This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

SV-DRR: High-Fidelity Novel View X-Ray Synthesis Using Diffusion Model

Chun Xie[1]✉[0000−0003−4936−7404], Yuichi Yoshii[2][0000−0003−1447−4664], and Itaru Kitahara[1][0000−0002−5186−789X]

[1] Center for Computational Sciences, University of Tsukuba, Japan
[2] Tokyo Medical University Ibaraki Medical Center, Japan
xiechun@ccs.tsukuba.ac.jp

Abstract. X-ray imaging is a rapid and cost-effective tool for visualizing internal human anatomy. While multi-view X-ray imaging provides complementary information that enhances diagnosis, intervention, and education, acquiring images from multiple angles increases radiation exposure and complicates clinical workflows. To address these challenges, we propose a novel view-conditioned diffusion model for synthesizing multi-view X-ray images from a single view. Unlike prior methods, which are limited in angular range, resolution, and image quality, our approach leverages the Diffusion Transformer to preserve fine details and employs a weak-to-strong training strategy for stable high-resolution image generation. Experimental results demonstrate that our method generates higher-resolution outputs with improved control over viewing angles. This capability has significant implications not only for clinical applications but also for medical education and data extension, enabling the creation of diverse, high-quality datasets for training and analysis. Our code is available at https://github.com/xiechun298/SV-DRR.

Keywords: Novel view synthesis · X-ray image generation · Diffusion models · Radiography · Medical image augmentation.

1 Introduction

X-ray imaging is a widely used, cost-effective technique for visualizing internal human anatomy. It utilizes high-energy X-rays to penetrate the body and capture residual radiation on a flat detector, enabling non-invasive diagnosis. However, conventional radiography is typically limited to a single-view projection, such as in chest radiography, where only a frontal image is acquired per session. While radiologists can infer 3D spatial relationships from these 2D images, this process relies heavily on subjective intuition, which is challenging to standardize and visualize explicitly. Furthermore, acquiring additional views requires multiple exposures, increasing radiation risk and complicating clinical workflows.

To address these challenges, novel view synthesis offers a promising solution [6, 19, 25] by generating multiple perspectives from a single X-ray image.

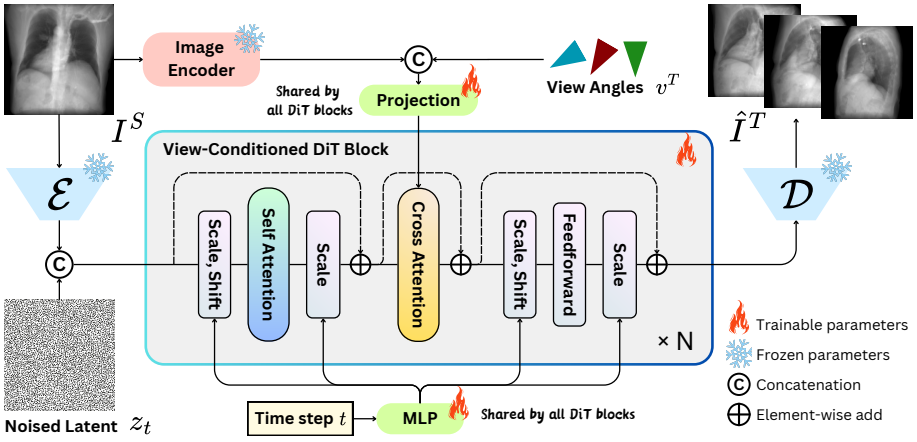

Fig. 1. Overview of SV-DRR. Given a source X-ray image I^S and relative target views v^T, SV-DRR synthesizes realistic X-ray projections $\hat{I}^T$ corresponding to v^T from Gaussian noise z_t via a denoising DiT. The denoising process takes place in the latent space of a VAE, where $\mathcal{E}$, and $\mathcal{D}$ denote the encoder and decoder, respectively.

This capability could facilitate more intuitive assessments, particularly in settings without access to CT imaging, reduce radiation exposure, and mitigate the effects of patient movement. Moreover, novel view synthesis is valuable for applications such as sparse-view CT reconstruction [12, 14], improving imaging efficiency and expanding the utility of X-ray imaging in diagnostic and educational settings.

