

ViCTr: Vital Consistency Transfer for Pathology Aware Image Synthesis

Onkar Susladkar^{1,3}, Gayatri Deshmukh¹, Yalcin Tur², Gorkem Durak¹, Ulas Bagci¹
¹Northwestern University ²Stanford University ³University of Illinois Ubrana-Champaign

onkarsus13@gmail.com, dgayatri9850@gmail.com, yalcintr@stanford.edu,
 gorkem.durak@northwestern.edu, ulas.bagci@northwestern.edu

Abstract

*Synthesizing medical images remains challenging due to limited annotated pathological data, modality domain gaps, and the complexity of representing diffuse pathologies such as liver cirrhosis. Existing methods often struggle to maintain anatomical fidelity while accurately modeling pathological features, frequently relying on priors derived from natural images or inefficient multi-step sampling. In this work, we introduce **ViCTr (Vital Consistency Transfer)**, a novel two-stage framework that combines a rectified flow trajectory with a Tweedie-corrected diffusion process to achieve high-fidelity, pathology-aware image synthesis. First, we pretrain **ViCTr** on the ATLAS-8k dataset using **Elastic Weight Consolidation (EWC)** to preserve critical anatomical structures. We then fine-tune the model adversarially with **Low-Rank Adaptation (LoRA)** modules for precise control over pathology severity. By **reformulating Tweedie’s formula** within a linear trajectory framework, **ViCTr** supports one-step sampling—reducing inference from 50 steps to just 4—without sacrificing anatomical realism. We evaluate **ViCTr** on BTCV (CT), AMOS (MRI), and CirrMRI600+ (cirrhosis) datasets. Results demonstrate state-of-the-art performance, achieving a Medical Fréchet Inception Distance (MFID) of 17.01 for cirrhosis synthesis—28% lower than existing approaches—and improving nnUNet segmentation by +3.8% mDSC when used for data augmentation. Radiologist reviews indicate that **ViCTr**-generated liver cirrhosis MRIs are clinically indistinguishable from real scans. To our knowledge, **ViCTr** is the first method to provide fine-grained, pathology-aware MRI synthesis with graded severity control, closing a critical gap in AI-driven medical imaging research.*

1. Introduction

The exponential growth in computer vision capabilities has been driven by significant advances in artificial intelligence models [7, 32]. However, medical imaging faces a fundamental tension between model complexity and data avail-

ability that limits the application of state-of-the-art techniques. While recent breakthroughs in generative AI have demonstrated remarkable capabilities in synthetic data creation, these advances demand training datasets of unprecedented scale—a requirement that poses unique challenges in the medical domain [2, 33].

Unlike general computer vision applications, medical imaging faces several critical constraints: privacy regulations necessitating complex deidentification, inherent data fragmentation across healthcare institutions, and fundamental interoperability constraints. These barriers have created a growing disparity between the rapid advancement of general computer vision and the relatively slower progress in medical imaging applications. Current approaches to bridging this gap face two major limitations: insufficient feature preservation across anatomical structures and inadequate handling of pathological variations.

This challenge is especially pronounced in **abdominal imaging**, where pathologies such as diffuse cirrhosis or multi-tissue disease processes manifest across multiple organ systems. Unlike the well-defined boundaries typical of tumors or simpler structures like bones in X-ray imaging, abdominal pathologies often involve **subtle and heterogeneous changes** in tissue characteristics, requiring more nuanced feature extraction and synthesis. The inherent complexity of MRI signals—spanning multiple sequences and high spatial resolution—presents additional computational hurdles, further distancing it from the more standardized nature of CT or simpler 2D radiographs.

To address these gaps, our work introduces **ViCTr (Vital Consistency Transfer)**, a two-stage framework that facilitates pathology-aware medical image synthesis with strong anatomical fidelity. We specifically target abdominal CT and MRI data, aiming to generate clinically relevant synthetic datasets that capture both normal anatomy and intricate pathological details. By providing robust augmentation material, our approach holds the potential to improve downstream tasks like segmentation and diagnosis, while alleviating issues of data scarcity and privacy constraints in medical imaging. Our contributions are:

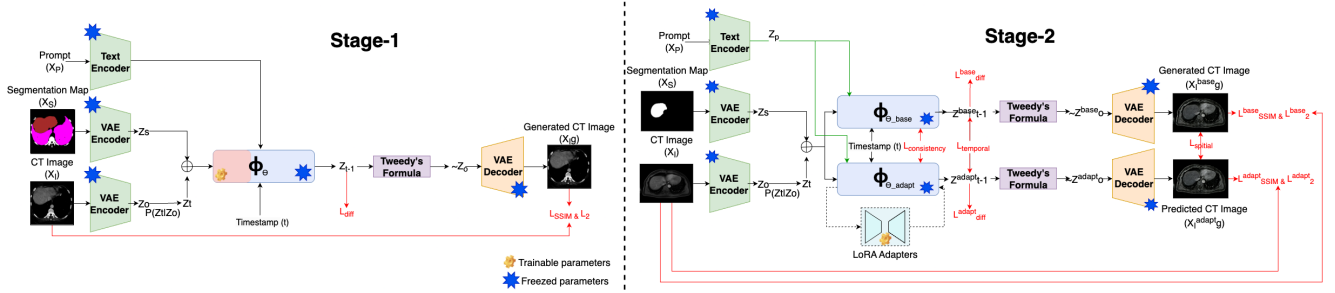


Figure 1. Overview of the proposed ViCTr methodology.

- **Novel Two-Stage Framework.** We propose ViCTr, a method that fuses anatomical consistency and pathological realism for CT and MRI synthesis with high diversity.
- **Tweedie's Formula in Rectified Flow.** We introduce a rectified flow trajectory reformulation of Tweedie's formula, ensuring accurate initialization and reducing sampling bias.
- **Wide Applicability.** ViCTr integrates seamlessly with various diffusion models, supporting a broad range of medical imaging tasks.
- **One-Step Sampling & Tweedie's Corrections.** We reduce computational overhead by cutting the number of diffusion steps from dozens to a handful, expediting inference without sacrificing quality.
- **Quantitative Improvements in Segmentation.** We demonstrate that adding ViCTr-synthesized CT/MRI images to real-world training sets significantly boosts segmentation performance, underscoring the practical benefits of pathology-focused data augmentation.
- **Enhanced Image Fidelity.** ViCTr achieves consistently lower FID (Fréchet Inception Distance) scores across multiple datasets, indicating superior structural and textural coherence.
- **First Abdominal MRI Pathology Synthesis.** To our knowledge, we are the first to generate abdominal MRI pathologies with progressive severity, offering new possibilities for research and clinical applications in medical imaging.

