Flow Matching for Medical Image Synthesis: Bridging the Gap Between Speed and Quality

Milad Yazdani[1][0009−0003−3922−102X], Yasamin Medghalchi[2][0009−0002−9415−2947], Pooria Ashrafian[2], Ilker Hacihaliloglu[3,4][0000−0003−3232−8193], and Dena Shahriari[2,5][0000−0002−3882−1977]⋆

[1] Department of Electrical and Computer Engineering, University of British Columbia (UBC), Vancouver, Canada
[2] School of Biomedical Engineering, UBC, Vancouver, Canada
[3] Department of Radiology, UBC, Vancouver, Canada
[4] Department of Medicine, UBC, Vancouver, Canada
[5] Department of Orthopaedics, Faculty of Medicine, UBC, Vancouver, Canada
dena.shahriari@ubc.ca

Abstract. Deep learning models have emerged as a powerful tool for various medical applications. However, their success depends on large, high-quality datasets that are challenging to obtain due to privacy concerns and costly annotation. Generative models, such as diffusion models, offer a potential solution by synthesizing medical images, but their practical adoption is hindered by long inference times. In this paper, we propose the use of an optimal transport flow matching approach to accelerate image generation. By introducing a straighter mapping between the source and target distribution, our method significantly reduces inference time while preserving and further enhancing the quality of the outputs. Furthermore, this approach is highly adaptable, supporting various medical imaging modalities, conditioning mechanisms (such as class labels and masks), and different spatial dimensions, including 2D and 3D. Beyond image generation, it can also be applied to related tasks such as image enhancement. Our results demonstrate the efficiency and versatility of this framework, making it a promising advancement for medical imaging applications. Code is available on: https://github.com/milad1378yz/MOTFM.

Keywords: Flow Matching · Diffusion Models · Image Generation.

1 Introduction

Over the past decade, artificial intelligence (AI), especially deep learning (DL), has significantly advanced disease detection and segmentation from medical images [28]. However, building reliable AI models for medical image analysis

⋆ Corresponding author

requires large, diverse datasets, which are hard to obtain due to privacy restrictions, rare diseases, and inconsistent diagnostic labels [21]. One solution is to generate synthetic data to augment existing datasets [6]. Deep generative models have shown promising results in various medical applications, enabling more robust training for machine learning models [16]. Early generative models like Generative Adversarial Networks (GANs) [7] have been widely used for medical image synthesis across modalities, including echocardiographic imaging [27]. Conditional GANs incorporate additional inputs for greater control, as in SPADE [15], which uses semantic layouts and has been applied to CT liver volumes, retinal fundus images, and cardiac cine-MRI [22]. Despite generating high-quality images, GANs often lack diversity and suffer from training instability and mode collapse without careful tuning [14]. To address these limitations, diffusion models, particularly Denoising Diffusion Probabilistic Models (DDPM) [9], have emerged as a powerful alternative. By formulating the image synthesis process as a continuous-time diffusion governed by stochastic differential equations (SDEs), DDPMs address some limitations of GANs, achieving superior performance in both image quality and diversity [4]. Recent studies show that these diffusion-based approaches even outperform GANs in generating medical images in various applications . Diffusion models incorporate various conditioning mechanisms for enhanced control. Prompt-based conditioning, like Latent Diffusion Models (LDM) [17], has been applied to Breast MRI and head CT synthesis [24], while mask- or image-based conditioning, as in ControlNet [33], has been used for colon polyp synthesis, showcasing their flexibility in medical imaging. Building on these advancements, diffusion models have also been optimized for efficiency. Deterministic variants like DDIM [23] reframe the stochastic process as an ordinary differential equation (ODE), significantly reducing inference steps while maintaining performance. However, despite their advantages, diffusion models still rely on iterative denoising through numerical ODE/SDE solvers, resulting in slow inference times. While DDIM mitigates this by reducing the number of steps, the underlying iterative solvers remain computationally demanding. In parallel, an alternative family of generative models, known as Flow Matching, emerged. Unlike diffusion models (which maximize a variational lower bound), flow matching directly approximates the transport between noise and data distributions [12]. Additionally, flow matching methods are flexible, allowing for diverse path definitions such as Gaussian, affine, and linear trajectories [12]. A key advancement in this family is flow matching with optimal transport [13], which provides a direct mapping between the source (typically noise) and the target distribution. Unlike diffusion models, which rely on iterative sampling, flow matching with optimal transport enables significantly faster sampling by constructing an efficient transport plan. Notably, optimal transport flow matching has shown remarkable performance in natural image generation [12] and enhancement [34], yet its potential for medical imaging remains largely unexplored. To the best of our knowledge, this is the first work leveraging flow matching with optimal transport for medical image synthesis. **Our key contributions are as follows: 1)** We present the first medical image synthesis framework leverag-

