Coarse-to-Fine Medical Image Translation by Incorporating Deterministic Guidance and Probabilistic Refinement

Hongnian Tian[1,†], Tianxu Lv[1,†], Jiansong Fan[1], Delin Pan[1], Lihua Li[2], and Xiang Pan[1,3,*]

[1] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China
[2] Institute of Biomedical Engineering and Instrumentation, Hangzhou Dianzi University, Hangzhou, China
[3] The PRC Ministry of Education Engineering Research Center of Intelligent Technology for Healthcare, Wuxi, Jiangsu 214122, China
[†] Equal contributions.*Corresponding author: xiangpan@jiangnan.edu.cn

Abstract. In clinical diagnosis and treatment, traditional enhanced imaging techniques often suffer from inherent limitations such as high time costs and radiation risks. Therefore, medical image translation technology provides an efficient and cost-effective alternative. However, images generated by existing medical image generation methods still face challenges, such as a lack of structural consistency and blurred local details. Most methods struggle to simultaneously integrate deterministic structural information, such as anatomical priors, and probabilistic dynamic variations, such as blood flow changes, to guide image generation. To address these challenges, we propose a Coarse-to-Fine Medical Image Translation (C2FMIT) model, which incorporates Deterministic Guidance and Probabilistic Refinement to balance generation controllability and fidelity. First, we design a Deterministic Guidance Branch (DGB) to extract coarse-grained features, such as organ contours, to provide global structural constraints. Then, these deterministic priors are fused into our Probabilistic Refinement Branch (PRB), where the Brownian Bridge diffusion is employed for fine-grained optimization, enhancing microvascular textures and dynamic enhancement regions. Notably, we designed a Coarse-to-Fine Guided Attention Module (C2FGAM) to achieve progressive optimization from global structure to local details. Experimental results demonstrate that our method achieves superior performance across multiple modalities of functionally contrast-enhanced medical imaging on both public and in-house datasets.

Keywords: Coarse-to-Fine Medical Image Translation · Deterministic Guidance · Probabilistic Refinement.

1 Introduction

Traditional functional imaging techniques, such as those relying on dynamic contrast enhancement or high-radiation-dose imaging modalities, are fundamental

to disease diagnosis [16]. However, their clinical application is limited by inherent drawbacks, including high time costs and radiation exposure risks [4]. To address these challenges, deep learning-based image-to-image translation techniques have been proposed. These methods learn the mapping from a source image domain to a target image domain, aiming to generate high-resolution functional images from non-enhanced or low-dose images, providing an efficient and low-risk alternative.

