ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

SDSC

Latent Diffusion Models for Domain Adaptation in Optoacoustic Imaging

Master Semester Project

Chia-Wen Chen

1 March 2025

Co-supervisor: Dr. Firat Ozdemir
Supervisor: Prof. Dr. Fernando Perez-Cruz

Department of Computer Science, ETH Zürich

Latent Diffusion Models for Domain Adaptation in Optoacoustic Imaging

Chia-Wen Chen
Department of Computer Science
ETH Zurich, Switzerland
chiachen@ethz.ch

Abstract

Optoacoustic (OA) imaging is a non-invasive biomedical imaging technique with high potential, providing optical contrast complementary to ultrasound by stimulating tissue with laser pulses. Synthetic OA data are easier to generate and annotate, but they often lack the detailed anatomical features and realistic noise patterns found in experimental OA images. This gap limits the usefulness of synthetic data in developing OA image processing methods. In this study, we investigate diffusion-based domain adaptation strategies, using diffusion, classifier-guided diffusion, latent diffusion, and conditional variational autoencoder (CVAE) approaches to transform synthetic OA images into experimental-like images. Our methods are evaluated on the standardized public OA dataset, focusing on adding key features such as skin gradients, tissue textures, vessel details, and acquisition-induced noise. Our image-space diffusion model [1] consistently produces high-fidelity adaptations that preserve anatomical structures and match experimental characteristics. These results offer a solid framework for enhancing the realism of synthetic OA data, potentially reducing the need for large clinical datasets for model training, and supporting the advancement of OA image processing algorithms. [2]

1 Introduction

Optoacoustic (OA) imaging provides optical contrast complementary to ultrasound through laser-induced tissue stimulation, making it a promising modality for both clinical and laboratory applications. Its ability to capture real-time vascular and tissue changes, along with its non-ionizing imaging method, has become increasingly important in biomedical research.

Recent efforts to standardize OA datasets have led to OADAT: Experimental and Synthetic Clinical Optoacoustic Data for Standardized Image Processing [1], which provides both experimental and synthetic forearm data under various tomographic setups. Synthetic data generation in OADAT is crucial for reducing patient recruitment and annotation costs, as well as mitigating displacement errors in multispectral OA images caused by sequential acquisitions.

However, a significant challenge remains: the synthetic data in OADAT still exhibit a distribution gap relative to real acquisitions, potentially limiting how well models trained on these synthetic images generalize to practical scenarios. This gap undermines the usefulness of synthetic data in evaluating new methods or developing models intended for clinical deployment.

To overcome this domain gap, we explore diffusion-based approaches that adapt synthetic OA images to closely resemble experimental OA images. Our methods aim to improve the fidelity of synthetic images, reduce the need for extensive real-world data collection, and preserve key anatomical details in the original synthetic data required for various OA processing tasks, such as segmentation.

[1] Repository for diffusion-based domain adaptaion: https://gitlab.renkulab.io/chiachen/oadat-ldm
[2] Msc semester project report; co-supervisor: Firat Ozdemir, supervisor: Prof. Fernando Perez-Cruz.

In this work, we explore three key aspects: (1) **Diffusion-Driven Domain Adaptation.** We propose diffusion-based methods, including diffusion [2], latent diffusion [3] and a conditional variational autoencoder (CVAE) [4] variant, to convert simulated OA images into realistic images that mirror real-world acquisitions. (2) **Systematic Evaluation on OADAT Data.** We compare these methods using publicly available OADAT datasets, focusing on their ability to preserve crucial anatomical structures (e.g., skin lines, vessels) and incorporate experimental-domain characteristics, while also assessing the diversity of the generated samples. (3) **Insights on Sparse Intensity Distributions.** We address the challenges posed by OA's sparse intensity distribution and demonstrate that diffusion-based methods with v prediction effectively maintain key intensity characteristics. These findings establish a robust framework for narrowing the synthetic-to-real gap in OA imaging and underscore the potential of diffusion models to enhance the realism of OA image generation.

2 Background

2.1 Optoacoustic imaging

OA imaging is a non-invasive imaging modality that uses short laser pulses to stimulate biological tissues, generating acoustic waves that carry optical absorption information detected by ultrasound transducers. Absorbers, such as hemoglobin, melanin, and other chromophores, absorb the laser energy and undergo thermoelastic expansion, emitting ultrasonic waves. Regions with higher optical absorption, including blood vessels or pigmented areas, produce stronger acoustic signals, creating contrast relative to surrounding tissues and providing diverse functional and anatomical information.

The high-contrast capability and non-ionizing nature of OA imaging have enabled a wide range of clinical research, such as skin cancer studies [5], inflammation analysis [6], and breast tumor detection [7]. In parallel, machine learning researchers have leveraged OA imaging data for tasks like limited-view reconstruction, artifact removal, and anatomical segmentation [1], aiming to improve diagnostic accuracy and quantitative analysis in biomedical imaging.

2.2 Experimental and Synthetic Clinical Optoacoustic DATA (OADAT)

OADAT [8] represents the first standardized, large-scale dataset of in vivo OA imaging acquired under diverse clinical settings. OADAT comprises 4 distinct datasets: two experimental datasets, the Multispectral Forearm Dataset (`MSFD`) and the Single Wavelength Forearm Dataset (`SWFD`); one simulated dataset, the Simulated Cylinders Dataset (`SCD`); and an additional fully annotated subset. Each dataset is further partitioned into subcategories. Notably, the experimental and synthetic datasets are unpaired, with ground truth anatomical annotations provided exclusively in the synthetic dataset. In this work, we focus on the semi-circle transducer array category from the experimental dataset (`SWFD_sc`) and the virtual circle transducer category from the synthetic dataset (`SCD_vc`).