ing *Optimal Transport Flow Matching*, significantly accelerating inference while outperforming diffusion-based models in image quality. **2)** We evaluate its effectiveness across Unconditional, Class-Conditional, and Mask-Conditional Image Generation, demonstrating its robustness, and versatility across diverse generative tasks. **3)** Our method adapts to different medical imaging modalities (e.g., ultrasound, MRI) and spatial dimensions (2D, 3D), ensuring broad applicability. **4)** The proposed approach supports end-to-end training, eliminating the need for complex post-processing steps and simplifying the learning pipeline. This study opens a new pathway for medical image generation, demonstrating that flow matching with optimal transport is an effective and efficient alternative to traditional generative models in the medical field, with fast inference enabling real-time and interactive clinical applications such as simulation, training, and point-of-care workflows.

2 Method

2.1 Preliminaries

Both flow matching with optimal transport and diffusion models [9] aim to generate data from a complex target distribution X_1 starting from a simple Gaussian prior X_0. Diffusion models achieve this by modeling the transformation as a continuous-time stochastic differential equation (SDE), where a neural network estimates the drift term. These SDEs can be reformulated as probability flow ordinary differential equations (ODEs) [23], preserving marginal distributions while enabling faster inference. In diffusion models, data gradually transitions from X_1 to X_0 by introducing Gaussian noise, ϵ, at each step, $x_t = \sqrt{\bar{\alpha}_t}\, x_1 + \sqrt{1-\bar{\alpha}_t}\, \epsilon$. such that at the final time step T the data distribution becomes pure noise, x_0, and the model learns to estimate ϵ. The training objective minimizes the noise prediction error, $\mathcal{L}_{\text{diff}} = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\, I)} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \right]$. Note that in practice the naming in diffusion models is often reversed (e.g., x_T is noise and x_0 is the target); for consistency, we denote x_0 as the noise (source) and x_1 as the target, as illustrated in Fig. 1. Since the diffusion process follows a highly non-linear trajectory (Fig. 1.a), inference requires multiple iterative steps. In contrast, flow matching with optimal transport addresses this inefficiency by defining a straight-line (or nearly straight) transformation between X_0 and X_1 that approximates the optimal transport map under a quadratic cost. Specifically, it models data as $x_t = t\, x_1 + (1-t)\, x_0$. and trains a neural network to estimate the velocity field $v_\theta(x_t, t)$ such that ideally $v_\theta(x_t, t) = x_1 - x_0$. The corresponding loss function is $\mathcal{L}_{\text{OTFM}} = \mathbb{E}_{x_0, x_1} \left[\|(x_1 - x_0) - v_\theta(x_t, t)\|_2^2 \right]$. From the perspective of optimal transport, this formulation seeks to minimize the transport cost between the source and target distributions by matching the learned velocity with the true optimal transport velocity. During inference, samples are generated by solving the differential equation, $\frac{dx_t}{dt} = v_\theta(x_t, t)$. This direct mapping allows the model to theoretically recover X_1 from X_0 in a single step. In practice, however, the inferencing path is not strictly linear, and a minimal number of steps is still required far fewer than in diffusion models [13].

