Mediffuse - Diffusion-Driven Translation for MRI Generation from CT Scans*

Saurav Dosi snd220002
Animesh Maheshwari axm230045
Pratiksha Aigal ppa220001
Varad Abhyankar vva230000

Abstract

Mediffuse is a diffusion based approach to generate MRI-like images from CT scans. CT scans, while fast and cost-effective, provide limited soft tissue detail, whereas MRI scans offer rich soft-tissue detail but are slow and expensive. We aim to combine the speed and availability of CT with the superior soft-tissue visualization of MRI. Our method uses Stable Diffusion as backbone, with ControlNet and InstructPix2Pix models trained on a paired CT-MRI dataset. Experimental results show promising initial results with scope for improvements in the future.

1. Introduction

Medical imaging is critical for accurate disease diagnosis and treatment. Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) are two of the most widely used tools in medical imaging, each having it's own advantages and disadvantages. CT scans use X-rays to capture detailed cross-sectional images of the body. A rotating X-ray beam and detectors generate multiple images from different angles, which are then processed by a computer to create a 3D representation of internal structures. MRI scans use strong magnetic fields and radio waves to create detailed images of organs and tissues. The magnetic field aligns hydrogen atoms in the body, and radio frequency pulses disrupt this alignment, emitting signals that are captured to form an image. CT scans are faster, widely available, and more cost-effective. However, CT scans rely on X-rays and thus offer poor soft-tissue contrast. In contrast, MRI provides superior soft-tissue visualization due to its use of magnetic resonance properties, enabling better imaging of brain structures, muscles, and other soft organs. The downside is that MRI is expensive and slower to acquire. This trade-off between CT and MRI brings forward the need to combine the speed of CT with the diagnostic power of MRI and bridge the gap between them while preserving key anatomical details.

In recent years, deep learning has enabled image-to-image transformations and diffusion models have shown remarkable success in generative tasks and image modification. This raises an interesting possibility of transforming a CT scan into an image that resembles an MRI, thus providing better diagnostic information without the time and cost of an actual MRI procedure. In this report, we discuss our approach for CT-to-MRI transformation which harnesses the capabilities of latent diffusion with a ControlNet [1] based model and an InstructPix2Pix [2] based model. We trained these models on a paired CT/MRI dataset of brain scans to generate MRI-like images and evaluated the results using metrics like SSIM and PSNR.

*GitHub Demo Video Slides

2. Related Work

Lyu et al. [3] proposed the use of denoising diffusion probabilistic models and score-matching techniques for converting MRI to CT images. Their experiments demonstrated that these models outperform CNN and GAN baselines in synthetic CT generation and use Monte Carlo sampling for uncertainty investigation. The study showed high-quality outputs, and competitiveness of diffusion-based methods for image synthesis.

Jha et al. [4] employed CycleGAN to translate CT scans into MRI-like images using unpaired data, aiming to reduce radiation exposure and imaging costs. The approach achieved strong results (MAE 0.5309, PSNR 52.344), demonstrating the effectiveness of CycleGANs in cross-modality translation without paired datasets.

Taqi et al. [5] presented a detailed review of deep learning methods for synthesizing MRIs from CT scans. They found that supervised learning models generate better results but are limited by the availability of paired datasets, while unsupervised models can mitigate this by generating synthetic data for supervised learning.

Pan et al. [6] introduced a 3D transformer-based denoising diffusion model (MC-IDDPM) for generating synthetic CT from MRI to support radiation therapy planning. Their method integrated Swin-VNet into the reverse diffusion process and achieved promising results across multiple datasets and evaluation metrics.

Nikbakhsh et al. [7] enhanced CycleGAN-based synthetic MRI generation by incorporating feature extraction mechanisms to improve mono-modal registration for radiotherapy. Using unpaired CT data, their method outperforms existing approaches in generating high-quality MRI images and demonstrates superior performance across evaluation metrics.