By emphasizing both anatomical fidelity and pathological variety, ViCTr presents a powerful step toward bridging the gap between limited medical imaging data and the demands of high-performing AI models.

2. Related Works

2.1. Generative Models for Medical Data Augmentation

Synthetic images have long been explored for data augmentation in medical imaging, with Generative Adversarial Networks (GANs) [25, 36, 50] initially offering promising results in tasks such as lesion synthesis and modality trans-

lation. However, GAN-based methods often struggle with mode collapse, requiring extensive architectural and training refinements to ensure sufficient diversity and realism in the synthesized images [12, 35].

2.2. Diffusion Models in Medical Imaging

Diffusion-based approaches have emerged as a powerful alternative to GANs for high-quality synthetic image generation [4, 5, 13, 49]. Their inherent noise-to-image paradigm provides a more stable training regime and can produce richer data variations. In medical imaging, diffusion models [11] have been leveraged to improve segmentation and classification performance by generating diverse training samples [1, 10, 37, 46, 47]. Recent specialized frameworks, including segmentation-guided diffusion [26] and text-driven generation built upon RadImageNet [34, 54], highlight the adaptability of diffusion models to specific clinical scenarios. Efforts such as MixUp-enhanced augmentation [6, 28] and domain adaptation methods [55] demonstrate that diffusion techniques can mitigate biases and improve generalization in tasks like classification and risk prediction.

2.3. General-Purpose Diffusion for Data Augmentation

Various general-purpose diffusion strategies have been applied to further enrich training datasets. DiffuseMix [21] employs conditional prompts to blend real and synthetic data while preserving labels, whereas DreamDA [18] combines diffusion-based perturbations with pseudo-labeling for semantic consistency. DetDiffusion [51] incorporates object-detection attributes into the diffusion process, while Effective Data Augmentation with Diffusion Models [48] explores diffusion-based techniques to boost performance in few-shot settings. These methods generally aim to enhance data diversity and improve the training of downstream models.

2.4. Rectified Flow Models

While traditional diffusion approaches rely on discretized noise schedules, Rectified Flow (ReFlow) models learn

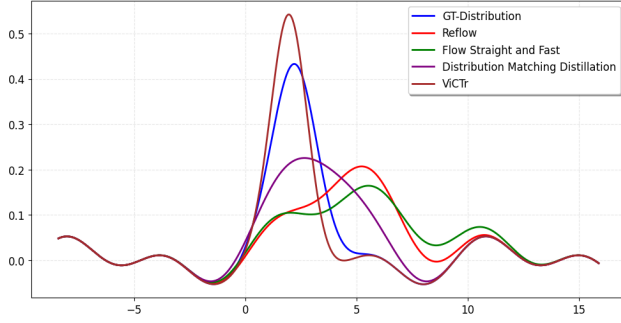


Figure 2. Comparison of synthetic data distributions from various methods against the ground-truth (GT) on the AMOS dataset. The GT distribution (blue) reflects real data, while Reflow (red), Flow Straight and Fast (green), and Distribution Matching Distillation (purple) show varying alignment. Our proposed method, ViCTr (brown) achieves the closest match to GT, demonstrating superior distribution alignment and fidelity.

smooth, near-linear trajectories to map between data distributions. This can result in more efficient sampling and reduced computational overhead. Recent works have leveraged rectified flows for improved convergence rates and lower numbers of function evaluations (NFE) [29, 30, 53], although most focus on general distribution transport rather than the strict alignment and pathology-aware realism needed in medical imaging. As shown in Figure 2, different ReFlow-based methods exhibit varying degrees of distribution matching, with our approach (ViCTr) offering closer alignment to the ground-truth distribution.

2.5. Domain-Specific Adaptation

Domain-focused adaptations continue to evolve. DiNO-Diffusion [24] introduces a self-supervised latent diffusion model to cope with limited annotated medical data, and Diverse Data Augmentation with Diffusions [17] enriches domain generalization using Stable Diffusion combined with cosine similarity filtering. Collectively, these methods illustrate the growing interest in customized diffusion pipelines for specialized domains, especially when high-quality annotated data are scarce.

Recent advancements have further showcased the potential of diffusion models for domain-specific adaptation. DiNO-Diffusion [24] addresses the challenge of limited annotated data in medical imaging by leveraging a self-supervised latent diffusion framework. Similarly, Diverse Data Augmentation with Diffusions [17] enhances domain generalization by integrating Stable Diffusion with cosine similarity-based filtering, enabling the generation of semantically diverse and high-quality data. However, these methods prioritize general-domain efficiency over the anatomical-pathological alignment critical for medical imaging, often failing to preserve fine-grained structures

like vascular networks or diffuse fibrosis patterns.

In this work, we address the critical need for efficient, high-fidelity pathology-aware synthesis in abdominal imaging by merging rectified flow concepts with a Tweedie-corrected diffusion process. Our approach (ViCTr) stands out from prior methods by introducing a linearized sampling framework that enables one-step generation while maintaining anatomical fidelity—a crucial requirement for robust data augmentation in clinical applications.

3. Methods

Our framework, ViCTr, uses **Rectified Flow** [31] and **Tweedie’s Formula** [43] for high-fidelity medical image synthesis with efficient one-step sampling. By rectifying the trajectory from a prior distribution p_0 to the target medical image distribution p_{target} , we reduce sampling bias and retain key anatomical and pathological details.

3.1. Rectified Flow Trajectory

Classical diffusion approaches gradually corrupt data samples $x \in \mathbb{R}^d$ into a Gaussian prior p_0 . While these methods have shown success, they often suffer from suboptimal sampling paths, cascading prediction errors, and diminished diversity in generated samples. We address these issues through *Rectified Flow optimization* [16, 31], which learns a continuous velocity field guiding the forward and reverse diffusion processes more directly.

