TCAM-Diff: Triplane-Aware Cross-Attention Medical Diffusion Model

Zhenkai Zhang, Krista A. Ehinger, Tom Drummond

School of Computing and Information Systems, The University of Melbourne
zhenkai.zhang@student.unimelb.edu.au, {kris.ehinger, tom.drummond}@unimelb.edu.au

Abstract

We introduce TCAM-Diff, a novel 3D medical image generation model that reduces the memory requirements to encode and generate high-resolution 3D data. This model utilizes a decoder-only autoencoder method to learn triplane representation from dense volume and leverages generalization operations to prevent overfitting. Subsequently, it uses a triplane-aware cross-attention diffusion model to learn and integrate these features effectively. Furthermore, the features generated by the diffusion model can be rapidly transformed into 3D volumes using a pre-trained decoder module. Our experiments on three different scales of medical datasets, BrainTumour $128 \times 128 \times 128$, Pancreas $256 \times 256 \times 256$, and Colon $512 \times 512 \times 512$, demonstrate outstanding results. We utilized MSE and SSIM to assess reconstruction quality and leveraged the Wasserstein Generative Adversarial Network (W-GAN) critic to assess generative quality. Comparisons with existing approaches show that our method gives better reconstruction and generation results than other encoder-decoder methods with similar sized latent spaces.

Introduction

The development of generative models has revolutionized numerous fields, ranging from natural language processing to image synthesis. These models not only enhance data diversity by generating synthetic examples of rare classes, improving classifier robustness, but also facilitate conditional generation. This allows for the creation of data constrained to match specific measurements, offering plausible interpretations and expanding the analytical capabilities across various applications.

Recently, diffusion models (Ho, Jain, and Abbeel 2020) have gained attention for their ability to generate high-quality realistic data by effectively modeling complex distributions. Their flexibility and robustness make them suitable for a wide range of applications. Building on this, latent diffusion models (LDM) (Rombach et al. 2022) extend these capabilities by operating in a lower-dimensional latent space, which reduces computational complexity and broadens their applicability to more diverse and complex tasks, such as image superresolution (Zhao et al. 2023), text-to-image generation (Rombach et al. 2022), 3D object generation (Shue et al. 2023; Ntavelis et al. 2023), medical image synthesis (Kwon, Han, and Kim 2019; Pinaya et al. 2022; Khader et al. 2022b), video generation (Wang et al. 2023), and speech synthesis (Liu, Guo, and Yu 2023).

However, despite substantial progress, existing generative techniques (Segato et al. 2020; Chong and Ho 2021; Pinaya et al. 2022; Khader et al. 2022b) often fail to deal with 3D medical data, such as CT and MRI, particularly when it comes to maintaining high resolution and anatomical accuracy due to the prohibitive memory footprint these entail. (Kwon, Han, and Kim 2019; Pinaya et al. 2022; Khader et al. 2022b) propose methods that utilize autoencoders to compress 3D medical volumes into a latent space, followed by training with latent generative models to reduce the memory footprint required for processing such data. However, due to the inherent limitations of the autoencoder architecture, including loss of critical detail during compression and the dual-step encoding-decoding process, these models still face challenges when dealing with high-resolution medical data.

To address this issue, we propose a novel approach that utilizes a latent diffusion generative model operating on a triplane representation (Chan et al. 2022) to learn complex structures within 3D medical data effectively. Additionally, in the process of learning triplane representations, we utilize the decoder-only architecture, which significantly reduces memory usage, allowing the model to handle higher-resolution medical data more efficiently. This is particularly advantageous because the diffusion model focuses solely on the decoding process. Furthermore, while previous applications of triplane representations have been limited to modeling the surfaces of 3D objects (Chan et al. 2022; Shue et al. 2023), our work demonstrates that this approach can also effectively represent dense volumes, such as medical datasets.

The main contributions of this paper can be summarized.

- We introduce a **decoder-only autoencoder** that decodes latent triplane representations (Chan et al. 2022) into dense 3D volumes using iterative backpropagation. Our approach outperforms baseline models like VAE-GAN (Pinaya et al. 2022) and VQ-GAN (Khader et al. 2022b) in reconstruction quality, with equivalent latent dimensions. Additionally, our model reduces memory usage during training, allowing for the processing of higher resolution images compared to baseline models.
- Our method introduces a novel diffusion model named

