Diffusion Transformer in Medical Image Translation

Group Project

Data Science

Thai Doan Kien

Nguyen Hoang Ha

Bui Dinh Lam

Dao Hoang Dung

Nguyen Van Phu

Hoang Minh Chi

Nguyen Phuc Minh

Contents

Abstract

Diffusion models(DMs) have been explored to deal with creating images using denoising autoencoders. They outperform traditional methods like GANs or CNNs by using a VAE encoder to compress images into a more straightforward, smaller latent space, effectively capturing their key features. We can apply these models to synthesizing CT images from MRI images. However, training these models requires substantial computational resources because they rely on pixel-level processing. Additionally, the image generation process can be slow due to the step-by-step approach. Moreover, it struggles to represent complex medical images accurately.

To address this issue, we propose a new method that combines medical Contrastive Language-Image Pretraining (CLIP) with multimodal VAEs. By leveraging the unique properties of medical images, like deterministic distribution in MRI images, our approach aims to generate higher-quality medical images while reducing computational costs.

In this project, we focus primarily on synthesizing high-resolution brain images. Experimental results on brain and prostate datasets demonstrate the accuracy and reliability of the proposed approach, establishing it as a robust solution for MRI-to-CT synthesis. These results emphasize the potential of our method to address challenges in medical imaging, delivering high-quality and reliable output for clinical applications.

Keywords: MRI-to-CT synthesis, CycleGAN, Deep learning, Diffusion, MRI-to-CT Denoising Diffusion Model (MC-DDPM)

1 Introduction

Magnetic Resonance Imaging (MRI) and Computed Tomography (CT) are essential in brain tumor diagnosis and radiation therapy[9, 27]. Magnetic resonance imaging provides superior soft tissue contrast[38], while CT provides electron density information essential for calculating the radiation dose[10]. Combining these modalities improves treatment planning accuracy, avoiding damaging nearby organs through excessive dose usage. Furthermore, combining both modalities has proven effective in treatment planning [38]. Using MRI, health practitioners can identify boundaries unseen in CT for brain tumors [24]. This combination enhances precision in targeting tumors while sparing critical structures, improving treatment outcomes. However, getting both scans is often expensive. It causes patient discomfort due to radiation, such as prolonged scan times, exposure to ionizing radiation, and the need for contrast agents in some instances. Also, hospitals and imaging centers may have long wait times for MRI and CT appointments. Patients can benefit from a quicker diagnosis when synthetic CT is derived from their MRI scan,

bypassing the need to schedule a second scan.

Their approach produced robust, high-quality synthetic CT (sCT) [39] images on brain and pelvic datasets within minutes. Zhao et al. introduced a method combining a hybrid convolutional neural network (CNN) and transformer architecture as a generator within the GAN framework. Their results showed that this approach could accurately create CT images from pelvic MRI and handle small differences between MR and CT images. However, GAN-based methods often face challenges like unstable training, collapsing to a single output (mode collapse), and producing overly similar results, which are common problems in adversarial training.[39]

GANs also face other disadvantages. For example, they can be challenging to train because the generator and discriminator networks must compete against each other, which can lead to imbalances or failures during training. They often require extensive tuning of hyperparameters and careful setup to achieve good results.

Additionally, GANs sometimes produce artifacts in their generated images, especially when dealing with complex or high-resolution data. They may also struggle to capture the full diversity of the data, resulting in limited variability in the generated outputs. These limitations make GANs less reliable and harder to use in certain applications, such as medical imaging, where stability and accuracy are critical.[17]

To address these limitations, diffusion, and score-matching models have emerged as alternative generative approaches inspired by nonequilibrium thermodynamics. These models use a Markov chain of diffusion steps to progressively add random noise to data, learning to reverse this process to generate samples from pure noise. Diffusion models typically leverage a neural network, often a U-shaped CNN, to perform denoising. Unlike GANs, diffusion models avoid adversarial training, leading to more stable training processes and generating higher-quality, more realistic images with greater semantic diversity. Several diffusion-based generative models have been proposed for medical image synthesis, consistently achieving state-of-the-art image quality that surpasses both CNN-based and GAN-based methods.[7]

Recent advances in machine learning have introduced deep learning models for sCT (synthetic Computed Tomography) generation. GANs, namely CycleGAN[29] and pix2pix[25], have synthesized MRI to CT, producing high-quality images. But they often suffer from training instability, mode collapse, and limited diversity in many cases[7].

Diffusion models are inspired by thermodynamics and offer a more stable alternative, gradually learning to reverse a noise diffusion process to generate realistic images. These models improve stability and produce higher-quality, more diverse outputs compared to GANs[7]. In this study, we attempt to implement Diffusion Transformer[40] to this task.

Integrating transformers into the diffusion framework increases overall scalability and performance. Trained in paired MRI-CT datasets from the SynthRAD2023 competition [2], Diffusion Transformer progressively generates high-quality CTs images with minimal artifacts. Experiments on brain datasets highlight its superior performance and scalability compared to models like pix2pix.

2 Related Work

2.1 Deep medical image synthesis.

Applying deep learning techniques to medical imaging, particularly for image synthesis, has yielded significant advancements. Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs) have been widely employed for tasks such as converting MRI to CT images. Among the prominent methods, U-Net architectures [51] has emerged as a leading approach, excelling in structural accuracy and detail preservation. However, U-Nets can struggle to handle diverse input data, limiting their adaptability to highly heterogeneous medical datasets.

CycleGANs [29], designed for unpaired image translation tasks, have proven effective in generating visually realistic synthetic images. Yet, they often fail to maintain structural consistency and can introduce artifacts, particularly in complex anatomical regions. Conditional GANs [6] offer improved control over the synthesis process through conditional inputs, enabling more targeted generation. However, their reliance on extensive labeled data restricts their performance in scenarios with limited annotations. Finally, ResNet-based GANs [11] leverage deep residual connections to enhance feature extraction and capture fine-grained details in medical images. Despite their strengths, these models are computationally intensive and demand careful hyperparameter tuning to achieve optimal results.

Comparative studies [30] indicate that the Pix2Pix approach produces synthetic MRI (sMRI) images that differ significantly from the ground truth, leading to a loss of anatomical details and an inability to predict cerebrospinal fluid accurately. For CycleGAN methods, while there is no noticeable difference between paired and unpaired sMRI images, the results, though realistic, often exhibit higher noise levels compared to those generated by U-Net networks[30]. They consistently demonstrate the superiority of U-Net-based models, which achieve lower Mean Absolute Error (MAE) and Mean Squared Error (MSE), along with higher Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR)[30]. However, the existing approaches face limitations in handling

noise, maintaining anatomical fidelity, or balancing computational efficiency with scalability.

To address these challenges, we propose Diffusion Transformers (DiT)[40], which combine the strengths of attention mechanisms from transformers with the progressive denoising capabilities of diffusion models. DiT leverages the transformer's ability to model long-range dependencies and adapt to diverse data modalities[5][4] while incorporating the iterative refinement process of diffusion models for generating high-fidelity synthetic images[19][22]. This hybrid approach is designed to overcome the shortcomings of existing methods, offering improved structural consistency, noise reduction, and adaptability in medical image synthesis tasks.

2.2 Transformers.

Transformers have applied domain-specific frameworks across diverse areas like language, computer vision[4], reinforcement learning[31][34], and meta-learning[50]. They exhibit exceptional scalability as model size, computational power, and data volume grow, especially within language tasks, autoregressive models, and Vision Transformers (ViTs)[40]. Beyond linguistic applications, transformers have been used to predict pixels autoregressively and in discrete codebook-based and masked generative models[40]. Furthermore, diffusion models have investigated transformers to generate non-spatial data, such as creating CLIP image embeddings in DALL·E 2 [3][1].

