Guidance-base Diffusion Models for Improving Photoacoustic Image Quality

Tatsuhiro Eguchi[1]
tatsuhiro.eguchi@human.ait.kyushu-u.ac.jp
Shunpei Takezaki[1]
shumpei.takezaki@human.ait.kyushu-u.ac.jp
Mihoko Shimano[2]
miho@nii.ac.jp
Takayuki Yagi[3]
yagi.takayuki@luxonus.jp
Ryoma Bise[1]
bise@ait.kyushu-u.ac.jp

[1] Kyushu University
Fukuoka, Japan
[2] National Institute of Informatics
[3] Luxonus Inc.

Abstract

Photoacoustic(PA) imaging is a non-destructive and non-invasive technology for visualizing minute blood vessel structures in the body using ultrasonic sensors. In PA imaging, the image quality of a single-shot image is poor, and it is necessary to improve the image quality by averaging many single-shot images. Therefore, imaging the entire subject requires high imaging costs. In our study, we propose a method to improve the quality of PA images using diffusion models. In our method, we improve the reverse diffusion process using sensor information of PA imaging and introduce a guidance method using imaging condition information to generate high-quality images.

1 Introduction

Photoacoustic(PA) imaging[16] is a technique that visualizes the fine vascular structures within the body. In PA imaging, blood vessels absorb laser energy from short-pulsed near-infrared light and convert the energy into heat, leading to the emission of ultrasonic waves. The structures of objects can be reconstructed by sensing the emitted photoacoustic waves. This technology is non-destructive and non-invasive and is used to understand vascular structures before surgery [4, 14, 32].

The issue with PA imaging is that single-shot images, which capture small local areas, are of low quality due to the limitations of the number of acoustic sensors. These images often contain significant noise, leading to a partial absence of foreground parts (such as blood vessels) (Figure 1, upper-right). Image-averaging techniques effectively reduce noise, which involves averaging multiple scans at the same position under the assumption that vessels are linearly correlated while noise is random [2, 5]. However, this approach requires increased acquisition time to cover a wide body area. Moreover, the data acquisition speed

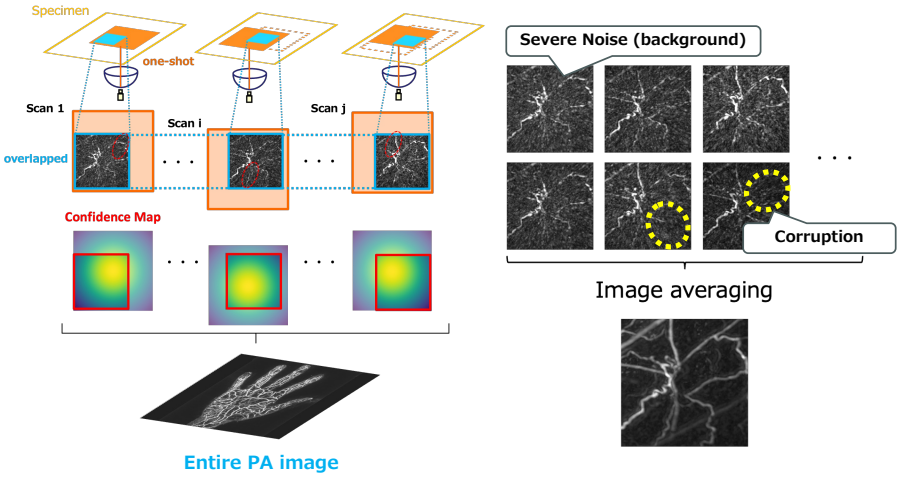

Figure 1: **Left**: Overall mechanisim of photoacoustic imaging. Heat map is the confidence map based on the position of light exposure. **Right**: The upper images are single-shot images (low-quality), and the bottom image is an averaged image (high-quality).

of photoacoustic imaging is limited by the laser repetition rate, and the number of samples is typically restricted to enable real-time imaging and minimize patient burden. This study aims to transform low-quality single-shot images into a high-quality image.