This paper presents SV-DRR(Fig.1), a novel view-conditioned diffusion model for multi-view X-ray synthesis, overcoming the limitations of existing methods in angular range, resolution, and stability. The name SV-DRR, short for Single-View DRR, is inspired by Digitally Reconstructed Radiography (DRR). Unlike DRR, which renders X-ray projections from a 3D CT volume, our method synthesizes novel views directly from a single 2D projection. Our approach formulates view synthesis as a view-conditioned image variation task, leveraging a Diffusion Transformer (DiT) [18] to enhance anatomical coherence and detail preservation. Additionally, a weak-to-strong training strategy stabilizes high-resolution synthesis, while a densely sampled dataset with view annotations provides a robust foundation for training and evaluation across varying viewpoints.

Our main contributions are summarized as follows:

– ***Advancing View Synthesis in X-ray Imaging:*** We propose a view-conditioned DiT model that overcomes previous limitations in angular coverage and resolution, significantly expanding the range of synthesizable viewpoints. This enables more flexible and clinically useful multi-view X-ray synthesis.

- ***Improving Stability and Resolution:*** Our weak-to-strong training strategy enhances synthesis quality, allowing for high-resolution X-ray image generation with improved stability and computational efficiency, making it more feasible for real-world applications.
- ***Enhancing Data Availability for Training and Evaluation:*** We create a densely sampled dataset with precise view annotations, enabling robust training for synthesizing multi-view X-ray images and facilitating research in sparse-view CT reconstruction and other related fields.

2 Related Work

GAN-based methods: GAN [8]-based methods have been widely explored for single-image view synthesis, particularly in natural image domains, where they infer missing structures to generate novel views [17]. In X-ray imaging, recent approaches leverage CT priors and geometric constraints [6, 19, 25] to enhance anatomical fidelity and structural coherence. However, these methods remain limited in view range and resolution, restricting their applicability for high-quality, multi-angle X-ray synthesis.

Diffusion-based methods: Recently, diffusion models have demonstrated exceptional performance in image generation by iteratively refining noisy inputs. Foundational works such as Denoising Diffusion Probabilistic Models (DDPM) [11] and Denoising Diffusion Implicit Models (DDIM) [26] established the basis for modern diffusion-based synthesis. These advancements have led to the development of powerful conditional image generation models [4, 5, 20, 23]. More recently, diffusion models have been applied to novel view synthesis [3,15,22,28], where camera parameters serve as conditioning inputs for generating new viewpoints. Building on these developments, we propose a view-conditioned diffusion model for X-ray synthesis, which incorporates explicit view embeddings and latent-space conditioning to achieve high-quality multi-view generation.

3 Proposed Method

3.1 Overview

X-ray view synthesis aims to generate novel projections from a given input X-ray image and its associated view parameters. Given an input source image I^S with corresponding view parameters v^S, the objective is to synthesize an X-ray image $\hat{I}^T$ from a target viewpoint defined by v^T. By modeling the structural relationships between different projections, our method estimates the conditional distribution:

$$\hat{I}^T \sim \mathcal{P}(I|I^S, v^S, v^T), \tag{1}$$

ensuring that the generated images accurately preserve anatomical structures of I^S.

3.2 View-Conditioned Diffusion Transformer

Our View-Conditioned Diffusion Transformer (VCDiT) follows the Latent Diffusion Model (LDM) framework [24], where the diffusion process is performed in VAE latent space [13], reducing computational complexity while preserving fine details. As shown in Fig.1, the denoising network ϵ_θ is based on DiT, enhanced with a cross-attention mechanism inspired by [5] to inject view-conditioning. To improve efficiency, we use a shared AdaLN-Single layer for timestep embedding t and a shared projection layer (a learnable linear layer) to efficiently map the view-conditioning signal to the VCDiT latent space.

To enable view-conditioned synthesis, VCDiT integrates conditioning through two complementary streams. First, the embedding of the source image I^S is concatenated with an encoded relative polar coordinate between the target view v^T and the source view v^S, forming a view embedding that provides spatial transformation information for controlled generation. Second, the source image latent is channel-concatenated with the noised target latent before denoising, reinforcing structural alignment and improving synthesis quality.