Attention mechanisms are integral to Transformer architectures, enabling models to focus on the most relevant parts of input sequences effectively. At its core, the scaled dot-product attention computes a weighted sum of values (V) based on the similarity between queries (Q) and keys (K). The attention formula is expressed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V,$$

Where d_k represents the dimensionality of the key vectors, providing scaling to ensure numerical stability [5]. The mechanism begins by calculating the similarity between Q and K using a dot product, followed by scaling by $\sqrt{d_k}$ to stabilize gradients [14]. These similarities are converted into probabilities through a softmax function, which is then used to weight the values (V). This process allows the model to focus selectively on essential elements within the input.

The benefits of attention mechanisms in deep learning are profound. By capturing long-range dependencies, attention mechanisms overcome the limitations of recurrent

neural networks (RNNs) and long short-term memory (LSTM) networks in processing lengthy sequences [45]. Additionally, attention facilitates parallelization, simultaneously processing all tokens in a sequence, unlike the sequential, step-by-step nature of RNNs [5]. This parallelization significantly enhances computational efficiency. Another advantage is the interpretability offered by attention weights, which highlight the input elements most relevant to the model's predictions [26]. Furthermore, attention mechanisms are highly versatile, seamlessly adapting to various modalities, including text, images, and more [4].

These properties make attention mechanisms a cornerstone of modern deep-learning models. They underpin the functionality of advanced architectures such as Transformers [5], BERT [18], GPT [12], and Vision Transformers (ViTs) [4], solving fundamental challenges in deep learning and enabling breakthroughs across diverse applications.

2.3 Denoising diffusion probabilistic models (DDPMs).

Diffusion[22] and score-based generative models[46] have achieved significant success in image generation, often surpassing the performance of generative adversarial networks (GANs)[17], which were previously considered state-of-the-art[40]. Recent advancements in DDPMs have been driven by sampling techniques[47] such as classifier-free guidance, reformulating models to predict noise instead of pixels[22], and utilizing cascaded pipelines with low-resolution base models and upsamplers trained in parallel[7]. Most diffusion models use convolutional U-Nets[37] as their backbone architecture. However, recent work has proposed efficient attention-based architectures for DDPMs, leading to the exploration of pure transformer-based designs[40].

Denoising diffusion probabilistic models (DDPMs) operate by modeling the data distribution $p(x)$ as a gradual transformation of a simple prior distribution, such as Gaussian noise [19], into a target data distribution through a forward and reverse diffusion process. The forward process incrementally corrupts the data x_0 by adding Gaussian noise over T timesteps, governed by

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{\alpha_t}x_{t-1}, (1-\alpha_t)\mathbf{I}) \tag{1}$$

Where α_t controls noise addition [22]. The reverse process learns to reconstruct the data using a parameterized model $p_\theta(x_{t-1}|x_t)$ to approximate $q(x_{t-1}|x_t)$, predicting the mean μ_θ and variance Σ_θ. Training minimizes the variational lower bound, which reduces

in practice to predict the noise ϵ added at each timestep with

$$L_{simple} = E_{x_0,\epsilon,t}\big[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \tag{2}$$

] [36]Architecturally, DDPMs often employ convolutional U-Nets [37] for efficient spatial feature processing. However, attention-based and transformer architectures [5] are gaining traction for their scalability and ability to capture long-range dependencies. Recent advancements, including classifier-free guidance [23] and cascaded pipelines [7], further enhance DDPM performance, enabling state-of-the-art results in image generation tasks.

3 Method

Denoising Diffusion Probabilistic Models (DDPMs) in translating images between MRI and CT have been shown to allow robust synthetic generation compared to CNN and GAN models. [33]. Such diffusion models provide a probabilistic scheme to match the generated image to the distribution of actual photos, enabling realistic image generation. However, these methods still require excessive evaluations in pixel space, leading to increased computation and energy costs.

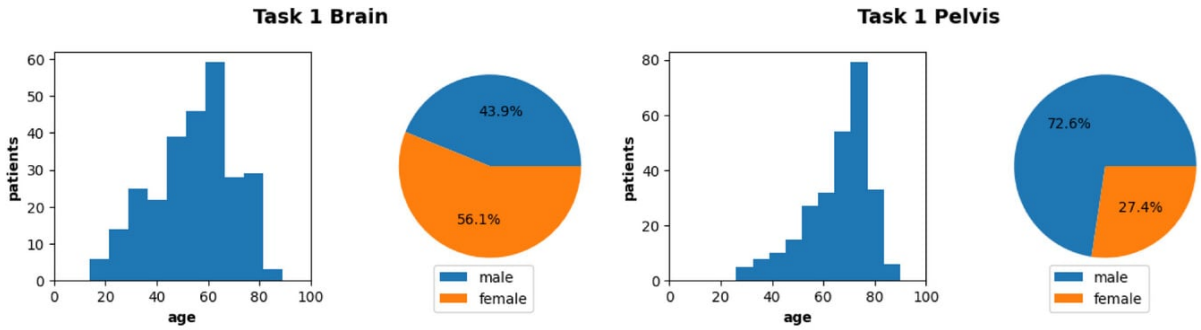
Latent Diffusion models [44] apply the VAE encoder to transform images into a smaller latent space, capturing a more intrinsic semantic meaning of the image. Nevertheless, multimodel VAE displays a limited joint representation of the latent space, especially with complex modalities such as images [49]. Therefore, more efficient ways have been developed to represent the original medical image as a latent space so that it preserves the underlying information in the image. For medical image generation tasks, each image type, such as CT and MRI, can be described by a deterministic distribution (e.g. Rician distribution for MRI image[15]). Such typical medical characteristics require a more robust handling of latent transformation to improve generating output quality.

We propose to circumvent this bottleneck by introducing the integration of medical Contrastive Language-Image Pretraining (CLIP), which is a biomedical vision-language foundation model pre-trained on 15 million paired caption-image data PMC-15M [54], into the based embedding provided by multimodel VAE. This approach aims to enhance medical interpretation within a targeted medical image type (MRI-CT in our single path translation), therefore addressing the ability to leverage the Diffusion Transformer model in medical image tasks by providing extended knowledge in the encoding phase.

3.1 Dataset

The dataset is a part of the SynthRAD2023[2] challenge, which focuses on synthesizing CT images from MRI scans. It includes paired MRI/CT images acquired between 2018 and 2022 from patients treated with external beam radiotherapy (photon or proton therapy) in the brain or pelvic regions. The data were collected from three Dutch university medical centers, anonymized as Center A, Center B, and Center C. The dataset consists of two subsets based on anatomical regions: Brain and Pelvis. MRI scans were primarily used for treatment planning, paired with corresponding CT scans. Inclusion criteria required patients to undergo radiotherapy and acquire CT and MRI data. This dataset represents diverse imaging protocols and patient characteristics, providing a robust foundation for developing and validating MRI-to-CT synthesis methods.

Case selection for the brain region was blind to clinical information regarding the primary tumor etiology, resulting in a random sample of tumor characteristics representative of clinical routine. In the pelvis region, cases of cervical, rectal, and prostate cancers were included with an approximately equal distribution among training, validation, and test sets at the institute level. Each subset generally contains an equal number of patients from each center, except for Brain, where no MR scans were available from Center B. To address this, Center A contributed twice the number of patients compared to other subsets. Imaging protocols varied both within and across centers, but only protocols with at least one-third of patients having comparable settings were included to ensure class balance and reduce variability. The dataset comprises 64% male and 36% female subjects, with a higher proportion of male patients in the pelvis subset due to the inclusion of prostate cancer cases (72.6% male in Pelvis). The patient population spans a wide age range, from 3 to 93 years, with a mean age of 65.



