CT^2-DiT: Enabling Abdominal Non-Contrast to Contrast-Enhanced CT Translation with Diffusion Transformers in the Low-Data Regime

Han Zhu, Yu Song, and Ikuko Nishikawa*

Ritsumeikan University, Japan
zhuhan0924@outlook.com, yusong@fc.ritsumei.ac.jp, nishi@ci.ritsumei.ac.jp

Abstract. In this paper, we propose CT^2-DiT, a simple yet powerful framework for abdominal non-contrast CT (NCCT) to contrast-enhanced CT (CECT) translation, with a particular focus on focal liver lesions (FLLs). CECT plays an important role in the clinical diagnosis of FLLs by providing a clear contrast between normal liver tissue and lesions. However, due to its invasive nature and potential risks, neural network-based translation from NCCT to CECT has become an attractive research direction. Diffusion Transformers (DiT) exhibit strong modeling capacity and effectively capture complex image structures, making them particularly promising for modeling abdominal CT images with complex liver anatomy. But their application to CECT synthesis is often limited by the need for large-scale training data, which is difficult to obtain in practice. To address this issue, we construct a medium-scale CT dataset using publicly available unlabeled data for pre-training, and improve generalization through manually synthesized CT-specific features and a spatial-intensity decoupled data augmentation strategy. We further enhance performance by incorporating deep supervision and improving data efficiency. With only 774 paired CT slices, CT^2-DiT generates realistic and diagnostically relevant CECT images, outperforming existing methods across multiple evaluation metrics and demonstrating the previously underexplored potential of DiT in CECT synthesis. Notably, CT^2-DiT trained on only 50% of the data still outperforms all baselines trained on the full dataset in structural fidelity.

Keywords: CECT Synthesis · DiT · Medical Image Translation.

1 Introduction

Among all cancer types, liver cancer shows the highest growth rate in mortality [27]. Contrast-enhanced computed tomography (CECT) is a critical imaging modality for early detection of liver cancer, particularly for the assessment of focal liver lesions (FLLs) [20, 27]. With contrast agents, CECT highlights the intensity differences between lesions and surrounding normal tissues, providing

* Corresponding author.

essential information for lesion classification and clinical decision-making. However, the acquisition of CECT comes with costs. The use of iodinated contrast agents poses potential health risks, such as nephrotoxicity and allergic reactions [8]. Furthermore, repeated scans increase radiation exposure for patients [31]. To mitigate these risks, synthesizing abdominal CECT images directly from non-contrast CT (NCCT) images has emerged as an attractive research direction.

Recent years' deep generative models have dominated medical image translation. Early approaches primarily relied on end-to-end Convolutional Neural Networks (CNNs) [24] or Generative Adversarial Networks (GANs) [11, 27, 28, 32]. While GANs can generate realistic textures with careful design of architectures and training strategies (e.g., combinations of modules and loss functions), they often suffer from training instability and mode collapse, leading to artifacts in the synthesized images. More recently, Denoising Diffusion Probabilistic Models (DDPMs) [10] have set new standards for image generation quality and diversity. In the medical domain, diffusion models (DMs) have shown superior performance over GANs in various translation tasks [31, 21, 9, 26].

Currently, most diffusion-based medical translation models rely on the U-Net architecture [3, 21, 31, 26, 9]. While U-Net is effective, its convolutional design limits its ability to model long-range dependencies and complex global structures, which are especially important for tasks such as CECT synthesis that require consistent anatomical appearance and intensity relationships across the entire slice. In the field of natural image generation, the Diffusion Transformer (DiT) [22, 18, 5] has demonstrated that replacing the U-Net backbone with a Transformer can significantly scale up performance. DiT exhibits a strong capacity to capture complex structural relationships [5], suggesting a potential advantage for modeling the intricate anatomy of abdominal CT images. In addition, DiT benefits from strong community support, allowing well-established and widely validated architectural and training improvements—originally developed for DiT and even large-scale NLP models—to be readily transferred to the medical imaging domain [13, 25]. However, applying DiT to the medical domain is not straightforward. DiT is known to be data-hungry [15]. Unlike the U-Net, which has strong inductive biases suitable for small datasets, Transformers typically require large-scale data to learn effective representations. In medical imaging, and specifically for NCCT-to-CECT translation, paired training data are scarce and difficult to collect. As a result, training a pure DiT on limited medical datasets often leads to overfitting. This challenge has deterred many researchers from exploring DiT for CT translation, leading them to stick with complex, heavily modified U-Net architectures instead [31, 6, 7].