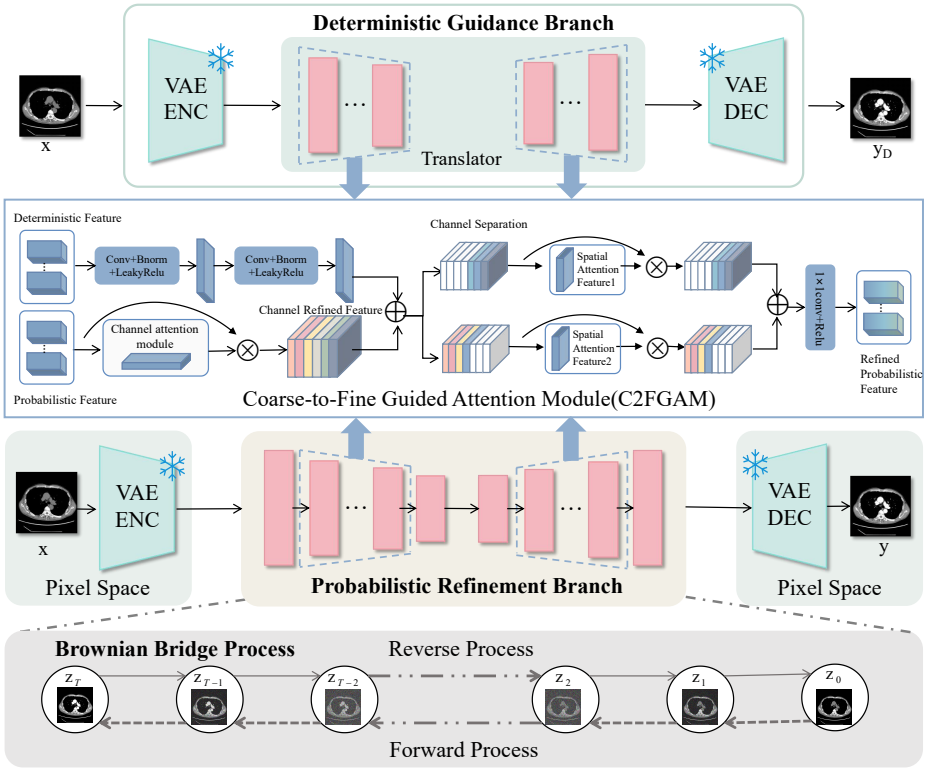

Fig. 1. Overview of the proposed method. First, coarse-grained features are obtained through the DGB. These features are then fused with the fine-grained features generated by the PRB via the C2FGAM module, resulting in refined probabilistic features. Finally, these features undergo the Brownian Bridge diffusion process in the PRB to generate the target image.

In recent years, generative models such as Generative Adversarial Networks (GANs) [6,5] and Variational Autoencoders (VAEs) [14,3,10] have shown great potential in medical image generation tasks. Conditional GAN-based medical image translation methods, such as Pix2Pix [11] and CycleGAN [24], can synthesize high-fidelity images, but their training instability and mode collapse is-

sues (such as vascular ruptures and artifact generation) limit the reliability of cross-domain medical image translation. VAEs generate images through latent variable modeling, but they have limitations in medical image tasks. VAEs tend to generate blurry images, lack explicit modeling of randomness during the generation process, and struggle to capture fine-grained details (such as microvascular textures). With the rise of denoising diffusion models, many studies, such as Denoising Diffusion Probabilistic Models (DDPMs) [7] and conditional diffusion models [23], have surpassed GANs in natural image generation through a stepwise denoising mechanism [2]. However, the multiple sampling steps during inference increase computational costs, and the lack of effective guidance from deterministic prior knowledge leads to structural biases in the generated images.

In summary, while existing medical image generation methods have improved image quality, they still suffer from structural inconsistency and blurred local details. Most methods struggle to simultaneously integrate deterministic structural information [21], such as anatomical priors, and probabilistic fine-grained variations, such as dynamic blood flow, to guide image generation.

To address these challenges, we propose a C2FMIT by incorporating Deterministic Guidance and Probabilistic Refinement to balance generation controllability and fidelity. Specifically, our contributions are summarized as: 1)Coarse-Grained Deterministic Guidance: We design the DGB, which explicitly extracts cross-domain coarse-grained features, such as organ contours and primary vascular topology, through adversarial feature disentanglement. 2)Fine-Grained Probabilistic Refinement: We construct the PRB, which models stochastic generation in the latent space through the Brownian Bridge diffusion process, enhancing microvascular textures and dynamic enhancement regions at the voxel level. 3)Progressive Feature Fusion: We propose that the C2FGAM integrate global structural information from the DGB with fine-grained details from the PRB. The coarse-grained deterministic features are embedded into each denoising step of the Brownian Bridge diffusion process, achieving a progressive optimization from global anatomical constraints to local texture synthesis. 4))Extensive experiments on the DCE-MRI (pre-contrast→post-contrast) and CT→CTA translation tasks demonstrate the superiority of our method. Evaluations on the public Duke-Breast-Cancer-MRI dataset and an in-house ChestCT-CTA dataset confirm state-of-the-art performance in both structural accuracy and functional plausibility.

2 Method

Our proposed framework is illustrated in Fig. 1 and consists of three core modules: The DGB, which extracts coarse-grained features such as cross-domain anatomical priors using a VAE-like structure; The PRB, which models and generates fine-grained features based on the Latent Brownian Bridge Diffusion Model (LBBDM) [15]; The C2FGAM, which dynamically fuses deterministic features with diffusion noise to achieve progressive optimization from global constraints to local details.