Liu et al. [8] presented a U-Net-based network for transforming brain CT images into MRI images, demonstrating its effectiveness on the SynthRAD2023 Grand Challenge dataset. Their approach incorporates tailored preprocessing and pretraining strategies to enhance the performance.

Wang et al. [9] proposed Diffusion Mamba (DiffMa), a novel CT-to-MRI image generation model that replaces traditional U-Net or Transformer backbones with a state-space model (Mamba) tailored for latent diffusion. By introducing soft-masked cross-sequence attention and spiral scanning, DiffMa better captures spatial continuity and token importance. Experimental results showed DiffMa achieves superior performance and scaling efficiency compared to existing benchmarks in medical image synthesis.

Tapp et al. [10] propose a novel cross-modality approach using 3D latent diffusion models to synthesize CT volumes from MRI scans. Their method leverages a shared MRI-CT latent space and enables both anatomically consistent and variably perturbed synthetic CT generation without retraining. Evaluations on public and private datasets demonstrate promising results, including a Fréchet Inception Distance of 3.28, perceptual similarity of 0.28 and 72% Dice similarity for skull segmentation, showcasing the potential of lightweight 3D cross-modal synthesis in medical imaging.

Bird et al. [11] validated a deep learning-based CycleGAN approach for generating dosimetrically accurate synthetic CT (sCT) images from MRI data across pelvic, brain, and head & neck cancer cases. Their model accommodates data from multiple MRI scanners and sequences and shows excellent results. It demonstrated the clinical feasibility of MRI-only radiotherapy planning using robust and flexible deep learning sCT generation.

Alkayyali et al. [12] provided a broad overview of recent AI advancements in medical imaging, emphasizing the impact of CNNs and deep learning on disease detection, segmentation, and diagnostic support across modalities like X-rays, MRIs, and CT scans. The paper also addressed key challenges including data limitations, model explainability, and ethical concerns.

3. Methodology

3.1. Data Preprocessing

We used the CT-MRI scans from the SynthRad2023 Challenge dataset [13] for our experiments. Each paired CT-MRI scan was processed to extract the central axial slice. The extracted slices were normalized to a fixed intensity range. For the ControlNet pipeline, Canny edge detection was applied on CT images to generate structural contour maps, which were then used as an additional conditioning input during training.

3.2. Model Architecture

We trained a ControlNet to generate MRI scans using spatial conditioning to guide the diffusion process, preserving anatomical structures. CT contours were used as a guidance to improve spatial fidelity during CT-to-MRI translation. We also fine-tuned InstructPix2Pix, an image modifying diffusion model, adapted to generate MRI-like output from CT input images.

3.2.1 Latent Diffusion models

Latent diffusion models are a type of generative deep learning model that generate images from text prompts, or other inputs like images or class labels, by learning to gradually denoise a latent representation of an image. They operate by mapping images into a lower-dimensional latent space and then performing iterative denoising in that space to generate realistic outputs. They are majorly composed of:

a) Autoencoders - To transform images into a lower-dimensional latent space, capturing essential features.

b) Denoising U-Net: Which learns to predict and remove noise from the latent space, effectively ”denoising” the image representation.

c) CLIP (Contrastive Language-Image Pre-training): Which creates rich embeddings that link text and image modalities, enabling the model to understand and generate images based on text descriptions.

Latent diffusion models excel in generating high-quality, diverse images by iteratively refining the latent space representation, making them highly efficient compared to traditional pixel-based models. By working in the lower-dimensional latent space, they can generate images faster, with less computational overhead, while still preserving fine details. Their ability to handle complex transformations has made them a powerful tool in tasks like image synthesis, inpainting, and cross-modal generation, including applications in medical imaging.

3.2.2 Stable Diffusion

Stable Diffusion [14] is a latent diffusion model which was originally designed for text to image generation, but can be adapted for image to image tasks by conditioning the generation on an input image along with an optional prompt. This allows the model to take a given image and transform it rather than generate from scratch. Stable Diffusion operates by encoding images into a lower-dimensional latent space, where it applies a denoising process to generate the desired output. Its flexibility allows for fine-tuning on specific tasks, making it suitable for applications like style transfer, inpainting, and cross-modal image generation. The model leverages a powerful UNet architecture that enables both high-quality image generation and refinement.