In this paper, we advocate for a different approach: **keeping the network simple while unleashing the power of DiT through effective data strategies.** We propose CT^2-DiT, a simple yet powerful framework for NCCT-to-CECT translation. Instead of designing cumbersome domain-specific modules, we address the data scarcity issue directly. First, we construct a medium-scale CT dataset using publicly available unlabeled data for pre-training. This allows the model to learn general abdominal CT representations without requiring

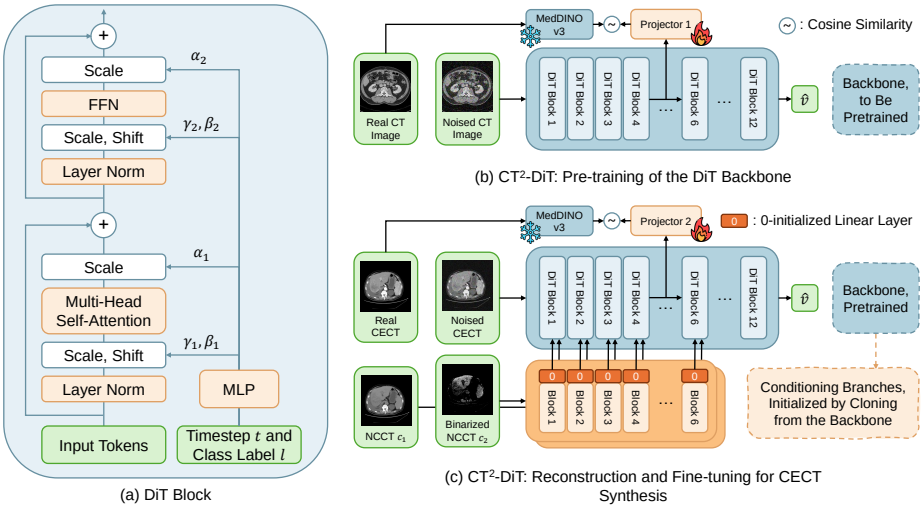

Fig. 1. CT^2-DiT Framework. (a) DiT Block structure. (b) Pre-training stage: the backbone reconstructs images with MedDINO deep supervision. (c) Fine-tuning stage: the model adapts to conditional translation with NCCT and texture-enhanced inputs.

paired labels. Second, to further improve generalization on limited paired data, we introduce manually synthesized CT-specific features and a spatial-intensity decoupled data augmentation strategy. Finally, we incorporate deep supervision to stabilize the training process.

Our approach demonstrates that with simple training strategies, DiT can outperform GANs and U-Net-based DMs even with limited data. We train our model on only 774 paired CT slices, yet it generates realistic and diagnostically relevant CECT images. Our main contributions are as follows: (1) We present CT^2-DiT, to our knowledge, the first successful application of a pure DiT to NCCT-to-CECT translation. (2) We address the data-hungry nature of DiT by introducing a pre-training strategy on public unlabeled data, complemented by CT-specific feature injection, decoupled data augmentation, and deep supervision. (3) We demonstrate that our simple framework outperforms existing GAN-based and U-Net-based diffusion methods across multiple metrics, providing a strong, easy-to-reproduce baseline for future CECT synthesis research.

2 Method

We propose CT^2-DiT, a framework for NCCT-to-CECT translation built on a standard DiT [22]. Our goal is to enable effective DiT training in the low-data regime through data-centric strategies rather than complex architectural modifications. We formulate the translation task as a conditional generative problem. Let x and y denote the source (NCCT) and target (CECT) images.

We operate in the latent space of a pre-trained Variational Autoencoder (VAE). We define $c = \mathcal{E}(x)$ as the condition latent and $z_0 = \mathcal{E}(y)$ as the target latent.