2.1 Deterministic Guidance Branch

The DGB aims to explicitly extract anatomical structural features (such as organ contours and vascular topology) from cross-domain images. To address issues of structural consistency and blurred local details, the outer layer of the DGB consists of a pretrained VAE Encoder $E(\cdot)$ and a pretrained VAE decoder, ensuring the features are extracted into latent space. These features then pass through a UNet-like [17] structure, the translator $tl(\cdot)$, to ensure consistency with features from the PRB and to extract cross-domain anatomical priors (such as vascular topology and organ contours). Given the input x, the loss function of the DGB is expressed as:

$$L_{DGB} = \|E(x) - tl(E(x))\|^2. \tag{1}$$

2.2 Probabilistic Refinement Branch

To ensure anatomical structural consistency while modeling fine-grained dynamic details in medical images, such as blood flow signals and microvascular textures, we designed the PRB, which focuses on extracting fine-grained features. This branch adopts the LBBDM, an extension of the classical Brownian Bridge Diffusion Model (BBDM) into the latent space. Unlike existing Denoising Diffusion Probabilistic Models, the Brownian Bridge process does not terminate at pure Gaussian noise but instead converges to a clean conditional input y. Following notation similar to DDPM, let (x, y) denote paired training data from domain A and domain B. The diffusion process operates in the latent space of a pre-trained VQGAN [22], accelerating both training and inference. For simplicity, we retain x and y to represent their latent features ($x := L_A(x), y := L_B(y)$). The forward Brownian Bridge diffusion process is defined as:

$$q_{BB}(x_t|x_0, y) = \mathcal{N}(x_t; (1 - m_t)x_0 + m_t y, \delta_t \mathbf{I}) \tag{2}$$

where $x_0 = x, x_t = y, m_t = t/T$ and the variance term $\delta_t = 2(m_t - m_t^2)$. The reverse process of the PRB aims to predict x_{t-1} based on x_t:

$$p_\theta(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \widetilde{\delta}_t \mathbf{I}) \tag{3}$$

where $\widetilde{\delta}_t$ is the variance of Gaussian noise at step t and $\mu_\theta(x_t, t)$ is the predicted mean value of the noise to be learned. The training objective of the PRB is optimizing the Evidence Lower Bound (ELBO):

$$L_{PRB} = E\left[c_{\epsilon t} \left\| m_t (y - x_0) + \sqrt{\delta_t}\epsilon - \epsilon_\theta(x_t, t) \right\|^2\right] \tag{4}$$

where $c_{\epsilon t}$ denotes the coefficient term of the estimated noise ϵ_θ in mean value term $\tilde{\mu}_t$. In summary, the objective of our jointly trained two branches is defined as follows:

$$L_{total} = L_{DGB} + L_{PRB} \tag{5}$$

2.3 Coarse-to-Fine Guided Attention Module

In order to integrate deterministic structural information, such as anatomical priors in medical imaging, to guide image generation, we design the C2FGAM. This module fuses global structural information from the DGB and detailed features from the PRB at different scales, thereby enhancing the structural consistency and detail fidelity of the generated images.

Given the features F_d from the DGB and F_p from the PRB, C2FGAM employs a two-stage feature fusion strategy. Firstly, at the channel level, a channel attention mechanism [19] is used to enhance the importance of the probabilistic features and to perform channel-wise weighted computation with the deterministic features:

$$F_c = Concat(\sigma(W_c * F_p) \odot F_p, LeakyReLU(BN(W_d * F_d))) \tag{6}$$

where W_c is the channel attention weight, $\sigma(\cdot)$ is the sigmoid activation function, $BN(\cdot)$ denotes Batch Normalization and $\odot$ represents element-wise channel weighting.A spatial attention mechanism [12] is then applied to enhance local details further. The fused feature F_c is decomposed into two separate spatial attention modules. After applying the attention mechanisms, the features are concatenated to obtain F_s:

$$F_s = Concat(F_{c1} \odot \sigma(W_s * F_{c1}), F_{c2} \odot \sigma(W_s * F_{c2})) \tag{7}$$

Finally, the fused feature F_s is mapped using a 1×1 convolution to generate the final optimized feature:

$$F_{refined} = LeakyReLU(W_f * F_s) \tag{8}$$

This feature contains global structural information while preserving rich local details, providing robust feature support for generating high-quality medical images.