To achieve this goal, we propose a method using diffusion models guided by imaging condition information. Diffusion models, a recently popular image generation model, enables the creation of diverse, high-quality images[7, 10, 12, 24, 28]. In image-to-image tasks, diffusion models have been widely used and achieved high accuracy [19, 23, 25].

In this study, we use diffusion models to generate high-quality PA images from single-shot images that include noise and missing foreground elements. Specifically, we utilize the guidance that uses multiple single-shot images (multi-shots) rather than a single-shot image (single-shot), resulting in higher-quality images. We guide the reverse diffusion model towards higher quality by using the vector from the noise estimated from the low-quality single-shot images to the noise estimated from the multi-shot images.

Moreover, this method introduces the unique property of image averaging in PA imaging into the noise estimation of the diffusion model. Specifically, as shown in Figure 1, the laser light scatters inside the body. It spreads with a Gaussian distribution from the irradiation position, weakening the signal strength as it moves away from this position. As shown in Figure 1, the relative positions of light irradiation in single-shot images vary, leading to areas with clear blood vessels and areas with missing details (i.e., regions of differing quality).

Therefore, we propose a method that combines the estimated noise from each single-shot image, considering the reliability of the signals based on the light irradiation positions. This allows for the estimation of high-quality images. We conducted experiments using actual PA images and confirmed the effectiveness of our method compared to traditional techniques.

2 Related Work

Supervised Denoising Methods. Numerous methods for image noise reduction have been developed to date. Research by Zhang [35], Zamir [34], and others have proposed methods using CNNs and Vision Transformers. Furthermore, Luo et al. [17] have achieved noise removal using diffusion models. Recently, diffusion models have been widely used for medical images. Particularly in denoising tasks, Gao et al. [9] proposed a method applying diffusion models for noise reduction in low-dose CT images. Methods using diffusion models for noise removal have also been proposed for positron emission tomography images and ultrasound images [26, 36]. All these methods focus on general images containing synthetic Gaussian noise, and to our knowledge, no studies have specifically targeted the quality improvement of photoacoustic images. In addition, no method uses the characteristic of image averaging for noise reduction. Therefore, our approach, which effectively utilizes imaging condition information, is considered superior in improving the quality of images.

Guidance Methods for Diffusion Models. As a method for conditional image generation based on specific classes using diffusion models, a Classifier-Guidance technique utilizes the classifier's gradients with the estimates of diffusion models[7, 30], enabling more stable image generation that reflects class information. In Classifier-free Guidance[11], vectors of the conditional and unconditional diffusion models are used instead of using classifier gradients. This method is not limited to class-conditional generation and has also been used in Text-to-Image tasks based on textual information [8, 20, 21]. Additionally, there exists guidance using a CLIP model and methods by guiding the internal representations of diffusion models, which are applicable to tasks such as segmentation and object detection[1, 3, 13, 18]. To the best of our knowledge, no studies have applied such guidance in denoising tasks. In this paper, we introduce a guidance mechanism that utilizes vectors from lower-quality to higher-quality images to enhance the generation of high-quality images.

3 Prepare a paired dataset (low and high-quality images)

Our method aims to train a diffusion model-based image transfer from a low-quality to a high-quality image. To achieve this, pairs of low-and-high-quality images are required for supervised training data. In this section, we explain how to prepare the dataset.

We employed an image-averaging technique with image alignment based on [5] to produce high-quality images. This method exploits the fact that the average of random noise in the background becomes a small constant while the linearly correlated foreground becomes more prominent. To apply the image-averaging technique, single-shot images taken while moving the light source and sensor positions are averaged over overlapping imaging areas (see Figure 1 left). Image averaging proves effective when the sample is static. To address issues arising from patient movement during scans, images are aligned using [5]. This method can produce high-quality images as the number of scans is increased.

For our training data, we captured specimens using many scans to produce high-quality images, and we decreased the number of scans to generate low-quality images. In this paper, we used an image generated by M images in the same location as a low-quality image and that generated by 20 to 40 images as a high-quality one. The paired images of low-and-high-quality images are denoted as $\{\{L^i_1,\ldots,L^i_M\},H^i\}^N_{i=1}$. L^i indicates a single-shot image in the same position i, M is a small number, and H^i indicates the corresponding high-quality image in the i-th position, which is generated using many scans.