Conditional Flow Matching: We adopt the flow matching framework [18, 16] to train the backbone. Instead of simulating the discrete diffusion steps, we learn a velocity field v_θ that drives a probability path from a Gaussian distribution $z_1 \sim \mathcal{N}(0, I)$ to the data distribution z_0. The training objective minimizes the regression loss between the model prediction and the ground-truth velocity:

$$\mathcal{L} = \mathbb{E}_{t,z_0,c,\epsilon}\left[\|v_\theta(z_t, t, c) - (\dot{\alpha}_t z_0 + \dot{\sigma}_t \epsilon)\|^2\right], \tag{1}$$

where $z_t = \alpha_t z_0 + \sigma_t \epsilon$ is the intermediate state, and $\dot{\alpha}_t, \dot{\sigma}_t$ are time derivatives of the schedule functions [18]. During inference, we generate the target latent $\hat{z}_0$ by solving the ordinary differential equation (ODE) $dz_t/dt = v_\theta(z_t, t, c)$ starting from pure noise, and reconstruct the image via the VAE decoder $\hat{y} = \mathcal{D}(\hat{z}_0)$.

2.1 CT^2-DiT Framework

Our framework involves two stages. (Fig. 1). **Pre-training Stage:** We first train the DiT backbone to reconstruct noised CT images conditioned on timestep t and class label l. To enhance representation learning, we employ deep supervision using MedDINO v3 [14], a foundation model pre-trained on large-scale CT data. We maximize the cosine similarity between the DiT's intermediate features and MedDINO representations (iREPA [25]), enabling the model to learn generalizable anatomical features from unlabeled data. **Fine-tuning Stage:** We convert the pre-trained backbone into a conditional translation model using a ControlNet-style architecture [29]. We initialize the conditioning branches by cloning the backbone. Unlike the original implementation [29], we remove the z_t input to the conditioning branches. Empirically, this simplification improved SSIM by 2%. The NCCT image and a binarized texture map are fed into two independent branches, injecting conditions into the first half of the DiT blocks. For deep supervision during fine-tuning, the DiT operates at 256×256 for tractable global self-attention, while MedDINO features are not token-aligned with the DiT backbone. We address this resolution mismatch with an asymmetric alignment layer (upsampling followed by a 3×3 convolution). During inference, we use only the fine-tuned backbone and conditioning branches.

2.2 Bridging the Data Gap

DiTs are data-hungry and prone to overfitting on small datasets [15]. We propose three strategies to bridge this gap. **Pre-training on Public Unlabeled Data:** We collect approximately 0.25M 2D CT slices from four public abdominal datasets (HCC-TACE-Seg [19], LiTS [1], RAOS [17], TCGA-LIHC [4]). Images are resized to 256×256 with standard windowing (center 40, width 400 for missing metadata). We generate pseudo-labels using K-Means clustering on image features to enable class-conditional pre-training. These clusters describe image-domain appearance in the unlabeled data and do not use lesion categories; fine-tuning uses only contrast phase labels, avoiding lesion-label leakage. This allows

Algorithm 1 PyTorch-style pseudocode for NCCT binarization.

```
def binarize_ct(x, t=0.05):
    roi = (x > t)
    u = x[roi].mean()
    roi_2 = roi & (x <= u)
    m = x[roi_2].mean()
    f = roi_2 & (x > m)
    return torch.where(f, 1.0, -1.0)
```

Notes:
1) $t = 0.05$: 50 HU liver prior, not tuned. For other organs, use organ-specific HU priors.
2) roi_2: excludes low/high CT-value voxels, retaining liver tissue.
3) f: final binary texture map used as model input.

the model to learn distribution statistics before fine-tuning on the target task. **CT-Specific Feature Injection:** VAE compression often blurs high-frequency details and subtle intensity changes crucial for lesion detection. To counteract this, we explicitly inject structural priors. We generate a binary texture map that highlights soft tissue boundaries while suppressing background noise and high-intensity bones. The process is detailed in Algorithm 1. Compared to manual tumor masks, whose boundaries can be uncertain and coarse, this map provides finer-grained guidance by encoding local texture information. **Spatial-Intensity Decoupled Data Augmentation:** To improve generalization without breaking the pixel-wise alignment required for translation, we employ a decoupled strategy. **Spatial Augmentation:** We apply random rotation and affine transformations identically to both the input NCCT and target CECT images. This preserves anatomical alignment. **Intensity Augmentation:** We apply Gaussian blur, HU shifting, and scaling **only** to the input NCCT images. Gaussian blur forces the network to learn robust edge reconstruction, while HU variations simulate scanner differences and contrast absorption variability.

