconstructed brain images fail to preserve structural integrity, and the vessel predictions exhibit distortions and discontinuities. By incorporating SFCAM, the model successfully restores the missing structures, leading to more accurate and reliable reconstructions. Similarly, removing Zero-SNR (w/o Zero-SNR) leads to a drop in performance across all metrics, confirming its effectiveness in enhancing image quality. It can be seen that both the Zero-SNR trick and SFCAM are beneficial.

2) Importance of Sample Prediction: We examine the importance of the sample prediction that is direct prediction of images instead of noise. To compare, we train an ϵ-Prediction model and a $\boldsymbol{x}^0$-Prediction model on brain dataset. The results are shown in Table IV and Fig. 10. It can be seen that

TABLE IV
ABLATION STUDY ON THE DESIGN OF SAMPLE PREDICTION. WE REPORT THE METRICS OF RMSE, PSNR, SSIM AND VOLUME IOU ON BRAIN DATASET

Methods	RMSE↓	PSNR↑	SSIM↑	Vol. IoU↑
w/ ϵ-Prediction	0.0392	31.191	0.931	0.809
w/ $\boldsymbol{x}^0$-Prediction	**0.0249**	**33.063**	**0.947**	**0.824**

$\boldsymbol{x}^0$-Prediction provides a more accurate reconstruction with a faster convergence. This is because sample prediction avoids the challenges of learning complex, non-Gaussian noise distributions in sparse-angle photoacoustic reconstruction, and instead focuses on structural restoration.

I. Employment in Molecular Photoacoustic Imaging

Quantitative photoacoustic imaging, which aims to extract molecular or functional information for each pixel or voxel from multiple frames of the same object, is crucial for various applications [45], [46], [47]. In such scenarios, pixel values across different frames are subject to physical constraints. For example, in bleaching-based temporal unmixing, the photoacoustic amplitude values of sequential frames are expected to exhibit an exponential decay [48], a characteristic that the network does not learn inherently. To assess whether D-STAR can address the inter-frame constraints using only spatial information, we applied it to the unmixing of a fluorescent protein expressed in a mouse brain. Initially, we designed an angular scan sequence consisting of 0°, 90°, 45°, and 135°, with a total of 24 cycles (Fig. 11 (a)). This configuration was intended to minimize errors in the fused photoacoustic amplitude due to anisotropic spatial frequency response and to ensure complete bleaching of the protein (Fig. 11 (b)). The imaging data were processed according to Section III-B, resulting in an initial reconstruction (Fig. 11 (c)). Subsequently, the bleaching curve for each voxel was determined using these frames (Fig. 11 (d)). This procedure was repeated for high-resolution images generated by D-STAR, using the initial reconstruction frames as input (Fig. 11 (e)). The RMSE value of the exponential fitting of the curves indicated minimal deviations from the expected exponential model, attributable to the D-STAR approach (Fig. 11 (f)). The fitting results were then employed to produce the unmixing outcomes, with iRFP visualized in red-yellow and the tissue background in grayscale. D-STAR effectively resolved the neuronal projections from the medial entorhinal cortex to the dentate gyrus (Fig. 11 (g, h)). To further validate spatial precision, brain slices of the same sample were prepared and imaged using confocal microscopy. The unmixed PAT images demonstrated consistent fluorescent signals with the confocal images, which served as a reliable reference (Fig. 11 (h, i)). Concurrently, the fitting RMSE showed a significant reduction with D-STAR, demonstrating its high fidelity in the temporal domain (Fig. 11 (j)).

V. CONCLUSION AND DISCUSSION

Multiangle imaging is a widely utilized methodology that trades off temporal resolution and excitation dose to achieve enhanced spatial resolution and imaging SNR. Although the priorities for these factors vary depending on specific applications, they all hold significant importance in the realm of bioimaging. In this work, we introduce D-STAR, a suite of deep generative model-based methods that substantially reduce the required number of angles while maintaining image quality comparable to full tomographic angular imaging. This approach provides an effective means to balance these imaging factors. We demonstrate the high fidelity of D-STAR, which is capable not only of recovering static structural information but also of extracting quantitative data from multiple frames.