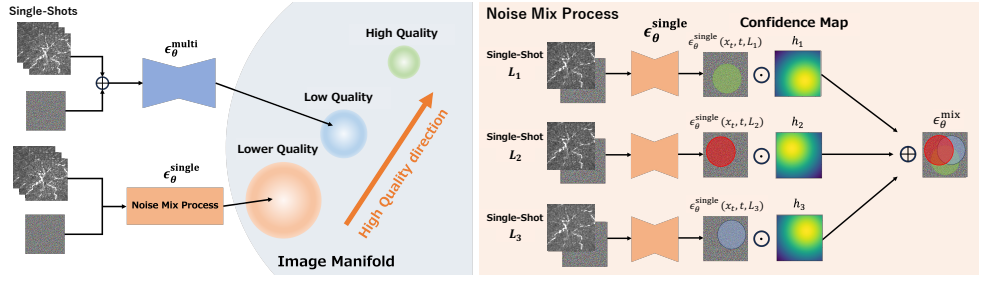

Figure 2: Overview of proposed method, which consists guidance toward higher quality images and Noise Mix Process with photoacoustic imaging condition

Our method utilizes the distribution map of scattered light in the skin, corresponding to L_m^i. This involves irradiating light into the body, which then scatters as it spreads throughout the body. As the absorbed light is large, the magnitude of acoustic signals becomes larger.

As described above, the laser irradiation locations of these M single-shot images $L_1^i, \ldots, L_M^i$ are different, where the center of the location is recorded in the imaging system as shown in Figure 1. We model the scattering light distribution as the Gaussian distribution, denoted by $\boldsymbol{h}_m^i$ for m-th scan at location i.

Note that generating high-quality images through multiple scans with image alignment is unsuitable for real-world applications because performing multiple scans and alignment for all images takes a long time, increasing the burden on patients. Therefore, we propose an image transfer method from low to high quality.

4 Guidance-based diffusion model for improving photoacoustic image quality

Given M single-shot images $L_1^i, \ldots, L_M^i$ capturing the same location i, the proposed method estimates the high-quality image H_i, in which the quality of H_i is comparable to that of an image generated by averaging over 20 images, and $M << 20$. In addition, we also use the corresponding scattering light distribution maps $\{\boldsymbol{h}_1^i, \ldots, \boldsymbol{h}_M^i\}$ for this task.

To achieve this, we integrate a guidance mechanism into the backbone method, Denoising Diffusion Probabilistic Models (DDPM) [12]. Our approach utilizes the low-quality images $\{L_1^i, \ldots, L_M^i\}$ as conditions and denoises the random noise to generate the corresponding high-quality image in a reverse process. Within this reverse process, we introduce guidance defined by the vector from the noise estimated by a lower quality (a single-shot image) to that estimated by the M single-shot images.

4.1 Denoising Diffusion Probabilistic Models

In this section, we explain the backbone method DDPM. In the diffusion process of DDPM, noise is progressively added to an original image $\boldsymbol{x}_0$, which is a high-quality image H, from timestep $t = 1$ to T according to the following equation:

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\varepsilon}, \tag{1}$$

where β_t indicates the intensity of the noise, and $\boldsymbol{\varepsilon} \sim N(0,I)$ represents random noise. The reverse diffusion process is defined from $t = T$ to 1 based on the following equation, utilizing a model $\boldsymbol{\varepsilon}_\theta$ with parameters θ:

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t)) + \sigma_t \boldsymbol{z}, \tag{2}$$

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$, $\sigma_t = \sqrt{\beta_t}$, and $\boldsymbol{z}$ is random noise following $N(0,I)$. For image transformation with DDPM, a conditional model $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t, c)$, where the input image c serves as the condition, is used. DDPM achieves image generation by estimating noise with the model $\boldsymbol{\varepsilon}_\theta$. Therefore, the model is trained to minimize the following loss function:

$$Loss = \mathbb{E}_{t,\boldsymbol{x}_0,\boldsymbol{\varepsilon}}[\| \boldsymbol{\varepsilon} - \boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t, c) \|]. \tag{3}$$

Based on this DDPM framework, our method inputs single-shot PA images as the condition and generates corresponding high-quality images. The condition is attached and provided at the network's input layer, i.e., single-shot images are concatenated to a noise image $\boldsymbol{x}_t$.