3.2.3 ControlNet

ControlNet is an extension of Stable Diffusion that adds explicit spatial control over the image generation process. It introduces additional input conditions like edges, depth maps, or annotated CT images, that guide the denoising steps in a parallel branch fused with the main UNet. ControlNet's architecture enables better control by utilizing additional conditioning inputs, such as edge maps, semantic segmentation, and keypoints, that guide the model's generation process. This allows for improved alignment of generated images with real-world structures and spatial properties. The approach ensures that high-level features like or-

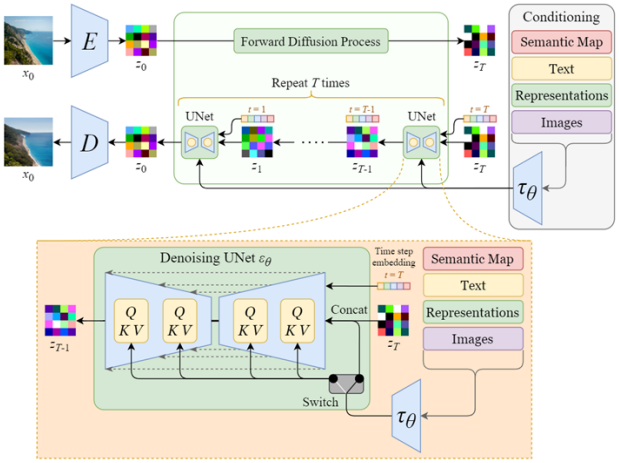

Figure 1. Stable Diffusion Architecture

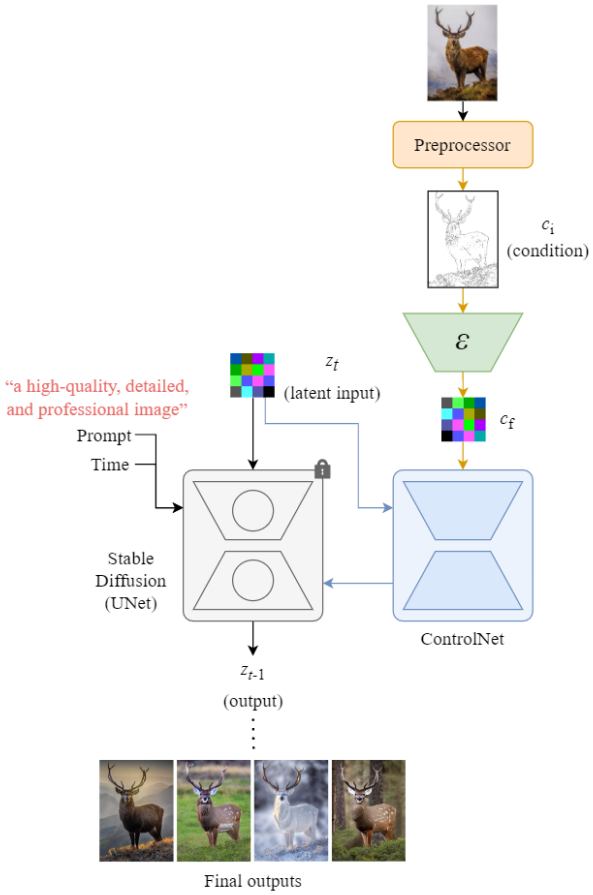

Figure 2. ControlNet Architecture

gan contours and spatial relationships are accurately transferred from the input CT to the output MRI, while preserving clinical relevance. ControlNet's ability to condition on multiple sources of information helps in creating images that have higher utility in medical applications where details are crucial.