To the best of our knowledge, we are the first to introduce diffusion model into 3D photoacoustic image reconstruction task. Our proposed method, D-STAR, has achieved optimal performance in both quantitative and qualitative comparisons. Previous U-shaped approaches, such as 3D-pU-Net [23], which uses a multi-stage progressive learning method to recover high-resolution images from low-resolution inputs, and 3DFD U-Net [32], specifically designed to improve the quality of vasculature photoacoustic images, have made notable progress in similar 3D super-resolution tasks. However, these networks have a common limitation, that is, they generate output in a single pass during inference, which makes it difficult to accurately reconstruct complex structures and often suffers for missed details. In contrast, our model performs multi-step denoising, allowing for a more organized reconstruction process that better captures intricate structures. Instead of requiring the model to remember the entire object and reconstruct it directly, we empirically believe it is more reasonable and efficient for the model to focus on what needs to be done at each step. Furthermore, in this conditional reconstruction task, we apply the condition at each timestep of the denoising process, guiding the model step-by-step. Given the complex data distribution in the STAR-PAT dataset, the diffusion model is trained by maximizing the data likelihood, which enhances its ability to generalize and handle these intricate distributions more effectively. To improve robustness against slight inconsistencies in the series of reconstructed images from the same sample, we added a small amount of Gaussian noise to the inputs during both the training and inference phases. The reliability of D-STAR is also evident in molecular imaging, which leverages a series of images of the same sample to extract additional information. Notably, the output from D-STAR adheres to the same exponential decay constraints as the input, yielding a more accurate estimation of bleaching extent. This accuracy is primarily attributed to the low noise output of D-STAR and the high-quality training dataset obtained from real-world conditions, which encapsulate the underlying physical laws. To further enhance the fidelity of D-STAR, integrating such physical priors into the model for multi-frame reconstruction could be beneficial.

Although theoretically capable of providing optimal imaging quality, temporal and spatial resolutions of 3D PAT systems using hemispherical ultrasound transducer arrays are still ultimately constrained by the channel count of the data acquisition system, which can be costly to augment [49], [50]. In contrast, D-STAR offers a cost-effective alternative that delivers high-performance imaging compatible with hemispherical systems. The cylindrical-focused ultrasound

transducer arrays used in D-STAR are established commercial products, which promotes wider adoption. Furthermore, the insensitivity of the approach to prior knowledge regarding the PSF alleviates some of the complexities associated with PAT, resulting in a computationally efficient reconstruction process. In addition, although D-STAR was designed and validated for the PAT modality, it can be generalized to various imaging techniques. The prerequisites for its application include: (1) the ability of the imaging system to generate 3D data stacks from multiple angles, (2) minimal distortion and adequate overlap among these angles to permit preprocessing and registration, and (3) a linear or otherwise resolvable response function for multi-image fusion.

We have demonstrated that D-STAR effectively eliminates the need to compromise on temporal resolution or SNR to achieve improvements in spatial resolution and sample health. This capability is particularly valuable in clinical settings, where maintaining high image quality while minimizing patient exposure to excitation doses is crucial for both diagnostic accuracy and safety. By leveraging existing hardware more efficiently, D-STAR reduces the need for costly equipment upgrades, making high-quality imaging more accessible and cost-effective in clinical environments. While D-STAR is not intended to compete with advanced full tomographic angular imaging methods, it provides additional flexibility in adjusting imaging parameters, allowing clinicians to tailor protocols to specific imaging needs or clinical scenarios. For instance, reduced excitation laser exposure can benefit imaging of light-sensitive monoclonal agents, such as photo-switchable proteins [54], [55], [56]. Additionally, enhanced temporal resolution afforded by D-STAR can facilitate the observation of rapid physiological phenomena such as cardiac rhythms [23], [51], vascular perfusion [52], and drug metabolism [53]. Consequently, integrating D-STAR into existing imaging frameworks, such as PAT, can expand the applicability and effectiveness of these techniques. By improving image quality while reducing excitation doses, D-STAR not only enhances patient safety and comfort but also opens up new possibilities for longitudinal studies in both clinical and research settings.

REFERENCES

[1] M. Lee, K. Kim, J. Oh, and Y. Park, "Isotropically resolved label-free tomographic imaging based on tomographic moulds for optical trapping," *Light, Sci. Appl.*, vol. 10, no. 1, p. 102, May 2021.

[2] P. J. Verveer, J. Swoger, F. Pampaloni, K. Greger, M. Marcello, and E. H. K. Stelzer, "High-resolution three-dimensional imaging of large specimens with light sheet–based microscopy," *Nature Methods*, vol. 4, no. 4, pp. 311–313, Apr. 2007.

[3] B. Chen et al., "Resolution doubling in light-sheet microscopy via oblique plane structured illumination," *Nature Methods*, vol. 19, no. 11, pp. 1419–1426, Nov. 2022.

[4] C. Demené et al., "4D microvascular imaging based on ultrafast Doppler tomography," *NeuroImage*, vol. 127, pp. 472–483, Feb. 2016.

[5] Y. Chen et al., "Photoacoustic tomography with temporal encoding reconstruction (PATTERN) for cross-modal individual analysis of the whole brain," *Nature Commun.*, vol. 15, no. 1, p. 4228, May 2024.