4.2 Guidance toward higher quality

In the generative process of diffusion models, guiding from lower (a single-shot) to low (few-shot images) quality can potentially result in higher quality outputs. In our approach, generation solely based on a single shot is deemed lower quality, while generation based on multiple shots is considered low quality. We offer guidance to produce higher-quality PA images.

Figure 2 (Left) shows the overview of our method. To introduce guidance, we prepare two models $\boldsymbol{\varepsilon}_\theta^{\text{single}}$ and $\boldsymbol{\varepsilon}_\theta^{\text{multi}}$, each trained with single-shot and multi-shots as conditions, respectively. Both models generate corresponding high-quality images based on the single-shot or few-shot images provided as conditions.

In the model $\boldsymbol{\varepsilon}_\theta^{\text{multi}}$, multiple-shot images are used as conditions. In contrast, in the model $\boldsymbol{\varepsilon}_\theta^{\text{single}}$, a single-shot image is used as a condition. This implies that the multi-shot model $\boldsymbol{\varepsilon}_\theta^{\text{multi}}$ can use richer information as conditions during the reverse diffusion process compared to the single-shot model $\boldsymbol{\varepsilon}_\theta^{\text{single}}$. Therefore, $\boldsymbol{\varepsilon}_\theta^{\text{multi}}$ is capable of higher quality generation than $\boldsymbol{\varepsilon}_\theta^{\text{single}}$. This is because, due to photoacoustic imaging, even single-shot images captured at the same location hold different structural information, and image averaging can reduce noises; thus, inputting more single-shot images results in higher accuracy.

We guide towards further quality improvement by using the difference in outputs from models trained with single-shot and multi-shots. Specifically, in each timestep of the reverse diffusion process, instead of $\boldsymbol{\varepsilon}_\theta(\boldsymbol{x}_t, t, c)$, we use $\tilde{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t, c)$ calculated based on the following:

$$\tilde{\boldsymbol{\varepsilon}}_\theta(\boldsymbol{x}_t, t, c) = (1+w)\boldsymbol{\varepsilon}_\theta^{\text{multi}}(\boldsymbol{x}_t, t, \{L_1^i, \ldots, L_M^i\}) - w\boldsymbol{\varepsilon}_\theta^{\text{mix}}(\boldsymbol{x}_t, t, c), \tag{4}$$

where $\boldsymbol{\varepsilon}_\theta^{\text{mix}}(x_t, t, c)$ is the output obtained through the Noise Mix Process using $\boldsymbol{\varepsilon}_\theta^{\text{single}}$, described in the next section. Additionally, w is a hyperparameter indicating the strength of guidance. Through equation (4), we can expand the difference between the consistent outputs of $\boldsymbol{\varepsilon}_\theta^{\text{single}}$ and $\boldsymbol{\varepsilon}_\theta^{\text{multi}}$ by w and guide towards generating higher quality photoacoustic images.

4.3 Noise mix process of single-shot image-based models

Our problem setup includes several single-shot images corresponding to H_i, each exhibiting different scattering light distribution maps as illustrated in Figure 1. When we randomly select one of these single-shot images, certain areas may exhibit high intensity, which has high reliability, depending on the light distribution map. In contrast, others show low intensity. This variability could potentially worsen the effectiveness of the guidance.

To address this issue, we introduce a noise-mixing technique to generate a higher-quality image by incorporating clear parts from each single-shot image. The noise mixing technique involves interpolating between two images [28] and controlling image generation from text [27]. Inspired by these methods, we propose a weighted noise mixing process to leverage the unique properties of image averaging.