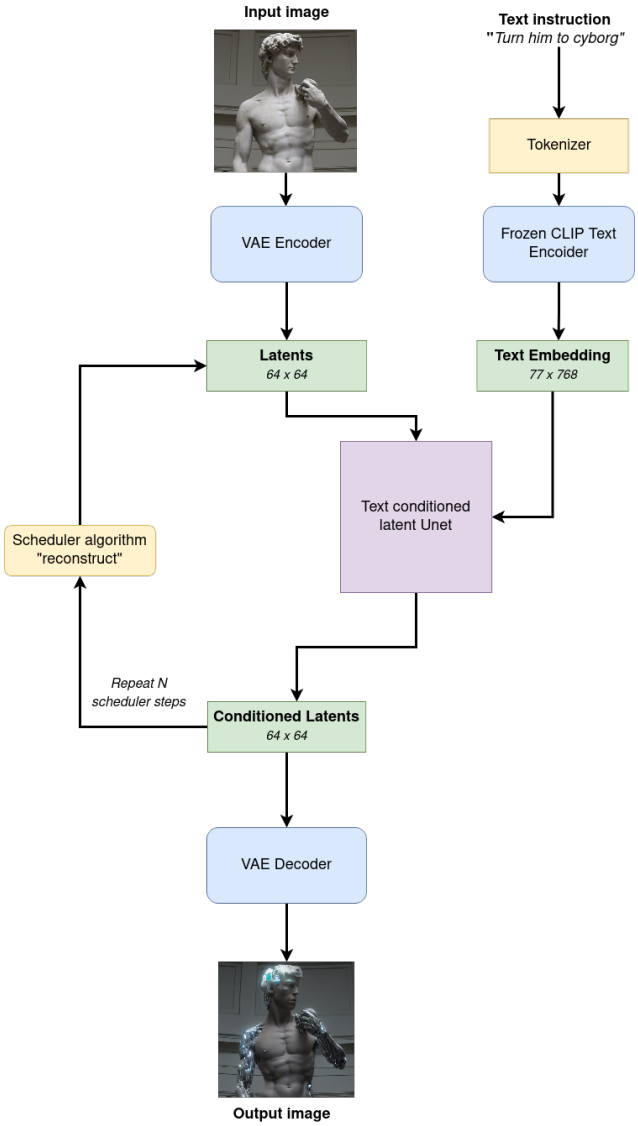

Figure 3. InstructPix2Pix Architecture

3.2.4 InstructPix2Pix

InstructPix2Pix extends Stable Diffusion by enabling instruction-based image editing using natural language. Trained on triplets of input image, instruction, output image, it learns to follow high-level editing commands. It supports image to image generation while responding dynamically to user-specified textual modifications. This introduces an interactive approach where users can specify detailed instructions for specific modifications, offering more control over the image generation process. This makes it particularly useful in medical applications where specific details, such as the intensity or shape of tissue types, need to be adjusted for different modalities. The model's ability to understand both visual and textual cues enables better transformations. Fine-tuning this model on medical images allows it to specialize in the CT→MRI conversion while leveraging the prior knowledge it has from general image editing training.

3.2.5 Model Training

Both the ControlNet-based model and the InstructPix2Pix model were initialized from pre-trained weights and then fine-tuned on our CT-MRI paired dataset. We started with Stable Diffusion v1.5 as the backbone for our models. For the ControlNet branch, we utilized a pre-existing ControlNet architecture configured for edge conditioning. The fine-tuning was done in a supervised manner: we provided the CT image (for InstructPix2Pix) or its edge map (for ControlNet model) as input and trained the diffusion model to reconstruct the corresponding MRI image. The ControlNet model was trained for 35,000 steps, and the InstructPix2Pix model for 15,000 steps.

3.3. Experimental Setup

1. Dataset - Central axial slices from paired CT-MRI scans from the SynthRad2023 Challenge dataset split into 2398 train image pairs and 126 test image pairs.
2. Preprocessing - Intensity normalized, contours generated using Canny edge detection for ControlNet.
3. Backbone - Stable Diffusion 1.5
4. ControlNet on Canny edges extracted from CT scans
5. InstructPix2Pix on CT scans
6. Training Steps - a) ControlNet - 35k b) InstructPix2Pix - 15k
7. Batch Size - 2-4
8. Hardware - 2x A5000 GPUs
9. Evaluation metrics - PSNR (db), SSIM, average inference time on test dataset.