3.2 Diffusion Transformer

Diffusion formulation. In order to understand our method, we review some basic concepts in Diffusion Probabilistic Models (DDPMs). These models leverage Gaussian diffusion processes as a mechanism to systematically apply forward noise to real data x_0, described by:

$$q(x_t|x_0) = \mathcal{N}\left(\sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)\mathbf{I}\right)$$

where $\overline{\alpha}_t := \prod_{s=1}^{t} a_s = \prod_{s=1}^{t} 1-\beta_s$, with β_s being the forward process variance which is being constant here as hyperparameters. It also can be learned through reparameterization trick [28], which enables us to sample at timestep t:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\epsilon_t$$

where $\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$. This forward diffusion process acts as a controlled diffusing mechanism that transforms structured data into noise over multiple timesteps, which is the foundation for the model's reverse generative process.

The core objective of diffusion models is to learn the reverse-time stochastic process that reverses the effects of forward noise corruption, represented by:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t), \Sigma_\theta(x_t))$$

where Neural Networks are utilized to approximate the parameters μ_θ and Σ_θ, efficiently predicting the statistical behavior of p_θ. Training this reverse process is represented as the optimization problem subjected to the variational lower bound (VLB) of the data log-likelihood. Specifically, the VLB is expressed as:

$$L(\theta) = -p(x_0|x_1) + \Sigma_t D_{KL}\left(q^*(x_{t-1}|x_t, x_0)||p_\theta(x_{t-1}|x_t)\right)$$

where the term $q*$ and reverse process p_θ are Gaussian, allowing the KL divergence term D_{KL} to be computed using the mean and the covariance.

To further simplify the implementation, the parametrization of μ_θ as a noise prediction network ϵ_θ transforms the training objective into the minimization of the mean-square error (MSE) between the predicted noise $\epsilon_\theta(x_t)$ and the true sampled Gaussian noise ϵ_t. This results in a simplified loss function

$$L_{simple}(\theta) = ||\epsilon_\theta(x_t) - \epsilon_t||_2^2$$

which is computationally efficient and forms the backbone of most diffusion models. However, to train diffusion models that incorporate a learned reverse process covariance Σ_θ, the complete VLB must still be considered for optimizing p_θ. Therefore, an improved Diffusion Probabilistic Models [36] is proposed, allowing ϵ_θ to be trained with L_{simple} and Σ_θ with full L. Then, new images can be sampled by initializing $x_{t_{max}} \sim \mathcal{N}(0, \mathbf{I})$ and sampling $x_{t-1} \sim p_\theta(x_{t-1}|x_t)$.

Transformer in Diffusion Transformers, specifically various attention blocks have been tested in the Diffusion concepts (e.g Multi-Head Self-Attention, Head-Attention) [7]. The attention block in this scheme plays the main role in transforming the conditioning output from our hybrid vision encoder, displaying the global relationships between a compressed range of tokens. By adding a transformer into Diffusion, the model captures all semantic information from the conditioning, aligning with the attention mechanism to ensure that the input tokens are transformed to generalize the conditioning data. In

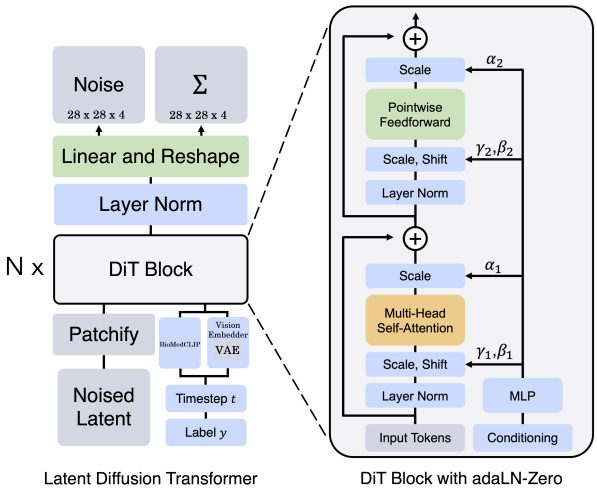


Figure 1: The Diffusion Transformer (DiT) architecture, with our hybrid encoder integrated

the context of Diffusion Transformer, we take advantage of the successor adaLN-Zero[40] (Adaptive layer norm Zero-out) architecture, which employs Multi-Head Self-Attention mechanisms incorporated with scale and shift parameter γ and β learned directly through the embedding vector of t (timesteps) and y (labeled classes). The zero-initialized modulation layer adaLN, which has proved to be beneficial in various medical image synthesis schemes [7] [41], is also reimplemented in our work. It adds the scale and shift parameters γ_1, β_1 and α_1 outside the Multi-Head Self-Attention layer. Same as the attention layer, the Feed-Forward Network also sits between scale and shift parameters γ_2, β_2, and α_2. The Patchify process transforms the spatial input into T tokens, each with dimension d, by linearly embedding each patch. For example, a patch size of $p \times p$ results in a sequence

length of $T = (I/p)^2$. Decreasing the patch size leads to a longer sequence and higher computational complexity in Gflops, although it does not significantly affect the total number of model parameters [40]. N layers of DiT blocks are placed one after another, then passed into a Norm Layer and linear layer, and decoded by the VAE decoder back to the original size.

The diffusion model is designed to learn the image distribution by gradually denoising a variable subjected to normal distribution, in other words, learn the reverse process of a fixed Markov Chain of length T 1. These models can be interpreted as equally weighted sequence of denoising autoencoders $\epsilon_\theta(x_t, t); t = 1...T$, which are trained to predict a denoised variant of their input x_t, where x_t is a noisy version of the input x 2. Our conditionals concatenate the results from both VAE and the pre-trained BioMedCLIP therefore, the conditional Diffusion Transformer is learned through:

$$L_{DiT} := E_{x_{mri}, \epsilon \sim \mathcal{N}(0,1)} \left[||\epsilon - \epsilon_\theta(e^{ve}, t, e^{clip})||_2^2 \right] \tag{3}$$

where e_{ve} 5 is fixed during training, e^{clip} represents pretrained model BioMedCLIP 4

Model size. For the experimental comparison between original VAE conditioning and our hybrid method, we create a model using 7 Diffusion Transformer blocks with path size 2x2. The original DiT model only uses the conditioning VAE from Stable Diffusion [44], whereas our model implements the Vision Embedder proposed in 3, concatenated with the BioMedCLIP Vision Encoder 6.

Following the original architecture, we designed our base Diffusion Transformer with 16 layers of DiT blocks of patch size 2x2, with a hidden size equal to 512, with the aim for stronger visualization of output results yet remaining comparable for experimental uses. DiT-L/2[40] not only suggests a higher efficiency of Gflops compared to the larger patch size implementation but it remains also important to improve performance.

Training. The model was trained using an MRI resolution of 224x224 before being compressed into the latent space of 28x28 by the VAE encoder. We apply the AdamW optimizer[32] with a learning rate of $1e-4$, without weight decay, learning rate or β_1, β_2 tuning.

3.3 Variational Autoencoder

We use Variational Autoencoders (VAEs) for the purpose of optimizing the Evidence Lower Bound (ELBO), balancing reconstruction loss for data fidelity and KL divergence to regularize the latent space, ensuring smooth and disentangled representations[8]. Vision Encoders play a vital role in mapping high-dimensional visual data[55], such as MRI

or CT scans, to lower-dimensional latent space[55]. This process helps to improve efficient representation learning and downstream tasks such as image synthesis and transformation on high-resolution images[40, 44]. The encoder maps input data to a latent space by approximating the posterior distribution $q_\phi(z|x)$, and outputs parameters for a Gaussian distribution[8]. Using the parametrization trick, a latent variable z is sampled and passed to the decoder, which reconstructs the input by modeling $p_\theta(x|z)$[8]. Choosing a fine-tuned model instead of training from scratch[44] not only saves time but also leverages pre-trained knowledge, which can lead to faster convergence and improved performance on specific tasks. The fine-tuned VAE models, such as ft-EMA and ft-MSE, offer optimized parameters that enhance the balance between reconstruction quality and latent space regularization[13].

