Medical Image Super Resolution

Saaim Raad
Student# 1010292853
saaim.raad@mail.utoronto.ca

Ayushi Agrawal
Student# 1010015145
ayushi.agrawal@mail.utoronto.ca

Pakhi Gupta
Student# 1010475807
pakhi.gupta@mail.utoronto.ca

Md Nazmus Saad
Student# 1010243194
md.saad@mail.utoronto.ca

Abstract

This project presents a latent diffusion-based super-resolution (SR) model to enhance low-resolution chest X-rays for improved diagnostic accuracy. The architecture combines a pretrained Variational Autoencoder (VAE), a custom U-Net-lite denoiser, and a noise scheduler to reconstruct high-resolution images with strong structural fidelity. For deployment, NVIDIA TensorRT optimization of the U-Net is applied, yielding over a $4\times$ inference speedup with negligible quality loss, enabling efficient deployment in resource-constrained clinical settings.
—-Total Pages: 10

1 Introduction

High-quality medical images are crucial for accurate diagnosis and treatment, especially when detecting subtle abnormalities in chest X-rays like early signs of infections, tumors, or cancer. However, imaging devices often suffer from hardware, environmental, and operational limitations that result in low-resolution (LR) outputs, obscuring critical details. This drop in image quality creates serious challenges for both human experts and Computer-Aided Diagnosis (CAD) systems, potentially compromising clinical decisions (Umirzakova et al., 2023).

Our project aims to develop a Super-Resolution (SR) system using a latent diffusion architecture that combines a pretrained Variational Autoencoder (VAE), a denoising U-Net, and a diffusion noise scheduler to efficiently reconstruct high-resolution images from low-resolution inputs. Our approach enables high-quality image enhancement while incorporating TensorRT-based inference optimization on NVIDIA GPUs, achieving significant speedups without substantially compromising output quality (Figure 1).

Deep learning, particularly using diffusion models with latent space encoders, is well-suited for our project as it can recover fine details missed by traditional methods, leverages large chest X-ray datasets effectively, and offers a scalable, practical solution for enhancing medical image quality in real-world settings.

2 Background And Related Work

2.1 A Brief History of Super Resolution (SR)

Super-resolution (SR) is the task of reconstructing high-resolution (HR) images from low-resolution (LR) counterparts. Early SR methods predominantly used traditional interpolation techniques such as bicubic interpolation, where weighted averages of surrounding pixels approximate unknown pixel values. This general-purpose resampling method works for both upsampling and downsampling but cannot capture fine detail and often introduces blur, as explained by the Data Processing Inequality, which states that information not present in the original image cannot be created (Cover & Thomas, 2012). As a result, there have been recent shifts to exploring Deep Neural Network based frameworks as a means to inject learned information from the Neural Networks, into low-resolution images in order to increase resolution in a manner that actually adds informative content.
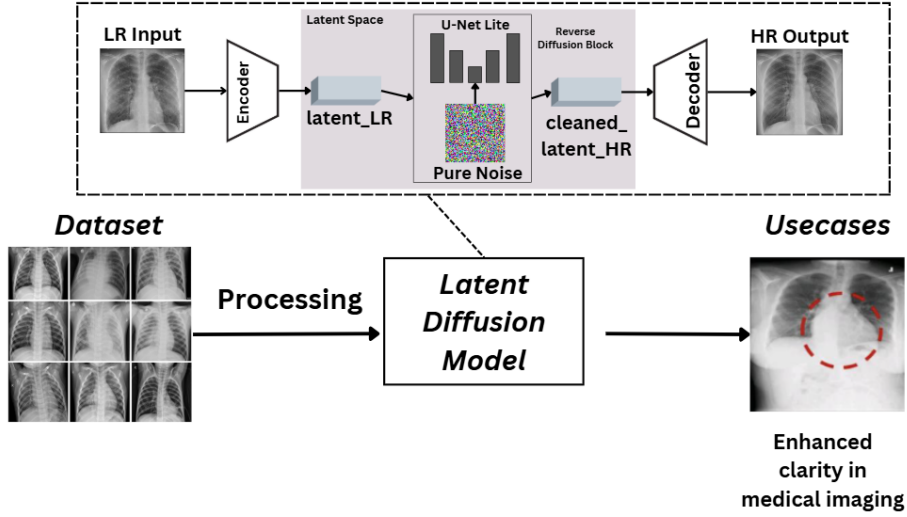
Figure 1: The proposed model.

Deep learning brought major advances, starting with SRCNN (Dong et al., 2015), which used convolutional neural networks (CNNs) to outperform traditional methods in PSNR. These CNN models, however, relied on pixel-wise losses (L1, L2) that often produced overly smooth outputs. SRGAN (Ledig et al., 2017) addressed this by introducing adversarial training to improve perceptual quality, and ESRGAN (Wang et al., 2019) further enhanced GAN stability and depth. Despite these gains, GAN-based SR remains challenging due to the delicate balance required between generator and discriminator for stable convergence.

U-Net (Ronneberger et al., 2015), originally for biomedical segmentation, became a popular backbone for image-to-image tasks. Its encoder–decoder with skip connections preserves fine detail while capturing global context, but operating in pixel space makes it memory-intensive, slow to train, and unsuitable for edge devices.