3 Experiments

3.1 Datasets and Implementation

To validate our approach, we conducted experiments on a publicly available DCE-MRI dataset (Duke-Breast-Cancer-MRI) and an in-house CT-CTA dataset (ChestCT-CTA).

Duke-Breast-Cancer-MRI. This dataset, released by Duke University Medical School in collaboration with the National Cancer Institute (NCI), contains DCE-MRI sequences from 922 breast cancer patients. For each case, the following MRI sequences are shared in DICOM format: non-fat-saturated T1-weighted sequences, fat-saturated gradient echo T1-weighted pre-contrast sequences, and most post-contrast sequences (3 to 4 sequences).

ChestCT-CTA. We selected the Chest Abdominal Aorta Angiography Image dataset to evaluate our method. This dataset consists of paired CT-CTA

images collected from a local hospital between May 2023 and March 2024. Each CT and CTA scan image was resampled to a volume of $0.67\times0.67\times1.25$ mm^3, consisting of 450-650 slices, each with a size of 512×512 pixels, totaling 1000 cases.

Implementation Settings. Our framework was implemented in PyTorch 2.0.0 with CUDA 12.1, and experiments were conducted on a computational platform equipped with four NVIDIA RTX A6000 GPUs to accelerate training. For the Duke-Breast-Cancer-MRI dataset, we selected 358 patients with contrast-enhanced MRI scans. Each patient's MRI volume consists of 60 slices, each with a size of 512×512. We chose pre-contrast and post-contrast phases as the source and target images, respectively. For the ChestCT-CTA dataset, 114 patients with paired CT/CTA scans were included, where each CT/CTA volume comprises 560 slices (512×512 pixels). Both datasets were split into train, validation, and test sets at a 7:1:2 case-level ratio. During preprocessing, all images were resized to 256×256 to meet the model's input requirements. In the training phase, we first pretrain the VQGAN using the collected dataset, with a downsampling factor set to 8. For the BBDM, we set the number of time steps to 1000 and used 200 sampling steps during the sampling phase to balance generation quality and efficiency.

3.2 Comparison with SOTA Methods

Quantitative Analysis. Our method was compared with six state-of-the-art synthesis methods, including Pix2Pix, CycleGAN, VQI2I [1], QS-Attn [9], BBDM and UNSB [13]. For a fair comparison, we retrained their networks using publicly available implementations to generate their best synthesis results. Quantitative comparisons were performed on the Duke-Breast-Cancer-MRI [18] and ChestCT-CTA datasets using Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR) [8] and Structural Similarity Index (SSIM) [20]. The results are summarized in Table 1 and Table 2, where the **best** and the second best results are highlighted.

Table 1. Quantitative comparison with the prediction baseline on the Duke-Breast-Cancer-MRI dataset (Mean ± Std).

Duke-Breast-Cancer-MRI			
Methods	MAE(Voxel)↓	PSNR(dB)↑	SSIM(%) ↑
Pix2Pix	6.73±1.35	23.86±1.53	70.88±4.47
CycleGAN	6.91±1.20	24.42±1.46	65.39±4.84
VQI2I	7.32±1.39	24.50±1.27	68.51±4.93
QS-Attn	9.29±1.43	24.11±1.65	42.69±4.65
BBDM	5.73±1.11	25.65±1.41	75.55±4.65
UNSB	7.68±2.12	23.54±2.16	71.98±6.05
Ours	**3.92±1.08**	**27.34±1.22**	**80.34±4.33**

Table 2. Quantitative comparison with the prediction baseline on the ChestCT-CTA dataset (Mean ± Std).