3.4 Vision Encoder

3.4.1 Multimodel Biomedical Contrastive Language-Image Pretraining

Recognizing the VAE alone lacks expression for downstream tasks, especially for image translation tasks in medical schemes. We seek another effective encoder for input MRI images, and Contrastive Language-Image Pretraining (CLIP) introduced by OpenAI is a feasible option to retain textual-spatial connection represented in images. CLIPs are trained so that vectors represented by similar text-image pairs are closed in the shared space while dissimilar pairs are far apart. The loss is represented as a symmetric cross-entropy loss over the similarity score:

$$-\frac{1}{N}\sum_i ln\frac{e^{v_i.w_i/T}}{\sum_j e^{v_i.w_j/T}} - \frac{1}{N}\sum_j ln\frac{e^{v_j.w_j/T}}{\sum_i e^{v_i.w_j/T}}$$

where $v_i.w_i$ represents the dot product between image and text vectors, and T is the temperature that can be parameterized. However, this CLIP is subjected to a generalized vocab, which is also trained on a general dataset like ImageNet [42]. This would raise a problematic limitation as our downstream task is aiming at medical data, with CT and MRI images as our focus. Then, we implemented BioMedCLIP [54] to regularize this nature. This CLIP model is pre-trained on a novel dataset PMC-15M introduced by the author, containing 15 million biomedical image-text pairs collected from scientific articles.

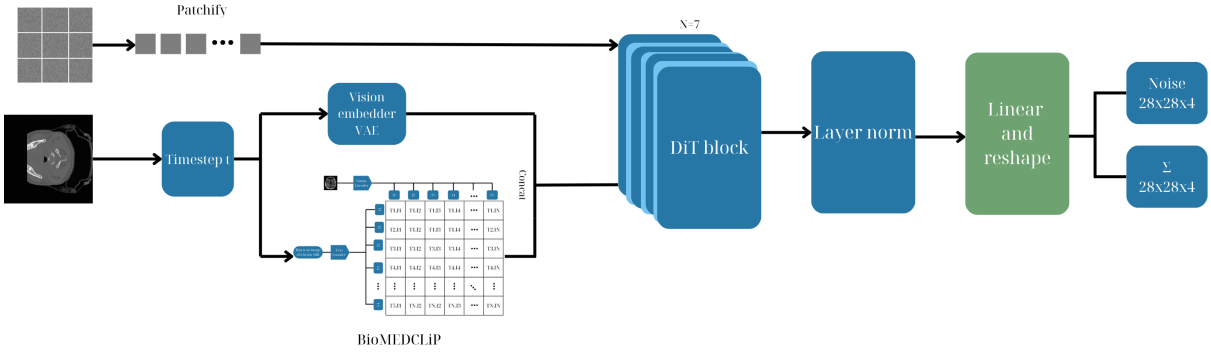


Figure 2: Model Architecture

3.4.2 Concatenate VAE encoder with BioMedCLIP

Biomedical data lies in diverse forms, from medical imaging, health records, and medical texts[35]. On the other hand, VAE is designed for unimodel data, therefore the pretrained VAE model itself [44] is unable to integrate the self-supervision advantages from cross-model correspondence [43]. Therefore, we decided to integrate a multimodel biomedical pre-trained model BioMedCLIP [54] along with the VAE model to incorporate additional information into the latent space. These models are represented as two specialized encoders: f^{VAE} and f^{CLIP}. In order to ensure the uniformity of two encoders, we integrate our pretrained Vision Embedder f^{ve} to map the output dimensions of VAE model available to concatenate with BioMedCLIP Vision Encoder. The formula is represented as below:

$$e^{clip} = f^{clip}(x_{mri}) \tag{4}$$

$$e^{ve} = f^{ve}(f^{VAE}(x_{mri})) \tag{5}$$

$$c = \text{concat}\left(\text{mean}(e^{ve}) + t, e^{clip} + t\right) \tag{6}$$

where $t, e^{clip} \in R^{B\times D}, e^{ve} \in R^{B\times L\times D}$ and $mean(.)$ denotes the function calculating the mean across the middle dimension.

The design of the Vision Embedder f^{ve} is displayed in **??**. Initially, the latent MRI obtained by VAE (4.2) is patchified into n by n patches identified by *latent size/patch size*, patch size must be divisible by latent size (28 in our case), hence available in 2x2, 4x4, 7x7, 14x14 in size respectively. Then, we employ a Convolutional Layer to extract low-level spacial features in the MRI image. The output is fed into both the Average Pooling layer and the Max Pooling layer since it is crucial to capture both the global context and local information from the image. Pooled features are feed-forward into a Neural Network, which contains a ReLU activation function between 2 linear transformations

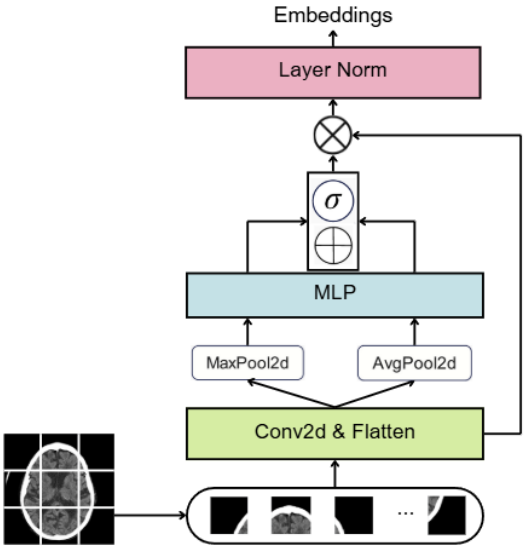


Figure 3: Vision Embedder

to introduce nonlinearity with economic focus, then are combined, activated, and reconnected to original features extracted in the Convolution by multiplication. Finally, the Layer Normalization is applied, ensuring the model is learning each image distribution individually, thereby keeping it generalizable through large examples.

3.4.3 Loss function

Self-supervised learning techniques allow models to be trained using the dataset's inherent structure or relationship within it, rather than being dependent on labeled data. Contrastive learning is a subset of self-supervised learning, where the model learns to extract meaningful representations by contrasting data points in the embedding space.
In order to train the vision embedder, contrastive learning was employed to allow the vision embedder to focus on the differences between MRI images. For this particular task, cross InfoNCE loss was utilized to optimize the vision embedder. Before this step, the embeddings generated from the vision embedder are flattened into a 1D vector $e^{-ve} \in R^{B\times(L.D)}$, then it is normalized using L2 normalization, where each row is normalized independently. The similarity matrix sim is then calculated as follows:

$$\hat{e}_k^{ve} = \frac{e_k^{-ve}}{\left\|e_k^{-ve}\right\|^2}, \quad k = 1, \cdots, B$$

$$sim = \frac{\hat{e}_k^{ve} \left(\hat{e}_k^{ve}\right)^{\mathrm{T}}}{\tau},$$

Where τ is a temperature hyperparameter that controls how sharp the softmax distribu-

tion is, a lower τ means that the models mainly focus on a similar pair, while a higher one means it also focuses on other different pairs as well. This means that a high τ will make the model focus on pulling similar pairs of samples together and negate how far or how close other data points are. A lower one would focus on both, pulling similar ones together, and pushing dissimilar ones away. Furthermore, $\left\|e_k^{-ve}\right\|^2$ stands for the magnitude or the L2 norms of the vector. The $(.)^T$ denotes the transpose operation. Finally, the InfoNCE loss is defined as follows:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp(\text{sim}(i,i))}{\sum_{i=1}^{N}\exp(\text{sim}(k,j))}$$

3.5 Evaluation

We decided to go with MAE, PSNR, and SSIM for measuring image similarity, similar to the ones used in SynthRAD2023 for evaluation and comparison. These metrics are widely used in assessing image quality in the image generation domain. They allow us to understand better how closely our generated images resemble the original images. Additionally, these metrics evaluate how closely the distribution of generated images matches that of the original images.