2.3 Discrepancy between experimental and synthetic data distributions

Experimental images from OADAT display distinct anatomical and acquisition-related features. Sample images from this dataset are shown on the left side of Figure 1.

Skin: The skin is typically represented as a bright, horizontal arc with the highest contrast at the center, gradually fading toward the sides (see Figure 1 A and 1 B). This high contrast is attributable to the significant optical absorption by melanin, which produces strong acoustic signals.

Tissues: The tissues immediately beneath the skin (i.e., the subcutaneous layer) often exhibit irregular intensities—appearing as white, yarn-like curves—and typically become gradually darker with increasing depth (see Figure 1 C). Additionally, other deeper regions may also display irregular intensity patterns (see Figure 1 D), which correspond to subtle variations in tissue composition.

Vessels: Beneath the skin, vessels appear as circular or oval structures (see Figure 1 E). Vessels are highlighted due to the absorption properties of blood, and their interiors exhibit irregular intensities that reflect the complex internal composition.

Background noise: The experimental images contain background noise in the form of concentric rings (see Figure 1 F and 1 G). These artifacts, which result from the acquisition and reconstruction processes, are an intrinsic part of the data and must be preserved in the target distribution. Similarly, ripple effects around the vessels (see Figure 1 H), though seemingly noisy, are routinely observed in clinical settings due to limited view and are therefore included in the model's target distribution.

In contrast, the synthetic data are designed to emulate these experimental features but with notable differences (see the right side of Figure 1):

Skin and tissues: In the synthetic images, the skin line, subcutaneous layer, and tissues are rendered uniformly without the gradual intensity variation observed in experimental images. There is no depiction of the skin's natural gradient, where the center is the brightest and the sides are progressively darker.

Vessels: The synthetic dataset includes a variety of vessel appearances, ranging from darker to brighter intensities, with some vessels displaying gradients (darker centers with brighter peripheries), and varying in size, shape (elongated versus circular), and clarity (sharp versus blurred). However, despite this variety, the synthetic vessels tend to be uniformly rendered and do not fully capture the complex textures observed in experimental images.

Background noise: Unlike the experimental images, the synthetic data lack realistic background noise, which is an important characteristic of real-world imaging due to the acquisition and reconstruction processes. These differences underscore the challenges of domain adaptation, as the synthetic images must be carefully adjusted to reflect the complexities and nuances present in real-world experimental data.

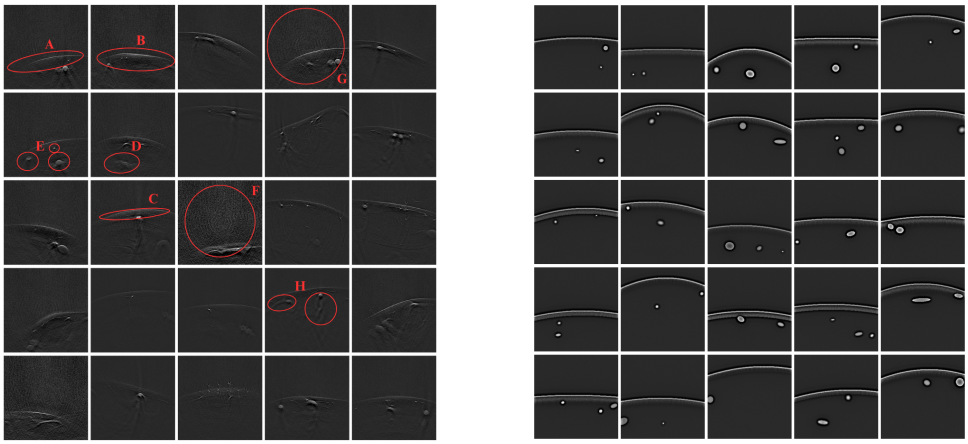

Figure 1: Samples from the OADAT dataset highlight the distribution gap between experimental and synthetic optoacoustic images. **Left:** Experimental images with natural skin gradients, diverse tissue and vessel textures, and inherent background noise. **Right:** Synthetic images with uniform rendering, simplified vessel appearances, and minimal noise.

2.4 Domain adaptation

Domain adaptation reduces the discrepancy between a source domain and a target domain by aligning their distributions. This enables models trained on the source domain to generalize to the target domain. Typically, the source domain has abundant labeled data, while the target domain has limited data. This situation is common in medical imaging, where clinical data are difficult and expensive to obtain. For example, acquiring experimental OA images requires specialized hardware and significant expertise. In contrast, synthetic images can be generated in large quantities with precise ground truth annotations. However, their distribution differs significantly from that of experimental images (see

Section 2.3). Domain adaptation is essential in this context because it translates synthetic images into realistic, experimental-like images. This process bridges the domain gap and supports the training of data-hungry models.