The noise mix process combines the estimated noise images $\boldsymbol{\varepsilon}_\theta^{\text{single}}(L_1),\ldots,\boldsymbol{\varepsilon}_\theta^{\text{single}}(L_M)$ using their corresponding light distribution maps $\{\boldsymbol{h}_1^i,\ldots,\boldsymbol{h}_M^i\}$. As mentioned above, in PA imaging, the reliability of the acquired signal changes depending on the distance from the light irradiation position, and these positions vary in each one-shot image. Therefore, a light distribution map $\boldsymbol{h}_m^i$ can be considered as a confidence map. Based on the assumption that the estimated noise in high-confidence areas is more reliable than that in low-confidence areas, our method combines the estimated noise images through a weighted average, where the weights are determined by the confidence maps as follows:

$$\boldsymbol{\varepsilon}_\theta^{\text{mix}}(\boldsymbol{x}_t,t,c)=\frac{1}{M}\sum_{m=1}^{M}\boldsymbol{h}_m^i\odot\boldsymbol{\varepsilon}_\theta^{\text{single}}(\boldsymbol{x}_t,t,L_m^i). \tag{5}$$

where $\boldsymbol{x}_t$ is the noise image at t step in diffusion process, $\boldsymbol{\varepsilon}_\theta^{\text{single}}(\boldsymbol{x}_t,t,L_m^i)$ represents the output of $\boldsymbol{\varepsilon}_\theta^{\text{single}}$ corresponding to each single-shot image L_m^i. $\boldsymbol{h}_m^i$ is the confidence map for the single-shot image L_m^i.

Mixing the noise estimated from individual single-shot images using equation 5 allows for the integration of their respective structural information. Furthermore, considering the confidence maps based on the light irradiation positions of each single-shot image enables more accurate completion of the foreground parts.

5 Experiments

Dataset. We used real PA images for evaluation. The images were taken of the lower limbs of two subjects, who were instructed to remain still during data collection. For the training, validation, and test data, we prepared paired images of low and high-quality images $\{\{L_1^i,\ldots,L_M^i\},H^i\}_{i=1}^N$ with their light distribution maps $\{\boldsymbol{h}_1^i,\ldots,\boldsymbol{h}_M^i\}$. The number of single-shot images M was 3. The training, validation, and test data contain 3988, 1328, and 907 pairs, respectively. For test data, we used different subjects' PA images with training and validation data.

Implimentation Details. we used a U-Net-like network [12] as the diffusion model $\boldsymbol{\varepsilon}_\theta^{single}$, $\boldsymbol{\varepsilon}_\theta^{multi}$. The U-Net-based diffusion model has the residual layer and self-attention to improve the model's representation performance. The training was conducted over 300,000 iterations with a batch size of 16. The optimization algorithm used was Adam, with a learning rate of 1.0×10^{-4}. The total number of timesteps T in sampling and the noise scheduler were the same as those in [12], with $T=1000$, $\beta_t=10^{-4}$, and $\beta_T=0.02$.

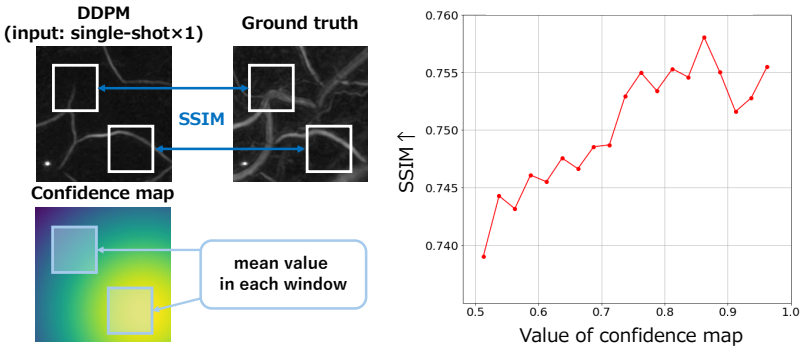

Figure 3: Preliminary experiments with confidence maps, **Left**: overview of evaluation method, **Right**: correlation of SSIM and confidence maps

There is a known issue that excessively large guidance scales w can degrade the quality of image generation[24]. In the generative process of diffusion models, semantic information is formed in the early stages, while finer image details are developed towards the end [6, 15]. Thus, especially towards the end, using a large guidance scale w can lead to deviations from the training data distribution of the diffusion model, resulting in degraded generation quality. Therefore, our method sets an interval $[T, t_{guide}]$ where guidance is used with a predefined w, and $[t_{guide}, 0]$ where $w = 0$. This t_{guide} was tuned using validation data.