[6] L. Lin and L. V. Wang, "The emerging role of photoacoustic imaging in clinical oncology," *Nature Rev. Clin. Oncol.*, vol. 19, no. 6, pp. 365–384, Jun. 2022.

[7] J. Park et al., "Clinical translation of photoacoustic imaging," *Nature Rev. Bioengineering*, vol. 3, no. 3, pp. 193–212, Sep. 2024.

[8] X. Li et al., "Three-dimensional structured illumination microscopy with enhanced axial resolution," *Nature Biotechnol.*, vol. 41, no. 9, pp. 1307–1319, Jan. 2023.

[9] F. A. Duck, "Medical and non-medical protection standards for ultrasound and infrasound," *Prog. Biophys. Mol. Biol.*, vol. 93, nos. 1–3, pp. 176–191, Jan. 2007.

[10] J. Swoger, P. J. Verveer, K. Greger, J. Huisken, and E. H. K. Stelzer, "Multi-view image fusion improves resolution in three-dimensional microscopy," *Opt. Exp.*, vol. 15, no. 13, p. 8029, Jun. 2007.

[11] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[12] M. Nikolova, "An algorithm for total variation minimization and applications," *J. Math. Imag. Vis.*, vol. 20, no. 1, pp. 89–97, Jan. 2004.

[13] A. Creswell et al., "Generative adversarial networks: An overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, Jan. 2018.

[14] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[15] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.

[16] C. Lu et al., "Diffusion-based deep learning method for augmenting ultrastructural imaging and volume electron microscopy," *Nature Commun.*, vol. 15, no. 1, p. 4677, Jun. 2024.

[17] M. Weigert et al., "Content-aware image restoration: Pushing the limits of fluorescence microscopy," *Nature Methods*, vol. 15, no. 12, pp. 1090–1097, Dec. 2018.

[18] C. Qiao et al., "Rationalized deep learning super-resolution microscopy for sustained live imaging of rapid subcellular processes," *Nature Biotechnol.*, vol. 41, no. 3, pp. 367–377, Mar. 2023.

[19] C. Qiao et al., "Evaluation and development of deep neural networks for image super-resolution in optical microscopy," *Nature Methods*, vol. 18, no. 2, pp. 194–202, Feb. 2021.

[20] H. Park et al., "Deep learning enables reference-free isotropic super-resolution for volumetric fluorescence microscopy," *Nature Commun.*, vol. 13, no. 1, p. 3297, Jun. 2022.

[21] L. V. Wang and J. Yao, "A practical guide to photoacoustic tomography in the life sciences," *Nature Methods*, vol. 13, no. 8, pp. 627–638, Aug. 2016.

[22] L. Lin et al., "High-speed three-dimensional photoacoustic computed tomography for preclinical research and clinical translation," *Nature Commun.*, vol. 12, no. 1, pp. 23–31, Feb. 2021.

[23] S. Choi et al., "Deep learning enhances multiparametric dynamic volumetric photoacoustic computed tomography in vivo (DL–PACT)," *Adv. Sci.*, vol. 10, no. 1, Jan. 2023, Art. no. 2202089.

[24] A. P. Jathoul et al., "Deep in vivo photoacoustic imaging of mammalian tissues using a tyrosinase-based genetic reporter," *Nature Photon.*, vol. 9, no. 4, pp. 239–246, Mar. 2015.

[25] L. Li et al., "Single-impulse panoramic photoacoustic computed tomography of small-animal whole-body dynamics at high spatiotemporal resolution," *Nature Biomed. Eng.*, vol. 1, no. 5, p. 0071, May 2017.

[26] K. S. Valluru, K. E. Wilson, and J. K. Willmann, "Photoacoustic imaging in oncology: Translational preclinical and early clinical experience," *Radiology*, vol. 280, no. 2, pp. 332–349, Aug. 2016.

[27] P. K. Upputuri and M. Pramanik, "Recent advances toward preclinical and clinical translation of photoacoustic tomography: A review," *J. Biomed. Opt.*, vol. 22, no. 4, Nov. 2016, Art. no. 041006.

[28] B. T. Cox, S. R. Arridge, and P. C. Beard, "Photoacoustic tomography with a limited-aperture planar sensor and a reverberant cavity," *Inverse Problems*, vol. 23, no. 6, pp. S95–S112, Dec. 2007.

[29] Y. Xu, L. V. Wang, G. Ambartsoumian, and P. Kuchment, "Reconstructions in limited-view thermoacoustic tomography," *Med. Phys.*, vol. 31, no. 4, pp. 724–733, Apr. 2004.

[30] H. Zhang et al., "Deep-E: A fully-dense neural network for improving the elevation resolution in linear-array-based photoacoustic tomography," *IEEE Trans. Med. Imag.*, vol. 41, no. 5, pp. 1279–1288, May 2022.