2.5 Diffusion model

Diffusion models are a class of generative models that learn to create data by gradually reversing a diffusion process [2], where noise is systematically added to the data. During training, the model learns to predict the noise component at each step of a forward process that progressively corrupts the data with Gaussian noise. The forward diffusion process can be mathematically described as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I) \tag{1}$$

During inference, this learned denoising process is reversed, starting from pure noise and iteratively refining it into a coherent image. This deterministic backward pass of the Denoising Diffusion Implicit Model (DDIM) [9] can be described by:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(x_t,t), \quad \text{where} \quad \bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s) \tag{2}$$

These models have recently gained attention for their ability to generate high-resolution, photorealistic images [10] and have been successfully applied to domain adaptation tasks [11]. Their strength lies in their capacity to capture complex data distributions and generate high quality outputs. Early studies have also demonstrated the potential of diffusion models in biomedical imaging modalities, including applications in OA imaging, where enhanced sparse-view image reconstruction is desired [12].

2.6 Classifier-guidance diffusion model

Classifier-guidance is a technique used within diffusion models to steer the generative process toward desired outcomes [10]. By incorporating gradients from a classifier trained to distinguish specific features or domains, the model can be conditionally guided to generate samples that align more closely with the target distribution. In our context, classifier-guidance allows us to direct the synthesis process so that the generated images resemble the distribution of the experimental data, ensuring that critical features and artifacts are appropriately captured. The forward pass of classifier guided diffusion is the same as Equation 1, while the classifier-guided DDIM backward pass is defined as:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{x_t - \sqrt{1-\bar{\alpha}_t}\,\tilde{\epsilon}_\theta(x_t,t|y)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\,\tilde{\epsilon}_\theta(x_t,t|y), \tag{3}$$

where the guided noise prediction is given by:

$$\tilde{\epsilon}_\theta(x_t,t|y) = \epsilon_\theta(x_t,t) - s\sqrt{1-\bar{\alpha}_t}\,\nabla_{x_t}\log p_\phi(y|x_t) \tag{4}$$

Here, s controls the classifier guidance strength, and $\nabla_{x_t}\log p_\phi(y|x_t)$ is the classifier gradients.

2.7 Latent diffusion model

Latent diffusion models (LDMs) extend the concept of diffusion models by operating in a lower-dimensional latent space rather than directly in the high-dimensional image space [3]. In this approach, images are first encoded into a compact latent representation $z = E(x)$ using an autoencoder. The diffusion process in the latent space is described by:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t I). \tag{5}$$

The deterministic backward process (denoising) in latent space is:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\left(\frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_t,t)}{\sqrt{\bar{\alpha}_t}}\right) + \sqrt{1-\bar{\alpha}_{t-1}}\,\epsilon_\theta(z_t,t) \tag{6}$$

Finally, the refined latent representation is decoded back to the image space using:

$$x \approx D(z_0). \tag{7}$$

Operating in this compressed latent space significantly reduces computational cost and provides easier manipulation of high-resolution images like OA images. Furthermore, the latent space often captures the most salient features of the data, facilitating a more accurate representation of complex patterns of OA data such as tissue structures, vessels, and background.

3 Datasets

In this study, we focus on two subcategories from the OADAT's SWFD and SCD datasets: the semi-circle transducer array dataset (`SWFD_sc`), representing experimental data, and the virtual circle transducer dataset (`SCD_vc`), representing synthetic data.

The `SWFD_sc` dataset contains OA images from 14 patients, each assigned a unique ID. We reserve data from 13 patients (IDs 1–13) for training and validation, using the first 80% for training and the remaining 20% for validation. The remaining patient (ID 14) serves as the test set. Each patient contributes 2,802 grayscale images at 256 $\times$ 256 pixels, yielding 29,140 training images, 7,286 validation images, and 2,802 test images.

The `SCD_vc` dataset is filtered to include only images with vessel sizes under 500 pixels, matching the typical vessel scale in `SWFD_sc` and retaining sufficient samples for training. In total, 7,903 grayscale images at 256 $\times$ 256 pixels are used; 6,322 (80%) for training, 1,581 (20%) for validation, and an additional 100 images for testing.

The diffusion and latent diffusion models are trained exclusively on `SWFD_sc`, while the classifier for classifier guidance and the variational autoencoders (VAEs) are trained on both `SWFD_sc` and `SCD_vc`. For clarity, we refer to `SWFD_sc` as experimental data and `SCD_vc` as synthetic data in the following sections.

4 Methodology

4.1 Diffusion model for OA image domain adaptation

The core of our approach leverages DDIM for domain adaptation of OA images [9]. The diffusion model is trained solely on the experimental data to learn the characteristics of real OA images and generate experimental-like images. Each image is normalized by its own maximum intensity, clipped at a minimum of -0.2, as described in the original paper [1], and then linearly scaled to $[-1, 1]$ before being fed into the diffusion model. Gaussian noise, sampled from $\mathcal{N}(0, 1)$, is added at each timestep according to a cosine noise schedule [13], which gradually decreases the signal-to-noise ratio (SNR) over 1,000 timesteps. During inference, we inject noise into a synthetic image at certain timesteps and then gradually denoise it using the trained diffusion model. The number of noise-injection steps should neither be too large, to ensure that the structural information of the original image is preserved, nor too small, to allow sufficient room for the image to gain realistic features during denoising, thus effectively achieving domain adaptation. Training details are shown in Appendix A.1.