MAE was used to measure the mean absolute difference between synthetic CT and CT. It is used in a number of researches[30, 16, 39] to measure image quality, the lower the MAE, the better the quality. Specifically, it calculates the average absolute difference between corresponding voxels in the synthetic CT and the ground truth CT. Having the following formula:

$$\text{MAE}(y,\hat{y}) = \frac{\sum_{i=0}^{N-1}|y_i - \hat{y}_i|}{N}$$

where N is the total number of voxels. With y being the voxels of the ground truth image, and $\hat{y}$ being the synthetic CT image and N is the total number of voxels in the image.

SSIM was used to find the structural similarity between the generated images and the ground truth. This particular metric is also utilized to assess generated images in various types of research [16, 39]. Unlike MAE, SSIM uses three key factors brightness, contrast and structure. This allowed SSIM to be a more comprehensive way to evaluate generated images [48]. This is defined by:

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

Where μ_x and μ_y are the means of the patches of the images, the two terms capture the brightness of both images. σ_{xy} is the covariance of CTs and CT within a patch. This term captures how the structure of ground truth and generated images varies. σ_x^2 and σ_y^2 are the variances of the patches of the images, these are to measure the contrast. C_1 and C_2 are small constants to avoid 0s in the division.

PSNR is used to measure the ratio between the image's maximum possible value and the amount of noise or error present in the generated image. It is defined as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{MAX^2}{\text{MSE}} \right)$$

Where MSE is the mean squared error between the two images

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

and MAX is the maximum possible pixel value. A higher PSNR value means better image quality.

4 Result

4.1 Harmonize the limitations of GAN and Pix2Pix

During training, we ran into one fairly notorious challenge of training GANs, which was the generator successfully tricked the discriminator in CycleGAN. This led to us having to implement WGAN-GP loss, but this cost much more resources and time. Even after that, the CycleGAN could not converge really well, since it already took more than a month as each time of training takes around 10 days, we abort the result as it was not working out well. Proving one of the main weaknesses of GANs

Furthermore, with training pix2pix, we made smoother progress as it converged and showed promising outcomes. However, there were still rare exceptions during sampling, where the model produced images with a lot of artifacts and noise. This has several aspects of mode collapse, which is another problem of GAN, where the model does well

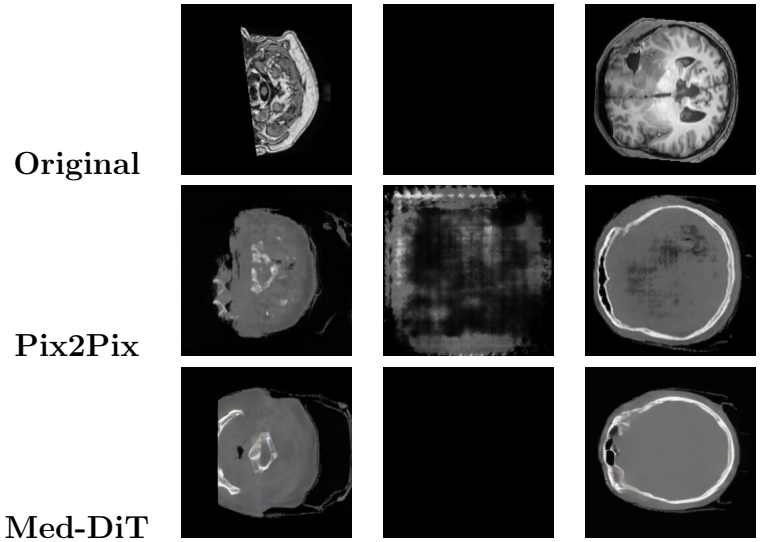


Table 1: Showcase of exceptions: Comparison of Ground Original and Generated.

on certain images, while some suffer.

The table showed that the diffusion transformer handles mapping really well, surpassing Pix2Pix in certain situations. With hardly any artifacts, it is not confusing when translating certain images.

4.2 Comparisons

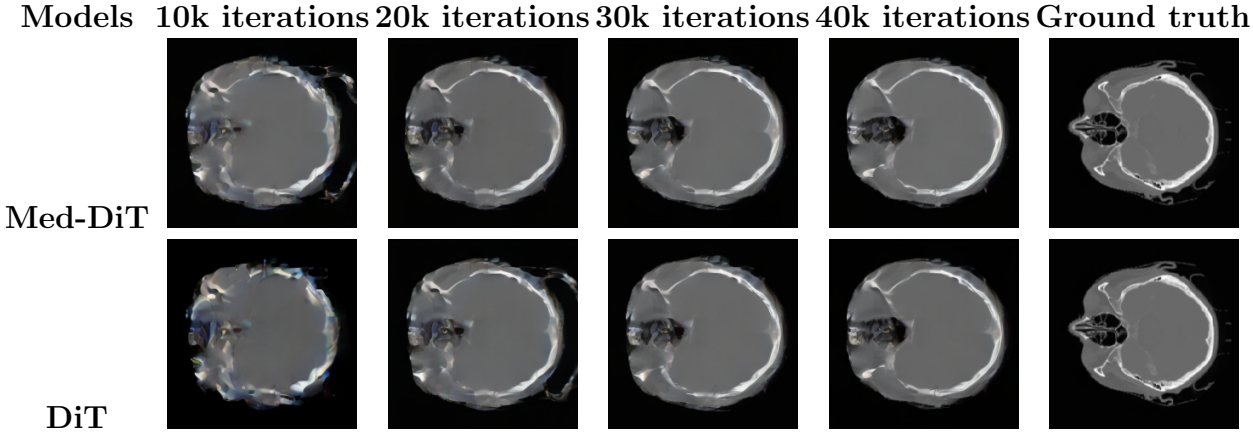


Table 2: Comparisons of CT images generated by hybrid DiT (denote by Med-DiT) and DiT in base DiT-SB/2

Effectiveness of proposed Vision Encoder. We started by training our based model with 7 AdaLN-Zero blocks, the patch size setting to 2, namely DiT-SB/2. The model is raised to quickly visualize the improvement of the selected method in comparison with the original Diffusion Transformer model. Changes made to the original Diffusion

Transformer model proposed in 3.4 will be selected and compared with the original model of the VAE encoder pre-trained by Stable Diffusion. Visualization of our comparison is presented in 2.

Within the same iterations trained, our method has the ability to converge faster, representing the most observable in reconstructing the frontal bone, parietal bone, and occipital bone (observation in 10k iterations). Moreover, in observation within 20k iterations, the generalization ability is further assessed by the ability to recognize the absence of outer soft tissue faster, while the original model still can only detect this difference within 30k iterations.