Comparative Methods. As the most naive approach, we used image averaging of three input images as the Baseline. For comparison, we employed a CNN-based method with U-Net [22] and DnCNN [35] for image transfer. As the state-of-the-art methods, a Transformer-based method with Restormer [34], and a standard diffusion model, DDPM [12] were evaluated. Here, DDPM refers to $\boldsymbol{\varepsilon}_\theta^{\text{multi}}$, given three single-shot images as conditions and performing conditional generation based solely on equation (2). Each deep neural network-based method was trained with the set of three one-shot images as input.

Evaluation Metrics. To compare the accuracy of the outputs from each method, Peak Signal to Noise Ratio(PSNR) and Structural Similarity Index Measure(SSIM) between the estimated high-quality image by each method and the ground truth were used, which have been widely used in image denoising tasks. The higher their value, the higher the similarity.

5.1 Correlation with Confidence Maps

First, we conducted preliminary experiments to quantitatively verify that the clarity of the foreground parts (blood vessels) changes depending on the distance from the light irradiation position in each shot image.

Figure 3 (Right) illustrates the relationship between the mean SSIM values of an image estimated by DDPM with a single-shot image in a local area (window size 50×50) and the mean confidence values of the light distribution map in the same area. To compute the average of the SSIM values, we divided the confidence values into 20 equal ranges, and calculated the means of SSIMs for all locations in all test data within each range, as plotted in Figure 3 (Right). As a result, the SSIM increased as the confidence values increased.

Figure 3 (Left) shows example images of the estimated high-quality image by DDPM, its ground truth, and the corresponding confidence map. The vessel clearly appears in the region with high confidence values (right-bottom area in the map). In contrast, the vessel is

Method	PSNR↑	SSIM↑
Baseline	20.63	0.3396
U-Net[22]	30.10	0.5055
DnCNN[35]	30.14	0.5210
Restormer[34]	30.48	0.5247
Ours	**30.63**	**0.5468**

Table 1: Comparative experiments with previous methods

Method		PSNR↑	SSIM↑
$w = 5$	**Ours**	30.27	0.5132
	w/o h	30.26	0.5144
$w = 10$	**Ours**	**30.63**	**0.5468**
	w/o h	30.61	0.5456
$w = 20$	**Ours**	30.29	0.5313
	w/o h	30.25	0.5301
$w = 30$	**Ours**	27.94	0.4611
	w/o h	27.70	0.4573

Table 2: Ablation experiments with different guidance scales

Method	PSNR↑	SSIM↑
DDPM[12]	29.39	0.4159
Classifier-Free Guidance[11]	29.38	0.5315
Ours w/o Noise Mix Process	30.38	0.5447
Ours	30.63	0.5468

Table 3: Ablation experiments with guidance condition

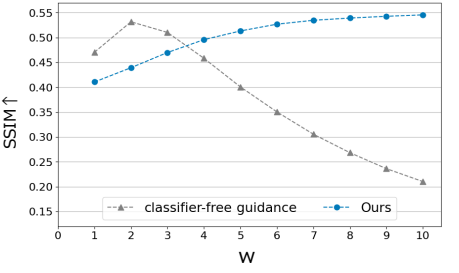

Figure 4: guidance scale-wise SSIM

unclear in the region with low confidence values (left-top).

These results suggest that considering the corresponding confidence maps for each can enable more accurate predictions when using multiple single-shot images as inputs.

5.2 Quantitative evaluation

Comparative study. Table 1 presents the average performances (PSNR and SSIM) based on evaluation metrics for each method's output results on the test data. Deep learning-based methods significantly improved image quality compared to the simple image averaging of the Baseline. It is confirmed that the proposed method shows the best results in both PSNR and SSIM, particularly in SSIM, which is a metric that assumes the similarity of image structures contributes to human perception of image quality degradation.

Hyper-parameter sensitivity. Table 2 shows the results of experiments conducted by varying the guidance scale w without using the confidence map h. First, comparing the variations in w within our method, a decline in accuracy is observed at $w = 30$. It is confirmed that excessively increasing the guidance scale can deteriorate the generative results.