4.1.1 V prediction for stability

In standard diffusion models, a typical approach is to use ϵ-prediction, where the model learns to estimate the noise added at each timestep. Specifically, each noisy sample x_t is generated from the original image x_0 and Gaussian noise ϵ according to:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\,\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{8}$$

However, common schedulers (including the cosine scheduler used in this research) and sampling steps are flawed [14]. Since the SNR is not strictly zero at the final noise-addition step, the diffusion model learns to predict the mean of the images by leveraging the leaked information from the last timestep. This creates a discrepancy between training and inference. During inference, we typically start with pure noise (i.e., SNR equal to 0), so a model that has learned to predict the mean at the final

step will treat pure Gaussian noise as a zero-mean image, resulting in images with medium brightness. This issue is particularly problematic when generating OA images, which are predominantly dark with sparse bright regions. See Appendix A.2 for an example that compares ϵ prediction and v prediction.

To address this, we adopt the fix proposed in [14], which involves rescaling the SNR to 0 at the final step (i.e., $\alpha_t = 0$) and substituting ϵ prediction with v prediction and v loss:

$$\mathbf{v}_t = \alpha_t \boldsymbol{\epsilon} - \sigma_t \mathbf{x}_0 \tag{9}$$

$$\mathcal{L} = \lambda_t \left\| \mathbf{v}_t - \tilde{\mathbf{v}}_t \right\|_2^2 \tag{10}$$

With the last step rescaled to zero, ϵ prediction becomes non-trivial. In contrast, v prediction can directly predict the image mean at the final step because it is composed of both the noise ϵ and the original image x_0. At the final step, with $\alpha_t = 0$ in Equation (9), the model's objective shifts to directly predicting x_0. Therefore, by employing v prediction and SNR rescaling during the final step, we can generate images with dark backgrounds that more closely align with the intensity distribution of OA images.

4.2 Classifier guidance on diffusion model

Classifier guidance on diffusion models is a technique that incorporates the gradient of a classifier into the denoising process, steering generated samples toward a specified class. In our case, we treat experimental and synthetic data as two distinct classes and guide the denoising process toward the experimental domain. The classifier is trained independently of the diffusion model using both experimental and synthetic images. The training procedure involves adding Gaussian noise at random timesteps to the input images and training the classifier to predict whether the noisy image belongs to the experimental or synthetic class. This training setup ensures that the classifier is robust to varying noise levels.

Since the dataset is imbalanced (29,140 experimental training images vs. 6,322 synthetic training images), the classifier may tend to overpredict the experimental class. To mitigate this, we assign higher weights to the synthetic samples in proportion to the class imbalance, ensuring a more balanced training process.

During inference, the classifier participates in the denoising loop of the diffusion model. At each step, the current image is classified as either experimental or synthetic, and the classifier's gradient is used to refine the diffusion model's noise prediction. The updated noise prediction then determines the denoised image at the previous timestep. This iterative feedback mechanism continuously pushes the image distribution away from the synthetic domain, potentially driving it toward the far end of the experimental manifold. As a result, the final output closely resembles experimentally acquired data.

4.3 Latent Diffusion Model for OA Image Domain Adaptation

Beyond the image-space diffusion model, we investigate latent diffusion models (LDMs) for domain adaptation in OA imaging. Given the high dimensionality and resolution of OA images, operating in a compact latent space can simplify and accelerate learning, while enabling classifier guidance to focus on salient features. In our framework, we first train a VAE and a CVAE on both experimental and synthetic data to learn compact latent representations. We then train the diffusion model exclusively on the latent representations derived from experimental data, ensuring that the generative process aligns with the experimental distribution.

However, autoencoding may exacerbate the distributional gap between experimental and synthetic data, potentially limiting the diffusion model's applicability to synthetic latents. To address this, we incorporate a domain classifier during VAE and CVAE training [15], to pull the latent distributions closer together. During inference, the LDM projects a synthetic image into the latent space, applies the noise process, and then decodes it. In the LDM + CVAE approach, the encoding and decoding steps remain unchanged, but we condition the decoder on the experimental class label. To further assess the impact of the CVAE, we also experiment with a direct CVAE-based adaptation method, where synthetic images are encoded and then decoded using the experimental label as a conditioning variable. Detailed experimental setups are provided in the Appendix A.1.