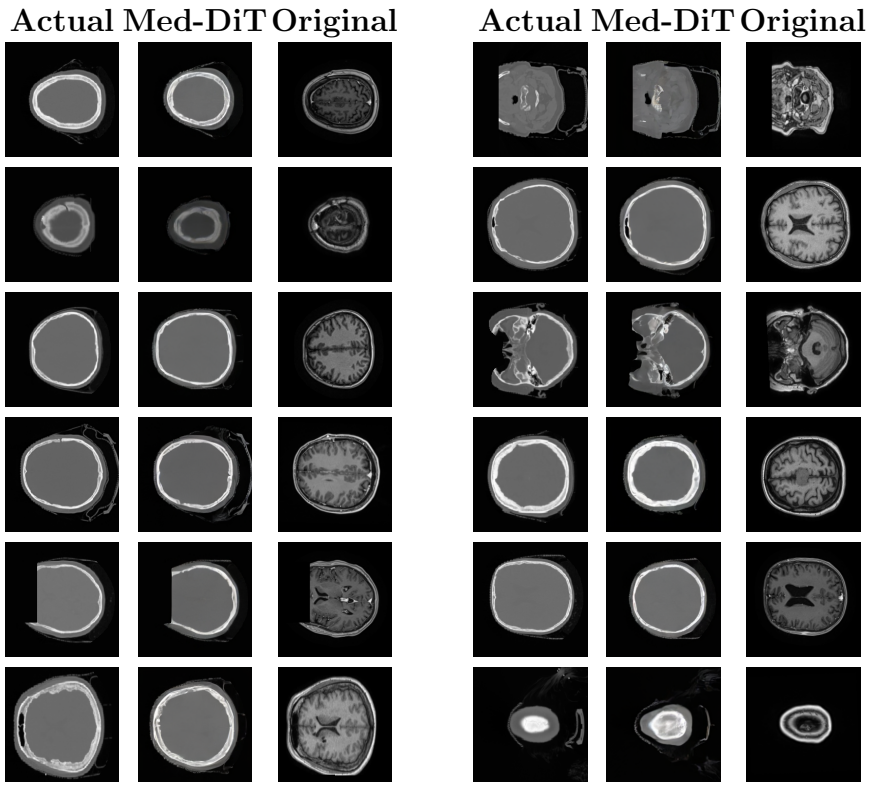


Table 3: Diffusion Transformer using BioMedCLIP, implementing DiT-L/2 architecture

More blocks, better model. We trained a more competitive model to the best computational ability with 16 DiT blocks represented in 1. Even though the soft tissue was not represented in MRI images, our Med-DiT has the ability to predict the structure of soft tissue in most cases, further suggesting that a deeper model with better-optimized parameters could further be scaled to improve proficiency and precision. In addition, a non-invasive imaging technique like pediatric MRI uses radio waves and magnetic fields to create detailed pictures of a child's body, with brain scans included. Because it doesn't use radiation, it is considered the preferred imaging technique for children. In those cases, our model can still perform relatively precisely with regard to the details of tissue and

the overall shape of skulls.

4.3 Metrics and samples

4.3.1 Metrics

Position	Model	MAE	SSIM	PSNR
1	SMU-MedVision	58.83 ± 13.41	0.885 ± 0.029	29.61 ± 1.79
2	FAYIU	61.72 ± 13.32	0.876 ± 0.030	28.83 ± 1.61
3	Elekta	62.76 ± 13.06	0.875 ± 0.030	28.80 ± 1.60
...				
14	**Diffusion Transformers**	85.55 ± 54.23	0.711 ± 0.088	19.10 ± 2.23
...				
16	reza.karinzadeh	113.38 ± 20.35	0.764 ± 0.034	24.71 ± 1.43
17	X-MAN	117.88 ± 45.08	0.774 ± 0.097	25.64 ± 2.20
18	thomashelfer	126.32 ± 17.01	0.756 ± 0.029	23.69 ± 0.94
19	Pix2pix	152.74 ± 53.819	0.6923 ± 0.1630	23.52 ± 2.52
20	CycleGAN	104.65 ± 14.12	0.4278 ± 0.059	11.99 ± 1.04

Table 4: Comparison of Metrics for Different Models ($\mu \pm \sigma$)

The plot illustrates the Structural Similarity Index (SSIM) scores for CT images generated at different training iterations (10k, 20k, 40k, and 70k). SSIM is a critical metric in image quality assessment, measuring the perceived similarity between the generated image and the ground truth. Higher SSIM values indicate better structural preservation and image fidelity.

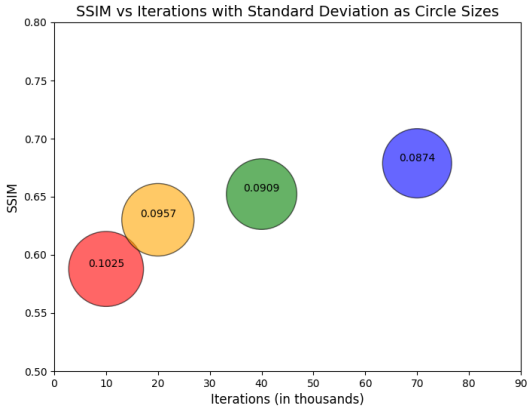


Figure 4: SSIM, an important metrics, improves significantly throughout training process. Radius of circle represents the standard deviation of each checkpoint

In this visualization, the x-axis represents the number of training iterations (in thousands), while the y-axis shows the SSIM values. The size of the circles corresponds to the standard deviation (std) of SSIM, highlighting the variability of the generated results.

As training progresses from 10k to 70k iterations, there is a noticeable improvement in SSIM, reflecting enhanced structural consistency in the generated CT images. Simultaneously, the standard deviation decreases, indicating greater stability and reliability of the model's outputs as training advances.