Latent Diffusion Models (LDMs) (Rombach et al., 2022) address these limitations by performing the diffusion process in a low-dimensional latent space produced by a Variational Autoencoder. This reduces the spatial size of U-Net inputs, lowering computation and memory costs while enabling state-of-the-art results in systems like Stable Diffusion. Recent work, such as Edge-SD-SR (Noroozi et al., 2025), demonstrates that latent diffusion-based SR can run efficiently on mobile and edge devices, greatly reducing model size and latency without compromising output quality.

2.2 EVALUATION METRICS IN SUPER-RESOLUTION

Super-resolution quality is evaluated using Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). PSNR quantifies pixel-wise differences between reconstructed and reference images, while SSIM measures structural similarity (Huynh-Thu & Ghanbari, 2008; Wang et al., 2003). LPIPS (Zhang et al., 2018) assesses perceptual similarity using deep features extracted from pretrained convolutional neural networks, capturing high-frequency details and aligning better with human perceptual judgment than PSNR or SSIM.

3 DATA PROCESSING

The dataset used is the publicly available NIH ChestX-ray14 dataset (Wang et al., 2017) consisting of 112,120 1024x1024 images of chest X-rays from 30,805 unique patients.

The Indiana University Chest X-Ray Collection of more than 7000 images accessible through Open-i (Demner-Fushman et al., 2016) is used as additional testing data. It contains both lateral and frontal

view X-rays, as opposed to the NIH set which contains only frontal X-rays. This will ensure that the model can be tested on images completely different from its training data.

3.1 DATA LOADING

The following steps were used to obtain and process the dataset, of which a few sample images are shown in Figure 2:

1. **Data downloading and organization:** A downloading script allows the user to download a specified number of batches from the dataset source, extracts them into the specified folder, and prevents overwriting of existing data.

2. **Generating HR and LR image pairs:** The `generate_downsampled_pairs` function downsamples the original 1024x1024 images twice using OpenCV's `resize` function with bicubic interpolation. The 256x256 downsampled image is saved into the `HR_256` folder and the 64x64 version is saved into the `LR_64` folder.

3. **Creating a custom PyTorch dataset:** The `PairDataset` class reads images from their folders, transforms them into tensors of type `float32` and normalizes them to [-1,1] for faster training convergence. It assembles mini-batches of dimension B×3×64×64 or B×3×256×256 for input images and labels (ground truths) respectively.

4. **Dataset splitting:** The `PairDataset` class also allows for configurable splits between train/validation/test data. A 70-20-10 split is currently in use, which provides 38,527 training, 10,907 validation, and 5,500 test images. One of the challenges here was to keep images from the same patient together, which was solved using careful parsing of filename while generating split indices.

The same steps will be followed for testing on the new data (Indiana University). The only required change will be to set the data directory (a configurable argument) to the directory where the new data has been extracted, and to ensure images are cropped to squares before resizing.
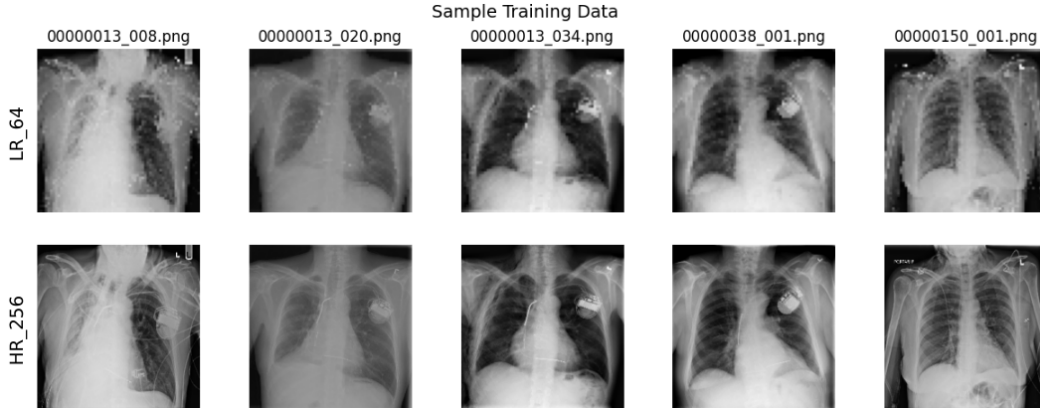

Figure 2: Some random samples of the training data without data augmentation.

3.2 DATA AUGMENTATION

To improve generalization and robustness, data augmentation was applied to 70% of the training samples. This percentage was chosen based on empirical studies suggesting that moderate augmentation (around 60–80%) helps prevent overfitting while maintaining data fidelity, which is particularly important in the medical imaging domain where fine-grained details are critical (Athalye & Arnaout, 2023; Sanaat et al., 2022). Each augmentation was implemented considering spatial alignment, ensuring that LR and HR image pairs remained synchronized. The augmentation techniques used are as follows:

- **Random Horizontal Flip (50%):** Flips images along the vertical axis to make the model invariant to left-right orientation.
- **Random Rotation (90°, 180°, 270°):** Simulates different acquisition angles while preserving anatomical structure.
- **Random Resized Crop (80%):** Crops and resizes the image to expose subregions, helping the model detect abnormalities anywhere in the X-ray.
- **Cutout (30%):** Masks a 20% square patch to simulate occlusions or artifacts, encouraging the model to rely on surrounding context.