5 Results

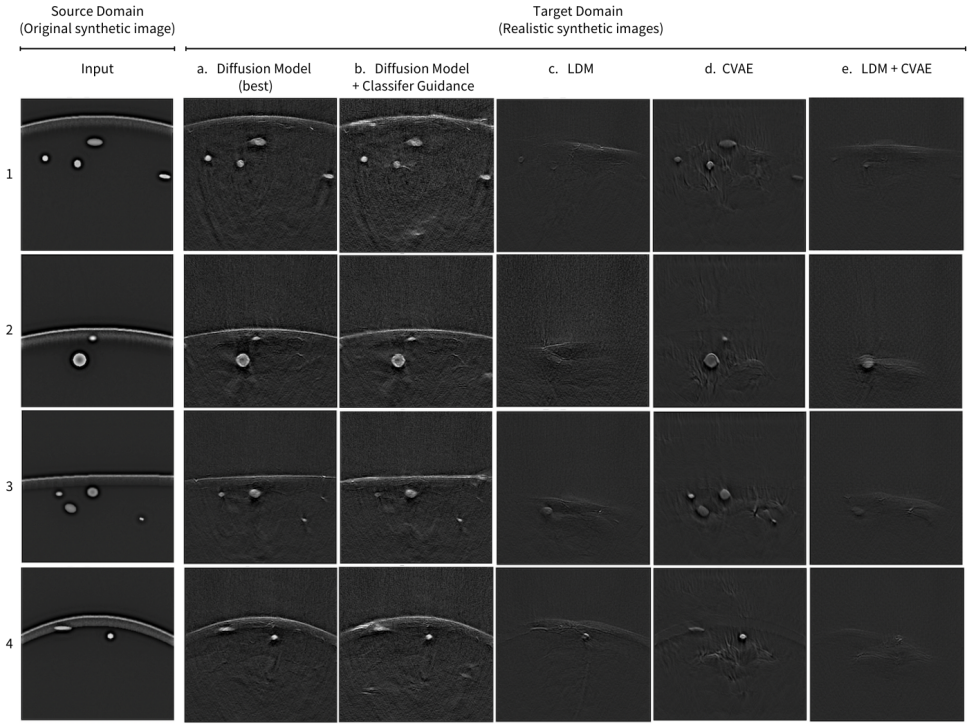

Figure 2: Domain adaptation results on four synthetic optoacoustic images using five methods: a. diffusion model, b. classifier-guided diffusion, c. latent diffusion, d. CVAE, and e. latent diffusion paired with CVAE. The diffusion model best preserves anatomical structures and experimental characteristics.

In this section, we evaluate the performance of the proposed methods[3] for adapting synthetic images to the experimental domain. We compare five methods: two image-space methods—the diffusion model and the classifier-guided diffusion model—and three latent-space methods—the latent diffusion model, the CVAE, and their combination. We focus on four key aspects in OA images: (1) skin lines, (2) tissues, (3) vessels, and (4) background noise. By comparing outputs from each method on our synthetic test set, we highlight the strengths and limitations of all diffusion-based adaptation strategies. Figure 2 shows the results of each approach across four example samples.

5.1 Skin Lines

Our diffusion model captures skin lines effectively. In Figure 2 column a, it transforms synthetic skin lines into thin bands at the top of each frame. These bands appear brightest in the center and fade toward the edges. The model sometimes introduces realistic irregularities, as seen in 3.a of Figure 2, where a small barb appears on the left tail of the skin.

In contrast, latent diffusion, the CVAE, and their combination struggle to preserve skin lines. They often fail to maintain the precise skin-line location and instead introduce a new, inconsistent boundary.

[3]Repository for diffusion-based domain adaptaion: https://gitlab.renkulab.io/chiachen/oadat-ldm

5.2 Tissues

The diffusion model converts the synthetic translucent layer located immediately below the skin surface into realistic tissue. For example, 1.a and 4.a in Figure 2 show that this region transitions from a lighter intensity to a darker intensity with depth, exhibiting short, white yarn-like curves. In deeper tissues, as seen in 2.a in Figure 2, irregularities appear that are not associated with vessels, indicating varied tissue composition.

5.3 Vessels

Our diffusion model excels at preserving vessel brightness from the synthetic input. Bright vessels remain bright, and dark vessels retain their lower intensity (see the four vessels in 1.a in Figure 2). The model also enriches the inner texture of each vessel with subtle variations, producing outputs that closely resemble experimental images. However, the balance between preserving dark or small vessel structures and adding realistic patterns depends on the noise-injection timestep. For example, the bottom-left vessel in 3.a is nearly erased due to its darkness. We find that injecting noise at 300 to 650 timesteps, followed by denoising, provides a good trade-off between structural preservation and fidelity adaptation.

When used independently, the CVAE retains the overall vessel layout (see column d in Figure 2). It produces vessels with darker intensities and bright tips reminiscent of real experimental data. On the other hand, both the latent diffusion model (see column c) and the latent diffusion model paired with the CVAE (see column e) often blur or remove vessels, especially smaller ones. This occurs even with minimal noise injection.

5.4 Background Noise

Beyond the skin and vessels, the diffusion model replicates background noise patterns effectively and generates realistic acquisition artifacts. For example, 2.a of Figure 2 shows concentric ring noise. Furthermore, it introduces realistic rippling effects around the vessels. In 1.a, the leftmost and rightmost vessels exhibit two distinct dark vertical lines beneath.

Coupling the diffusion model with classifier guidance further amplifies experimental noise characteristics (see column b). This can produce extreme samples with elongated vessels, uneven skin lines, or intensified background noise. Notably, such outlier cases occur in real experimental data and broaden the diversity of the generated images.

In contrast, outputs from the CVAE alone often contain repetitive and unrealistic textures in the central regions, which suggests pattern memorization of local patches in the latent space (see column d). Although adding latent diffusion can remove these artifacts, accurately reconstructing vessels and skin lines in their original positions remains challenging for latent-space methods (see column e).

5.5 Overall Performance

Among all approaches, the diffusion model trained directly in the image space provides the most consistent domain adaptation results. It preserves anatomical structures, aligns with experimental brightness distributions, replicates realistic noise patterns, and most importantly enhances the realism of skin, tissue, and vessel texture. Classifier guidance can further steer generated samples toward rare but valid modes of the experimental distribution, providing a way to sample extreme cases and thus increase variety. Meanwhile, the latent diffusion, the CVAE, and their combination often struggle with small-scale anatomical fidelity, resulting in blurred or missing vessels, incomplete skin lines, or repetitive background artifacts. In general, our findings suggest that diffusion models operating in the image space best balance structural preservation and realistic image fidelity for OA images, offering strong potential for adaptation and enhancement in this setting. See Appendix A.4 for comprehensive test set results for all domain adaptation approaches.