Next, w/o h in Table 1 refers to averaging the outputs based on each single-shot image in our method's noise mix process without utilizing the corresponding confidence map h. The results show that using the confidence map h achieves higher PSNR values for each w, and although SSIM values are higher without the confidence map at $w = 5$, using the confidence map at $w = 10$ consistently shows the best results. Here, the best $w = 10$ can be selected using validation data. This suggests that considering the confidence of each single-shot image contributes to improved accuracy.

Ablatoin study. Table 3 also presents the results of the ablation study of our method. Here,

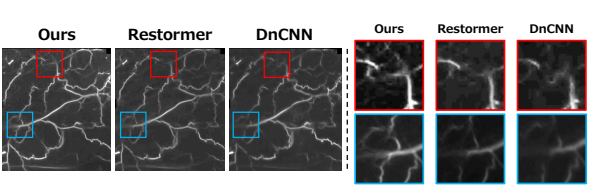

Figure 5: **Left:** Wide-view PA image obtained by ours and other methods, **Right:** Enlarged images.

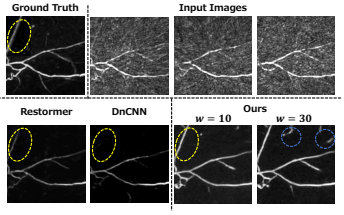

Figure 6: Example of complementation for corruption

variations in the guidance methods were tested: DDPM was executed without guidance, using three single-shot images as conditions, classifier-free guidance was based on unconditional noise, and w/o Noise Mix Process directly used the output from $\boldsymbol{\varepsilon}_\theta^{\text{single}}$ as the basis for guidance. The simplest DDPM output was the least accurate, and accuracy improved with each guidance method. Notably, the proposed method of introducing the noise mix process based on imaging conditions achieved the highest performance.

Additionally, in classifier-free guidance and the proposed method, the results of varying the guidance scale from 1 to 10 are shown in Figure 4. In classifier-free guidance, accuracy declined starting from $w = 3$, whereas in the proposed method, accuracy consistently improved from $w = 1$ to 10. It is considered that the guidance method of the proposed approach, using correlated low-quality and high-quality outputs, allows for guidance toward higher-quality image generation.

5.3 Qualitative evaluation

Figure 5 shows example wide-view images of the results for our method, Restormer [34], and DnCNN [35]. The small images on the right in Figure 5 are enlarged areas of the red and blue boxes in the wide-view images. In these results, some blood vessels were missing in the comparison methods. In contrast, the proposed method made the blood vessels clearer, as shown in the enlarged images.

Figure 6 shows examples of ground truth, single-shot images(input images), and output results of each method. In the results, noise in the background parts contained in the single-shot images was removed using all methods. Focusing on the parts highlighted in yellow, the comparative methods do not sufficiently complete the missing blood vessels, whereas our method successfully completed these missing vessels.

The qualitative evaluation was conducted for different guidance scales w. Regarding the output results at $w = 30$ in our method, as indicated by the blue frames, blood vessels are complemented in areas where, according to ground truth, they should not exist. This suggests that excessively increasing the guidance scale can lead to over-detection of blood vessels, adversely affecting the generation results. In contrast, the suitable scale of w, which is automatically selected using validation data, shows a good result.

6 Discussion and Conclusion

In this study, we proposed a method for enhancing the quality of photoacoustic(PA) images from a small number of single-shot images to reduce imaging costs in PA imaging.

We introduce a guidance approach that considers the reliability of the signals obtained at the time of imaging, and it improves the reverse diffusion process in diffusion models. The structural information from multiple single-shot images is effectively reflected in the generation results by defining and utilizing a confidence map based on the light irradiation position during imaging. Moreover, we introduced a guidance technique for diffusion models that leads to higher-quality generation results. Experiments using real data from PA images demonstrated the effectiveness of our method compared to traditional techniques.

As with most applications with diffusion models, our method is limited primarily by slow inference times. However, our main goal is to achieve reconstruction with a few one-shot images, which, if successful, could reduce the long scan times typically required in PA imaging. Also, our method can explore acceleration techniques as in [21, 29, 31, 33].