The training objective follows standard conditional diffusion models, minimizing the loss:

$$\mathcal{L} = \mathbb{E}_{z_0,c,\epsilon,t}\|\epsilon - \epsilon_\theta(z_t, c(I^S, v^S, v^T), t)\|_2^2 \tag{2}$$

where $\epsilon \sim \mathcal{N}(0, I)$ represents Gaussian noise added to the latent, and z_0 and z_t denote the original and noised latents of the target image I^T, respectively. The function c represents the conditioning mechanism.

By incorporating these conditioning mechanisms, VCDiT enables the synthesis of high-resolution novel X-ray views while maintaining anatomical accuracy and spatial coherence.

3.3 Weak-to-Strong Training Strategy

Our training strategy follows a progressive resolution refinement approach, where the model is initially trained on lower-resolution(LR) images before gradually increasing the resolution. This method allows the model to efficiently learn fundamental X-ray image features at a reduced computational cost before adapting to high-resolution(HR) image synthesis.

Previous studies have reported a decline in performance during the transition from LR to HR training due to inconsistencies in positional embeddings across different resolutions. To address this issue, we follow the approach of [4, 5, 20] and adopt the positional embedding interpolation trick described in [29], where the HR model's positional embeddings are initialized by interpolating those from the LR model. This approach mitigates discrepancies between resolutions, significantly improving the stability and efficiency of fine-tuning at HR X-ray images while maintaining structural coherence and detail accuracy.

4 Experiment and Analysis

To validate the superiority of SV-DRR, we compare it with multiple state-of-the-art view synthesis approaches on simulated X-ray images generated from CT volumes. The experiments assess model performance across different resolutions and view sets, evaluating both fidelity and robustness in synthesizing novel X-ray views.

4.1 Dataset

LIDC-IDRI-DRR The Lung Image Database Consortium image collection (LIDC-IDRI) [2] consists of 1,012 chest CT scans. To ensure higher image fidelity and better reconstruction quality, we exclude CT scans with a slice thickness greater than 2.5 mm, resulting in a total of 889 CT volumes. We randomly select 16 volumes for evaluation and use the rest for training. The dataset will be released along with our code.

For preprocessing, we use 3D Slicer [1] to remove CT tables. Chest X-ray images are synthesized from 1,500 views for each CT scan using DiffDRR [9] to capture a wide range of perspectives, enhancing the robustness of our training dataset. We render X-ray images separately for different resolutions to avoid any loss in image quality caused by resizing. The view orientations are directed toward the center of the CT volumes. View positions are sampled on a hemisphere with a radius of 1.8 meters using Fibonacci lattice sampling. The first view is set to the standard Posterior-Anterior (PA) perspective and is always used as the source image in our experiment.

4.2 Implementation Details

For latent encoding, we employ the VAE [13] used in SDXL [20]. We utilize CLIP [21] as the image encoder to generate image embeddings, which serve as conditioning inputs for the VCDiT model's cross-attention mechanism, ensuring improved feature alignment and synthesis quality. Additionally, our VCDiT model is initialized using PixArt-Σ-256 [4] weights, leveraging its robust feature extraction capabilities.

Our model is trained using the AdamW optimizer, with learning rates set to 5×10^{-6}, 3×10^{-6}, and 1×10^{-6} for 256×256, 512×512, and 1024×1024 resolutions, respectively. The corresponding batch sizes are 64, 32, and 8. Training is conducted on a single H100 GPU: 200K steps ($\sim$2 days) for 256×256, and 100K steps for 512×512 and 1024×1024 ($\sim$1.5 and $\sim$2 days, respectively). For sampling, we use DPM-Solver [16] with 20 inference steps and a guidance scale of 3. In our experiments, the generation time for a single image is approximately 1.5 seconds for 256×256, 2 seconds for 512×512, and 5 seconds for 1024×1024.