5.6 Variety of the diffusion model's generation

In the previous section, we demonstrated that the diffusion model can generate images with high fidelity. In this section, we investigate its diversity.

Varied adaptations from a single input. We take one synthetic image and apply nine different noise seeds. We then add noise for 650 out of 1,000 steps and subsequently denoise the image. Figure 3 (left) shows that using different noise seeds at inference produces multiple realistic variants while preserving the original anatomical structure. This variability broadens the range of adapted images and can help augment training data.

Pure noise generation. Instead of adding noise for only 650 time steps, we generated nine samples from pure Gaussian noise with different noise seeds and denoise them. Figure 3 (right) shows that the diffusion model has learned the patterns of experimental data and can generate a variety of entirely new realistic images from pure noise. This capability may support unsupervised learning tasks.

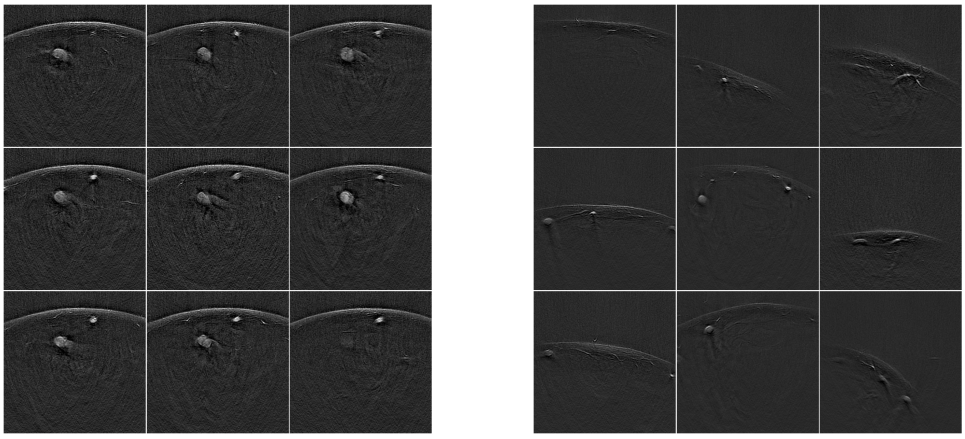

Figure 3: Diffusion model variability. **Left:** Multiple realistic adaptations generated from a single synthetic input using nine different noise seeds with noise added for 650 timesteps. **Right:** New realistic images generated entirely from pure noise (1,000 timesteps), demonstrating the model's ability to maintain variability while preserving image fidelity.

6 Discussion

In this work, we focus on two subcategories from the experimental and synthetic image domains. Our initial results for domain adaptation are promising, but we do not know if our method generalizes to other subcategories. Future work should evaluate a wider range of experimental and synthetic data. In addition, quantitative assessments of the performance of domain adaptation are needed.

Our experiments show that image-space diffusion models outperform latent diffusion models. One explanation is that the autoencoders map latent representations to widely separated regions. Even with a domain classifier to mitigate this effect, it remains difficult to create a smooth latent space. Therefore, a diffusion model trained on experimental latents cannot easily generalize to synthetic latents.

The classifier guidance added experimental-like background noise to the generated images during inference. However, key structures such as vessels and skin textures did not show further improvement. This may be due to the high-dimensional nature of the original images. The image-space classifier may not capture all the complexity.

The diffusion model successfully reduced the gap between synthetic and experimental data. In particular, when sampling from pure Gaussian noise, the generated images appeared very realistic. Nevertheless, some images generated with fewer noise injection steps—aimed at preserving all of the original anatomical structures—still exhibit certain limitations. For example, the skin lines are thicker than those in experimental data. In addition, the vessels are generally brighter and show a relatively uniform brightness across their circular structures. These observations suggest that improvements

can be made in the original synthetic image generation process. Adjustments such as darkening the vessels while preserving a bright tip, thinning the skin lines, and increasing the angular variety of the skin could narrow the gap between synthetic and experimental data from the root and enhance the performance of OA imaging domain adaptation.

7 Conclusion

In summary, this study demonstrates that diffusion models effectively bridge the gap between synthetic and experimental optoacoustic (OA) images. Our experiments indicate that image-space diffusion models successfully preserve critical anatomical features, such as skin lines and vessel structures, while replicating realistic noise patterns and complex textures inherent in experimental data. Incorporating classifier guidance further enhances the model's capability to generate realistic experimental features, although it occasionally introduces variations reflecting natural diversity rather than uniform improvements. In contrast, latent diffusion models and CVAE-based approaches show potential but often struggle to preserve fine-scale anatomical details, such as delicate vessel structures and consistent skin boundaries. These findings underscore the promise of diffusion-based methods in adapting synthetic data, potentially reducing dependence on extensive real-world datasets while maintaining high-quality imaging.