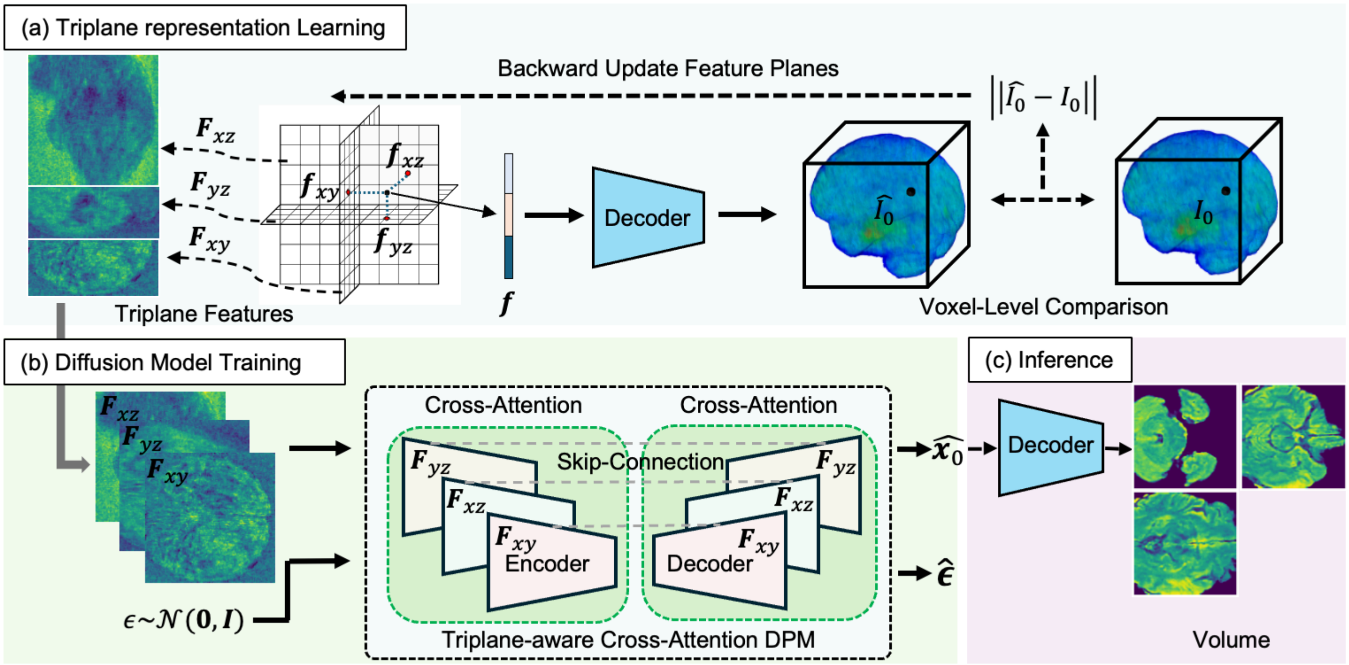

Figure 1: The overview of our two-stage model architecture. (a). Triplane representation learning: triplane features $\{\mathbf{F}_{xy}, \mathbf{F}_{yz}, \mathbf{F}_{xz}\}$ are learned using a decoder-only method with voxel-level comparison. (b). Diffusion model training: triplane features are processed with triplane-aware cross-attention and skip connections in the diffusion model. (c). Inference: the diffusion model's outputs are decoded by the pre-trained decoder to reconstruct 3D volumes.

TCAM-Diff (**T**riplane-aware **C**ross-**A**ttention **M**edical **Diff**usion model) that uses convolutional layers to operate on each of the three orthogonal feature planes independently and specially designed transformer layers to integrate shared information across them.

- To objectively compare our method with others, we present an approach that uses a **W-GAN critic** (Arjovsky, Chintala, and Bottou 2017) for evaluating the relative performance of generative models in novel domains. This approach estimates how closely the models' samples mimic real data, offering a robust metric for assessing generative quality.

Related Work

Diffusion models have achieved state-of-the-art results in non-medical fields, showcasing their advanced capabilities in image generation and data synthesis across various applications (Yang et al. 2023; Phung, Dao, and Tran 2023), especially, the advent of latent diffusion models (Rombach et al. 2022) has furthered this progress by optimizing within the latent space to increase the resolution of generated images. This success has sparked efforts to apply these models within the medical field, aimed at enhancing the diversity of medical data (Khader et al. 2022b; Qin et al. 2023; Pinaya et al. 2022), facilitating the conditional generation of medical data (Qin et al. 2023), and improving the performance of downstream medical tasks (Amirhossein et al. 2022).

Triplane representation (Chan et al. 2022) is a method that encodes 3D data by projecting it onto three orthogonal planes, effectively reducing the complexity of 3D structures into manageable 2D representations. This approach captures the essential features of a 3D object by learning from its projections on the $\{XY, YZ, ZX\}$ planes, enabling efficient spatial information modelling. The benefits of triplane representation include reduced computational complexity, as it simplifies 3D data processing into lower-dimensional spaces, and the ability to preserve essential structural information while minimizing the memory footprint. Consequently, triplane representations (Chan et al. 2022) have been widely employed in facial 3D reconstruction (Chan et al. 2022) and 3D object modelling (Shue et al. 2023), primarily used to capture surface information. However, research on applying triplane representation to dense volumes, such as medical data, is noticeably lacking. This gap likely exists because dense volumes contain significantly more complexity and detailed features compared to the surface representations that triplane-related methods typically address. Our work demonstrates that triplane representations can also effectively represent dense volumes.

Medical image data, such as CT and MRI, presents greater challenges compared to two-dimensional image data, due to its volumetric attributes. The intuitive approaches are modelling the medical data directly with GAN (Kwon, Han, and Kim 2019; Hong et al. 2021; Özbey et al. 2020; Cirillo, Abramian, and Eklund 2021), or diffusion models (Dorjsembe, Odonchimed, and Xiao 2022). However, these generative models in the medical field are often limited by memory constraints, and struggle to handle large-scale or high-resolution medical datasets. Recently, several advanced models (Pinaya et al. 2022; Khader et al. 2022b) have worked on utilizing the 3D autoencoder models to extract the latent dimensions from medical data, as the input of the 3D latent diffusion models, to generate the high-resolution images. Despite the advancements, such models still suffer the limitations of 3D autoencoder models, such as high computational load, which takes longer training times and higher demands on hardware, memory intensity, complexity in training, etc. In contrast, our approach utilizes a decoder-only architecture which effectively reduces memory usage, and computational load during the training and inference. This method not only simplifies the model structure but also enhances efficiency, particularly when handling large-scale or complex datasets.

Methodology

Our model, shown in Figure 1, consists of two stages. First, a decoder-only autoencoder compresses 3D medical images into a triplane representation (Chan et al. 2022) using iterative backpropagation for encoding. This approach is advantageous because, for generative models, decoding speed is critical, while slower encoding is acceptable during training. In the second stage, a triplane-aware cross-attention 2D diffusion model learns from the triplane features, effectively integrating them into the diffusion process.