Let (x_0, x_1) be a sample pair with $x_0 \sim p_0$ and $x_1 \sim p_{\text{target}}$. We define an interpolated point $x_t = (1-t)x_0 + tx_1$ where $t \in [0, 1]$. A velocity model $v_\theta : \mathbb{R}^d \times [0, 1] \rightarrow \mathbb{R}^d$ then predicts how to transition from x_t toward x_1 . We train v_θ by minimizing:

$$\hat{\theta} = \arg \min_{\theta} \mathbb{E}_{t \sim \text{Uniform}(0,1)} \left[\|(x_1 - x_0) - v_\theta(x_t, t)\|^2 \right], \quad (1)$$

ensuring the predicted flow $v_\theta(x_t, t)$ aligns with the true path $(x_1 - x_0)$. Once trained, the rectified flow is realized by solving the ODE:

$$dx_t = v_{\hat{\theta}}(x_t, t)dt, \quad (2)$$

leading $x_0 \sim p_0$ toward $x_1 \sim p_{\text{target}}$. To enable one-step sampling, we distill this multi-step process into a neural network $\hat{\mathcal{T}} : \mathbb{R}^d \rightarrow \mathbb{R}^d$, such that $\hat{\mathcal{T}}(x_0) = x_0 + v(x_0, 0)$, is trained to directly predict x_1 from x_0 . The loss function,

$$\mathcal{L} = \mathbb{E} \left[\|(x_1 - x_0) - v(x_0, 0)\|^2 \right], \quad (3)$$

drives $\hat{\mathcal{T}}(x_0)$ to approximate the final, rectified state x_1 .

3.2. Rectified Flow with Tweedie’s Formula

Though rectified flow refines sampling, medical data distributions often need extra bias correction. Tweedie’s Formula [43] addresses this by adjusting noisy observations to better approximate the posterior mean. For Gaussian variables $z \sim \mathcal{N}(\mu_z, \Sigma_z)$, Tweedie’s formula indicates:

$$\mathbb{E}[\mu_z | z] = z + \Sigma_z \nabla_z \log p(z)$$

, where $\nabla_z \log p(z)$ is the gradient of the log probability. In typical diffusion, Tweedie’s formula estimates x_0 from noisy x_t , guiding predictions toward the data manifold.

ViCTr incorporates this correction into the rectified flow ODE:

$$dx_t = v_\theta(x_t, t)dt + (1 - \bar{\alpha}_t) \nabla_{x_t} \log p(x_t)dt,$$

and updates the training objective to:

$$\hat{\theta} = \arg \min_{\theta} \mathbb{E}_{t \sim \text{Uniform}(0,1)} \left[\left\| (x_1 - x_0) - v_\theta(x_t, t) - (1 - \bar{\alpha}_t) \nabla_{x_t} \log p(x_t) \right\|^2 \right].$$

Here, $\bar{\alpha}_t$ denotes the cumulative product of variance decay factors in the diffusion schedule. The additional Tweedie term $(1 - \bar{\alpha}_t) \nabla_{x_t} \log p(x_t)$ corrects for sampling bias, driving x_t more accurately toward the target distribution.

We similarly extend the one-step distillation to integrate Tweedie’s correction:

$$\hat{\mathcal{T}}(x_0) = x_0 + v(x_0, 0) + (1 - \bar{\alpha}_0) \nabla_{x_0} \log p(x_0),$$

and optimize

$$\mathcal{L} = \mathbb{E} \left[\left\| x_1 - \hat{\mathcal{T}}(x_0) \right\|^2 \right].$$

This single-step approach balances computational efficiency with high-quality, pathology-aware generation—critical in medical applications where datasets are limited and synthetic realism is paramount.

In summary, **ViCTr** unifies rectified flow and Tweedie’s correction to deliver anatomically consistent, pathology-aware sampling in a single forward pass. The resulting framework exhibits reduced inference costs, minimized sampling bias, and improved fidelity for generating high-resolution medical images.

3.3. Pathology Aware Image Synthesis

Our proposed **ViCTr** method adopts a two-stage training paradigm: **Stage 1** establishes a foundational diffusion model tailored to medical imaging data, and **Stage 2** fine-tunes this pre-trained model for downstream tasks, including semantic-guided multi-modal generation and counterfactual pathology synthesis.

Pre-training on ATLAS-8k Dataset (Stage 1)

In the absence of large-scale, domain-specific pre-trained diffusion models, we begin by training on the ATLAS-8k [40] dataset, which comprises abdominal CT scans and their segmentation annotations. We leverage these annotations as conditional guidance—enabling precise control over anatomical structures—and integrate textual prompts (e.g., “create image having $\langle\langle\text{organs}\rangle\rangle$ ”) to further refine semantic consistency.

Latent Representations. Following Figure 1, we feed raw CT images (X_I) and segmentation masks (X_S) into a frozen VAE encoder, producing latent embeddings (Z_o and Z_s). Simultaneously, textual prompts (X_p) are processed by a pre-trained text encoder (details in Table 1) to yield prompt embeddings (Z_p). These latent representations guide the generative diffusion backbone ϕ_θ during both forward and reverse diffusion.

Forward Diffusion. We progressively inject noise into Z_o forming a noisy representation Z_t as

$$P(Z_t | Z_o) = (1 - t) \cdot Z_o + t \cdot \epsilon_{\text{true}},$$

where ϵ_{true} is the true noise at step t . We concatenate Z_t with Z_s and feed them into the diffusion model ϕ_θ . Crucial layers are selectively unfrozen based on elastic weight consolidation, maintaining model stability while adapting to medical domain specifics.

Reverse Diffusion. In the reverse process,

$$P(Z_{t-1} | Z_t, Z_s, Z_p, t) = Z_t + \delta T \times \phi_\theta(Z_t, Z_s, Z_p, t), \quad (4)$$

iteratively removes noise to reconstruct Z_{t-1} conditioned on the segmentation map (Z_s) and text prompt (Z_s). The diffusion loss L_{diff} between Z_{t-1} and ϵ_{true} is:

$$L_{\text{diff}} = -|\phi_\theta(Z_t, Z_s, Z_p, t) - (\epsilon_{\text{true}} - Z_o)|^2. \quad (5)$$

After refining Z_{t-1} via Tweedie’s formula, the denoised latent Z_o is fed into the frozen VAE decoder (matching ϕ_θ ; see Table 1) to obtain a reconstructed CT image.