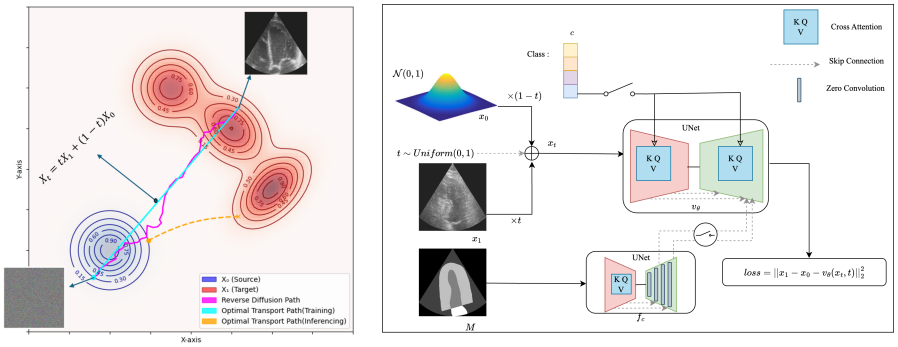

Fig. 1: a) The figure illustrates transitions from source x_0 (blue) to target x_1 (red). Diffusion models map noisy samples to targets (magenta), while flow matching provides a more efficient path (cyan for training, dashed orange for inference). The contour map represents probability density. b) MOTFM framework with different conditioning strategies.

Mathematically, by leveraging the structure of optimal transport, flow matching minimizes the discrepancy between the learned and optimal velocity fields, effectively bridging the two distributions with near-optimal efficiency.

2.2 Medical Optimal Transport Flow Matching (MOTFM)

Our framework, Medical Optimal Transport Flow Matching (MOTFM), generates synthetic images using a UNet backbone [18] with attention layers [29] to estimate velocity in flow matching. To improve memory efficiency, we employ flash attention [3] instead of standard attention. Flash attention is a GPU-optimized technique that reduces memory usage by loading queries, keys, and values only once and performing block-wise computations, thereby accelerating training and inference. As shown in Fig. 1.b, training involves progressively adding Gaussian noise to images with optimal transport. The UNet learns to estimate the velocity field that maps noisy inputs to the original image. The loss between the predicted and actual velocity fields is used to optimize network parameters. This approach is generalizable to both unconditional and conditional image generation, as described below.

a) Unconditional Image Generation As shown in Fig. 1.b, when no conditioning is provided, all switches are off, generating an image unconditionally. **b) Class-conditional Image Generation** The generation process can be guided by incorporating a one-hot class vector, indicating class number, into the UNet via the cross-attention mechanism, enabling generating images with a specific class. **c) Mask-Conditional Image Generation** Our method can optionally be guided by masks to generate images, which is particularly important if the downstream task is image segmentation. We introduce an additional UNet, f_c

(Fig. 1.b), to encode the mask. The decoder of f_c employs zero convolutions, and skip connections are established between the decoder of f_c and v_θ to incorporate mask-based guidance. Notably, during training, both UNets are updated end-to-end in contrast with [33]. Furthermore, this mask-conditioning approach can be seamlessly combined with class-conditioning. This pipeline can be extended to other image-to-image translation tasks.

