Cross-Channel Consistent Mamba for Extreme Limited-View Photoacoustic Tomography Reconstruction

Shicheng Xu[1,2,3], Jincheng Li[1,2,3], Hengrong Lan[1,2,3⋆], and Fei Gao[1,2,3,4*]

[1] School of Biomedical Engineering, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, Anhui, China
[2] Hybrid Imaging System Laboratory, Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, Jiangsu, China
[3] Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou, Jiangsu 215123, China
[4] School of Engineering Science, University of Science and Technology of China, Hefei, Anhui, China
{xushicheng,jc.lee}@mail.ustc.edu.cn
{hrlan,fgao}@ustc.edu.cn

Abstract. Extreme limited-view photoacoustic tomography reconstruction suffers from severely missing features and dominant strong artifacts, which often leads to structural discontinuities, blurred boundaries, and loss of fine details. Moreover, channel availability can vary in real acquisitions, making methods that rely on early multi-channel concatenation and fixed input configurations less stable. To this end, we propose S^2-Mamba, which decouples the multi-channel input into multiple single-channel branches to preserve weak but informative cues within each channel. This design naturally accommodates missing channels and changes in acquisition configurations. Furthermore, we introduce a Cross-Channel Constrained Loss that explicitly aligns multi-channel representations at the latent distribution level and mitigates representation collapse through a uniform prior constraint. We also propose an Adaptive Gradient-Gated Loss that uses progressive hard mining to emphasize structurally relevant difficult regions and enables joint optimization via uncertainty-adaptive weighting. Under an extreme limited-view setting on an *in vivo* mice dataset, we train the model using 4-channel (11.25°) observations. The results show that our method outperforms a range of representative networks in terms of SSIM, PSNR, FID, and KID, and remains superior even when the channel configuration differs between training and testing, demonstrating its robustness and practical deployability.

Keywords: Photoacoustic tomography · extreme limited-view reconstruction · gradient gating · cross-channel alignment.

⋆ Corresponding authors: Hengrong Lan and Fei Gao.

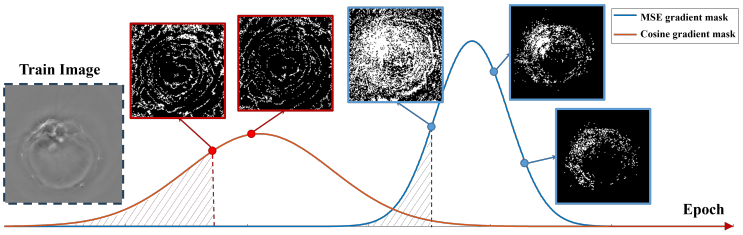

Fig. 1. Adaptive Gradient-Gated Loss that progressively suppresses gradients from easy pixels to shift optimization toward structurally hard locations.

1 Introduction

Photoacoustic tomography (PAT) is a hybrid modality that converts optical absorption into ultrasound via nanosecond pulsed excitation, allowing high-resolution imaging of tissue absorption [17]. In practical systems, due to constraints in probe geometric coverage, spatial accessibility, and clinical operation, detection angles are often incomplete, leading to missing acoustic information under limited-view acquisition [21]. This in turn produces striated artifacts, structural discontinuities, and blurred boundaries, substantially degrading geometric fidelity and quantitative reliability [28]. Conventional reconstruction methods, such as Delay-and-Sum [25], Filtered Backprojection [3], Time Reversal [19], and prior-guided iterative regularization frameworks [16], can be effective under full sampling; however, when views are missing, they struggle to compensate for the intrinsic loss of information. Moreover, iterative approaches are typically computationally expensive and sensitive to hyperparameters, and are prone to over-smoothing or structural misinterpretation [4,2]. In recent years, deep learning has been widely used for limited-view compensation, spanning image-domain artifact suppression, end-to-end reconstruction, and strategies for handling missing angles [7,11]. Despite these advances, existing methods largely focus on relatively mild limited-view settings and commonly rely on stable multi-channel input configurations and sufficiently reliable supervised references, leaving a gap to real-world scenarios with more extreme view loss, artifact-dominated measurements, and varying channel availability.