Triplane Representation Learning

The triplane representation optimally integrates implicit and explicit modelling advantages by maintaining three 2D feature planes in memory, rather than an entire voxel grid. This configuration enhances memory efficiency and accelerates the prediction process (Chan et al. 2022). Here, our model uses the three orthogonal 2D feature planes $\mathbf{F}_{xy,\theta_{xy}} \in \mathbb{R}^{C\times H\times W}$, $\mathbf{F}_{yz,\theta_{yz}} \in \mathbb{R}^{C\times W\times D}$, $\mathbf{F}_{xz,\theta_{xz}} \in \mathbb{R}^{C\times H\times D}$ to represent a 3D dense medical volume, where the (H, W, D) is the shape size of the 3D medical volume, C is the feature channels for each feature plane, and $\theta_{xy}, \theta_{yz}, \theta_{xz}$ mean these three planes are parametric planes that are learned during the training process, and a compact decoder module, typically an MLPs (Multilayer Perceptron), is used to convert features sampled from these planes into corresponding spatial intensities.

We predefined multiple parametric triplane representations,

$$\left\{\left(\mathbf{F}_{xy}^{(0)}, \mathbf{F}_{yz}^{(0)}, \mathbf{F}_{xz}^{(0)}\right), \cdots, \left(\mathbf{F}_{xy}^{(i)}, \mathbf{F}_{yz}^{(i)}, \mathbf{F}_{xz}^{(i)}\right), \cdots\right\}$$

based on the number of objects in the training set. Given a three-dimensional point $\mathbf{x}^{(i)} = (x, y, z)$, this point is projected onto each axis-aligned plane of the i-th predefined triplane, the corresponding feature values are queried as $\mathbf{f}_{xy}^{(i)} = \mathbf{F}_{xy}^{(i)}(x, y),\ \mathbf{f}_{yz}^{(i)} = \mathbf{F}_{yz}^{(i)}(y, z),\ \mathbf{f}_{xz}^{(i)} = \mathbf{F}_{xz}^{(i)}(x, z)$ and concatenated as $\mathbf{f}(\mathbf{x}^{(i)}) = \text{concat}\left(\mathbf{f}_{xy}^{(i)}, \mathbf{f}_{yz}^{(i)}, \mathbf{f}_{xz}^{(i)}\right)$, and then decoded through an MLP (multi-layer perceptron) network with parameters θ, to estimate the intensity value $\widehat{\mathbf{I}}(\mathbf{x}^{(i)})$ at point $\mathbf{x}^{(i)}$ of the i-th object.

$$\widehat{\mathbf{I}}(\mathbf{x}^{(i)}) = \text{MLP}_\theta(\mathbf{f}(\mathbf{x}^{(i)})) \quad (1)$$

During the training process of the triplane decoder-only model, we optimize both the MLP network and the parameters of the triplane feature planes.

In our experiments, all objects in the same dataset share the same MLP network. Since all objects within the same medical dataset contain similar types of data and features, sharing the same MLP network among these objects leverages this similarity to enhance learning and generalization. This commonality ensures that the network can efficiently learn a generalized model capable of handling variations across similar instances without overfitting specific features of individual objects, and ensure that the outputted triplane features are consistent and of high quality across the entire dataset.

Loss Functions and Regularization Approaches

The foundational training objective for our model is the SmoothL1Loss (Girshick 2015). This loss function, akin to the Mean Squared Error (MSE) Loss but less sensitive to outliers, calculates the discrepancy between the estimated intensity value $\widehat{\mathbf{I}}(\mathbf{x}^{(i)})$, and ground truth intensity value $\mathbf{I}(\mathbf{x}^{(i)})$. The formulation of the loss function is expressed as follows:

$$\mathcal{L}_{basic} = \sum_{i}^{N} \sum_{j}^{M} Smooth_{L1}\left(\widehat{\mathbf{I}}\left(\mathbf{x}_j^{(i)}\right) - \mathbf{I}\left(\mathbf{x}_j^{(i)}\right)\right) \quad (2)$$

in which

$$Smooth_{L1}(x) = \begin{cases} \frac{x^2}{2\beta} & if\ |x| < \beta \\ |x| - \frac{1}{2\beta} & \text{otherwise} \end{cases} \quad (3)$$

where N is the total object number, M is the total points sampled from an object, i represents the i-th object, j indicates the point within a single object, and in our experiments, the hyper-parameter β was set to 0.3.

While this approach is effective in reducing the average error across pixel values, it often overlooks aspects critical to human perception, such as texture, contrast, and structural integrity. Through our experiments, we found that images optimized only for SmoothL1Loss may appear blurry or lack fine details and visual fidelity perceptible to the human eye.

To address these shortcomings, we augment our loss function framework with Perceptual Loss (Johnson, Alahi, and Fei-Fei 2016), which contributes significantly to enhancing the quality of the reconstructed images beyond mere pixel accuracy. In addition, it also can increase generalization by focusing on high-level features rather than low-level pixel details.

$$\mathcal{L}_{\text{feat}}(x, y) = \sum_{l=1}^{L} \frac{1}{N_l} \|\phi_l(x) - \phi_l(y)\|^2 \quad (4)$$

where $\phi_l(x), \phi_l(y)$ are the activations of the l-th layer of a pre-trained neural network (using the pre-trained 'squeeze' network in our experiments), N_l is the number of elements in the l-th layer, and L is the total number of layers used for the loss calculation.

In addition to employing loss functions aimed at enhancing the performance of image reconstruction, (Shue et al.