3 Experiments

We evaluate our method on a private clinical dataset of 86 patients containing four focal liver lesion (FLL) types: hepatocellular carcinoma (HCC), focal nodular hyperplasia (FNH), metastases (METS), and cysts. Three phases are available: Non-Contrast (NC), Arterial (ART), and Portal Venous (PV). We use lesion-bearing slices matched across phases at the largest tumor cross-section, yielding 774 paired slices for training under an 80%/20% patient-level split. This avoids healthy slices that provide little signal for FLL assessment and gives direct pathological correspondence for our 2D setting. We use a 2D formulation because clinical scans have varying slice thickness and our 0.25M-slice pre-training pool includes data available only as 2D slices. We employ the DiT-B/2 backbone with a sine-cosine Generalized Variance-preserving (GVP) interpolant [18].

3.1 Qualitative and Quantitative Comparison

We compare CT^2-DiT against three GAN-based methods (Pix2pix [11], AttentionGAN [2], PPT [30]) and three DMs (LDM [23], SynDiff [21], BBDM [12])

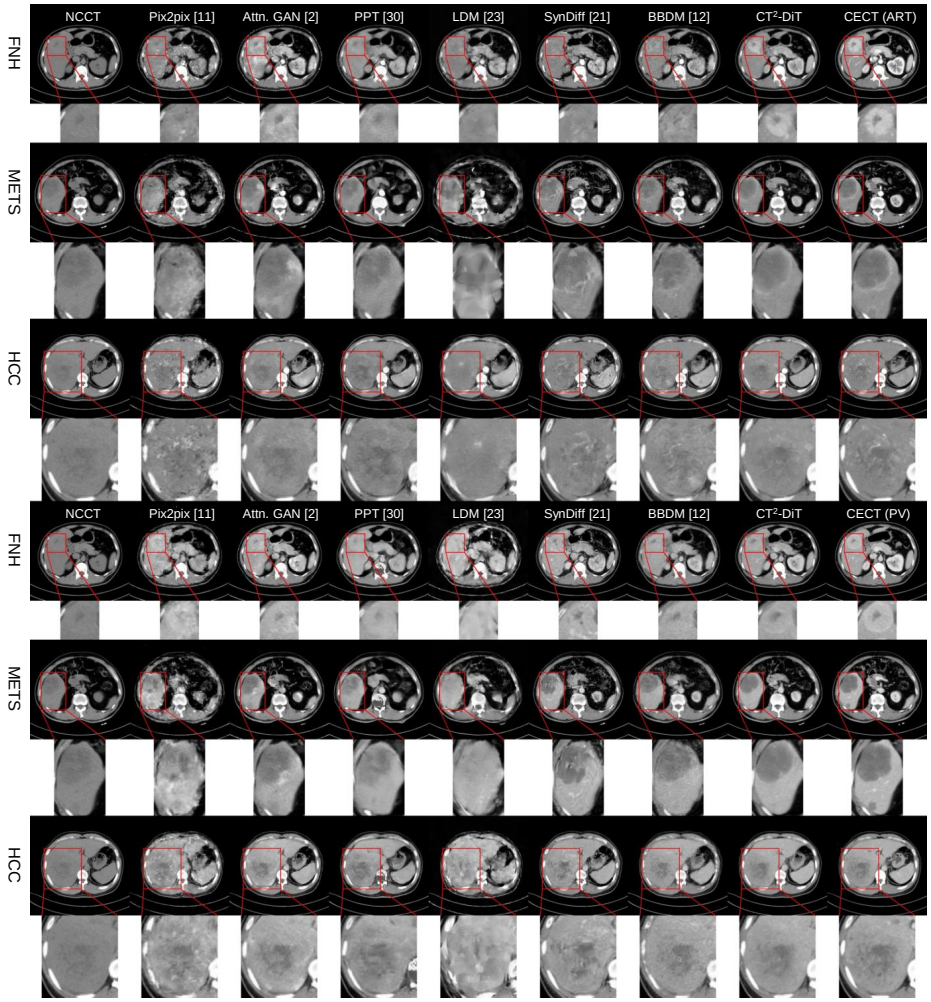

Fig. 2. Qualitative results on CECT synthesis for ART (top three rows) and PV (bottom three rows) phases. The figure compares the input NCCT, six competing methods, our CT^2-DiT, and the ground truth. Red boxes highlight the pathological regions.