3 Results

Experiment Settings. All generative models were trained with the Adam optimizer with learning rate 1e-4 for 200 epochs on an NVIDIA RTX 4090. For classification and segmentation, the same optimizer was used for 50 epochs, selecting the best model based on validation performance. For 3D generation tasks, such as MRI synthesis, the overall approach remains consistent, with the primary modifications being the use of 3D layers in place of their 2D counterparts.
Datasets. We utilized two datasets, one of which is the **CAMUS echocardiography (echo) dataset** [11], which contains 2D apical views of both two-chamber (2CH) and four-chamber (4CH) perspectives from 450 patients, covering both end-diastole (ED) and end-systole (ES) phases. Initially, a subset of images from 50 patients was randomly selected for the test split, while the remaining 400 patients were designated for training. Following the baseline approach [25], we further partitioned the dataset by assigning the first 50 patients to the validation set, leaving the remaining 350 patients for training. This resulted in a total of 1400 images for training and 200 images for validation. As a second dataset, we evaluate our pipeline on a different modality and dimension using the **3D MSD MRI Brain Dataset** [1]. The brain tumor dataset from the Medical Segmentation Decathlon (MSD) challenge [1] comprises 750 multiparametric MRI scans from patients diagnosed with glioblastoma or lower-grade glioma. Each scan includes T1-weighted (T1), post-Gadolinium contrast T1 (T1-Gd), T2-weighted (T2), and T2-FLAIR sequences. These images, collected from 19 institutions, were acquired using diverse clinical protocols and imaging equipment. For simplicity, we focus only on T1-weighted images in this study.
Experiments. We evaluate our pipeline on two mentioned datasets, comparing it to baselines, including our framework with DDPM (under various conditioning settings), SPADE, and ControlNet as a mask-guided approach. For efficiency, we used the DDIM scheduler [14] during DDPM sampling. The echocardiography dataset is used for all conditioning methods, while the MRI dataset is used for unconditional generation.
Qualitative Visualization. Fig. 2 presents image generation examples for echocardiography (first two rows) and MRI (last row) using different conditioning strategies for DDPM and MOTFM. In the echocardiography dataset, the mask is used solely for conditioning in the mask-based generation approach. Across both modalities, DDPM introduces a noticeable increase in brightness, deviating from the real data distribution, whereas MOTFM maintains a pixel distribution more consistent with real images. A similar brightness shift is also

observed in the ControlNet-generated echocardiography images. To quantify this effect, Fig. 3 shows the kernel density estimation (KDE) of pixel intensities for real, DDPM-generated, and MOTFM-generated echocardiographic images. The KDE plot highlights a higher density in the bright intensity range [150, 250] for DDPM, indicating a shift toward brighter outputs. This brightness shift, also observed in natural image generation [2], highlights the limitations of diffusion models in maintaining realistic intensity distributions. Furthermore, MOTFM outperforms DDPM, ControlNet, and SPADE in echocardiographic image quality. In MRI synthesis, 50-step MOTFM produces higher-quality images than DDPM, and even its 10-step output outperforms 50-step DDPM, highlighting its efficiency.

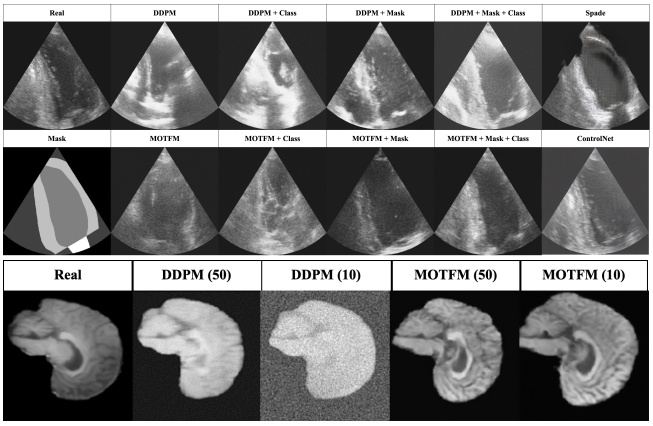

Fig. 2: Comparison of echocardiographic and brain MRI synthesis using DDPM and MOTFM, with SPADE and ControlNet applied only to echocardiography. The first two rows show echocardiographic images, while the last row presents brain MRI synthesis, with numbers in parentheses indicating inference steps.