Composite Loss. To ensure both local accuracy and global perceptual realism, we combine: Diffusion Loss (L_{diff}), pixel-level reconstruction loss (L_2) evaluating the gap between generated X_{ig} and ground-truth X_I , and structural similarity (SSIM) loss (L_{SSIM}) emphasizing higher-order textural correspondence. Optimizing this composite loss enables the diffusion backbone ϕ_θ to capture essential anatomical features, establishing a versatile medical foundation model ready for fine-tuning.

Fine-tuning on downstream tasks (Stage 2)

Building on the Stage1 pre-trained model, Stage2 targets specialized tasks like semantic-guided CT/MRI generation and counterfactual pathology synthesis. As shown in Figure 1, a dual-network setup is adopted.

1. ϕ_{base} remains frozen, retaining the robust anatomical knowledge acquired in Stage 1.
2. ϕ_{adapt} is augmented with LoRA [19] modules within the previously fine-tuned layers. These modules are the only trainable parameters, ensuring targeted and stable adaptation.

Training Input Consistency Loss. Similar to Stage 1, each training sample includes prompts, segmentation maps, and either CT or MRI images (depending on the task). A frozen VAE encoder extracts latent representations Z_s , Z_o , and Z_p . Following forward diffusion of Z_o into noisy Z_t , both ϕ_{base} and ϕ_{adapt} process $[Z_t, Z_s]$ with text embedding Z_p . We introduce a consistency loss $L_{consistency}$ to align intermediate outputs of ϕ_{base} and ϕ_{adapt} , ensuring stable adaptation without drifting from core representations.

Loss Components. Diffusion loss (L_{diff}^{base} , L_{diff}^{adapt}) encourages accurate noise prediction. Temporal consistency loss $L_{temporal}$ preserves smooth transitions across time steps in the reverse diffusion. Spatial consistency loss $L_{spatial}$ ensures alignment in the output reconstructions (Z_{t-1}^{base} , Z_{t-1}^{adapt}). Pixel-level losses (L_2 and L_{SSIM}) compare generated results to ground-truth scans, enforcing both local fidelity and structural coherence.

CT and MRI Generation. We adapt to BTCV [15] (CT) and AMOS [23] (MRI) datasets, using organ segmentation masks for prompts. A standard 80/10/10 partition ensures robust evaluation.

Pathology Generation. For liver cirrhosis, we employ CirrMRI600+ [22] dataset (T1/T2 MRIs) with segmentation masks marking liver regions. Prompts specify “low,” “mild,” or “severe” cirrhosis intensity. As in the CT/MRI setting, the training/validation/testing split is 80/10/10. The inference details and inference diagram are in the supplementary material.

This selective LoRA-based fine-tuning allows **ViCTr** to handle complex tasks—ranging from normal organ generation to pathology synthesis—while preserving the anatomical and text-driven guidance learned in Stage 1. The experiments (Section 4) demonstrate that this two-stage framework consistently yields high-quality, clinically realistic images in both CT and MRI settings, including nuanced liver cirrhosis modeling.

Diffusion Methods	Denoiser	Text Encoder
Stable Diffusion [44]	UNet [45]	Clip-B/16 [41]
Pixart-alpha [9]	Transformer (DiT) [38]	T5-XXXL [42]
Stable Diffusion XL [39]	Dual Unet [39]	Clip-L/14
Flux [27]	MultiModal Transformer [14]	T5-XXXL + Clip-L/15
Stable Diffusion-3 [14]	MultiModal Transformer	T5-XXXL + Clip-B/16 + Clip-L/14

Table 1. Selection of denoisers and text encoders for different diffusion methods ϕ_θ

4. Experiments and Results

The implementation details for training and sampling are provided in the supplementary material.

4.1. Quantitative Results

Synthetic Data Generation Results. Table 2 summarizes our comparative evaluation of synthetic medical image generation under vanilla fine-tuning versus the proposed **ViCTr** pipeline, spanning multiple datasets (BTCV, AMOS, CirrMRI600+) and diffusion backbones (Stable Diffusion, SD-XL, SD-3, Pixart-alpha, Flux). While vanilla fine-tuning applies a direct end-to-end approach after ATLAS-8k pre-training, it generally yields higher Fréchet Inception Distance (FID) and Medical FID (MFID) values across all datasets. Conversely, **ViCTr** integrates domain-specific loss functions and specialized architectural settings, consistently achieving lower FID/MFID scores and producing more realistic medical images.

To address the well-known limitation of FID (which uses ImageNet features), we adopt M3D-CLIP [3] to compute a more domain-specific MFID. As shown in the table, **ViCTr** provides notable gains: for example, Stable Diffusion under **ViCTr** achieves FID/MFID of 21.98/19.02 (BTCV), 20.37/19.11 (AMOS), and 25.57/21.46 (CirrMRI600+). These results represent substantial improvements over vanilla fine-tuning and demonstrate consistent performance boosts across all tested architectures. Such gains highlight the robustness of **ViCTr**’s two-stage framework in generating clinically meaningful synthetic images for both CT and MRI modalities.

Segmentation Results. To assess the quality and utility of **ViCTr**-generated synthetic data, we performed segmentation experiments on three datasets—BTCV, AMOS, and CirrMRI600+—tracking mean Dice Similarity Coefficient (mDSC) and mean Hausdorff Distance 95 (mHD95). Higher mDSC and lower mHD95 respectively indicate improved overlap accuracy and spatial precision. We examined four training configurations: (1) baseline using original data only, (2) standard augmentation (Random Crop, Rotate, Blur, Affine, Geometric Distortion at 0.4 probability), (3) 30% synthetic data via vanilla fine-tuning, and (4) 30% synthetic data from **ViCTr**. Table 3 highlights consistent performance gains from **ViCTr** across all datasets and backbone architectures. Notably, CirrMRI600+—a liver cirrhosis dataset—exhibited the most significant improvements, reflecting **ViCTr**’s capacity to capture complex pathological features. These results confirm that **ViCTr**-generated images significantly boost segmentation metrics (mDSC and mHD95), highlighting the practical value of our two-stage framework for medical data augmentation.