We believe that our guidance technique, which utilizes vectors from noise estimated in lower-quality images to that in higher-quality images for diffusion models, can be applied to any denoising tasks, not just PA imaging.

Acknowledgment

This work was supported by SIP-JPJ012425, AMED JP19he2302002, JSPS KAKENHI Grant JP23K18509 and JP24KJ1805.

References

[1] Nichol Alex, Dhariwal Prafulla, Ramesh Aditya, Shyam Pranav, Mishkin Pamela, McGrew Bob, Sutskever Ilya, and Chen Mark. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, 2022.

[2] Takanori Asanomi, Kazuya Nishimura, Heon Song, Junya Hayashida, Hiroyuki Sekiguchi, Takayuki Yagi, Imari Sato, and Ryoma Bise. Unsupervised deep non-rigid alignment by low-rank loss and multi-input attention. In *Medical Image Computing and Computer Assisted Intervention*, pages 185–195, 2022.

[3] Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

[4] Ryoma Bise, Sato Imari, Kajiya Kentaro, and Yamashita Toyonobu. 3d structure modeling of dense capillaries by multi-objects tracking. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVMI)*, pages 29–37, 2016.

[5] Ryoma Bise, Yingqiang Zheng, Imari Sato, and Masakazu Toi. Vascular registration in photoacoustic imaging by low-rank alignment via foreground, background and complement decomposition. In *Medical Image Computing and Computer-Assisted Intervention*, pages 326–334, 2016.

[6] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 11472–11481, 2022.

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

[8] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene:scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106, 2022.

[9] Qi Gao, Zilong Li, Junping Zhang, Yi Zhang, and Hongming Shan. Corediff: Contextual error-modulated generalized diffusion model for low-dose ct denoising and generalization. *IEEE Transactions on Medical Imaging*, 43(2):745–759, 2024.

[10] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. In *Advances in Neural Information Processing Systems*, 2022.

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[13] Vincent Tao Hu, David W. Zhang, Yuki M. Asano, Gertjan J. Burghouts, and Cees G. M. Snoek. Self-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18413–18422, 2023.

[14] Ryo Kikkawa, Hiroyuki Sekiguchi, Itaru Tsuge, Susumu Saito, and Ryoma Bise. Semi-supervised learning with structured knowledge for body hair detection in photoacoustic image. In *ISBI*, 2019.

[15] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. In *The Eleventh International Conference on Learning Representations*, 2023.

[16] Changhui Li and Lihong V Wang. Photoacoustic tomography and sensing in biomedicine. *Physics in Medicine & Biology*, 54(19):R59, 2009.

[17] Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Image restoration with mean-reverting stochastic differential equations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 23045–23066, 2023.

[18] Andrea Vedaldi Minghao Chen, Iro Laina. Training-free layout control with cross-attention guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5331–5341, 2024.

[19] Aimon Rahman, Jeya Valanarasu, Jose Maria, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11546, 2023.

[20] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, pages 234–241, 2015.

[23] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022.

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[25] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022.

[26] Chenyu Shen, Ziyuan Yang, and Yi Zhang. Pet image denoising with score-based diffusion probabilistic models. In *Medical Image Computing and Computer Assisted Intervention*, pages 270–278, 2023.

[27] Takahiro Shirakawa and Seiichi Uchida. Noisecollage: A layout-aware text-to-image diffusion model based on noise cropping and merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8921–8930, 2024.

[28] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[29] Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024.

[30] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[31] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

[32] Saito Susumu, Bise Ryoma, Yoshikawa Aya, Sekiguchi Hiroyuki, Tsuge Itaru, and Toi Masakazu. Digital artery deformation on movement of the proximal interphalangeal joint. *Journal of Hand Surgery (European Volume)*, 44(2):187–195, 2019.

[33] Liu Xingchao, Gong Chengyue, and Liu Qiang. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

[34] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[35] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[36] Yuxin Zhang, Clément Huneau, Jérôme Idier, and Diana Mateus. Ultrasound image reconstruction with denoising diffusion restoration models. In *Deep Generative Models: Third MICCAI Workshop*, page 193–203, 2023.