4. Experiments

4.1. Datasets

We utilized the SynthRad2023 Challenge dataset [13], which comprises of paired CT and MRI scans. The dataset provides paired CT-MRI images of the brain and pelvic regions for a large number of patients (Only the brain images were selected for this project). From this dataset, central axial slices were extracted to form a set of 2398 training image pairs and 126 testing image pairs. Each CT slice was paired with its corresponding MRI slice to enable supervised learning of the CT-to-MRI translation task.

4.2. Preprocessing

The original dataset stores data as NIfTI (Neuroimaging Informatics Technology Initiative) compressed files, and from each 3D scan volume the central axial slices of the CT and the corresponding MRI were extracted and stored as PNG files. Each extracted slice image was then resized to a uniform resolution of 512×512 pixels and intensity normalization was applied. In addition to the base images, contour maps were generated using Canny edge detection which were used to guide the ControlNet based model.

4.3. Models Trained

Two primary backbones were investigated:

- **Stable Diffusion 1.5**: Used as the base model for ControlNet experiments, where CT-derived contour maps conditioned the MRI generation process.
- **InstructPix2Pix**: Fine-tuned directly on CT slices to learn the transformation to MRI images without explicit edge guidance.

4.4. Training Details

ControlNet models were trained for **35,000 steps**, while InstructPix2Pix models were trained for **15,000 steps**. The batch size was set between **2 to 4** depending on GPU memory constraints. All training was conducted on **two NVIDIA A5000 GPUs**.

4.5. Evaluation Metrics

- **Structural Similarity Index Measure (SSIM)**: - SSIM was used to measure the perceptual similarity between the generated MRI image and the ground truth MRI image by comparing luminance, contrast, and structural information. Unlike PSNR which treats each pixel independently, SSIM is computed on local regions and combines comparisons of brightness, contrast, and the correlation of structural patterns between images. It's value ranges from 0 to 1 with 1 indicating a perfect stuctural similarity between the generated image and the ground truth image.

$$\text{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (1)$$

- **Peak Signal-to-Noise Ratio (PSNR)**- PSNR quantifies the reconstruction quality by measuring the pixel-wise error between the generated I_{gen} and real MRI images I_{gt}. It is expressed in decibels (dB) and is derived from the mean squared error. A higher PSNR indicates that the generated image is closer to the reference image, with fewer pixel-level errors.

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(I_{\text{gen}}(i) - I_{\text{gt}}(i)\right)^2 \quad (2)$$

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right) \quad (3)$$

- **Inference Time**: The average time to generate an MRI image from a CT input on the test set was recorded to assess model efficiency.

4.6. Results

The experiments yielded promising results and the models produced MRI-like images from CT scans that were visually convincing and aligned with the ground truth MRIs. Figure 9 shows an example result from the ControlNet-guided diffusion model, while Figure 10 shows results from

Model	#Params	PSNR (dB)		SSIM		Time (s)		Train Steps	GPU-RAM (GB)	
		train	test	train	test	train	test		train	inference
ControlNet (CT-contours→MRI)	1.68B	9.523	8.5162	0.1732	0.1318	—	10.95	35k	~**22**	12
InstructPix2Pix (CT→MRI)	**983M**	28.12	26.805	0.743	0.715	—	**3.5**	**15k**	~40	**4**
InstructPix2Pix (MRI→CT)	**983M**	**33.328**	**31.975**	**0.799**	**0.8328**	—	**3.5**	**15k**	~40	**4**

Figure 4. Quantitative Results

the InstructPix2Pix-based model on a CT input. Quantitatively, the InstructPix2Pix based model performed much better than the ControlNet guided diffusion model on the test set as shown in Figure 4. The reasons for this and planned improvements are discussed later in the Future Scope section.