To address the scarcity of informative cues and artifact-dominated representations under extreme limited-view conditions, we propose a task-oriented reconstruction framework. First, we propose a Shared Single-Channel Mamba (S^2-Mamba), which preserves weak but informative features while naturally accommodating missing channels and changes in acquisition configurations at inference time. Second, we introduce a Cross-Channel Consistency Loss to explicitly align multi-channel representations at the latent distribution level, together with a uniform-prior regularization to mitigate representation collapse. Finally, as illustrated in Fig. 1, we propose an Adaptive Gradient-Gated Loss that progressively suppresses gradients from easy-to-fit pixels via hard mining, shifting the optimization focus toward structurally hard locations, and uses homoscedastic-

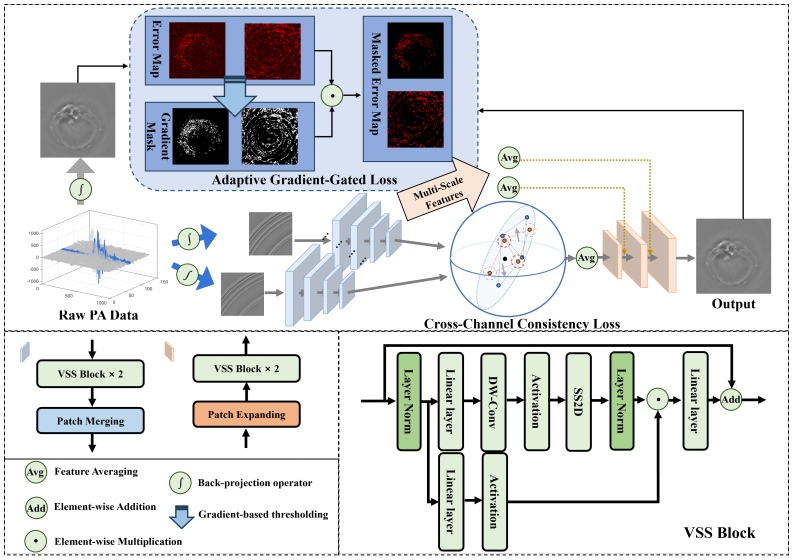

Fig. 2. S^2-Mamba uses a shared single-channel encoder and scale-wise feature averaging for multi-channel fusion. The Cross-Channel Consistency Loss aligns latent feature distributions with a uniform prior, and the Adaptive Gradient-Gated Loss progressively suppresses gradients from easy pixels to focus learning on structurally hard locations.

uncertainty-based adaptive weighting to balance amplitude fidelity and structural consistency.

2 Method

2.1 S^2-Mamba: Shared Single-Channel Mamba Reconstruction Network

Under extreme limited view sampling, observations are sparse and artifacts often dominate the measurements. If multi-channel inputs are concatenated and encoded early, artifacts can be embedded into low level features and overwhelm the weak but informative cues in each channel, which limits boundary and fine detail recovery. To mitigate this, we propose S^2-Mamba, a shared single channel reconstruction network that splits the input into per channel branches and applies a weight shared encoder to extract features independently for each channel, then maps them into an alignable feature space for stable fusion. Because the encoder parameters are shared and fusion does not depend on a fixed channel dimension, the model naturally supports missing channels and changing acquisition configurations at inference time. We adopt Mamba [1] as the core building block due to its linear complexity and strong global modeling ability, which has been shown effective for preserving global structures and boundary continuity

in medical images [15,24] and for recovering slender vessels [20]. As illustrated in Fig. 2, we build the network using the VSS block [15].