[31] W. Zheng et al., "Deep-E enhanced photoacoustic tomography using three-dimensional reconstruction for high-quality vascular imaging," *Sensors*, vol. 22, no. 20, p. 7725, Oct. 2022.

[32] W. Zheng et al., "Deep learning enhanced volumetric photoacoustic imaging of vasculature in human," *Adv. Sci.*, vol. 10, no. 29, Aug. 2023, Art. no. 2301277.

[33] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.

[34] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4713–4726, Apr. 2022.

[35] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.
[36] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[37] S. Lin, B. Liu, J. Li, and X. Yang, "Common diffusion noise schedules and sample steps are flawed," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 5404–5411.
[38] J. Yao, L. Wang, C. Li, C. Zhang, and L. V. Wang, "Photoimprint photoacoustic microscopy for three-dimensional label-free subdiffraction imaging," *Phys. Rev. Lett.*, vol. 112, no. 1, Jan. 2014, Art. no. 014302.
[39] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2016, pp. 424–432.
[40] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," 2020, *arXiv:2010.02502*.
[41] D. He, J. Zhou, X. Shang, X. Tang, J. Luo, and S.-L. Chen, "De-noising of photoacoustic microscopy images by attentive generative adversarial network," *IEEE Trans. Med. Imag.*, vol. 42, no. 5, pp. 1349–1362, May 2023.
[42] K. Guo et al., "Score-based generative model-assisted information compensation for high-quality limited-view reconstruction in photoacoustic tomography," *Photoacoustics*, vol. 38, Aug. 2024, Art. no. 100623.
[43] D. Kong and Y. Chen. (2024). *Open3DFDUNet: Open-Source 3D Fully-Dense U-Net*. [Online]. Available: https://github.com/thu-bplab/Open3DFDUNet
[44] J. Yao et al., "High-speed label-free functional photoacoustic microscopy of mouse brain in action," *Nature Methods*, vol. 12, no. 5, pp. 407–410, May 2015.
[45] L. Tan et al., "Non-invasive optoacoustic imaging of glycogen-storage and muscle degeneration in late-onset pompe disease," *Nature Commun.*, vol. 15, no. 1, p. 7843, Sep. 2024.
[46] V. Ntziachristos and D. Razansky, "Molecular imaging by means of multispectral optoacoustic tomography (MSOT)," *Chem. Rev.*, vol. 110, no. 5, pp. 2783–2794, May 2010.
[47] G. Diot et al., "Multispectral optoacoustic tomography (MSOT) of human breast cancer," *Clin. Cancer Res.*, vol. 23, no. 22, pp. 6912–6922, 2017.
[48] J. Weber, P. C. Beard, and S. E. Bohndiek, "Contrast agents for molecular photoacoustic imaging," *Nature Methods*, vol. 13, no. 8, pp. 639–650, Aug. 2016.
[49] S. Gottschalk et al., "Rapid volumetric optoacoustic imaging of neural dynamics across the mouse brain," *Nature Biomed. Eng.*, vol. 3, no. 5, pp. 392–401, Mar. 2019.
[50] R. Cao et al., "Single-shot 3D photoacoustic computed tomography with a densely packed array for transcranial functional imaging," 2023, *arXiv:2306.14471*.
[51] Ç. Özsoy, A. Özbek, M. Reiss, X. L. Deán-Ben, and D. Razansky, "Ultrafast four-dimensional imaging of cardiac mechanical wave propagation with sparse optoacoustic sensing," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 45, Nov. 2021, Art. no. e2103979118.
[52] J. Ahn, J. Y. Kim, W. Choi, and C. Kim, "High-resolution functional photoacoustic monitoring of vascular dynamics in human fingers," *Photoacoustics*, vol. 23, Sep. 2021, Art. no. 100282.
[53] Z. Yang et al., "Stimuli-responsive nanotheranostics for real-time monitoring drug release by photoacoustic imaging," *Theranostics*, vol. 9, no. 2, pp. 526–536, 2019.
[54] L. Li et al., "Small near-infrared photochromic protein for photoacoustic multi-contrast imaging and detection of protein interactions *in vivo*," *Nature Commun.*, vol. 9, no. 1, p. 2734, Jul. 2018.
[55] J. Yao et al., "Multiscale photoacoustic tomography using reversibly switchable bacterial phytochrome as a near-infrared photochromic probe," *Nature Methods*, vol. 13, no. 1, pp. 67–73, Jan. 2016.
[56] K. Mishra et al., "Multiplexed whole-animal imaging with reversibly switchable optoacoustic proteins," *Sci. Adv.*, vol. 6, no. 24, Jun. 2020, Art. no. eaaz6293.