Table 1. Quantitative comparison of novel view synthesis on LIDC-IDRI-DRR

	Simple views				Hemisphere views			
	SSIM↑	PSNR↑	LPIPS↓	FID↓	SSIM↑	PSNR↑	LPIPS↓	FID↓
XraySyn [19]	0.4634	8.7670	0.4671	1.1902	0.2226	3.4804	0.2484	0.8450
Zero123 [15]	0.2933	8.2895	0.4849	0.9060	0.1217	3.6117	0.2714	1.0777
Zero123-XL [7]	0.5092	13.3156	0.3205	0.5412	0.2146	5.6295	0.1941	0.6832
SV-DRR 256	0.7374	22.5136	0.1170	**0.1880**	0.3600	10.8452	0.0640	**0.1493**
SV-DRR 512	**0.7509**	**23.3984**	**0.1073**	0.2040	**0.3680**	**11.2855**	**0.0588**	0.1693
SV-DRR 1024	0.7336	22.7290	0.1191	0.1955	0.3600	10.9678	0.0644	0.1753

4.3 Comparison with State-of-the-art Methods

Table1 presents a quantitative comparison of our method against state-of-the-art methods, including XraySyn [19], Zero123 [15], and Zero123-XL [7], using PSNR, SSIM [27], LPIPS [30], and FID [10] metrics.

We evaluate performance using two distinct view sets. The first set *Simple* involves varying the azimuth angle from -90° to 90° in 5° increments, resulting in 36 novel views. This setup resembles typical chest CT rotations and aligns with XraySyn's testing conditions. The second set *Hemisphere* is more challenging, consisting of 1,499 views distributed on a hemisphere with positions sampled using Fibonacci lattice sampling. This setup evaluates the flexibility of each method in terms of the range of synthesizable viewpoints, which is essential for future adaptation to anatomically diverse regions such as the knee, shoulder, or pelvis.

Our method achieves superior performance across all evaluation metrics in both view sets, demonstrating its capability to synthesize high-fidelity novel X-ray views even for large angular displacements. This improvement is driven by our densely sampled LIDC-IDRI-DRR dataset, which provides diverse training data, and our VCDiT model, which effectively captures the relationship between view conditioning and the preservation of fine details as well as structural coherence in synthesized images.

Furthermore, the minimal variance in evaluation metrics across different output resolutions highlights the stability of our approach in generating high-resolution X-ray images. This robustness is attributed to our weak-to-strong training strategy, which ensures consistency and reliability in high-resolution synthesis.

4.4 Visualization

Fig.2 presents a comparative analysis of view synthesis results generated by our method at resolutions of 256, 512, and 1024, alongside Zero123 [15], Zero123-XL [7], XraySyn [19], and MedNeRF [6], as well as the ground truth images from LIDC-IDRI-DRR.