The network uses an encoder-decoder architecture. Both encoder and decoder are composed of VSS blocks, with downsampling and upsampling implemented by patch merging and patch expanding. The i-th encoder stage is denoted as $E_i(\cdot)$ and the j-th decoder stage as $D_j(\cdot)$. Given N single channel inputs $X_{n=1}^{N}$, we construct one branch per channel while sharing the same encoder parameters across branches. Let f_i^n be the feature of channel n at scale i, then $f_i^n = E_i(f_{i-1}^n), n = 1, ..., N$. At each scale, we fuse features by mean aggregation to obtain a scale consistent representation $\bar{f}_i = \frac{1}{N}\sum_{n=1}^{N} f_i^n$ The decoder reconstructs the output through a single pathway and uses skip connections for detail recovery. Different from common designs, skip features are also cross channel fused. At decoding scale j, we update $g_j = D_j\left(g_{j-1} + \bar{f}_{i(j)}\right)$, where g_{j-1} is the previous decoding feature and $i(j)$ aligns encoder and decoder scales. This scale wise fusion and additive skip fusion provide structure consistent priors throughout decoding, improving robustness to noise and enhancing continuity and boundary stability for slender vascular structures.

2.2 Cross-Channel Consistency Loss

Multi-channel photoacoustic (PA) observations provide complementary angle dependent information [12], which is crucial under extreme limited view sampling. In S²-Mamba, we fuse channel wise features by scale wise aggregation, and we further add an explicit latent prior to make this fusion stable under noise and artifacts. Existing multi-channel approaches [7,11] often rely on deterministic aggregation or early feature concatenation, so cross channel alignment is learned implicitly and can be unstable. We therefore propose a Cross-Channel Consistency Loss that explicitly aligns channel wise latent feature distributions. Following [22], we compute an augmentation-invariant cross entropy between channels. Let $p^n = \mathrm{softmax}(f^n)$ denote the channel-wise probability, and we compute the pairwise symmetric cross entropy:

$$L_{in\nu} = -\frac{2}{N(N-1)} \sum_{1 \le n_i < n_j \le N} \left(p^{n_i} \log p^{n_j} + p^{n_j} \log p^{n_i}\right) \tag{1}$$

Minimizing cross channel discrepancies alone may induce latent-space collapse, where features from different channels are overly aggregated into near-identical representations, thereby diminishing their complementarity. To prevent such collapse, we further introduce a prior-matching term that drives the overall feature distribution toward a uniform prior. This regularization encourages balanced utilization of the latent space, ensuring that channel-wise features can be drawn closer for alignment without degenerating into excessive aggregation.

$$L_{prior} = -\sum_{c=1}^{C} u_c \log\left(\frac{1}{N}\sum_{n=1}^{N} p_c^n\right) \tag{2}$$

Here, u_c corresponds to a prior probability vector, which is chosen to be uniform in all our experiments, i.e., $u_c = [\frac{1}{C}, \frac{1}{C}, \ldots, \frac{1}{C}]$. The overall objective of the Cross-Channel Consistency Loss is given by: $L_{3C} = L_{inv} + L_{prior}$

2.3 Adaptive Gradient-Gated Loss

PA images are locally smooth yet globally sparse [26]. Under extreme limited view sampling, undersampling artifacts can dominate the error, and optimization may be driven by large easy regions, which biases learning toward smooth backgrounds and low frequency components and weakens boundary and fine detail recovery. Inspired by [9], we introduce an Adaptive Gradient-Gated Loss that progressively suppresses gradients from easy pixels so that training focuses on structurally relevant hard locations. We use pixel wise MSE to enforce amplitude accuracy:

$$L_{mse}(X, Y) = \| X - Y \|_2^2, \tag{3}$$

Here X is the reconstruction and Y is the ground truth. Since amplitude matching is not equivalent to structural matching, MSE alone tends to smooth high frequency details and blur boundaries [23,27]. We therefore add a global cosine similarity term to stabilize the overall energy distribution and relative intensity pattern:

$$L_{\cos}(X, Y) = 1 - \frac{\langle \mathcal{F}(X), \mathcal{F}(Y) \rangle}{\|\mathcal{F}(X)\| \|\mathcal{F}(Y)\|}, \tag{4}$$