Despite these promising outcomes, several challenges remain. This work has explored only one subcategory of experimental and synthetic data. To fully assess the generalizability of our domain adaptation approach, future studies should evaluate additional subcategories, such as limited-view and multisegment data available in the OADAT dataset. Moreover, our evaluations were primarily qualitative; comprehensive quantitative assessments are essential. Future work should include quantitative evaluations, such as training models on synthetic data adapted by our method and measuring their performance on downstream tasks or assessing the distributional similarity between experimental and adapted synthetic datasets. Overall, this study establishes a solid foundation for applying diffusion models to domain adaptation in OA imaging, enhancing synthetic data realism and providing abundant data to facilitate the development of OA image processing algorithms.

References

[1] F. Ozdemir, B. Lafci, X. L. Dean-Ben, D. Razansky, and F. Perez-Cruz, "OADAT: Experimental and synthetic clinical optoacoustic data for standardized image processing," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=BVi6MhKO0G

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *CoRR*, vol. abs/2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, vol. abs/2112.10752, 2021. [Online]. Available: https://arxiv.org/abs/2112.10752

[4] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf

[5] X. L. Deán-Ben and D. Razansky, "Optoacoustic imaging of the skin," *Experimental Dermatology*, vol. 30, no. 11, pp. 1598–1609, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/exd.14386

[6] A. P. Regensburger, E. Brown, G. Krönke, M. J. Waldner, and F. Knieling, "Optoacoustic imaging in inflammation," *Biomedicines*, vol. 9, no. 5, 2021. [Online]. Available: https://www.mdpi.com/2227-9059/9/5/483

[7] A. Oraevsky, B. Clingman, J. Zalev, A. Stavros, W. Yang, and J. Parikh, "Clinical optoacoustic imaging combined with ultrasound for coregistered functional and anatomical mapping of breast tumors," *Photoacoustics*, vol. 12, pp. 30–45, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213597918300168

[8] B. Lafci, F. Ozdemir, X. L. Dean-Ben, D. Razansky, and F. Perez-Cruz, "Oadat: Experimental and synthetic clinical optoacoustic data for standardized image processing," Zurich, 2022-02-24.

[9] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *CoRR*, vol. abs/2010.02502, 2020. [Online]. Available: https://arxiv.org/abs/2010.02502

[10] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *CoRR*, vol. abs/2105.05233, 2021. [Online]. Available: https://arxiv.org/abs/2105.05233

[11] C. Saharia, W. Chan, H. Chang, C. A. Lee, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," *CoRR*, vol. abs/2111.05826, 2021. [Online]. Available: https://arxiv.org/abs/2111.05826

[12] X. Song, G. Wang, W. Zhong, K. Guo, Z. Li, X. Liu, J. Dong, and Q. Liu, "Sparse-view reconstruction for photoacoustic tomography combining diffusion model with model-based iteration," *Photoacoustics*, vol. 33, p. 100558, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213597923001118

[13] A. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," *CoRR*, vol. abs/2102.09672, 2021. [Online]. Available: https://arxiv.org/abs/2102.09672

[14] S. Lin, B. Liu, J. Li, and X. Yang, "Common diffusion noise schedules and sample steps are flawed," 2024. [Online]. Available: https://arxiv.org/abs/2305.08891

[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," 2016. [Online]. Available: https://arxiv.org/abs/1505.07818

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CoRR*, vol. abs/1801.03924, 2018. [Online]. Available: http://arxiv.org/abs/1801.03924

A Appendix

A.1 Training details

For all training experiments, we used an NVIDIA A100 GPU with shuffled training sets, a cosine scheduler, and an initial learning rate of 0.0001 with a 5-epoch warmup and without any weight decay. Other common parameters included 4 workers, seed 42, and the AdamW optimizer. Preprocessing involved normalizing each image by its maximum value, clipping at -0.2, and scaling to the range -1 to 1, while outputs were postprocessed back to -0.2 to 1.0. All diffusion models were trained using 1,000 steps, with image dimensions of 256×256, latent size of 32×32, and 3 latent channels. Training epochs and batch sizes varied by model: CVAE and VAE were trained for 200 epochs with a batch size of 92; the diffusion model for 250 epochs with a batch size of 32; LDM and LDM-CVAE for 200 epochs with a batch size of 128; and the classifier for 200 epochs with a batch size of 100.

For VAE and CVAE training, we incorporated LPIPS loss [16] and a discriminator to enhance detail reconstruction, adding the discriminator loss only after 10 epochs [4]. A sigmoid function was applied to the latents to constrain values between 0 and 1, given the diffusion model's sensitivity to intensity changes. During LDM training, images were first encoded, passed through a sigmoid to ensure latent values were within 0 to 1, then linearly scaled to -1 to 1 before being fed into the diffusion model; the denoised outputs were directly input to the decoder. In the CVAE, the labels were first projected into an embedding, and then separate scale and shift networks are trained to modulate the latent representations via a feature-wise affine transformation. The U-Net implementations for diffusion models and VAEs were based on the Diffusers library developed by Hugging Face [5]. The implementation of the classifier for classifier-guidance is based on the repository developed by OpenAI [6]. Additionally, for VAE training, a domain classifier was used to align the latent distributions of the two domains. During training, the domain classifier loss is added to the total loss with a scaling factor of 5000 to match the magnitude of the other loss terms. Moreover, its gradients are reversed by multiplying with a negative scaler (which gradually increases from 0 to 1 over the course of training) to encourage closer alignment of the latent distributions.