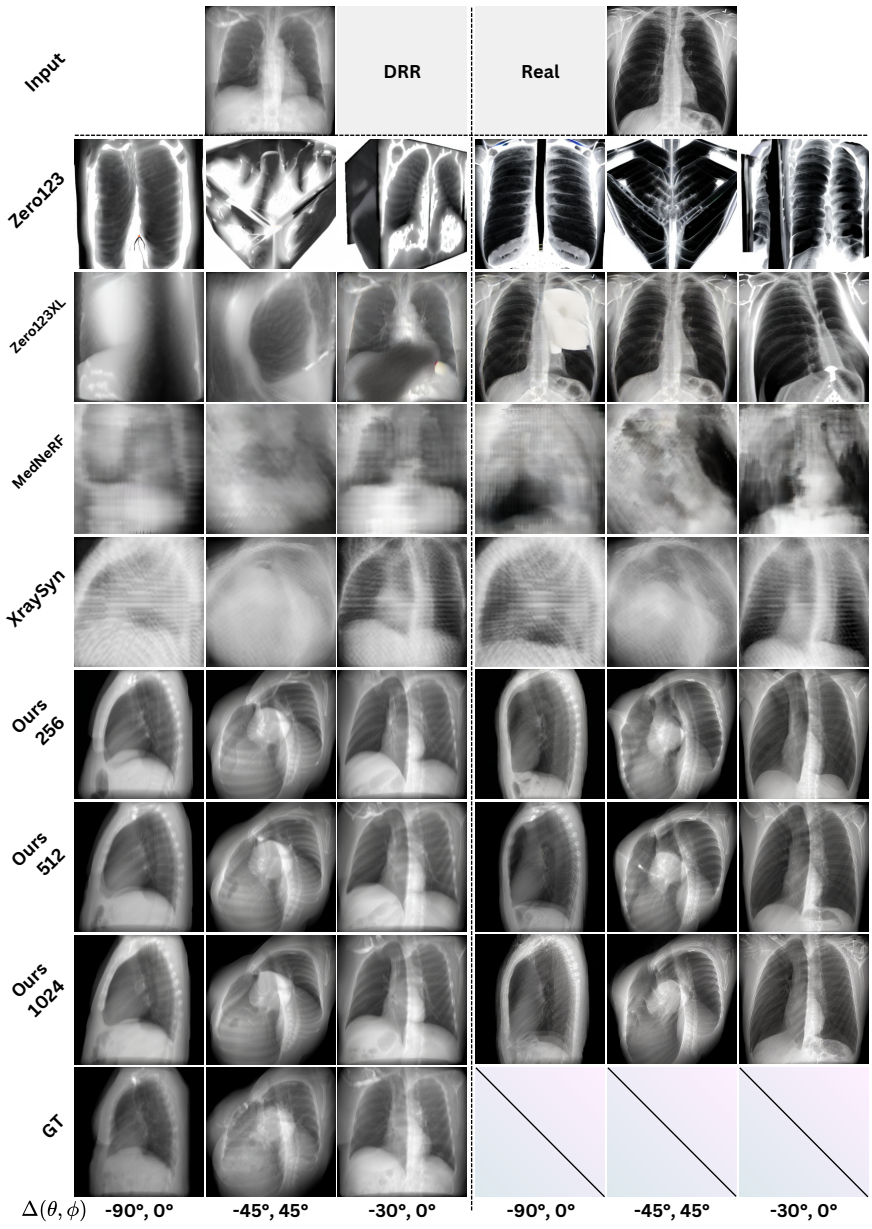

Fig. 2. Comparison of synthesized X-ray images at different resolutions using our method (256, 512, 1024) against baselines. Our method achieves superior fidelity in structure and orientation for both simulated DRR(left) and real X-ray(right) input. θ and ϕ represent the azimuthal and elevation angles of the synthesized view, respectively. Our method's different resolutions are normalized for better visualization. Additionally, the brightness and contrast of XraySyn images were manually adjusted for better comparison without introducing any negative effects on the results.

We evaluate our method using both simulated and real X-ray images as inputs, where the real X-ray images are obtained from publicly available resources such as Radiopaedia. Across all tested resolutions, our approach consistently outperforms state-of-the-art methods, demonstrating superior fidelity in preserving anatomical structures and maintaining spatial consistency. XraySyn and MedNeRF retain the main anatomical structures only within a limited angular displacement, with increasing noise at larger angles. Zero123 produces transformations that resemble rigid rotations of a cubic object, leading to unrealistic projections. Interestingly, Zero123-XL, despite its enhanced zero-shot learning capabilities, struggles to incorporate the given conditions effectively, resulting in random distortions of the input image.

5 Subjective Evaluation of Image Realism

5.1 Experimental Design

To evaluate the realism of the generated chest X-ray images, we conducted a user study involving medical experts. The objective was to determine whether experts could reliably distinguish between two sets of generated images: one produced using our proposed model and the other synthesized by applying DiffDRR to the LIDC-IDRI dataset to simulate real multi-view X-ray images.

Fifteen board-certified medical experts and clinical practitioners evaluated 50 pairs of images. Each pair consisted of one image generated using DiffDRR and one image generated by our method, both from the same view angle. Experts were asked to determine which image in each pair appeared more real. The classification was conducted using a web-based interface, and each expert completed the evaluation independently.

5.2 Results

The average classification accuracy across participants was **48.6%** (range: **40.0%–54.8%**), indicating performance close to random chance (50%). A **one-sample t-test** against the null hypothesis of 50% accuracy yielded a **t-statistic of** -1.06 with a **p-value of 0.306**, showing no significant difference from chance. A **binomial test** on the total correct classifications (**358 out of 750**) produced a **p-value of 0.228**, further supporting this conclusion.

The results strongly suggest that the images generated by our method are indistinguishable from those simulated using DiffDRR, even by medical experts. The absence of statistical significance in classification performance implies that the generated images possess high realism and fidelity. These findings highlight the potential of our method for data augmentation, simulation, and training applications in radiology.