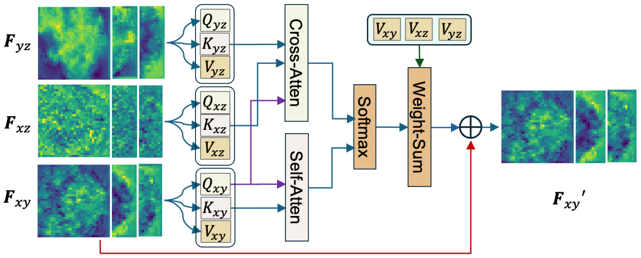

Figure 2: The attention block for $\mathbf{F}_{xy}$, incorporating both self-attention and cross-attention mechanisms. The same structure is applied to the feature planes $\mathbf{F}_{yz}$ and $\mathbf{F}_{xz}$. The structure ensures the effective integration of information across different planes.

2023) propose using the total variation (TV) regularization terms can simplify the data manifold by eliminating unnecessary high-frequency information from feature planes, aligning their distribution with natural images.

$$\mathcal{L}_{TV} = \mathbf{TV}(\mathbf{F}_{xy}^{(i)}) + \mathbf{TV}(\mathbf{F}_{yz}^{(i)}) + \mathbf{TV}(\mathbf{F}_{xz}^{(i)}) \qquad (5)$$

We employ root mean square normalization (RMSNorm) (Zhang and Sennrich 2019) to ensure each feature plane maintains unit energy during training, providing consistent scaling across planes. Compared to direct scaling to fixed ranges like $[0, 1]$ or $[-1, 1]$, RMSNorm (Zhang and Sennrich 2019) preserves relative feature dynamics without imposing rigid bounds, preventing bias from varying input magnitudes and promoting stable, uniform learning in the diffusion model.

$$\bar{a}_i = \frac{a_i}{\mathbf{RMS}(a)}, \ \mathbf{RMS}(a) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^2} \qquad (6)$$

To improve MLP generalization, we propose MLP Noise Regularization. This involves adding controlled noise to features during training. Since features are normalized to maintain unit energy, noise is carefully managed using the signal-to-noise ratio (SNR).

Then, the normalized and generalized features become,

$$\mathbf{f}(\mathbf{x})' = \text{RMSNorm}(\mathbf{F})(\mathbf{x}) + C * \epsilon \qquad (7)$$

where $C = 0.32$, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ can achieve better generalization in our experiments.

Our loss functions for training the triplane decoder-only model are described as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{basic} + \lambda_1 \mathcal{L}_{\text{feat}} + \lambda_2 \mathcal{L}_{TV} \qquad (8)$$

in which, λ_1 =1e$-$2, λ_2 =1e$-$3 in our experiments.

Triplane-aware Cross-Attention Diffusion Model

Upon acquiring meaningful triplane features, the subsequent phase involves leveraging a diffusion model to learn from these data representations. The conventional method (Shue et al. 2023) concatenates three orthogonal 2D feature planes $(\mathbf{F}_{xy}, \mathbf{F}_{yz}, \mathbf{F}_{xz})$ as the input of the diffusion model. While this concatenation technique simplifies data handling and reduces computational requirements, it introduces significant drawbacks. It does not preserve the intrinsic 3D spatial relationships inherent in the data, as the model processes the concatenated planes without recognizing their mutual dependencies. This limitation often leads to diminished accuracy in the reconstructed 3D structures and impairs the model's ability to comprehend the complexity of 3D spatial dynamics fully.

In Figure 1 (b), we present a novel approach where each of the three mutually orthogonal feature planes is trained independently within the diffusion model. Although these planes are segmented, they originate from the same object and therefore share intrinsic information. To effectively leverage this shared data, we incorporate a cross-attention layer that facilitates bidirectional information flow between the planes. The structure of our newly designed attention blocks is depicted in Figure 2. This innovative strategy allows for the deep learning of each plane's unique characteristics while simultaneously integrating knowledge across all planes, thereby significantly enhancing the model's overall performance.

Our diffusion model is based on simultaneously estimating both the image and the noise (Zhang, Ehinger, and Drummond 2023), the training objectives are:

$$\min_{\theta} \mathbb{E}\left[\|\mathbf{R}_\theta(\mathbf{x}_t, t) - \mathbf{x}_0\| + \|\epsilon_\theta(\mathbf{x}_t, t) - \epsilon\|\right] \qquad (9)$$

where $\widehat{\mathbf{x}_0} = \mathbf{R}_\theta(\mathbf{x}_t, t)$, $\widehat{\epsilon} = \epsilon_\theta(\mathbf{x}_t, t)$, and $\mathbf{x}_0$ is the triplane representations with $\{\mathbf{F}_{xy}, \mathbf{F}_{yz}, \mathbf{F}_{xz}\}$.

Experiments Setup

Dataset

We conduct the experiments on three publicly available datasets with varying scales: the BraTS dataset includes 750 4D MRI volumes (484 training, 266 testing) with dimensions of $(240 \times 240 \times 155)$, covering four modalities: Fluid-Attenuated Inversion Recovery (FLAIR), T1-weighted (T1w), T1 with gadolinium contrast (T1gd), and T2-weighted (T2w) (Menze et al. 2014b; Simpson et al. 2019). The Pancreas Tumour dataset contains 420 3D CT volumes (282 training set, and 139 testing set) with varying dimensions (Simpson et al. 2019), The Colon Cancer dataset includes 190 3D CT volumes with 126 training and 64 testings, with diverse shapes of each scan (Simpson et al. 2019).

Preprocessing

Given that current 3D medical generative models (Pinaya et al. 2022; Khader et al. 2022b; Liu et al. 2023; Friedrich et al. 2024) are trained on single-channel datasets, we processed three datasets following VAE-GAN (Pinaya et al. 2022) preprocessing methods to extract single-channel data for training. Using the MONAI toolkit (The MONAI Consortium 2020), we further preprocessed the datasets to validate the model's capabilities at different scales: the BraTS dataset was center-cropped to $128 \times 128 \times 128$, the Pancreas dataset was resized to $256 \times 256 \times 256$, and the Colon dataset to $512 \times 512 \times 512$. Intensity values were scaled to the $[0, 1]$ range across all datasets (Simpson et al. 2019).