Efficiency of ViCTr. Table 2 summarizes the number of diffusion steps and average inference times (in seconds) for vanilla fine-tuning versus the **ViCTr** framework. Across all

Baselines	BTCV Dataset (CT Generation)		AMOS Dataset (MRI Generation)		CirrMRI600+ (Pathology Generation)		Diffusion Steps Inference Time	
	Fine-tuned Vanilla	ViCTr	Fine-tuned Vanilla	ViCTr	Fine-tuned Vanilla	ViCTr	Fine-tuned Vanilla	ViCTr
Stable Diffusion [44]	25.44 / 19.67	21.98 / 19.02	25.43 / 21.76	20.37 / 19.11	28.34 / 23.43	25.57 / 21.46	40 13.76	4 3.12
Stable Diffusion XL [39]	23.47 / 18.21	20.33 / 17.44	24.11 / 20.23	19.44 / 18.45	27.34 / 22.11	24.02 / 20.76	30 14.55	4 3.45
Stable Diffusion-3 [14]	19.07 / 16.22	17.37 / 16.02	22.32 / 19.76	18.02 / 19.08	24.49 / 21.78	21.28 / 19.34	50 18.98	3 2.78
Pixart-alpha [9]	21.32 / 17.09	19.22 / 16.96	23.78 / 20.04	18.76 / 18.56	26.06 / 20.07	23.04 / 18.92	25 10.67	3 1.74
Flux [27]	15.52 / 15.01	13.28 / 14.08	19.02 / 18.28	15.55 / 16.58	22.46 / 18.88	19.96 / 17.01	30 15.66	3 2.87

Table 2. Quantitative results on CT, MRI, and Cirrhosis generation. All baselines were pre-trained on ATLAS-8k and fine-tuned on target datasets. Metrics reported are Fr chet Inception Distance (FID) and Medical FID (MFID), shown as FID/MFID, with inference time in seconds. Lower values indicate better performance.

Baselines	Original Dataset (Org.)		Augmentation		Org. + Synth. by FLUX 30% (vanilla Fine-tuning)		Org. + Synth. by FLUX 30% (ViCTr)	
	mDSC (�)	mHD95 (�)	mDSC (�)	mHD95 (�)	mDSC (�)	mHD95 (�)	mDSC (�)	mHD95 (�)
<i>BTCV dataset Segmentation results</i>								
UNet [45]	76.72	34.42	78.45	33.17	79.32	32.37	81.22	30.17
TransUnet [8]	85.52	32.33	87.01	31.43	87.54	30.11	89.78	29.12
nnUnet [20]	80.48	30.19	82.54	31.02	83.37	29.01	85.19	27.77
nnFormer [56]	83.47	30.01	83.98	30.88	85.88	28.78	87.72	26.22
MedSegDiff [52]	87.91	27.67	88.65	26.72	89.78	25.52	91.92	23.31
<i>AMOS dataset Segmentation results</i>								
UNet [45]	68.92	34.57	70.55	32.32	71.02	31.78	73.34	29.11
TransUnet [8]	71.33	33.12	72.56	31.09	73.43	29.19	77.54	27.56
nnUnet [20]	73.33	32.32	74.11	30.19	75.68	28.75	78.29	26.32
nnFormer [56]	75.78	31.19	76.44	29.76	77.77	27.57	81.32	24.41
MedSegDiff [52]	76.83	29.92	78.38	26.99	79.03	25.45	84.02	22.18
<i>CirrMRI600+ dataset Segmentation results</i>								
UNet [45]	68.74	36.73	69.38	35.43	70.12	35.11	73.39	32.11
TransUnet [8]	70.77	35.42	70.98	34.01	71.78	33.92	74.56	31.09
nnUnet [20]	71.02	34.35	72.49	32.78	73.56	31.27	78.89	30.25
nnFormer [56]	74.88	33.78	75.23	31.88	76.45	31.09	79.44	29.78
MedSegDiff [52]	76.92	30.79	77.11	30.34	78.03	29.89	81.37	27.34

Table 3. Segmentation results on BTCV, AMOS, and CirrMRI600+ using various training settings and baselines. Metrics reported are mDSC (%) and mHD95 (mm). Top three results per setting are highlighted: **blue** (1st), **green** (2nd), and **orange** (3rd).

evaluated models, ViCTr consistently reduces the required steps, thereby shortening inference duration without compromising image fidelity. This efficiency stems from two key innovations: (1) Tweedy’s formula to approximate Z_0 and streamlines the reconstruction process, and (2) Two-Stage Training, enabling the model to learn both anatomical consistency and pathology-specific variations in a targeted manner. By minimizing extraneous steps, ViCTr can generate high-quality synthetic images more rapidly, making it a viable solution for resource-constrained clinical or research environments where computational overhead is a critical consideration.

4.2. Qualitative Results

We conducted extensive qualitative evaluations of ViCTr across multiple modalities (CT and MRI) and tasks, covering both anatomical and pathological image synthesis.

Anatomical and Pathology-Driven Image Generation. Unlike prior methods that primarily generate normal anatomical scans or isolated tumors, ViCTr extends synthesis capabilities to complex pathologies, such as liver cirrhosis. Figure 3 shows pairs of non-cirrhotic MRIs alongside

their segmentation masks and the corresponding ViCTr-generated cirrhotic images. Notably, the synthesized images retain essential anatomical structures while introducing realistic cirrhotic texturing, an advancement that enables more diverse data augmentation and supports research into disease progression. These high-fidelity synthetic samples offer potential to enrich clinical training datasets, advance diagnostic algorithms, and support detailed analyses of pathology progression.

Qualitative Segmentation Outcomes. To further showcase ViCTr’s impact on downstream tasks, we visually assessed segmentation quality on BTCV and AMOS (Figure 6). Focusing on regions like the left kidney and pancreas, models trained with ViCTr-augmented data show sharper boundaries and more precise delineation than standard fine-tuning, indicating improved spatial learning.