Here, $\mathcal{F}(\cdot)$ denotes the flattening operator. This term is more robust to global amplitude scaling, helping to stabilize the overall energy distribution and suppress drifts in structural patterns. Building on this, we perform hard-mining-based gating on the per-pixel loss map M and block gradient flow from easy-to-learn locations:

$$M' = M + (sg(M) - M)H(\frac{M - \mu(M)}{\sigma(M)} - \alpha). \tag{5}$$

Here, sg$(\cdot)$ denotes the stop-gradient operator, $\mu(\cdot)$ and $\sigma(\cdot)$ are the mean and standard deviation, respectively. α is the threshold and $H(\cdot)$ is the Heaviside step function. We denote the hard-mining-modulated losses as L'_{mse} and L'_{cos}. Since the optimal weighting between these two terms is not constant across training stages and is difficult to tune manually, we adopt the homoscedastic-uncertainty-based adaptive weighting method proposed by Kendall et al. [5] to automatically learn their weights:

$$L_{AGG} = \frac{1}{2\varepsilon_1^2} L'_{mse} + \frac{1}{2\varepsilon_2^2} L'_{cos} + \log(\varepsilon_1 \varepsilon_2), \tag{6}$$

Here, ε_1 and ε_2 are learnable parameters. Finally, the overall loss is defined as: $L_{total} = L_{3C} + L_{AGG}$

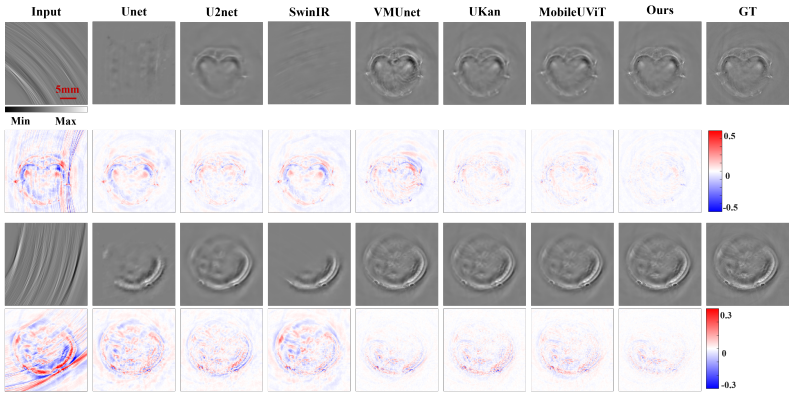

Fig. 3. Visual comparison of reconstruction results. Rows 1 and 3: 4-channel input, 512-channel ground truth (GT), and reconstructions by different methods. Rows 2 and 4: corresponding error maps with respect to the 512-channel GT.

3 Experiment

3.1 Datasets and Implementation Details

We evaluate on a private *in vivo* mice dataset, following the protocol in [6]. The data was collected from four healthy mice using a 512-channel panoramic system (SIPPACT-512), with three mice for training and one for testing, resulting in 3,000 training slices and 400 test slices. More acquisition details can be found in [6]. Although the training and testing animals are independent, the test set comes from one mouse, and broader multi-subject and human PAT validation will be pursued in future work. All experiments were conducted on Ubuntu 20.04.5 LTS with three NVIDIA GeForce RTX 4090 GPUs. All methods were implemented in Python 3.8.20. Unless otherwise specified, we consider a 4-channel (11.25°) limited-view setting, which introduces severe structural missingness and strong undersampling artifacts. During training, the batch size was set to 24 and models were trained for 200 epochs. For fair comparison, all methods were optimized with the same objective and trained using an MSE loss. We report SSIM and PSNR to assess pixel-wise fidelity and structural preservation, and FID and KID to measure perceptual distribution consistency. We schedule the threshold α to increase linearly from -3 to 1 over the 200 training epochs.