Additionally, utilizing current mainstream 3D medical segmentation models, which are trained on multi-channel data, to verify that our model's reconstructed results do not lose crucial information, we have also used the 4-channel multi-modal BraTS dataset (Menze et al. 2014b; Simpson et al. 2019) as the fourth dataset in our experiments. This dataset contains three segmented labels: GD-enhancing Tumor (ET — Label 4), Peritumoral Edema (ED — Label 2), and Necrotic and Non-Enhancing Tumor Core (NCR/NET — Label 1) (Menze et al. 2014a; Bakas et al. 2018).

Baseline Models

Our experiment utilized two baseline models based on architectures that integrate autoencoder with the latent diffusion model (LDM). One is a VAE-GAN combined with a 3D LDM (Rombach et al. 2021; Pinaya et al. 2022), which is reproduced by MONAI (The MONAI Consortium 2020) based on the content of the paper, and the other is a VQ-GAN coupled with a 3D LDM (medical diffusion) (Khader et al. 2022a). We conduct comparative analyses on the quality and performance of autoencoder and generative models separately. Since the baseline model hasn't released its pre-trained models, the results were obtained by following the steps in the papers and retraining with the publicly available code.

Evaluation Metrics

Autoencoder reconstruction quality was evaluated using two metrics: MSE for pixel-level assessment and 3D SSIM (Wang et al. 2004) for structural-level evaluation. We also used SegResNet (Myronenko 2019) to compare the performance of our model's reconstructions with baseline models in downstream 3D medical segmentation tasks.

To verify the performance of the generative models, we utilized critic from W-GAN (Arjovsky, Chintala, and Bottou 2017), to assess the distance between the real data distribution and the generated distribution. It measures how "far" the generated data is from the real data in terms of distribution, instead of just classifying whether data is real or fake in traditional GAN (Goodfellow et al. 2020). W-GAN critic (Arjovsky, Chintala, and Bottou 2017) assigns a score that represents the "earth mover's distance" (also known as the Wasserstein distance) between the distribution of the generated data and the distribution of the real data. $W(\mathbb{P}_r, \mathbb{P}_\theta) = \sup_{\|f\|_L \leq 1} (\mathbb{E}_{x\sim\mathbb{P}_r}[f(x)] - \mathbb{E}_{x\sim\mathbb{P}_\theta}[f(x)])$ where $\mathbb{P}_r$ is real data distribution and $\mathbb{P}_\theta$ from generative model. In practice, the lower Wasserstein distance means the closer the generated data's distribution to the real distribution, which indicates the model's ability to generate high-fidelity outputs.

(Pinaya et al. 2022) used **Fréchet Inception Distance (FID)** (Heusel et al. 2017) to measure the performance of the generated model. However, they mentioned utilizing a pre-trained Med3D model (Chen, Ma, and Zheng 2019) for feature extraction from 3D medical data but did not make the associated code publicly available. Moreover, there are issues with loading the provided pre-trained models, which will be questionable whether the extracted features are meaningful. Additionally, useful FID value must be computed in a large enough dataset. Due to the size of our training dataset, obtaining convincing results is challenging, leading us to decide against using FID as an evaluation metric.

Implementation Details

Our experiments were conducted on A100 GPUs, each with 80GB of GPU RAM. We trained the baseline models on 4 GPUs with default settings. Our decoder-only model was trained using only 1 GPU, and we utilized 4 GPUs to train the generative model. The hyperparameters in our experiments included an Adam optimizer with a $3\mathrm{e}{-5}$ learning rate for the BraTS dataset and a $1\mathrm{e}{-5}$ learning rate for higher-resolution datasets. The loss function weights were set to $\lambda_1 = 1\mathrm{e}{-2}$ and $\lambda_2 = 1\mathrm{e}{-3}$ for BraTS, and $\lambda_1 = 0$ for higher-resolution datasets. To ensure stability, we applied gradient clipping with a max norm of 1.0. Additional hyperparameters, architectures, and experimental details are provided in the appendix.

Experimental Results

In the following, VAE-GAN and VQ-GAN refer to the autoencoder process, while VAE-GAN-LDM (Pinaya et al. 2022) and VQ-GAN-LDM (Khader et al. 2022b) refer to the complete models, including both the autoencoder and generative components.

Autoencoder Models

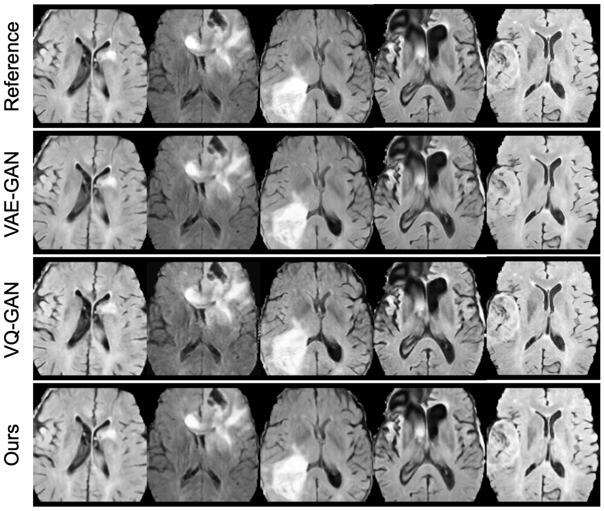

Figure 3: Qualitative comparison of reconstruction quality between autoencoder models: VAE-GAN (Pinaya et al. 2022), VQ-GAN (Khader et al. 2022b) and our model.