[4] Repository developed by the CompVis for latent diffusion: https://github.com/CompVis/latent-diffusion
[5] Hugging Face Diffusers library: https://huggingface.co/docs/diffusers/index
[6] Repository developed by OpenAI for guided-diffusion: https://github.com/openai/guided-diffusion

A.2 ϵ prediction vs. v prediction visualization

Figure 4 presents the denoising results using ϵ prediction and v prediction. When denoised from the final step, the generated sample becomes overly bright with ϵ prediction, whereas v prediction restores the normal intensity.

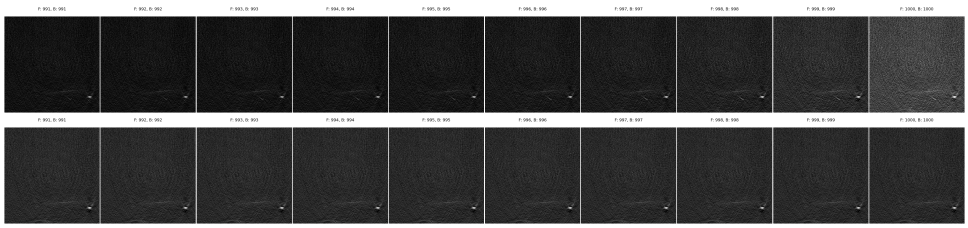

Figure 4: Comparison of ϵ prediction (top) versus v prediction (bottom). F: number of forward pass steps; B: number of backward pass steps; total timesteps: 1,000.

A.3 Different classifier guidance scales

Figure 7 illustrates the impact of various classifier guidance scales on the generated images, offering flexibility in tuning the effect—higher values lead to more pronounced guidance.

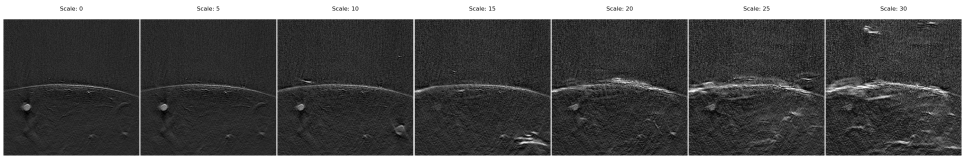

Figure 5: The impact of various classifier guidance scales

A.4 Comprehensive test set results for all domain adaptation approaches

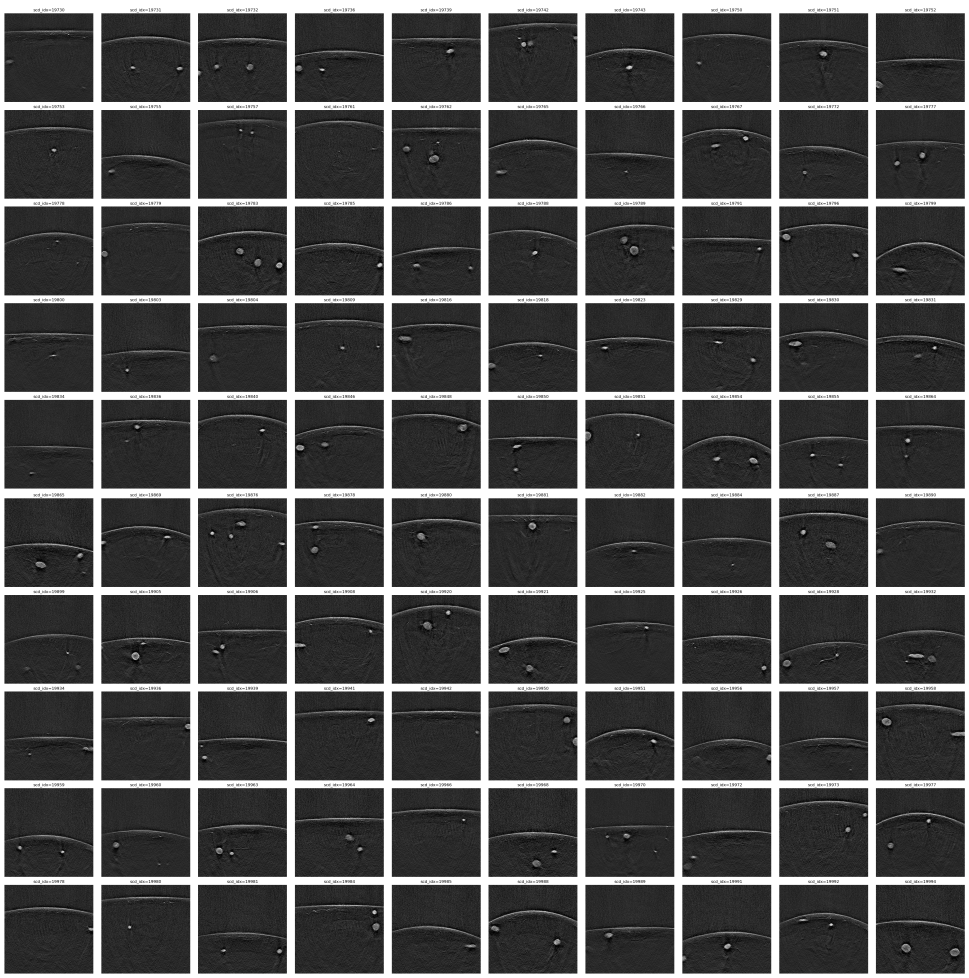

Figure 6: Test set results of the diffusion model