3.2 Qualitative and Quantitative Analysis

Qualitative Analysis: As illustrated in Fig. 3, we visually compare our method with six representative reconstruction networks. Under the extreme limited view setting, limited measurements lead to strong directional streak artifacts in the inputs. Unet and SwinIR reduce the overall error but still show structural breaks and missing fine details, while VMUNet tends to introduce pseudo structures.

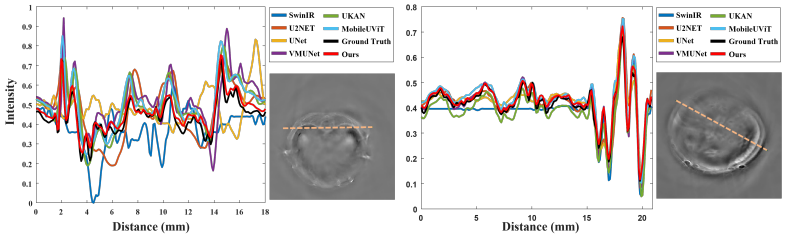

Fig. 4. Intensity profile comparisons along the orange dashed line. Profiles are compared with the 512-channel GT to assess boundary fidelity and internal structural recovery. Broader profiles indicate boundary smoothing or structural spread caused by severe limited-view artifacts, rather than resolution beyond the GT. The scale bar and color bar are consistent with those in Fig. 3.

Table 1. Quantitative comparison results of different methods. The best results are highlighted in bold.

Method	**Single-image metrics**		**Dataset-level metrics**	
	SSIM↑	PSNR↑	FID↓	KID↓
Unet [14]	0.8222	31.79	120.9	0.1111
U2net [13]	0.8626	32.51	182.9	0.1564
SwinIR [10]	0.8215	31.82	99.00	0.0788
VMUNET [15]	0.9493	38.61	56.25	0.0275
UKan [8]	0.9430	36.58	120.7	0.0995
MobileUViT [18]	0.9385	36.90	221.8	0.2238
Ours	**0.9590**	**38.83**	**50.31**	**0.0239**

UKan and MobileUViT yield visually appealing reconstructions, but the error maps reveal remaining deviations. In contrast, our method more reliably restores continuous contours and key details across both examples while suppressing streak artifacts. The error maps further show lower and more dispersed residuals, indicating improved artifact mitigation and structure preservation under severe information loss.

As shown in Fig. 4, we plot intensity profiles along the orange dashed line to compare structural recovery and spatial resolution. In both examples, our profile follows the ground truth more closely, with better matched peak and valley positions and amplitudes, indicating more accurate recovery of boundary and internal structure variations.

Quantitative Analysis: As shown in Table 1, we compare our method with six representative reconstruction networks under the extreme limited view setting. Our method achieves the best structural fidelity, reaching 0.959 SSIM and 38.83 dB PSNR, and also delivers superior perceptual and distributional quality with 50.31 FID and 0.0239 KID. These results indicate that the proposed approach not only improves pixel level accuracy but also better suppresses streak artifacts and preserves realistic structural details, whereas competing methods

Table 2. Ablation study of different components.

Method	Components			Metrics	
	S^2-Mamba	L_{3C}	L_{AGG}	SSIM↑	PSNR↑
A1	✓			0.8798	34.10
A2	✓	✓		0.9016	35.43
A3	✓	✓	✓	**0.9590**	**38.83**

Table 3. Quantitative comparison for channel-robust reconstruction under mismatched channel configurations. All models are trained with 4-channel measurements and evaluated using 2-channel or 3-channel measurements as input.