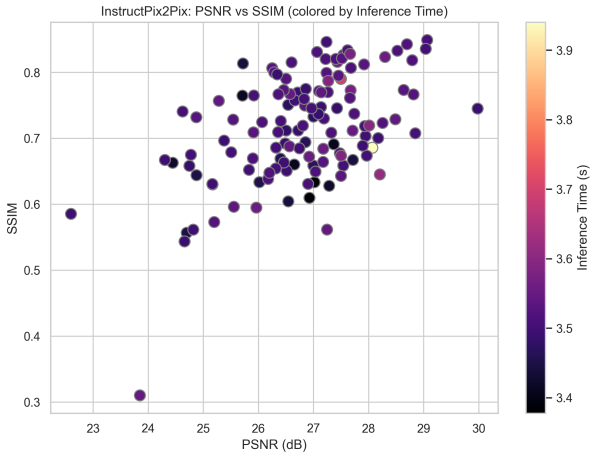

Figure 5. InstructPix2Pix: PSNR vs SSIM (colored by Inference Time)

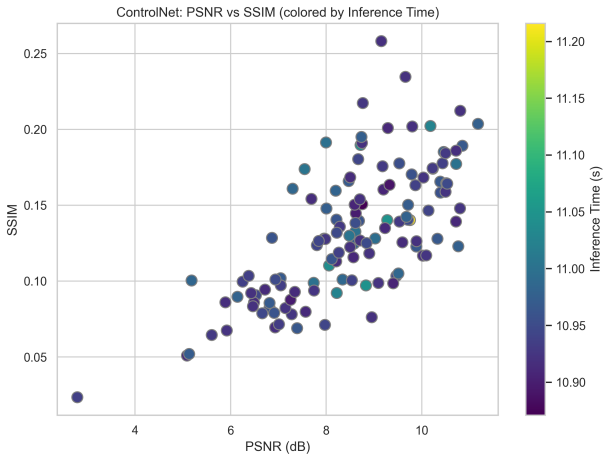

Figure 6. ControlNet: PSNR vs SSIM (colored by Inference Time)

The scatter plots in Figures 6 and 5 depict the relationship between PSNR and SSIM for each test sample, with point colors representing inference time. InstructPix2Pix achieves significantly higher PSNR and SSIM values compared to ControlNet, while also demonstrating faster inference times. This indicates that InstructPix2Pix not only produces higher-quality MRI outputs but is also more computationally efficient.