ChestCT-CTA			
Methods	MAE(Voxel)↓	PSNR(dB)↑	SSIM(%)↑
Pix2Pix	9.74±2.13	21.22±1.99	76.60±4.71
CycleGAN	8.57±2.11	22.56±2.29	81.34±4.82
VQI2I	18.95±2.29	16.06±0.85	58.63±3.71
QS-Attn	10.66±3.33	20.99±2.36	80.14±5.30
BBDM	12.70±1.95	19.31±1.10	72.80±4.61
UNSB	10.00±3.90	21.33±3.26	78.78±6.71
Ours	**6.99±2.07**	**24.30±1.29**	**84.41±4.83**

Table 1 and Table 2 present the average MAE, PSNR, and SSIM scores for all methods on the Duke-Breast-Cancer-MRI and ChestCT-CTA datasets. When comparing the latest generative models with our method, our approach demonstrates superior quantitative performance. For the former, our average MAE is 3.92, the average PSNR is 27.34, and the average SSIM is 80.34%. For the latter, the average MAE is 6.99, the average PSNR is 24.30, and the average SSIM is 84.41%. These results indicate that our model generates the highest-quality images, closely matching the ground-truth distribution and achieving better human-perceived visual fidelity than other methods.

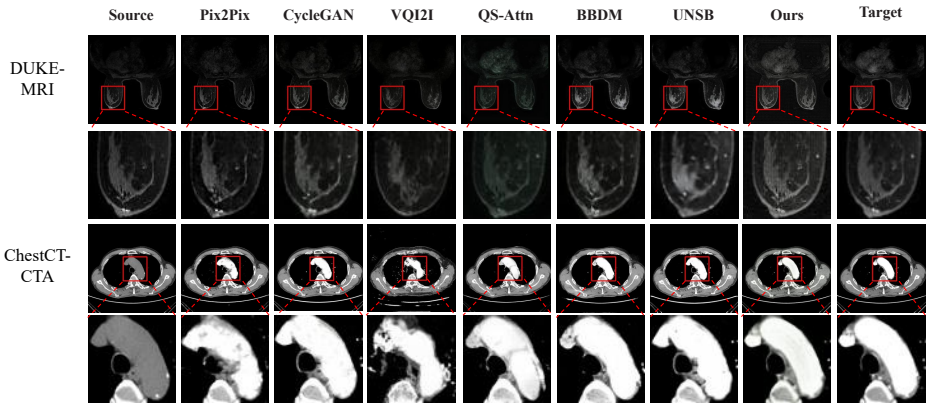

Fig. 2. Visual comparisons of proposed methods and other state-of-the-art methods.

Qualitative Analysis. Fig. 2 shows a qualitative comparison between our method and previous state-of-the-art techniques. Compared to these models, the images generated by our method exhibit the highest overall quality. Our approach effectively preserves fine-grained details while ensuring the structural consistency of the generated images. Specifically, our model excels in maintaining important anatomical features, such as organ contours and vascular topology,

while also capturing subtle dynamic details like microvascular textures and blood flow signals. In contrast to other models, our method avoids blurring critical details, especially in complex structures. This is achieved by combining deterministic guidance from anatomical priors with the probabilistic refinement that simulates the stochastic variations in medical images. As a result, our model ensures the preservation of both global structure and local details, leading to more accurate and clinically relevant image synthesis.

3.3 Ablation Study

We conducted an ablation study to investigate the effectiveness of the DGB, PRB, and C2FGAM. We performed ablation analysis on variants of our proposed method using these two datasets. The results are shown in Table 3, where "w/o" denotes our method without a certain module:

Table 3. Ablation study of designed components in our methods.