6 Conclusion

We introduced SV-DRR, a novel diffusion-based X-ray view synthesis method, addressing the limitations of previous works in terms of angular range, resolution, and stability. Using DiT and a weak-to-strong training strategy, our method generates high-resolution, anatomically coherent X-ray images. Additionally, our densely sampled dataset with view annotations enhances training and evaluation. Experimental results demonstrate that our method surpasses state-of-the-art approaches in quality and flexibility. Future work will focus on improving cross-view consistency to enhance anatomical alignment and visual realism, as well as adapting the model to diverse anatomies, including cases involving disabilities or pediatric patients. These efforts aim to further advance the applicability of multi-view X-ray synthesis in clinical and diagnostic settings.

Acknowledgments. This work was supported in part by JSPS KAKENHI (grant number JP25100497), AMED (grant number JP250126803), and the Multidisciplinary Cooperative Research Program in CCS, University of Tsukuba.

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

References

1. 3d slicer. https://www.slicer.org/, accessed: 2024-10-18
2. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics **38**(2), 915–931 (2011)
3. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., De Mello, S., Karras, T., Wetzstein, G.: GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
4. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-Σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation (2024)
5. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023)
6. Corona-Figueroa, A., Frawley, J., Taylor, S.B., Bethapudi, S., Shum, H.P.H., Willcocks, C.G.: Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). pp. 3843–3848 (2022). https://doi.org/10.1109/EMBC48229.2022.9871757
7. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., VanderBilt, E., Kembhavi, A., Vondrick, C., Gkioxari, G., Ehsani, K., Schmidt, L., Farhadi, A.: Objaverse-xl: A universe of 10m+ 3d objects. arXiv preprint arXiv:2307.05663 (2023)

8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (eds.) Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014)
9. Gopalakrishnan, V., Golland, P.: Fast auto-differentiable digitally reconstructed radiographs for solving inverse problems in intraoperative imaging. In: Workshop on Clinical Image-Based Procedures. pp. 1–11. Springer (2022)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems. pp. 6626–6637 (2017)
11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
12. Jeong, Y.S., Yoo, H.B., Chun, I.Y.: Dx2ct: Diffusion model for 3d ct reconstruction from bi or mono-planar 2d x-ray(s) (2025), https://arxiv.org/abs/2409.08850
13. Kingma, D.P., Welling, M.: An introduction to variational autoencoders. Foundations and Trends in Machine Learning **12**(4), 307–392 (2019). https://doi.org/10.1561/2200000056
14. Kyung, D., Jo, K., Choo, J., Lee, J., Choi, E.: Perspective projection-based 3d ct reconstruction from biplanar x-rays. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 1–5 (2023)
15. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object (2023)
16. Lu, C., Zhou, Y., Bao, F., Chen, J., LI, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 5775–5787. Curran Associates, Inc. (2022)
17. Park, E., Yang, J., Yumer, E., Ceylan, D., Berg, A.C.: Transformation-grounded image generation network for novel 3d view synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 702–711 (2017). https://doi.org/10.1109/CVPR.2017.82
18. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4195–4205 (2023). https://doi.org/10.1109/ICCV53048.2023.00420
19. Peng, C., Liao, H., Wong, G.G., Luo, J., Zhou, S.K., Chellappa, R.: Xraysyn: Realistic view synthesis from a single radiograph through ct priors. In: AAAI Conference on Artificial Intelligence (12 2020). https://doi.org/10.1609/aaai.v35i1.16120
20. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023), https://arxiv.org/abs/2307.01952
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021)
22. Rockwell, C., Fouhey, D.F., Johnson, J.: PixelSynth: Generating a 3d-consistent experience from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
23. 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. pp. 10684–10695 (2022)

24. 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)
25. Shen, L., Yu, L., Zhao, W., Pauly, J., Xing, L.: Novel-view x-ray projection synthesis through geometry-integrated deep learning. Medical Image Analysis **77**, 102372 (2022). https://doi.org/10.1016/j.media.2022.102372
26. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2021)
27. 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)
28. Watson, D., Chan, W., Martin-Brualla, R., Ho, J., Tagliasacchi, A., Norouzi, M.: Novel view synthesis with diffusion models. In: International Conference on Learning Representations (2023)
29. Xie, E., Yao, L., Shi, H., Liu, Z., Zhou, D., Liu, Z., Li, J., Li, Z.: Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4230–4239 (October 2023). https://doi.org/10.1109/ICCV51070.2023.00390
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)