Generation Evaluations (CAMUS Dataset). To evaluate our framework, we compare MOTFM with baselines (mentioned in "**Experiments**") in echo image generation. Performance is evaluated using FID, CMMD, and KID (distribution distance), IS (sample diversity), and SSIM (structural similarity) [20, 19, 10, 32, 30]. Table 1 presents the results for the CAMUS dataset, computed over 2000 generated images. MOTFM consistently outperforms baselines across most metrics. Notably, one-step MOTFM outperforms 10-step DDPM and achieves the same order of performance as 50-step DDPM, demonstrating superior efficiency without compromising quality. While its slightly lower Inception Score suggests a trade-off between realism and diversity, MOTFM's substantial gains in FID, SSIM, and CMMD establish it as a more effective and efficient alternative for echocardiographic image synthesis. We also compared inference times:

10-step MOTFM finishes in 1.47s, much faster than 50-step DDPM (6.20s) and approaching SPADE's 0.68s, further underscoring its efficiency.

Generation Evaluations (MSD MRI Brain Dataset). To evaluate 3D image generation on the 3D MSD brain dataset, we compare MOTFM with DDPM for unconditional synthesis using 3D-FID, MMD (distribution distance), and MS-SSIM (structural similarity), following prior work on 3D brain MRI generation [26, 8, 31, 5]. Table 2 reports results over 2000 generated samples, showing that MOTFM consistently outperforms DDPM. Notably, one-step MOTFM surpasses 50-step DDPM in MS-SSIM and MMD and achieves competitive results in the 3D-FID, demonstrating its efficiency and adaptability for 3D medical image synthesis.

Table 1: Evaluation of Echocardiography Image Generation. D, M, and SD-M represent DDPM, MOTFM, and ControlNet, respectively. "-C", "-M", and "-CM" denote class, mask, and class + mask conditionings. The columns 1, 10, and 50 indicate the number of inference steps. The best results for each conditioning are highlighted in **bold**.

	FID ↓			**SSIM ↑**			**KID ↓**			**CMMD ↓**			**IS ↑**		
	1	10	50	1	10	50	1	10	50	1	10	50	1	10	50
D	1.9e2	22.83	1.84	.00	.08	0.29	1.6e3	62.2	1.29	5.38	1.30	1.52	1.07	**3.16**	2.07
M	3.04	.16	**.04**	**.70**	.65	.62	4.61	0.19	**.03**	1.95	.96	**.50**	1.16	1.42	1.39
D-C	1.9e2	15.21	4.01	.00	.10	0.27	1.6e3	22.1	5.36	5.38	1.61	1.59	1.07	**2.84**	1.92
M-C	1.93	.08	**.06**	.62	.64	**0.65**	2.93	**.07**	.08	2.15	1.73	**.76**	1.34	1.35	1.34
Spade	.46			.54			.73			.46			1.73		
D-M	1.9e2	14.58	1.61	.00	.14	.35	1.6e3	24.1	.75	5.39	1.62	1.57	1.07	**2.36**	1.72
SD-M	5.67	2.25	1.82	.57	.56	.63	8.87	2.37	1.99	3.42	0.42	.39	1.70	1.76	1.67
M-M	3.91	.58	**.22**	**.72**	.67	.66	5.81	.75	**.23**	1.29	.16	**.12**	1.30	1.28	1.23
D-CM	1.9e2	25.66	7.98	.00	.15	.28	1.6e3	83.8	17.05	5.39	1.37	1.23	1.07	**2.67**	2.13
M-CM	3.21	**.07**	**.07**	.64	.69	**.70**	5.01	.05	**.03**	1.94	.72	**.64**	1.34	1.34	1.33