Method	Channel 2		Channel 3	
	SSIM↑	PSNR↑	SSIM↑	PSNR↑
UKan	0.8972	34.34	0.9227	35.61
VMUNET	0.8965	35.24	0.9260	36.99
MobileUViT	0.8885	34.33	0.9163	35.75
Ours	**0.9219**	**37.75**	**0.9295**	**38.20**

may still present discontinuities, missing fine structures, or pseudo structures under severe information loss.

3.3 Ablation Studies

Component Ablation: Table 2 summarizes the ablation study. The backbone alone provides basic reconstruction through a shared encoder and scale wise mean aggregation, but performance is limited, reaching 0.8798 SSIM and 34.10 dB PSNR. Adding the L_{3C} improves cross view feature alignment by reducing distribution gaps and preventing representation collapse, raising the metrical scores 0.9016 and 35.43 dB. With the L_{AGG} further applied, progressive hard mining emphasizes artifact related difficult regions and uncertainty adaptive weighting balances the reconstruction and consistency objectives. This yields the best results with 0.9590 SSIM and 38.83 dB PSNR.

Channel Robustness: To assess robustness to channel mismatch, we train the model with 4-channel inputs and evaluate reconstruction using only Channel 2 or Channel 3. This setting is challenging for most baselines that require consistent channel numbers across training and inference. Our method remains stable under reduced channels and achieves the best results on both test channels. As shown in Table 3, on Channel 2 we obtain an SSIM of 0.9219 and a PSNR of 37.75 dB, indicating improved structural fidelity and lower reconstruction error.

4 Conclusion

We study extreme limited-view PAT reconstruction, where severe measurement loss causes strong artifacts, and real deployments face varying channel availability. We propose S^2-Mamba, which extracts features from each channel using

a weight-shared Mamba encoder and performs reconstruction via multi-scale mean aggregation and level-wise additive skip fusion, thereby preserving weak but informative cues and enabling stable inference under missing channels or changing acquisition configurations. We further introduce a cross-channel distribution alignment loss with a uniform prior, and an Adaptive Gradient-Gated Loss for progressive hard mining of structurally challenging regions. Experiments on *in vivo* mice data showed consistent and significant gains over representative baselines under extreme limited-view settings, including cross-channel train–test mismatches, demonstrating improved reconstruction quality and robustness. This study focuses on extreme limited-view PAT with severe angular information loss and channel mismatch. Extending the proposed framework to other reconstruction tasks and backbone architectures will be explored in future work.

Acknowledgments

This work was supported in whole, or in part, by the National Natural Science Foundation of China (62401323) and the Suzhou Innovation and Entrepreneurship Leading Talent Program (ZXL2025307).

Disclosure of Interests

The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)
2. Gutta, S., Kalva, S.K., Pramanik, M., Yalavarthy, P.K.: Accelerated image reconstruction using extrapolated tikhonov filtering for photoacoustic tomography. Medical Physics **45**(8), 3749–3767 (2018)
3. Haltmeier, M., Schuster, T., Scherzer, O.: Filtered backprojection for thermoacoustic computed tomography in spherical geometry. Mathematical Methods in the Applied Sciences **28**(16), 1919–1937 (2005)
4. Hsu, K.T., Guan, S., Chitnis, P.V.: Fast iterative reconstruction for photoacoustic tomography using learned physical model: Theoretical validation. Photoacoustics **29**, 100452 (2023)
5. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7482–7491 (2018)
6. Lan, H., Huang, L., Wei, X., Li, Z., Lv, J., Ma, C., Nie, L., Luo, J.: Masked cross-domain self-supervised deep learning framework for photoacoustic computed tomography reconstruction. Neural Networks **179**, 106515 (2024)
7. Lan, H., Yang, C., Gao, F.: A jointed feature fusion framework for photoacoustic image reconstruction. Photoacoustics **29**, 100442 (2023)

8. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Liu, Y., Chen, Z., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4652–4660 (2025)
9. Li, K., Yang, J., Liang, W., Li, X., Zhang, C., Chen, L., Wu, C., Zhang, X., Xu, Z., Wang, Y., et al.: O-press: Boosting oct axial resolution with prior guidance, recurrence, and equivariant self-supervision. Medical Image Analysis **99**, 103319 (2025)
10. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1833–1844 (2021)
11. Linger, C., Atlas, Y., Winter, R., Vandebrouck, M., Faure, M., Lucas, T., Bridal, S.L., Gateau, J.: Volumetric and simultaneous photoacoustic and ultrasound imaging with a conventional linear array in a multiview scanning scheme. IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control **70**(12), 1607–1620 (2023)
12. Miao, J., Zhang, S., Sun, G., Li, L.: Multiview linear-array photoacoustic computed tomography with improved elevational resolution in 3d. Biomedical Optics Express **16**(10), 4051–4062 (2025)
13. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern Recognition **106**, 107404 (2020)
14. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
15. Ruan, J., Li, J., Xiang, S.: Vm-unet: Vision mamba unet for medical image segmentation. ACM Transactions on Multimedia Computing, Communications and Applications (2024)
16. Shaw, C.B., Prakash, J., Pramanik, M., Yalavarthy, P.K.: Least squares qr-based decomposition provides an efficient way of computing optimal regularization parameter in photoacoustic tomography. Journal of Biomedical Optics **18**(8), 080501–080501 (2013)
17. Susmelj, A.K., Lafci, B., Ozdemir, F., Davoudi, N., Deán-Ben, X.L., Perez-Cruz, F., Razansky, D.: Signal domain adaptation network for limited-view optoacoustic tomography. Medical Image Analysis **91**, 103012 (2024)
18. Tang, F., Nian, B., Ding, J., Ma, W., Quan, Q., Dong, C., Yang, J., Liu, W., Zhou, S.K.: Mobile u-vit: Revisiting large kernel and u-shaped vit for efficient medical image segmentation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3408–3417 (2025)
19. Treeby, B.E., Zhang, E.Z., Cox, B.T.: Photoacoustic tomography in absorbing acoustic media using time reversal. Inverse Problems **26**(11), 115003 (2010)
20. Wang, H., Chen, Y., Chen, W., Xu, H., Zhao, H., Sheng, B., Fu, H., Yang, G., Zhu, L.: Serp-mamba: Advancing high-resolution retinal vessel segmentation with selective state-space model. IEEE Transactions on Medical Imaging (2025)
21. Wang, R., Zhu, J., Xia, J., Yao, J., Shi, J., Li, C.: Photoacoustic imaging with limited sampling: a review of machine learning approaches. Biomedical Optics Express **14**(4), 1777–1799 (2023)
22. Wang, T., Isola, P.: Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In: International Conference on Machine Learning. pp. 9929–9939. PMLR (2020)

23. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
24. Xing, Z., Ye, T., Yang, Y., Liu, G., Zhu, L.: Segmamba: Long-range sequential modeling mamba for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 578–588. Springer (2024)
25. Xu, M., Wang, L.V.: Universal back-projection algorithm for photoacoustic computed tomography. Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **71**(1), 016706 (2005)
26. Zhang, Y., Wang, Y., Zhang, C.: Total variation based gradient descent algorithm for sparse-view photoacoustic image reconstruction. Ultrasonics **52**(8), 1046–1055 (2012)
27. Zhao, H., Gallo, O., Frosio, I., Kautz, J.: Loss functions for image restoration with neural networks. IEEE Transactions on Computational Imaging **3**(1), 47–57 (2016)
28. Zhu, J., Huynh, N., Ogunlade, O., Ansari, R., Lucka, F., Cox, B., Beard, P.: Mitigating the limited view problem in photoacoustic tomography for a planar detection geometry by regularized iterative reconstruction. IEEE Transactions on Medical Imaging **42**(9), 2603–2615 (2023)