We compare our decoder-only autoencoder model's performance with VAE-GAN (Pinaya et al. 2022) and VQ-GAN (Khader et al. 2022b). Figure 3 shows our model achieves remarkable reconstruction results, closely approximating the ground truth. Table 1 indicates lower MSE and 3D SSIM values, with faster training and less computational resource use, highlighting our model's efficiency and effectiveness in preserving image details. We gener-

Method	MSE ↓	3D SSIM ↑	Latent Dimension	Parameters	GPUs	Time
VAE-GAN	0.0015	0.8720	2 × (8, 32, 32, 32)	524,288	4	40h
VAE-GAN*	0.0014	0.8818	2 × (8, 32, 32, 32)	524,288	4	38.5h
VQ-GAN	0.0079	0.8393	(16, 32, 32, 32)	524,288	4	37h
VQ-GAN*	0.0045	0.8785	(16, 32, 32, 32)	524,288	4	36h
Ours	**0.0008**	**0.9311**	3 × (42, 64, 64)	516,096	1	8h

Table 1: Quantitative comparison of reconstruction quality between autoencoder models: VAE-GAN (Pinaya et al. 2022), VQ-GAN (Khader et al. 2022b) and our model. VAE-GAN* (Pinaya et al. 2022) and VQ-GAN* (Khader et al. 2022b) impose the same constraint [0,1] on the output as our model. The number of latent dimensions has been approximately matched across models to ensure fairness.

Method	Labels	Precision ↑	Recall ↑	Specificity ↑	Accuracy ↑	F1 Score ↑
VQ-GAN	ED	0.8140±2e-3	0.7498±3e-2	0.9950±1e-3	0.9880±1e-3	0.7815±1e-2
	NCR/NET	0.6545±3e-2	0.6849±3e-2	0.9972±2e-3	0.9948±1e-3	0.6682±2e-2
	ET	0.7127±3e-2	0.7557±2e-2	0.9968±1e-3	0.9944±1e-3	0.7333±2e-2
Ours	ED	**0.9430**±3e-3	**0.9540**±2e-3	**0.9983**±2e-5	**0.9970**±1e-4	**0.9500**±2e-3
	NCR/NET	**0.9250**±1e-2	**0.9321**±1e-2	**0.9993**±1e-4	**0.9988**±1e-4	**0.9273**±1e-2
	ET	**0.9183**±1e-2	**0.9455**±5e-3	**0.9992**±1e-4	**0.9987**±1e-4	**0.9350**±2e-3

Table 2: Downstream segmentation task: quantitative performance comparison in 3D medical segmentation using the BraTS dataset, compared to the VQ-GAN model (Khader et al. 2022b). This table presents Precision, Recall, Specificity, Accuracy, and F1 Score for each segmented label, based on four repeated experiments. Training the VAE-GAN model (Pinaya et al. 2022) on this dataset encounters out-of-memory issues.

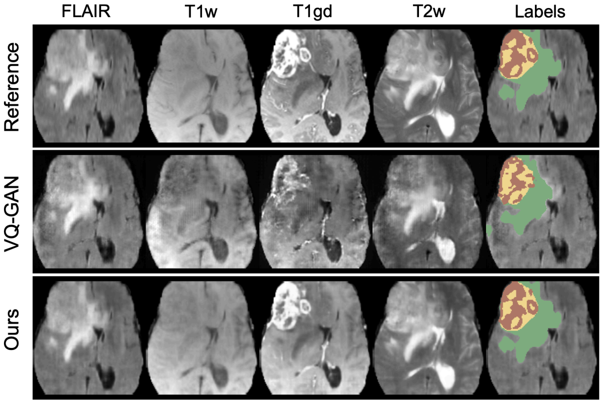

Figure 4: Downstream segmentation task: evaluation of reconstructed image performance in multi-modal 3D segmentation, comparing our model to the VQ-GAN model (Khader et al. 2022b). ET label is the brown area, the ED label is the green area, and NCR/NET label is the yellow area (Menze et al. 2014a; Bakas et al. 2018).

ate multi-modal data using the same triplane representations to ensure our decoder-only autoencoder retains essential medical information. Due to out-of-memory errors, VAE-GAN (Pinaya et al. 2022) couldn't be trained on the 4-channel BraTS dataset (Menze et al. 2014b; Simpson et al. 2019), so it was excluded from segmentation comparisons. Figure 4 and Table 2 show that our model preserves critical information for medical segmentation and scales more efficiently from single-channel to multi-channel configurations compared to VQ-GAN (Khader et al. 2022b) while retaining crucial information.

Generative Models

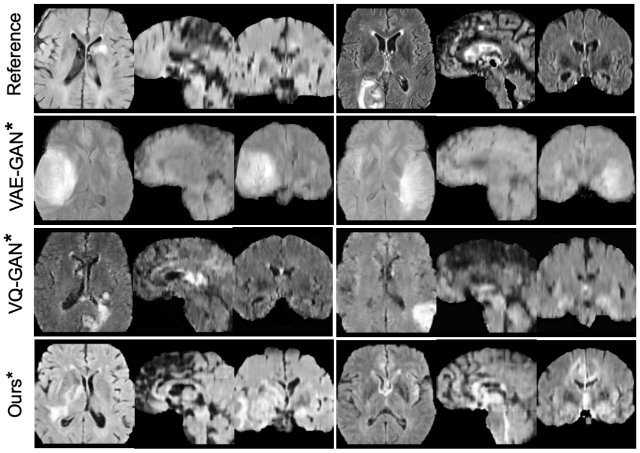

Figure 5: Qualitative comparison of generative quality between VAE-GAN-LDM (VAE-GAN*) (Pinaya et al. 2022), VQ-GAN-LDM (VQ-GAN*) (Khader et al. 2022b), and our triplane-aware diffusion model in BraTS dataset.