Comparative Performance on BTCV and AMOS. Figures 4 (BTCV) and 5 (AMOS) compare diffusion-based image generation results using Stable Diffusion v3 (SD-3) and Flux backbones. On BTCV—a dataset with limited samples and larger spatial dimensions—vanilla fine-tuning with SD-3 often yields poor mask adherence, sug-

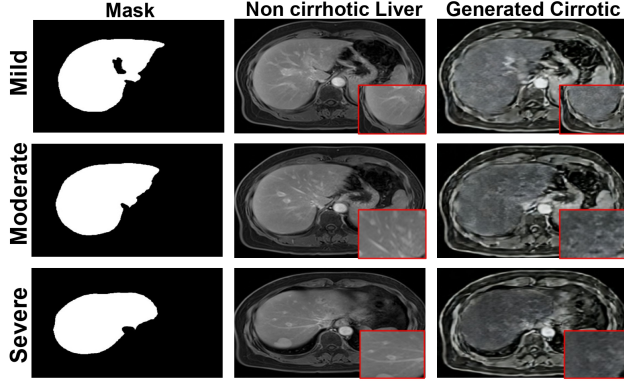


Figure 3. Examples of synthetic cirrhotic images with different severity levels generated from non-cirrhotic scans using ViCTr. For T1-MRI, we show the segmentation mask, the original non-cirrhotic image, and the corresponding synthetic cirrhotic image.

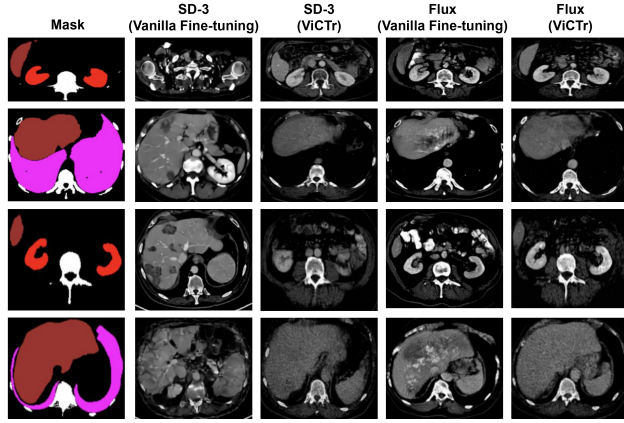


Figure 4. Diffusion-Based Image Generation on BTCV MRI Dataset. From left to right: segmentation mask, SD-3 with standard fine-tuning, FLUX with standard fine-tuning, SD-3 with ViCTr, FLUX with ViCTr.

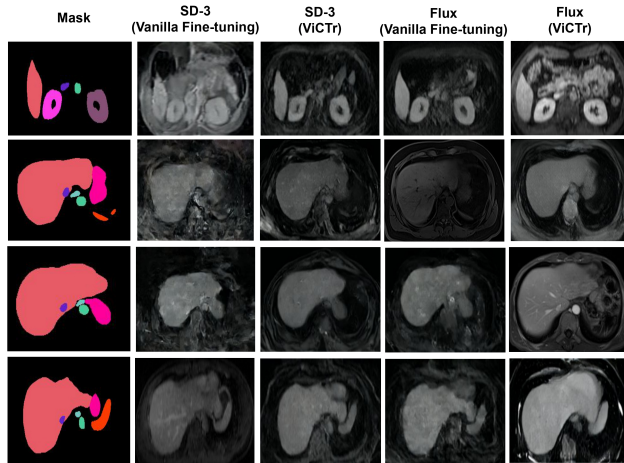


Figure 5. Diffusion-Based Image Generation on AMOS Dataset. From left to right: segmentation mask, SD-3 with Vanilla fine-tuning, SD-3 with our method, FLUX with Vanilla fine-tuning, FLUX with our method.

gesting difficulties in learning accurate spatial constraints. In contrast, the same model shows markedly improved mask alignment on the AMOS dataset, attributed to its larger training set and more manageable image dimensions.

Regardless of these dataset-specific variations, ViCTr achieves consistently high-quality outputs with robust mask adherence and anatomical accuracy across both BTCV and AMOS, underscoring its enhanced spatial learning and resilience. By maintaining performance even under data-sparse conditions, ViCTr demonstrates a clear advantage over conventional fine-tuning approaches, which can exhibit greater sensitivity to dataset size and complexity.

Second, we evaluate the impact of ViCTr on segmentation tasks, illustrated in Figure 6. Both BTCV and AMOS datasets encompass various abdominal structures (e.g., left kidney, pancreas), and our results highlight how ViCTr-based training yields notably more precise organ boundaries compared to alternative methods. This improvement is particularly evident in challenging regions where standard segmentation often struggle, highlighting ViCTr’s strength in enhancing spatial delineation and anatomical accuracy.

4.3. Ablation Studies

Ablation Based on Pre-training of Model		
ViCTr (Without Stage-1)	17.33	
Ablation Based on Rectified Flow Algorithms		
ViCTr (Without Proposed Tweedies Formula)	18.78	
ViCTr (With Reflow)	20.19	
ViCTr (With Flow Straight and Fast)	21.37	
ViCTr (With Distribution Matching Distillation)	22.33	
Ablation Study Based on Loss Functions		Ablation Based on LoRA Rank
L_{diff}	18.77	r = 8 18.46
$L_{diff} + L_{spatial}$	18.21	r = 16 17.52
$L_{diff} + L_{consistency}$	17.02	r = 32 16.66
$L_{diff} + L_{spatial} + L_{consistency}$	15.55	r = 64 15.55

Table 4. Ablation study for the effect of model pretraining, loss functions, and LoRA rank settings are shown.

To rigorously validate our ViCTr framework, we conducted a series of ablation experiments on the BTCV dataset (Table 4), focusing on four key aspects: pretraining, rectified flow algorithms, loss function components, and LoRA rank selection. We used FID (Fréchet Inception Distance) as our primary metric to gauge how each architectural choice affects both anatomical fidelity and pathological realism.

Impact of Pretraining. We first examined the necessity of Stage 1 pretraining by comparing our full, two-stage pipeline to a variant that proceeds directly to downstream fine-tuning. Omitting Stage 1 degrades the model’s FID score from 15.55 to 17.33 (Table 4), underscoring the vital role of an anatomical prior. These findings affirm that establishing robust structural representations in Stage 1 is crucial for achieving high-quality medical image synthesis.