	Duke-Breast-Cancer-MRI			ChestCT-CTA		
Methods	MAE↓	PSNR↑	SSIM↑	MAE↓	PSNR↑	SSIM↑
Ours	3.92	27.34	80.34	6.99	24.30	84.41
w/o DGB	5.73	25.65	75.55	12.70	19.31	72.80
w/o C2FGAM	6.30	25.48	75.71	14.80	18.59	70.02
w/o PRB	6.33	24.36	73.11	15.25	17.95	68.87

We found that omitting the DGB, PRB, or C2FGAM leads to a decline in generation quality, further emphasizing the importance of combining anatomical priors and other deterministic structural information with probabilistic dynamic details such as blood flow to guide the image generation.

4 Conclusion

In this work, we propose a coarse-to-fine medical image translation framework that harmonizes deterministic guidance and probabilistic refinement to balance controllability and fidelity in generation. Specifically, the framework extracts anatomical priors (coarse-grained features) via the DGB, integrates them with fine-grained probabilistic optimization through the PRB, and guides the synthesis via the C2FGAM. Ablation studies validate the necessity of each component, demonstrating that removing any module degrades synthesis quality. Experiments on Duke-Breast-Cancer-MRI and ChestCT-CTA datasets show that our method achieves state-of-the-art performance, offering a robust solution for clinical scenarios requiring both structural precision and dynamic detail preservation.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China under grants W2411054, U21A20521 and 62271178, the Postgraduate Research & Practice Innovation Program of Jiangsu Province KYCX23_2524, National Foreign Expert Project of China under Grant G2023144009L, Zhejiang Provincial Natural Science Foundation of China (LR23F010002), Wuxi Health Commission Precision Medicine Project (J202106), Jiangsu Provincial Six Talent Peaks Project (YY-124), and Major Projects of Wuxi Health Commission (Z202324).

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

References

1. Chen, Y.J., Cheng, S.I., Chiu, W.C., Tseng, H.Y., Lee, H.Y.: Vector quantized image-to-image translation. In: European Conference on Computer Vision. pp. 440–456. Springer (2022)
2. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
3. Dong, H., Neekhara, P., Wu, C., Guo, Y.: Unsupervised image-to-image translation with generative adversarial networks. arXiv preprint arXiv:1701.02676 (2017)
4. Faucon, A.L., Bobrie, G., Clément, O.: Nephrotoxicity of iodinated contrast media: From pathophysiology to prevention strategies. European journal of radiology **116**, 231–241 (2019)
5. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., Tao, D.: Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2427–2436 (2019)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
8. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
9. Hu, X., Zhou, X., Huang, Q., Shi, Z., Sun, L., Li, Q.: Qs-attn: Query-selected attention for contrastive learning in i2i translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18291–18300 (2022)
10. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
12. Jaderberg, M., Simonyan, K., Zisserman, A., et al.: Spatial transformer networks. Advances in neural information processing systems **28** (2015)
13. Kim, B., Kwon, G., Kim, K., Ye, J.C.: Unpaired image-to-image translation via neural schr\" odinger bridge. arXiv preprint arXiv:2305.15086 (2023)
14. Kingma, D.P., Welling, M., et al.: Auto-encoding variational bayes (2013)

15. Li, B., Xue, K., Liu, B., Lai, Y.K.: Bbdm: Image-to-image translation with brownian bridge diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition. pp. 1952–1961 (2023)
16. Liu, L., Aviles-Rivero, A.I., Schönlieb, C.B.: Contrastive registration for unsupervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems (2023)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
18. Saha, A., H.M.R.G.L.J.W.J.C.E.H.K.C.E.G.S.V.W.R..M.M.A.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations. The Cancer Imaging Archive **28** (2021)
19. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
20. 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)
21. Yoon, D., Seo, M., Kim, D., Choi, Y., Cho, D.: Deterministic guidance diffusion model for probabilistic weather forecasting. arXiv preprint arXiv:2312.02819 (2023)
22. Yu, J., Li, X., Koh, J.Y., Zhang, H., Pang, R., Qin, J., Ku, A., Xu, Y., Baldridge, J., Wu, Y.: Vector-quantized image modeling with improved vqgan. arXiv preprint arXiv:2110.04627 (2021)
23. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3836–3847 (2023)
24. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)