4.3.2 Samples

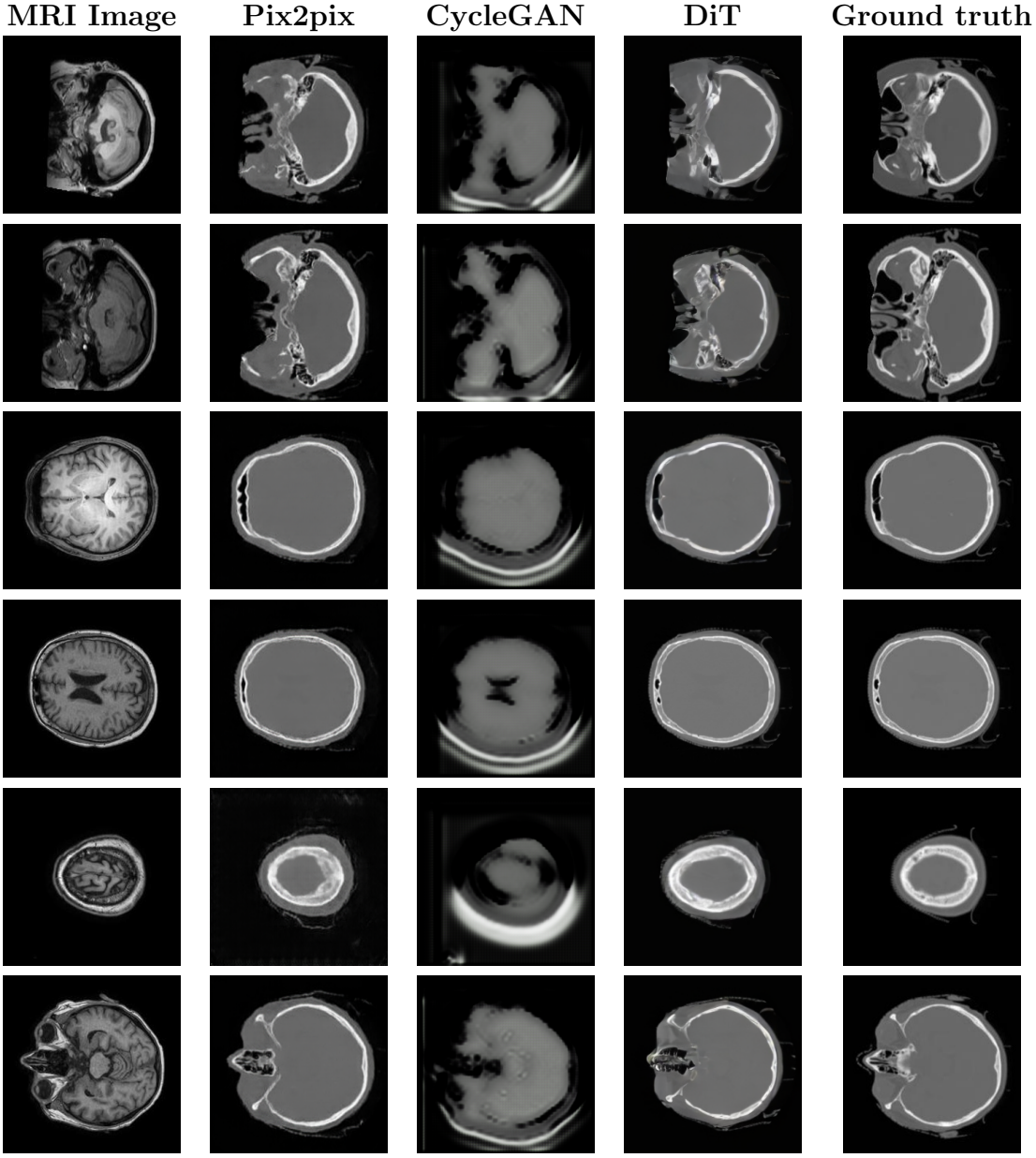


Table 5: Comparison of Outputs from Different Models (Note: DiT being Diffusion Transformer

5 Analysis and Discussion

The results of this study reveal key differences in performance between GAN-based models (Pix2Pix, CycleGAN, StarGAN) and the Diffusion Transformer for medical image synthesis. These findings provide valuable insights into the strengths and limitations of each method and their potential applications.

The Diffusion Transformer demonstrated strong overall performance with lower Mean Absolute Error (MAE) (85.56 ± 54.23) compared to Pix2Pix (152.74 ± 53.81) and CycleGAN (104.65 ± 14.12). The Diffusion Transformer offered more consistent results, as reflected in its Structural Similarity Index (SSIM) (0.7110 ± 0.0885) and Peak Signal-to-Noise Ratio (PSNR) (19.10 ± 2.23). These metrics suggest that the Diffusion Transformer produces more reliable and realistic images with fewer artifacts.

In comparison, Pix2Pix performed well during training and achieved convergence, but it struggled with occasional artifacts and noise in some outputs. This indicates a known issue in GAN-based models called mode collapse, where the model performs well on certain inputs but fails on others.

CycleGAN, the model encountered considerable challenges during the training phase despite its notable flexibility in managing unpaired datasets. Specifically, the generator frequently succeeded in deceiving the discriminator with excessive power, necessitating the integration of Wasserstein GAN with Gradient Penalty (WGAN-GP) loss to achieve training stability. However, this remedial measure significantly increased resource consumption and extended the training duration to over a month, yet it failed to yield satisfactory results. This scenario underscores a critical limitation of CycleGAN: its pronounced sensitivity to adversarial training dynamics coupled with a high computational overhead. Such constraints highlight the model's diminished practicality, especially in scenarios demanding efficient and scalable solutions.

Delving deeper into the mechanics of GANs, these models depend fundamentally on adversarial training, including a generator and a discriminator engaging in a competitive process. While this dynamic can facilitate the production of high-fidelity images, it is frequently accompanied by instability, mode collapse, and heightened sensitivity to hyperparameter configurations, as evidenced in both Pix2Pix and CycleGAN implementations. These challenges can impede the reliability and consistency of GAN-based models, particularly in complex applications like medical imaging, where precision is paramount.

In contrast, Diffusion Transformers adopt a progressive refinement methodology that circumvents the adversarial dynamics intrinsic to GANs. This approach ensures more stable training processes and yields consistent results, which is critically important in

medical applications where accuracy and reliability cannot be compromised. The absence of adversarial competition mitigates common GAN-related issues, resulting in lower artifact rates and enhanced image quality. Consequently, the Diffusion Transformer emerges as a robust and scalable solution for generating high-quality medical images, effectively addressing many challenges that hinder GAN-based models.

The implications of these findings are significant for the medical imaging domain. The superior stability and consistent performance of the Diffusion Transformer render it a promising tool for applications such as MRI-to-CT or CT-to-MRI synthesis. Its ability to produce reliable and high-quality images with reduced computational demands positions it as a more viable option than GAN-based models, which, despite their capability to generate impressive results under controlled conditions, scalability and reliability concerns restrict their broader applicability.

In summary, the Diffusion Transformer outperforms GAN-based methods in robustness and reliability, making it a more practical choice for real-world medical imaging tasks. Future work could further enhance the Diffusion Transformer's architecture to improve efficiency and extend its applicability to a broader range of datasets.

6 Conclusions

Based on the results obtained in this study, we draw significant conclusions regarding the performance and applicability of the Diffusion Transformer in handling large-scale medical image datasets. However, our model is trained using a default configuration aimed at research purposes, which leaves substantial room for enhancing its efficiency, scalability, and real-world applicability. Below are the research directions we intend to pursue in the future.

To enhance the performance of the Diffusion Transformer, we plan to implement new Transformer designs and advanced optimization methods. Specifically, we aim to utilize the Swin Transformer [53], an architecture that organizes data hierarchically and employs small attention windows, thereby making the model more flexible and easily scalable. Additionally, the Focal Transformer [21] will be integrated to dynamically adjust attention regions, enabling the model to focus on critical areas of the image while preserving finer details. To reduce computational power requirements, we will also explore efficient models such as the Efficient Former [52], and Grouped Self-Attention Mechanism [20]. These enhancements are expected to make the model faster and more practical for handling sizeable medical image datasets.

Furthermore, we will undertake a comprehensive hyperparameter optimization process

to fine-tune the model's performance. This involves adjusting the learning rate to identify the most suitable level, facilitating faster convergence and optimal performance. We will also optimize the number of attention heads to balance the model's ability to learn complex relationships with computational efficiency. Additionally, adjusting the model depth will ensure that the model is sufficiently complex to capture essential features without causing overfitting.

We will expand our experiments to include diverse datasets to ensure that the Diffusion Transformer performs effectively across various real-world scenarios. This includes utilizing different imaging modalities, such as MRI, CT, and ultrasound, and datasets related to multiple organs. We aim to evaluate the model's generalization capabilities and ensure its applicability in diverse medical applications by testing the model on a wide range of data.

In addition to directly improving the model's architecture and optimization methods, we will allocate resources to explore supplementary utilities that support the Diffusion Transformer. Specifically, we plan to implement advanced data augmentation techniques to enhance the model's learning capabilities and reduce the risk of overfitting. Simultaneously, we will continuously research and incorporate the latest deep learning optimization techniques to improve the model's training speed and efficiency.

References

[1] Alex Nichol Casey Chu Aditya Ramesh, Prafulla Dhariwal and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[2] Arthur Galapon Jr Joost J. C. Verhoeff Johannes A. Langendijk Stefan Both Cornelis (Nico) A. T. van den Berg Matteo Maspero Adrian Thummerer, Erik van der Bijl. Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. *Medical Physics*, 50(7):4664–4674, June 2023.

[3] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark et al Alec Radford, Jong Wook Kim. Learning transferable visual models from natural language supervision, 2021.

[4] Alexander Kolesnikov Dirk Weissenborn Xiaohua Zhai Thomas Unterthiner Mostafa Dehghani Matthias Minderer Georg Heigold Sylvain Gelly Jakob Uszkoreit Neil Houlsby † Alexey Dosovitskiy†, Lucas Beyer. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[5] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N. Gomez Lukasz Kaiser Illia Polosukhin Ashish Vaswani, Noam Shazeer. Attention is all you need, 2023.