Rectified Flow Algorithms. We next assessed our rectified flow formulation, augmented with Tweedie’s formula,

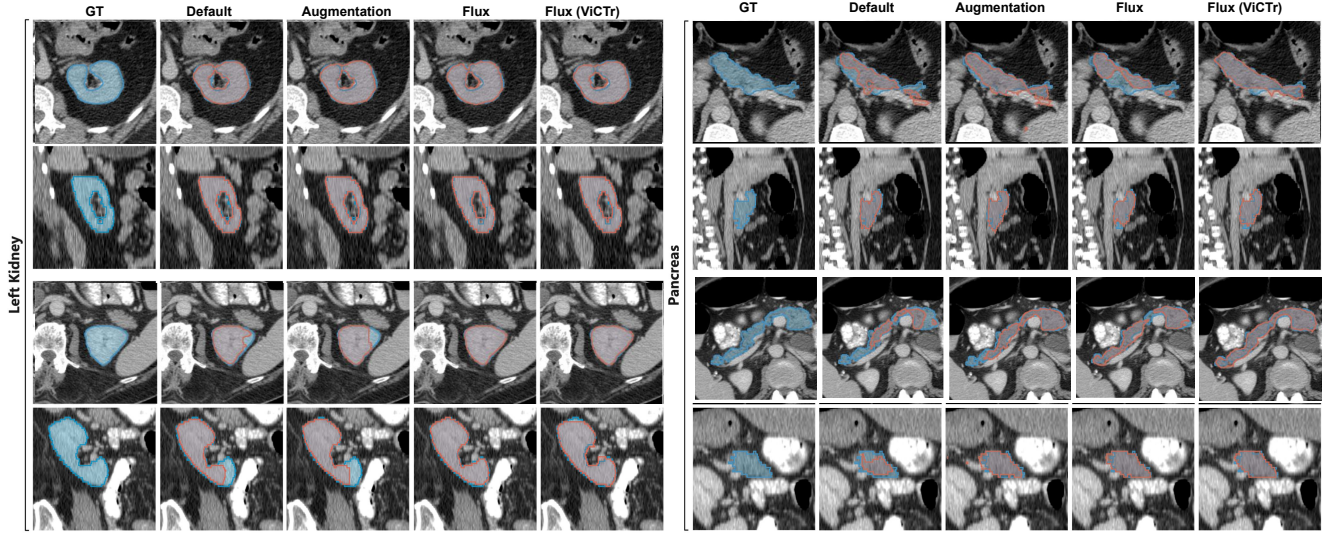


Figure 6. Segmentation results on BTCV and AMOS datasets’ left kidney and pancreas. We compare the ground truth (blue), models trained on the default dataset, augmented dataset, default dataset plus 30% synthetic data generated by FLUX with standard fine-tuning, and default dataset plus 30% synthetic data generated by FLUX with ViCTR (red).

against alternative rectified flow–based models, including ReFlow, Flow Straight and Fast, and Distribution Matching Distillation. Removing Tweedie’s correction alone increased the FID to 18.78, highlighting its significance for trajectory alignment. Across all baselines, we observed FID deteriorations, ranging from 13.28 to as high as 22.33, thereby demonstrating that our proposed method provides tighter distribution matching and superior synthesis fidelity.

Ablation Study Based on Loss Functions. We then investigated each ViCTR loss component—diffusion loss, spatial consistency, and consistency loss—to determine their relative contributions: the baseline model, utilizing only diffusion loss, produced an FID of 18.77 (Table 4). Incorporating spatial consistency loss yielded a moderate improvement to 18.21, highlighting its role in maintaining structural integrity. The addition of consistency loss further enhanced performance, reducing the FID to 17.02 by promoting image coherence across generations. When combining all three terms, we achieved an FID of 15.55, indicating that these components collectively offer significant improvements in anatomical accuracy and visual realism.

Ablation Based on LoRA Rank. Finally, we explored how varying the LoRA rank (r) influences generative performance. Starting at $r = 8$ and incrementing upwards, we found $r = 64$ to yield the best FID at 15.55. This result suggests that higher-dimensional adaptation spaces enable more nuanced parameter updates during fine-tuning, thereby improving visual quality and fidelity.

Visual Turing Tests. To further validate clinical plausibility, three radiologists participated in Visual Turing Tests using 15 randomly generated MRI scans depicting vary-

ing levels of liver cirrhosis (mild, moderate, severe). All scans were uniformly judged to be clinically realistic, with identical outcomes observed when using additional random samples. Notably, the synthetic cirrhotic images correctly exhibited surface nodularity and textural irregularities, consistent with radiologic findings in mild-to-severe cirrhosis (Figure 3). These results confirm ViCTR’s ability to synthesize pathology-specific features, enabling applications in training, algorithm development, and clinical research.¹

5. Conclusion

We presented ViCTR, a novel two-stage framework designed to generate high-fidelity medical images by integrating robust anatomical pre-training with precise pathology-specific fine-tuning. By aligning Tweedie’s formula with linear projection methods used in flow matching, ViCTR maintains accurate initial distribution estimates even amid diffusion processes. This setup is further enhanced by LoRA adapters, which preserve essential anatomical information while flexibly adapting to diverse pathologies. Experimental results demonstrate that ViCTR consistently surpasses conventional fine-tuning strategies, reducing FID scores and producing clinically realistic outputs. Beyond data augmentation for segmentation and classification, ViCTR can be extended to tasks such as modality translation, contrast synthesis, and professional training, offering a powerful and versatile tool for advancing AI-driven medical imaging.

¹Code: <https://github.com/Onkarsus13/ViCTR-2D>
Weights: <https://huggingface.co/onkarsus13/ViCTR-2D>

Acknowledgments

This research is supported by the following NIH grants: R01-HL171376 and U01-CA268808.