on NC→ART and NC→PV tasks. Fig. 2 visualizes the results. Generally, GAN-based methods suffer from blurring and artifacts, while standard DMs improve texture but occasionally hallucinate incorrect enhancement patterns. In contrast, CT^2-DiT excels in preserving lesion-specific features. For **FNH** (top row), which exhibits arterial hyper-enhancement, GAN-based methods fail to produce sufficient lesion-to-parenchyma contrast, whereas CT^2-DiT generates high-intensity lesions with sharp boundaries matching the ground truth. For **METS** (second row), our model synthesizes the characteristic rim enhancement and heterogeneous internal textures, avoiding the over-smoothing seen in AttentionGAN and

Table 1. Quantitative comparison on different-sized datasets. The "Half" dataset denotes training with 50% of the data randomly removed. Best results are bolded.

Dataset	Model	NC → ART				NC → PV			
		MAE↓	SSIM↑	PSNR↑	MI↑	MAE↓	SSIM↑	PSNR↑	MI↑
Full	Pix2pix [11]	9.60	0.7599	21.87	1.10	10.14	0.7542	21.48	1.12
	Attn. GAN [2]	8.20	0.8246	22.61	1.18	8.40	0.8194	22.37	1.20
	PPT [30]	9.52	0.7840	21.70	1.14	10.63	0.7773	20.34	1.15
	LDM [23]	12.40	0.6402	20.64	1.09	15.07	0.6124	19.35	1.10
	SynDiff [21]	8.31	0.8094	22.67	1.17	8.46	0.8077	22.47	1.21
	BBDM [12]	7.42	0.8313	**23.58**	1.21	8.02	0.8052	22.92	1.23
	CT^2-DiT	**7.42**	**0.8378**	23.48	**1.24**	**7.91**	**0.8317**	**22.92**	**1.26**
Half	Pix2pix [11]	9.15	0.7715	22.27	1.12	9.84	0.7722	21.74	1.14
	Attn. GAN [2]	8.73	0.8097	22.16	1.15	8.98	0.8021	21.88	1.16
	PPT [30]	9.91	0.7763	21.33	1.13	10.30	0.7702	20.99	1.15
	LDM [23]	23.72	0.4624	17.67	1.03	25.74	0.4220	16.82	1.04
	SynDiff [21]	8.28	0.8110	22.72	1.17	8.46	0.8079	22.43	1.21
	BBDM [12]	10.99	0.7587	20.90	1.12	10.22	0.7710	21.15	1.15
	CT^2-DiT	**7.68**	**0.8317**	**23.19**	**1.22**	**8.20**	**0.8270**	**22.65**	**1.25**

Table 2. Lesion type classification accuracy. The downstream classifier was trained on synthetic data generated by each model and evaluated on real data. NC-only and Real denote the NCCT-only baseline and real NCCT+CECT upper bound.

Input	NC-only	Pix2pix	Attn. GAN	PPT	LDM	SynD.	BBDM	**Ours**	Real
Acc.	81.86%	83.41%	87.43%	86.64%	85.83%	90.53%	90.55%	**92.91%**	96.07%

BBDM. For **HCC** (third row), we reproduce the critical wash-in and wash-out dynamics. Although VAE compression slightly limits high-frequency detail compared to pixel-space models like BBDM, CT^2-DiT achieves superior structural fidelity and separates lesions from the parenchyma more effectively.

We quantitatively evaluate synthesis quality using MAE, PSNR, SSIM, and Mutual Information (MI). Table 1 reports results on two datasets. **Full Dataset Performance:** CT^2-DiT achieves the best structural fidelity across tasks. For NC → ART, it attains an SSIM of 0.8378 and MI of 1.24, outperforming all baselines. Although BBDM achieves a marginally higher PSNR (23.58 dB vs. 23.48 dB), our method matches the best MAE (7.42). For NC → PV, CT^2-DiT achieves the best performance across all metrics (e.g., lowest MAE of 7.91). The consistent superiority in SSIM and MI indicates that our method preserves anatomical structures and lesion details more faithfully than GANs or U-Net-based DMs. **Low-Data Regime:** To validate data efficiency, we trained models on only 50% of the data. While GANs and other DMs' performance degrade

Table 3. Ablation study of the proposed strategies, we gradually add each component to the baseline DiT model.