Downstream Tasks. To further evaluate the realism of our generated images, we assess their performance in downstream tasks, specifically classification and segmentation. We first train models (Table 3) on real training images and evaluate them on the real test set of the CAMUS dataset. For segmentation, UNet-ResX refer to UNet models that use ResNetX as their encoder backbones. To compare, we train classifiers on only class-conditioned synthetic images and segmentors on only mask-conditioned synthetic images, ensuring dataset sizes match the real training set. All downstream models trained on synthetic data were also tested exclusively on real ultrasound data. MOTFM generates images in just 10 steps, whereas DDPM requires 50 steps to achieve comparable quality. As shown in Table 3, our 10-step MOTFM method outperforms the 50-step DDPM, yielding results closer to the original distribution. This demonstrates

Table 2: Evaluation of Brain MRI Unconditional Image Generation. D and M represent DDPM and MOTFM respectively. The columns 1, 10, and 50 indicate the number of inference steps. MMD values in the tables are divided by 1000 for readability. The best results are highlighted in **bold**.

	3D-FID ↓			MS-SSIM ↑			MMD ↓		
	1	10	50	1	10	50	1	10	50
D	146.47	51.68	29.67	.06	.51	.59	39.8	26.1	4.28
M	32.10	9.27	**7.93**	.66	.77	**.77**	.51	.25	**.22**

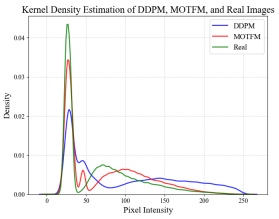

Fig. 3: KDE Plot of Pixel Intensity Distributions for Generated and Real Echo Images.

MOTFM's improved fidelity and efficiency, making it a more effective approach for medical image synthesis.

Table 3: Classification and Segmentation Metrics for Echo

Classification				Segmentation				
Model	**Data**	**ACC ↑**	**F1 ↑**	**Model**	**Dice ↑**	**IOU ↑**	**HD ↓**	**ASD ↓**
ResNet18	Real	0.89	0.89	UNet-Res18	0.92	0.86	16.67	3.04
	DDPM (50)	0.78	0.78		0.82	0.71	88.03	10.94
	MOTFM (10)	**0.82**	**0.82**		**0.91**	**0.84**	**21.2**	**3.65**
ResNet50	Real	0.88	0.88	UNet-Res50	0.92	0.87	15.02	2.84
	DDPM (50)	0.78	0.78		0.73	0.59	133.18	25.66
	MOTFM (10)	**0.81**	**0.81**		**0.91**	**0.85**	**20.89**	**3.64**

Generalizability to Other Medical Imaging Tasks. Our pipeline extends beyond image generation, demonstrating versatility in tasks like denoising. We apply it to speckle noise removal on the CAMUS dataset by adding speckle noise with different power to clean images and training the model to recover the original data. Instead of sampling from Gaussian noise, we initialize from noisy images, learning a transformation between noisy and clean domains (Fig.1.b). While denoising is not our main focus, Fig.4 shows its promising performance across inference steps. Based on validation data, the denoised images achieve average metrics of PSNR: 32.76, SSIM: 0.8401, and SNR: 23.02, compared to the noisy images' PSNR: 21.61, SSIM: 0.5802, and SNR: 11.87, highlighting its broader potential in medical imaging.

4 Conclusion

This study introduces Optimal Transport Flow Matching framework for medical image synthesis with diverse conditioning strategies, adaptable across modalities

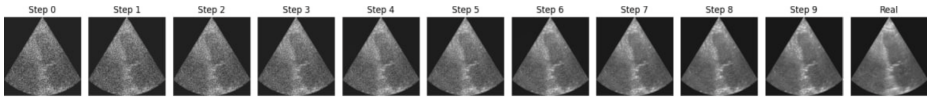

Fig. 4: Denoising Example: From a Noisy Image to a Denoised Image in 10 Steps

and dimensions. Our approach surpasses diffusion-based baselines with fewer inference steps, achieving superior image quality and efficiency. Beyond synthesis, it can be extended to tasks such as image-to-image translation and denoising. Future work will focus on improving sample diversity with this pipeline.

Acknowledgments. This work was supported by the Canadian Foundation for Innovation - John R. Evans Leaders Fund (CFI-JELF) and the Natural Sciences and Engineering Research Council of Canada (NSERC). Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG).