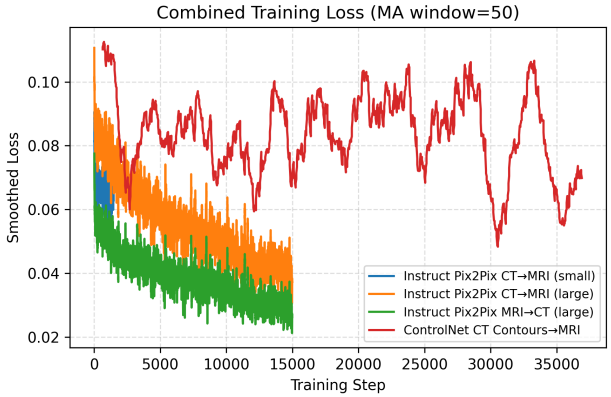

Figure 7. Training Loss Plots

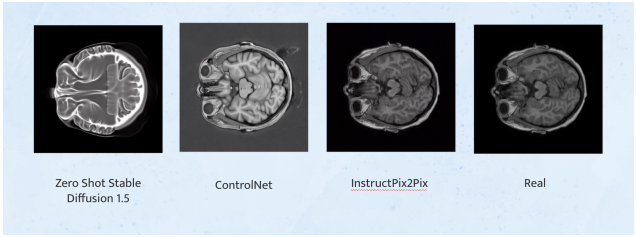

Figure 8. Qualitative Comparison - Zero Shot vs Fine-tuned vs Ground-Truth

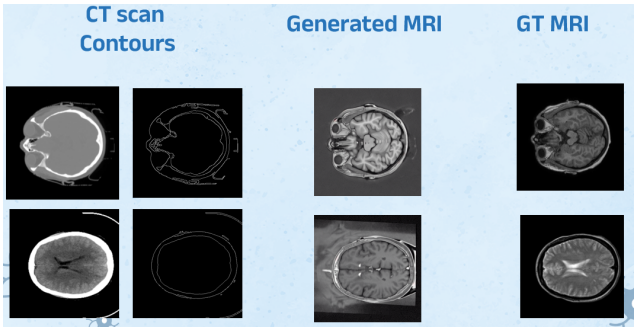

Figure 9. Sample results - ControlNet

In addition to the primary CT-to-MRI translation task, we also conducted experiments on the reverse direction, i.e., MRI-to-CT translation. This task is often considered more

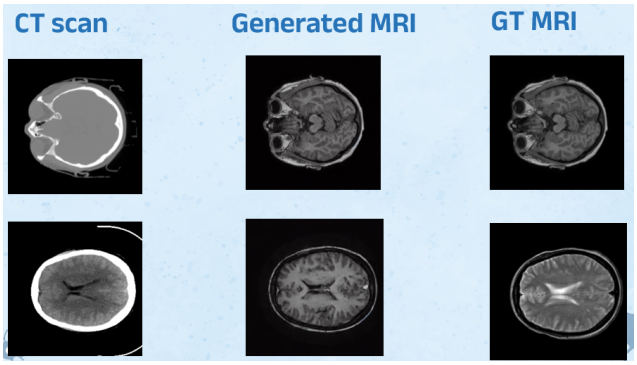

Figure 10. Sample results - InstructPix2Pix

reliable and carries less clinical risk, as CT imaging generally captures bony structures and anatomical boundaries more consistently. The MRI-to-CT models demonstrated better accuracy compared to the CT-to-MRI models. Figure 11 illustrates the the MRI-to-CT translation results.

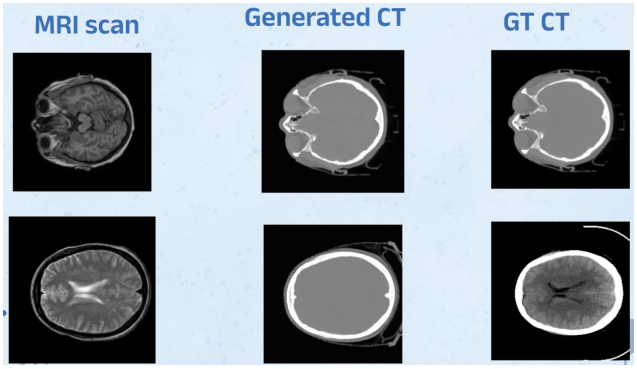

Figure 11. Sample results - MRI to CT InstructPix2Pix

We show qualitative results on a test sample and on an in the wild image pair from a completely different data distribution that uses different CT and MRI machines. InstructPix2Pix impressively outperforms ControlNet on both test and in the wild sample pairs.

4.7. Demo with Gradio

Gradio is an open-source Python package that allows us to quickly build a demo or web application for a machine learning model, API, or any arbitrary Python function. The link to the demo or web application can be shared using Gradio's built-in sharing features. We used Gradio for giving a quick demo of Mediffuse for presentation purposes, Figure 12.

5. Conclusion

We presented Mediffuse, an approach for MRI generation from CT scans using diffusion models. Our approach used ControlNet for spatial conditioning and InstructPix2Pix for instruction-driven image translation, both of which were fine-tuned on paired CT-MRI data. Our results show that these models can learn the complex mapping between CT and MRI domains, yielding outputs that are visually and quantitatively close to true MRIs. Additionally, the reverse translation from MRI to CT was found to be more reliable and of lesser risk of missing the anomalies in the soft tissues in the CT to MRI translation. The promising results of our experiments encourage further research into use of generative models for medical imaging and suggests some future improvements (discussed below).