We evaluate the performance of the generative model, between the VAE-GAN-LDM (Pinaya et al. 2022), the VQ-GAN-LDM (Khader et al. 2022b) and our triplane-aware

Method	Wasserstein distance ↓
VAE-GAN-LDM	120±20
Ours	**36**±4
VQ-GAN-LDM	160±20
Ours	**16**±5

Table 3: Quantitative comparison of generative quality measured by Wasserstein distance using a W-GAN critic (Arjovsky, Chintala, and Bottou 2017) for two baseline models and our generative model. Experiments were conducted three times using different random seeds to ensure reliability.

cross-attention diffusion model. Figure 5 demonstrates that the volumes generated by our model possess more detail and are closer to real images, compared to the results from the VAE-GAN-LDM (Pinaya et al. 2022) and VQ-GAN-LDM (Khader et al. 2022b). From Table 3, our model achieved a lower Wasserstein distance in the experiments, compared to these baseline models. This demonstrates that the generated data distribution from our model is closer to the real data distribution, surpassing these two baseline models' ability to create realistic samples. See more results in the appendix.

Performance in High-Resolution Data

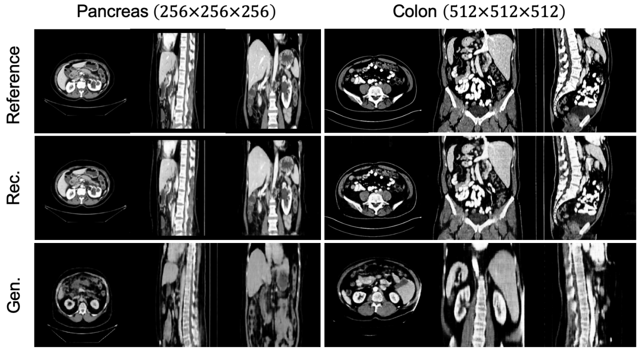

Figure 6: Reconstruction quality and generative quality of our model in pancreas dataset (Simpson et al. 2019) with $256 \times 256 \times 256$ resolution, and colon dataset (Simpson et al. 2019) with $512 \times 512 \times 512$ resolution. $Rec.$ means the reconstruction output. $Gen.$ means the generative output.

When handling the high-resolution dataset, VAE-GAN-LDM(Pinaya et al. 2022) and VQ-GAN-LDM (Khader et al. 2022b) suffer the out-of-memory problem due to significant increases in the number of parameters, more gradient storage, etc. In contrast, our model uses the decoder-only autoencoder model, which can significantly reduce the memory usage during training and inference to allow handling higher-resolution datasets than VAE-GAN (Pinaya et al. 2022) or VQ-GAN (Khader et al. 2022b). Figure 6 present the reconstruction and generation quality of our model when handling high-resolution datasets. See more results in the appendix.

Ablation Experiments

Autoencoder Ablation. We decreased the channel size for the triplane representation to utilize fewer parameters in the latent dimension to achieve similar results compared to the VAE-GAN diffusion model (Pinaya et al. 2022) in terms of the performance for the autoencoder model. From Table 4, our model achieves comparable reconstruction quality to the VAE-GAN (Pinaya et al. 2022) while utilizing only its 48% of the latent dimension parameters.

Method	Params	Ratio	MSE ↓	3D SSIM ↑
VAE-GAN*	524,288	1.0	**0.0014**	**0.8818**
Ours	221,184	0.41	0.0015	0.8914
	245,760	**0.48**	**0.0014**	**0.8985**
	294,912	0.56	0.0012	0.9117
	516,096	0.98	0.0008	0.9311

Table 4: Performance changes with decreasing channel size. Ratio refers to the latent dimensions of our models compared to the VAE-GAN model (Pinaya et al. 2022).

Diffusion Ablation. In Figure 7, we conducted a comparison of diffusion models with and without the integration of cross-attention, revealing that cross-attention significantly improves the extraction of information from 3D volumes. Additionally, it enhances the consistency of the sampled triplanes, thereby contributing to more reliable and accurate reconstructions in our models.

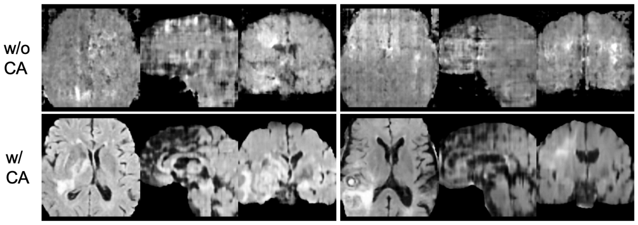

Figure 7: Performance between the diffusion model with (second row) and without (first row) cross-attention layers.

Conclusion

In this paper, we introduce a novel and effective diffusion model called TCAM-Diff, triplane-aware cross-attention mechanisms for generating 3D medical datasets. We demonstrate how TCAM-Diff surpasses existing encoder-decoder methods by delivering superior reconstruction and generation quality with similar-sized latent spaces. Its decoder-only design enables efficient handling of high-resolution datasets without encountering memory issues. Extensive qualitative and quantitative experiments validate the feasibility and effectiveness of TCAM-Diff in generating high-quality 3D medical data.

Acknowledgments

This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

References

Amirhossein, K.; Khodapanah, A. E.; Moein, H.; Reza, A.; Mohsen, F.; Ilker, H.; and Dorit, M. 2022. Diffusion models for medical image analysis: a comprehensive survey. *arXiv preprint arXiv*, 2211.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *ArXiv*, abs/1701.07875.

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; et al. 2018. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. *arXiv preprint arXiv:1811.02629*.

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16123–16133.

Chen, S.; Ma, K.; and Zheng, Y. 2019. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*.

Chong, C. K.; and Ho, E. T. W. 2021. Synthesis of 3D MRI brain images with shape and texture generative adversarial deep neural networks. *IEEE Access*, 9: 64747–64760.