Disclosure of Interests. The authors declare no competing interests for this paper.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
3. Dao, T., Fu, D., Ermon, S., Rudra, A., Ré, C.: Flashattention: Fast and memory-efficient exact attention with io-awareness. Advances in Neural Information Processing Systems **35**, 16344–16359 (2022)
4. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
5. Dorjsembe, Z., Pao, H.K., Odonchimed, S., Xiao, F.: Conditional diffusion models for semantic 3d brain mri synthesis. IEEE Journal of Biomedical and Health Informatics (2024)
6. Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., Greenspan, H.: Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. Neurocomputing **321**, 321–331 (2018)
7. 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)
8. Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.: A kernel two-sample test. The Journal of Machine Learning Research **13**(1), 723–773 (2012)
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
10. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking fid: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9307–9315 (2024)

11. Leclerc, S., Smistad, E., Pedrosa, J., Østvik, A., Cervenansky, F., Espinosa, F., Espeland, T., Berg, E.A.R., Jodoin, P.M., Grenier, T., et al.: Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. IEEE transactions on medical imaging **38**(9), 2198–2210 (2019)
12. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
13. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
14. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021)
15. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2337–2346 (2019)
16. Reynaud, H., Qiao, M., Dombrowski, M., Day, T., Razavi, R., Gomez, A., Leeson, P., Kainz, B.: Feature-conditioned cascaded video diffusion models for precise echocardiogram synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 142–152. Springer (2023)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
18. 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)
19. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016)
20. Seitzer, M.: pytorch-fid: Fid score for pytorch. `https://github.com/mseitzer/pytorch-fid` (08 2020), version 0.3.0
21. Singh, R.P., Hom, G.L., Abramoff, M.D., Campbell, J.P., Chiang, M.F., et al.: Current challenges and barriers to real-world artificial intelligence adoption for the healthcare system, provider, and the patient. Translational Vision Science & Technology **9**(2), 45–45 (2020)
22. Skandarani, Y., Jodoin, P.M., Lalande, A.: Gans for medical image synthesis: An empirical study. Journal of Imaging **9**(3), 69 (2023)
23. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
24. Song, W., Jiang, Y., Fang, Y., Cao, X., Wu, P., Xing, H., Wu, X.: Medical image generation based on latent diffusion models. In: 2023 International Conference on Artificial Intelligence Innovation (ICAII). pp. 89–93 (2023)
25. Stojanovski, D., Hermida, U., Lamata, P., Beqiri, A., Gomez, A.: Echo from noise: synthetic ultrasound image generation using diffusion models for real image segmentation. In: International Workshop on Advances in Simplifying Medical Ultrasound. pp. 34–43. Springer (2023)
26. Sun, L., Chen, J., Xu, Y., Gong, M., Yu, K., Batmanghelich, K.: Hierarchical amortized gan for 3d high resolution medical image synthesis. IEEE journal of biomedical and health informatics **26**(8), 3966–3975 (2022)
27. Tiago, C., Gilbert, A., Beela, A.S., Aase, S.A., Snare, S.R., Šprem, J., McLeod, K.: A data augmentation pipeline to generate synthetic labeled datasets of 3d echocardiography images using a gan. IEEE Access **10**, 98803–98815 (2022)

28. Ting, D., Liu, Y., Burlina, P., Xu, X., Bressler, N., Wong, T.: Ai for medical imaging goes deep. Nature Medicine **24**(5), 539–540 (2018)
29. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
30. 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)
31. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
32. Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., Weinberger, K.: An empirical study on evaluation metrics of generative adversarial networks. arXiv preprint arXiv:1806.07755 (2018)
33. 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)
34. Zhu, Y., Zhao, W., Li, A., Tang, Y., Zhou, J., Lu, J.: Flowie: Efficient image enhancement via rectified flow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13–22 (2024)