[6] Vanquin L Wagner A Lacornerie T Pasquier D Reynaert N Brou Boni KND, Klein J. Mr to ct synthesis with multicenter data in the pelvic area using a conditional generative adversarial network. *Phys Med Biol*, 65, 2020 Apr 2.

[7] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021.

[8] Max Welling Diederik P Kingma. Auto-encoding variational bayes, 2013.

[9] Piet Dirix, Karin Haustermans, and Vincent Vandecaveye. The value of magnetic resonance imaging for radiotherapy planning. *Seminars in Radiation Oncology*, 24(3):151–159, 2014. Magnetic Resonance Imaging in Radiation Oncology.

[10] H Jane Dobbs, Robert P. Parker, N. J. Hodson, Pauline Hobday, and Janet E. S. Husband. The use of ct in radiotherapy treatment planning. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*, 1 2:41–133, 1983.

[11] Nejad-Davarani SP Glide-Hurst CK Emami H, Dong M. Generating synthetic cts from magnetic resonance images using generative adversarial networks. *Phys Med Biol*, 2018.

[12] Radford et al. Language models are unsupervised multitask learners, 2019.

[13] Hugging Face. Stable diffusion fine-tuned vae - ft-mse. https://huggingface.co/stabilityai/sd-vae-ft-mse-original, 2025.

[14] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[15] HáKon Gudbjartsson and Samuel Patz. The rician distribution of noisy mri data. *Magnetic Resonance in Medicine*, 34(6):910–914, 1995.

[16] Evi M. C. Huijben, Maarten L. Terpstra, Arthur Jr. Galapon, Suraj Pai, Adrian Thummerer, Peter Koopmans, Manya Afonso, Maureen van Eijnatten, Oliver Gurney-Champion, Zeli Chen, Yiwen Zhang, Kaiyi Zheng, Chuanpu Li, Haowen Pang, Chuyang Ye, Runqi Wang, Tao Song, Fuxin Fan, Jingna Qiu, Yixing Huang, Juhyung Ha, Jong Sung Park, Alexandra Alain-Beaudoin, Silvain Bériault, Pengxin Yu, Hongbin Guo, Zhanyao Huang, Gengwan Li, Xueru Zhang, Yubo Fan, Han Liu, Bowen Xin, Aaron Nicolson, Lujia Zhong, Zhiwei Deng, Gustav Müller-Franzes, Firas Khader, Xia Li, Ye Zhang, Cédric Hémon, Valentin Boussot, Zhihao Zhang, Long Wang, Lu Bai, Shaobin Wang, Derk Mus, Bram Kooiman, Chelsea A. H. Sargeant, Edward G. A. Henderson, Satoshi Kondo, Satoshi Kasai, Reza Karimzadeh, Bulat Ibragimov, Thomas Helfer, Jessica Dafflon, Zijie Chen, Enpei Wang, Zoltan Perko, and Matteo Maspero. Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report, 2024.

[17] Mehdi Mirza Bing Xu David Warde-Farley Sherjil Ozair Aaron Courville Ian Goodfellow, Jean Pouget-Abadie and Yoshua Bengio. Generative adversarial nets, 2014.

[18] Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[19] Niru Maheswaranathan Surya Ganguli Jascha Sohl-Dickstein, Eric A. Weiss. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[20] Mengxing Huang Siling Feng Jian Liu, Wenlong Feng and Yu Zhang. Grouped multilayer practical byzantine fault tolerance algorithm: A practical byzantine fault tolerance consensus algorithm optimized for digital asset trading scenarios, 2023.

[21] Xiyang Dai Jianfeng Gao Jianwei Yang, Chunyuan Li. Focal modulation networks, 2021.

[22] Ajay Jain Jonathan Ho and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[23] Tim Salimans Jonathan Ho. Classifier-free diffusion guidance, 2022.

[24] M Just, H P.H. Roesler, H P Higer, J Kutzner, and M Thelen. Mri-assisted radiation therapy planning of brain tumors–clinical experiences in 17 patients. *Magnetic Resonance Imaging; (USA)*, 9:2, 01 1991.

[25] Bodo Kaiser and Shadi Albarqouni. Mri to ct translation with gans, 2019.

[26] Ryan Kiros Kyunghyun Cho Aaron Courville Ruslan Salakhutdinov Richard Zemel Yoshua Bengio Kelvin Xu, Jimmy Ba. Show, attend and tell: Neural image caption generation with visual attention, 2015.

[27] Danial Khan, Khalil Abbas, and Sharjeel Nawaz. The role of ct scan in modern radiology: From diagnosis to treatment planning. 05 2023.

[28] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[29] Yang Lei, Joseph Harms, Tonghe Wang, Yingzi Liu, Hui-Kuo Shu, Ashesh B. Jani, Walter J. Curran, Hui Mao, Tian Liu, and Xiaofeng Yang. Mri-only based synthetic ct generation using dense cycle consistent generative adversarial networks. *Medical Physics*, 46(8):3565–3581, 2019.

[30] Qin W et al Li W, Li Y. Magnetic resonance image (mri) synthesis from brain computed tomography (ct) images based on deep learning methods for magnetic resonance (mr)-guided radiotherapy. *Quant Imaging Med Surg*, 10:1408–1419, 2020.

[31] Aravind Rajeswaran Kimin Lee Aditya Grover Misha Laskin Pieter Abbeel Aravind Srinivas Lili Chen, Kevin Lu and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling, 2021.

[32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

[33] Qing Lyu and Ge Wang. Conversion between ct and mri images using diffusion and score-matching models, 2022.

[34] Qiyang Li Michael Janner and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem, 2021.

[35] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, Apr 2023.

[36] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.

[37] Philipp Fischer Olaf Ronneberger and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[38] Amir Owrangi, Peter Greer, and Carri Glide-Hurst. Mri-only treatment planning: Benefits and challenges. *Physics in Medicine and Biology*, 63, 02 2018.

[39] Shaoyan Pan, Elham Abouei, Jacob Wynne, Tonghe Wang, Richard L. J. Qiu, Yuheng Li, Chih-Wei Chang, Junbo Peng, Justin Roper, Pretesh Patel, David S. Yu, Hui Mao, and Xiaofeng Yang. Synthetic ct generation from mri using 3d transformer-based denoising diffusion model, 2023.

[40] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.

[41] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017.

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[43] Khadija Rais, Mohamed Amroune, Abdelmadjid Benmachiche, and Mohamed Yassine Haouam. Exploring variational autoencoders for medical image generation: A comprehensive study, 2024.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022.

[45] Hochreiter Schmidhuber. Long short-term memory, 1997.

[46] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020.

[47] Timo Aila Tero Karras, Miika Aittala and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022.

[48] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[49] Daniel Wesego and Amirmohammad Rooshenas. Revising multimodal vaes with diffusion decoders, 2024.

[50] Tim Brooks Alexei Efros William Peebles, Ilija Radosavovic and Jitendra Malik. Learning to learn with generative models of neural network checkpoints, 2022.

[51] Han X. Mr-based synthetic ct generation using a deep convolutional neural network method. *Medical Physics*, pages 1408–1419, 2017.

[52] Yang Wen Ju Hu Georgios Evangelidis Sergey Tulyakov Yanzhi Wang Jian Ren Yanyu Li, Geng Yuan. Efficientformer: Vision transformers at mobilenet speed, 2022.

[53] Yue Cao Han Hu Yixuan Wei Zheng Zhang Stephen Lin Baining Guo Ze Liu, Yutong Lin. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[54] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024.

[55] Lituan Wang Zhenwei Zhang Zhenbin Wang, Lei Zhang. Soft masked mamba diffusion model for ct to mri conversion, 2024.