Cirillo, M. D.; Abramian, D.; and Eklund, A. 2021. Vox2Vox: 3D-GAN for brain tumour segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 6th International Workshop, BrainLes 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers, Part I 6*, 274–284. Springer.

Dorjsembe, Z.; Odonchimed, S.; and Xiao, F. 2022. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical Imaging with Deep Learning*.

Friedrich, P.; Wolleb, J.; Bieder, F.; Durrer, A.; and Cattin, P. C. 2024. WDM: 3D Wavelet Diffusion Models for High-Resolution Medical Image Synthesis. *arXiv preprint arXiv:2402.19043*.

Girshick, R. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 1440–1448.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, S.; Marinescu, R.; Dalca, A. V.; Bonkhoff, A. K.; Bretzner, M.; Rost, N. S.; and Golland, P. 2021. 3D-StyleGAN: A style-based generative adversarial network for generative modeling of three-dimensional medical images. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections: First Workshop, DGM4MICCAI 2021, and First Workshop, DALI 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, October 1, 2021, Proceedings 1*, 24–34. Springer.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711. Springer.

Khader, F.; Mueller-Franzes, G.; Arasteh, S. T.; Han, T.; Haarburger, C.; Schulze-Hagen, M.; Schad, P.; Engelhardt, S.; Baessler, B.; Foersch, S.; Stegmaier, J.; Kuhl, C.; Nebelung, S.; Kather, J. N.; and Truhn, D. 2022a. Medical Diffusion - Denoising Diffusion Probabilistic Models for 3D Medical Image Generation.

Khader, F.; Mueller-Franzes, G.; Arasteh, S. T.; Han, T.; Haarburger, C.; Schulze-Hagen, M. F.; Schad, P.; Engelhardt, S.; Baessler, B.; Foersch, S.; Stegmaier, J.; Kuhl, C.; Nebelung, S.; Kather, J. N.; and Truhn, D. 2022b. Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation.

Kwon, G.; Han, C.; and Kim, D.-s. 2019. Generation of 3D brain MRI using auto-encoding generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 118–126. Springer.

Liu, Y.; Dwivedi, G.; Boussaid, F.; Sanfilippo, F.; Yamada, M.; and Bennamoun, M. 2023. Inflating 2D convolution weights for efficient generation of 3D medical images. *Computer Methods and Programs in Biomedicine*, 240: 107685.

Liu, Z.; Guo, Y.; and Yu, K. 2023. Diffvoice: Text-to-speech with latent diffusion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; Burren, Y.; Porz, N.; Slotboom, J.; Wiest, R.; et al. 2014a. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE transactions on medical imaging*, 34(10): 1993–2024.

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; Farahani, K.; Kirby, J.; et al. 2014b. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10): 1993–2024.

Myronenko, A. 2019. 3D MRI brain tumor segmentation using autoencoder regularization. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II 4*, 311–320. Springer.

Ntavelis, E.; Siarohin, A.; Olszewski, K.; Wang, C.; Gool, L. V.; and Tulyakov, S. 2023. Autodecoding latent 3d diffusion models. *Advances in Neural Information Processing Systems*, 36: 67021–67047.

Özbey, M.; Yurt, M.; Dar, S. U. H.; and Çukur, T. 2020. Three dimensional mr image synthesis with progressive generative adversarial networks. *arXiv preprint arXiv:2101.05218*.

Phung, H.; Dao, Q.; and Tran, A. 2023. Wavelet diffusion models are fast and scalable image generators. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10199–10208.

Pinaya, W. H. L.; Tudosiu, P.-D.; Dafflon, J.; Costa, P. F. D.; Fernandez, V.; Nachev, P.; Ourselin, S.; and Cardoso, M. J. 2022. Brain Imaging Generation with Latent Diffusion Models. *ArXiv*, abs/2209.07162.

Qin, Y.; Zheng, H.; Yao, J.; Zhou, M.; and Zhang, Y. 2023. Class-balancing diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18434–18443.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10674–10685.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Segato, A.; Corbetta, V.; Di Marzo, M.; Pozzi, L.; and De Momi, E. 2020. Data augmentation of 3D brain environment using deep convolutional refined auto-encoding alpha GAN. *IEEE Transactions on Medical Robotics and Bionics*, 3(1): 269–272.

Shue, J. R.; Chan, E. R.; Po, R.; Ankner, Z.; Wu, J.; and Wetzstein, G. 2023. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20875–20886.

Simpson, A. L.; Antonelli, M.; Bakas, S.; Bilello, M.; Farahani, K.; van Ginneken, B.; Kopp-Schneider, A.; Landman, B. A.; Litjens, G. J. S.; Menze, B. H.; Ronneberger, O.; Summers, R. M.; Bilic, P.; Christ, P. F.; Do, R. K. G.; Gollub, M. J.; Golia-Pernicka, J.; Heckers, S.; Jarnagin, W. R.; McHugo, M.; Napel, S.; Vorontsov, E.; Maier-Hein, L.; and Cardoso, M. J. 2019. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *ArXiv*, abs/1902.09063.

The MONAI Consortium. 2020. Project MONAI. Zenodo.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13: 600–612.

Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; and Yang, M.-H. 2023. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 56(4): 1–39.

Zhang, B.; and Sennrich, R. 2019. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32.

Zhang, Z.; Ehinger, K. A.; and Drummond, T. 2023. Improving Denoising Diffusion Models via Simultaneous Estimation of Image and Noise. In *Asian Conference on Machine Learning*.

Zhao, K.; Hung, A. L. Y.; Pang, K.; Zheng, H.; and Sung, K. 2023. PartDiff: image super-resolution with partial diffusion models. *arXiv preprint arXiv:2307.11926*.