Model	NC → ART				NC → PV			
	MAE↓	SSIM↑	PSNR↑	MI↑	MAE↓	SSIM↑	PSNR↑	MI↑
Baseline	8.99	0.7517	22.52	1.17	9.35	0.7452	22.18	1.20
+ Pre-training	8.49	0.8109	22.36	1.19	8.80	0.8036	22.02	1.22
+ Binarize Map	8.27	0.8160	22.60	1.20	8.67	0.8062	22.15	1.23
+ Data Aug.	7.50	0.8348	23.42	1.23	8.09	0.8273	22.81	1.25
+ iREPA	**7.42**	**0.8378**	**23.48**	**1.24**	**7.91**	**0.8317**	**22.92**	**1.26**

significantly, CT^2-DiT remains robust. Notably, our model trained on half the data still outperforms all baselines trained on the *full* dataset in terms of SSIM and MI, demonstrating its suitability for data-scarce medical scenarios.

3.2 Downstream Task Evaluation and Ablation Study

To assess whether the synthesized images preserve diagnostically relevant patterns, we perform a lesion classification task using the "Train on Synthetic, Test on Real" protocol. We choose classification rather than segmentation because lesion localization is usually available in this setting, while the key question is whether contrast patterns for FLL typing are preserved. We train a ConvNeXt-T classifier to distinguish four lesion types (Cyst, FNH, HCC, METS) using concatenated NCCT and synthetic CECT images, split into 3 folds. Table 2 also reports an NCCT-only baseline and a real NCCT+CECT upper bound. As shown in Table 2, the classifier trained on CT^2-DiT data achieves the highest accuracy of 92.91% on real patient data, outperforming the closest competitor (BBDM) by 2.36%. It improves over the NCCT-only baseline by 11.05 points (81.86% to 92.91%) and approaches the real-data upper bound of 96.07%. This suggests that our generated images preserve features relevant to lesion classification, rather than merely optimizing pixel-level similarity.

We analyze the contribution of each component in Table 3. **Baseline:** Training a standard DiT from scratch leads to overfitting and suboptimal structure (0.7517 SSIM). **Pre-training:** Initializing with weights learned from our unlabeled dataset significantly improves structural consistency. **Feature Injection:** Injecting the binarized texture map via the ControlNet-style branch recovers high-frequency lesion boundaries often blurred by VAE compression. **Decoupled Augmentation:** This strategy boosts generalization (increasing PSNR to 23.42 dB) by preventing the model from memorizing intensity mappings while preserving spatial alignment. **Deep Supervision:** Aligning intermediate features with the pre-trained MedDINO encoder yields the best overall performance, ensuring the model learns semantically meaningful medical representations. Because lesions occupy only a small fraction of each slice, global metrics are dominated by background and parenchyma. Thus, the binarized map mainly improves

boundary recovery, while iREPA gives a consistent gain through representation alignment rather than an easily isolated visual effect.

4 Conclusion and Future Work

We proposed CT^2-DiT for low-data abdominal NCCT-to-CECT translation. With unlabeled pre-training and CT-specific strategies, it outperforms existing GANs and U-Net-based DMs while preserving diagnostically relevant patterns. This study is limited to a single-center dataset with four lesion types; validation on other institutions, scanners, and lesion subtypes remains necessary. Future work will explore data-efficient pixel-space DiTs, which may require more data but avoid VAE limits on high-frequency fidelity [13], and radiologist studies.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Number JP24K15118. We thank Prof. Hongjie Hu (Department of Radiology, Sir Run Run Shaw Hospital, Zhejiang University, China) and Prof. Yen-Wei Chen (College of Information Science and Engineering, Ritsumeikan University, Japan) for providing the multi-phase CT data.

Disclosure of Interests. The authors declare no competing interests.