References

- [1] Hazrat Ali, Shafaq Murad, and Zubair Shah. Spot the fake lungs: Generating synthetic medical images using neural diffusion models. In *Artificial Intelligence and Cognitive Science*, pages 32–39, Cham, 2023. Springer Nature Switzerland. 2
- [2] Alhanoof Althnian, Duaa AlSaeed, Heyam Al-Baity, Amani Samha, Alanoud Bin Dris, Najla Alzakari, Afnan Abou El-wafa, and Heba Kurdi. Impact of dataset size on classification performance: an empirical evaluation in the medical domain. *Applied Sciences*, 11(2):796, 2021. 1
- [3] Fan Bai, Yuxin Du, Tiejun Huang, Max Q. H. Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models, 2024. 5
- [4] Hritik Bansal and Aditya Grover. Leaving reality to imagination: Robust classification via generated datasets. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*, 2023. 2
- [5] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fuzzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023. 2
- [6] Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *Journal of Machine Learning Research*, 23(325):1–31, 2022. 2
- [7] Junyi Chai, Hao Zeng, Anming Li, and Eric W.T. Ngai. Deep learning in computer vision: A critical review of emerging techniques and application scenarios. *Machine Learning with Applications*, 6:100134, 2021. 1
- [8] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021. 6
- [9] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024. 5, 6
- [10] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11147–11158, 2024. 2
- [11] Gayatri Deshmukh, Onkar Kishor Susladkar, Debesh Jha, Elif Keles, Halil Ertugrul Aktas, Daniela P Ladner, Amir A Borhani, Gorkem Durak, Ulas Bagci, et al. Meddelinea: Scalable and efficient medical image segmentation via controllable diffusion transformers. In *Medical Imaging with Deep Learning*, 2025. 2
- [12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2
- [13] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E. Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2
- [14] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yan-nik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. 5, 6
- [15] Xi Fang and Pingkun Yan. Multi-organ segmentation over partially labeled datasets with multi-scale feature abstraction. *IEEE Transactions on Medical Imaging*, 39(11):3619–3629, 2020. 5
- [16] Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. Flux that plays music, 2024. 3
- [17] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 3
- [18] Yunxiang Fu, Chaoqi Chen, Yu Qiao, and Yizhou Yu. Dreamda: Generative data augmentation with diffusion models. *arXiv preprint arXiv:2403.12803*, 2024. 2
- [19] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5
- [20] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 6
- [21] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27621–27630, 2024. 2
- [22] Debesh Jha, Onkar Kishor Susladkar, Vandan Gorade, Elif Keles, Matthew Antalek, Deniz Seythanoglu, Timurhan Cebeci, Halil Ertugrul Aktas, Gulbiz Dagoglu Kartal, Sabahattin Kaymakoglu, et al. Cirrmri600+: Large scale mri collection and segmentation of cirrhotic liver. *arXiv preprint arXiv:2410.16296*, 2024. 5
- [23] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35:36722–36732, 2022. 5
- [24] Guillermo Jimenez-Perez, Pedro Osorio, Josef Cersovsky, Javier Montalt-Tordera, Jens Hooze, Steffen Vogler, and Sadegh Mohammadi. Dino-diffusion. scaling medical diffusion via self-supervised pre-training. *arXiv preprint arXiv:2407.11594*, 2024. 3

- [25] Sunho Kim, Byungjai Kim, and HyunWook Park. Synthesis of brain tumor multicontrast mr images for improved data augmentation. *Medical Physics*, 48(5):2185–2198, 2021. 2
- [26] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024. 2
- [27] Black Forest Labs. black-forest-labs/FLUX.1-dev · Hugging Face. <https://huggingface.co/black-forest-labs/FLUX.1-dev>, 2024. [Online; accessed 05-October-2024]. 5, 6
- [28] Hansang Lee, Haeil Lee, and Helen Hong. Genmix: Combining generative and mixture data augmentation for medical image classification. *arXiv preprint arXiv:2405.20650*, 2024. 2
- [29] Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in Neural Information Processing Systems*, 37:63082–63109, 2025. 3
- [30] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 3
- [31] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. 3
- [32] Supriya V. Mahadevkar, Bharti Khemani, Shruti Patil, Ketan Kotecha, Deepali R. Vora, Ajith Abraham, and Lubna Abdelkareim Gabralla. A review on machine learning styles in computer vision—techniques and future directions. *IEEE Access*, 10:107293–107329, 2022. 1
- [33] Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M. Alvarez, Zhiding Yu, Sanja Fidler, and Marc T. Law. How much more data do i need? estimating requirements for downstream tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 275–284, 2022. 1
- [34] Xueyan Mei, Zelong Liu, Philip M Robson, Brett Marinelli, Mingqian Huang, Amish Doshi, Adam Jacobi, Chendi Cao, Katherine E Link, Thomas Yang, Ying Wang, Hayit Greenspan, Timothy Deyer, Zahi A Fayad, and Yang Yang. RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. *Radiol Artif Intell*, 4(5):e210315, 2022. 2
- [35] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarbuerger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, Jakob Nikolas Kather, and Daniel Truhn. A multi-modal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023. 2
- [36] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 417–425. Springer, 2017. 2
- [37] Kaliprasad Pani and Indu Chawla. Synthetic mri in action: A novel framework in data augmentation strategies for robust multi-modal brain tumor segmentation. *Computers in Biology and Medicine*, 183:109273, 2024. 2
- [38] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 5
- [39] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. 5, 6
- [40] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 36, 2024. 4
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5
- [42] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 5
- [43] Herbert E. Robbins. *An Empirical Bayes Approach to Statistics*, pages 388–394. Springer New York, New York, NY, 1992. 3, 4
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 5, 6
- [45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 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*, pages 234–241. Springer, 2015. 5, 6
- [46] Daniel G. Saragih, Atsuhiko Hibi, and Pascal N. Tyrrell. Using diffusion models to generate synthetic labeled data for medical image segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 19(8):1615–1625, 2024. 2
- [47] Vanshali Sharma, Abhishek Kumar, Debesh Jha, M.K. Bhuyan, Pradip K. Das, and Ulas Bagci. Controlpolyp-net: Towards controlled colon polyp synthesis for improved polyp segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2325–2334, 2024. 2
- [48] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2
- [49] Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffu-

- sion models. In *The Twelfth International Conference on Learning Representations*, 2024. [2](#)
- [50] Toan Tran, Trung Pham, Gustavo Carneiro, Lyle Palmer, and Ian Reid. A bayesian data augmentation approach for learning deep models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. [2](#)
- [51] Yibo Wang, Ruiyuan Gao, Kai Chen, Kaiqiang Zhou, Yingjie Cai, Lanqing Hong, Zhenguo Li, Lihui Jiang, Dit-Yan Yeung, Qiang Xu, et al. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7246–7255, 2024. [2](#)
- [52] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024. [6](#)
- [53] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024. [3](#)
- [54] Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. Emit-diff: Enhancing medical image segmentation via text-guided diffusion model. *arXiv preprint arXiv:2310.12868*, 2023. [2](#)
- [55] Yuan Zhong, Suhan Cui, Jiaqi Wang, Xiaochen Wang, Ziyi Yin, Yaqing Wang, Houping Xiao, Mengdi Huai, Ting Wang, and Fenglong Ma. *MedDiffusion: Boosting Health Risk Prediction via Diffusion-based Data Augmentation*, pages 499–507. [2](#)
- [56] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021. [6](#)