Figure 12. Gradio Demo Screenshot

6. Future Scope

6.1. Improving ControlNet Performance

As displayed in the results earlier, in this scenario, the ControlNet based model lags way behind InstructPix2Pix based model in quantitative results. However, better contour synthesis for CT to MRI in the future for ControlNet could lead to improved results. Perhaps a dedicated model for generating contours, could also be considered. Structural conditioning of contour to image method not only improves the realism and consistency of the MRI outputs but also mitigates mode collapse and hallucinations common in standard generative models. Hence with better contour generation methods we intend to enhance clinical reliability even further.

6.2. Building a Smaller Dedicated Model

Diffusion models are resource-intensive, typically requiring a GPU and several seconds per image to sample. For any such model to be practical in a clinical setting, the computational load needs to be reduced. One approach is to employ knowledge distillation to train a lighter "student" model dedicated to this task that can emulate the behavior of a large diffusion model.

6.3. Evaluation of Diagnostic Value and Anomaly Detection

An extensive validation of the synthetic MRIs is needed to ensure that the synthetic MRIs are diagnostically useful for certain tasks (e.g., tumor detection). Currently the dataset this model was trained on does not have any data related to patient diagnostics. However, additional data can be encoded using Multi-layer Perceptron layers or CLIP encoders for text data as additional conditioning heads in the InstructPix2Pix architecture.

6.4. Application to Other Modalities

Future work may explore applying Mediffuse to other anatomical regions or modalities. For example, translating MRI to CT could be useful for generating synthetic CT for radiation therapy planning (the reverse of the current project). Additionally, integrating patient-specific information could further tailor the synthetic images.

References

[1] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023. 1

[2] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. 1

[3] Qing Lyu and Ge Wang. Conversion between ct and mri images using diffusion and score-matching models, 2022. 1

[4] Anamika Jha and Hitoshi Iima. Ct to mri image translation using cyclegan: A deep learning approach for cross-modality medical imaging. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence - Volume 3: ICAART*, pages 951–957. INSTICC, SciTePress, 2024. 1

[5] Makki Taqi and Pintu Kumar Ram. Synthesizing mris from ct scans using deep learning techniques: A comprehensive review. In *2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN)*, pages 189–197, 2024. 1

[6] 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. 1

[7] Saba Nikbakhsh, Lachin Naghashyar, Morteza Valizadeh, and Mehdi Chehel Amirani. Enhanced synthetic mri generation from ct scans using cyclegan with feature extraction, 2023. 1

[8] Haoyang Liu. Brain ct to mri medical image transformation based on u-net. *Journal of Physics: Conference Series*, 2824(1):012002, aug 2024. 1

[9] Zhenbin Wang, Lei Zhang, Lituan Wang, and Zhenwei Zhang. Soft masked mamba diffusion model for ct to mri conversion, 2024. 2

[10] Austin Tapp, Abhijeet Parida, Can Zhao, Van Lam, Natasha Lepore, Syed Muhammad Anwar, and Marius George Linguraru. Mr to ct synthesis using 3d latent diffusion. *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2024. 2

[11] David Bird, Richard Speight, Sebastian Andersson, Jenny Wingqvist, and Bashar Al-Qaisieh. Deep learning mri-only synthetic-ct generation for pelvis, brain and head and neck cancers. *Radiotherapy and Oncology*, 191:110052, 12 2023. 2

[12] Zakaria K. D. Alkayyali, Ashraf M. H. Taha, Qasem M. M. Zarandah, Bassem S. Abunasser, Alaa M. Barhoom, and Samy S. Abu-Naser. Advancements in ai for medical imaging: Transforming diagnosis and treatment. *International Journal of Academic Engineering Research(Ijaer)*, 8(8):8–15, 2024. 2

[13] A. Thummerer, E. van der Bijl, and M. Maspero. SynthRAD2023 Grand Challenge dataset: synthetizing computed tomography for radiotherapy (0.1), 2023. 2, 4

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2