References

1. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical image analysis **84**, 102680 (2023)
2. Chen, X., Xu, C., Yang, X., Tao, D.: Attention-gan for object transfiguration in wild images. In: Proceedings of the European conference on computer vision (ECCV). pp. 164–180 (2018)
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
4. Erickson, B.J., Kirk, S., Lee, Y., Bathe, O., Kearns, M., Gerdes, C., Rieger-Christ, K., Lemmerman, J.: The cancer genome atlas liver hepatocellular carcinoma collection (tcga-lihc) (2016)
5. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
6. Feng, L., Guo, L., He, W., Ye, X., Pan, Y., Gao, S., Zhang, R.: Diffcta: Diffusion model-based non-contrast ct angiography with contrast-enhanced mask guidance. Biomedical Signal Processing and Control **112**, 108645 (2026)
7. Ferreira, V.R.S., de Paiva, A.C., Silva, A.C., Almeida, J., et al.: Diffusion model for generating synthetic contrast enhanced ct from non-enhanced heart axial ct images. In: International Conference on Enterprise Information Systems (2024)
8. Gleeson, T.G., Bulugahapitiya, S.: Contrast-induced nephropathy. American Journal of Roentgenology **183**(6), 1673–1689 (2004)
9. He, Q., Summerfield, N., Wang, P., Glide-Hurst, C., Dong, M.: Deterministic medical image translation via high-fidelity brownian bridges. arXiv preprint arXiv:2503.22531 (2025)

10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
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. 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)
13. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
14. Li, Y., Wu, Y., Lai, Y., Hu, M., et al.: Meddinov3: How to adapt vision foundation models for medical image segmentation? arXiv preprint arXiv:2509.02379 (2025)
15. Liang, Z., He, H., Yang, C., Dai, B.: Scaling laws for diffusion transformers. arXiv preprint arXiv:2410.08184 (2024)
16. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
17. Luo, X., Li, Z., Zhang, S., Liao, W., Wang, G.: Rethinking abdominal organ segmentation (raos) in the clinical scenario: A robustness evaluation benchmark with challenging cases. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–541. Springer (2024)
18. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., et al.: Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
19. Moawad, A.W., Fuentes, D., Morshid, A., Khalaf, A.M., Elmohr, M.M., Abusaif, A., Hazle, J.D., Kaseb, A.O., Hassan, M., Mahvash, A., Szklaruk, J., Qayyom, A., Elsayes, K.: Multimodality annotated hcc cases with and without advanced imaging segmentation (2021)
20. Oliva, M.R., Saini, S.: Liver cancer imaging: role of ct, mri, us and pet. Cancer imaging **4**(Spec No A), S42 (2004)
21. Özbey, M., Dalmaz, O., Dar, S.U., Bedel, H.A., Özturk, Ş., Güngör, A., Cukur, T.: Unsupervised medical image translation with adversarial diffusion models. IEEE Transactions on Medical Imaging **42**(12), 3524–3539 (2023)
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
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. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
25. Singh, J., Leng, X., Wu, Z., Zheng, L., Zhang, R., Shechtman, E., Xie, S.: What matters for representation alignment: Global information or spatial structure? arXiv preprint arXiv:2512.10794 (2025)
26. Xing, Z., Yang, S., Chen, S., Ye, T., Yang, Y., Qin, J., Zhu, L.: Cross-conditioned diffusion model for medical image to image translation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 201–211. Springer (2024)
27. Yang, Y., Chen, Q., Li, Y., Wang, F., Han, X.H., Iwamoto, Y., Liu, J., Lin, L., Hu, H., Chen, Y.W.: Segmentation guided crossing dual decoding generative adversarial network for synthesizing contrast-enhanced computed tomography images. IEEE Journal of Biomedical and Health Informatics **28**(8), 4737–4750 (2024)

28. Yang, Y., Liu, J., Zhan, G., Chen, Q., Wang, F., Li, Y., Kumar Jain, R., Lin, L., Hu, H., Chen, Y.W.: OA-GAN: Organ-aware generative adversarial network for synthesizing contrast-enhanced medical images. Biomedical Physics & Engineering Express **10**(3), 035012 (2024)
29. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3813–3824 (2023). https://doi.org/10.1109/ICCV51070.2023.00355
30. Zhang, W., Hui, T.H., Tse, P.Y., Hill, F., Lau, C., Li, X.: High-resolution medical image translation via patch alignment-based bidirectional contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 178–188. Springer (2024)
31. Zheng, K., Huang, M., Li, X., Ma, J., Feng, Q., Yang, W., Zhong, L.: Contrast flow pattern and cross-phase specificity-aware diffusion model for ncct-to-multiphase cect synthesis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 89–99. Springer (2025)
32. Zhong, L., Xiao, R., Shu, H., Zheng, K., Li, X., Wu, Y., Ma, J., Feng, Q., Yang, W.: Ncct-to-cect synthesis with contrast-enhanced knowledge and anatomical perception for multi-organ segmentation in non-contrast ct images. Medical Image Analysis **100**, 103397 (2025)