Demonstrations of these augmentations are shown in Figure 3.

While implementing these techniques, care was taken to maintain alignment between LR and HR image pairs, particularly during cropping, which required scaling coordinates appropriately. Unrealistic augmentations such as color jitter or geometric warping were avoided to preserve the medical validity of the X-rays. Applying augmentation to roughly 70% of the training data balanced introducing useful variation with retaining enough original examples for accurate learning.
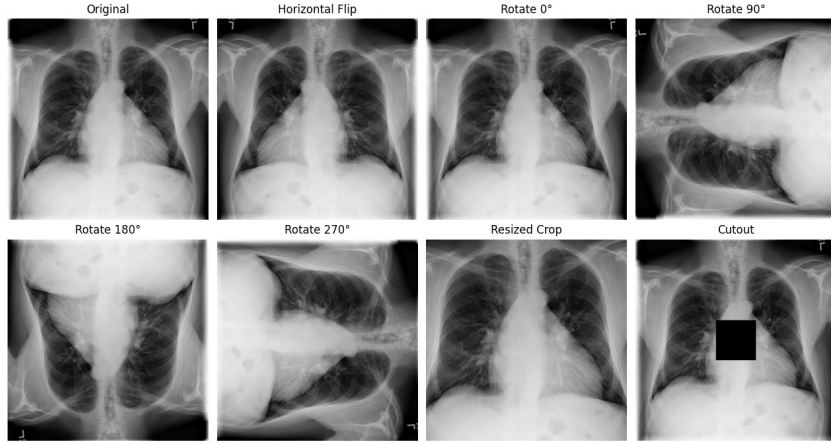

Figure 3: Demonstration of augmentation techniques used in model training.

4 ARCHITECTURE

The primary model performs super-resolution on low-resolution chest X-rays using a latent diffusion-based architecture. As shown in Figure 4, the system includes a pretrained VAE, custom diffusion schedulers, and a custom U-Net, totaling 813,220 trainable parameters. All training and inference was conducted on an NVIDIA GeForce RTX 4090.

4.1 LATENT DIFFUSION-BASED SUPER-RESOLUTION

The model takes paired LR and HR images (3×64×64 and 3×256×256). LR images are first upsampled to 3×256×256 via bicubic interpolation, then both LR and HR images are encoded to latent space separately using Hugging Face's pretrained AutoencoderKL VAE, yielding 4×32×32 latents by reducing spatial resolution 8× and expanding channels from 3 to 4.

4.2 FORWARD DIFFUSION AND TIMESTEP CONDITIONING

Training uses the forward diffusion process, where Gaussian noise is progressively added to the HR latent over multiple time steps. Rather than sequentially applying noise at each step, we use the closed-form DDPM formula for efficiency:

$$z_t = \sqrt{\bar{\alpha}_t} \cdot z_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon$$
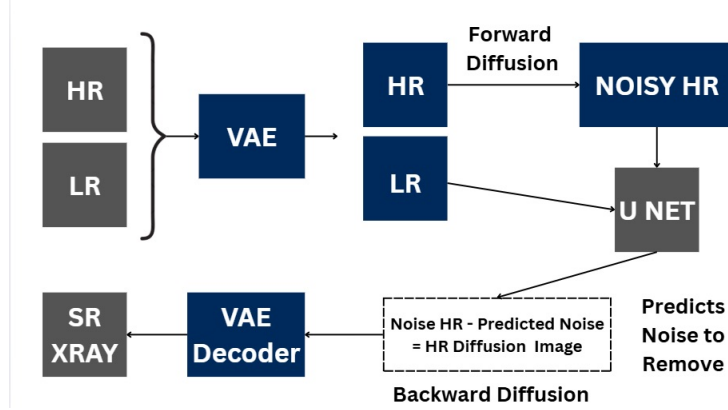
Figure 4: High level block diagram of architecture used for both training and inference.

This closed form calculates the final noisy HR latent given some number of timesteps $t \in [0, 300)$. The timestep t controls the noise level, as a higher t implies more steps that add noise. z_0 is the clean HR latent, $\epsilon \sim \mathcal{N}(0, 1)$, and

$$\bar{\alpha}_t = \prod_{i=1}^{t}(1 - \beta_i)$$

with β linearly scheduled from 0.0001 to 0.02, as per the original DDPM paper (Ho et al., 2020). The noisy HR latent is concatenated with the LR latent to form an 8-channel tensor passed to the U-Net, alongside the timestep t.

U-Net uses an encoder-decoder structure with skip connections (Figure 5) (Ronneberger et al., 2015). Through conditioning available from the LR latent and t, the U-Net learns to predict the noise added to the HR latent, trained using LPIPS loss against the true ϵ added during the forward diffusion process.
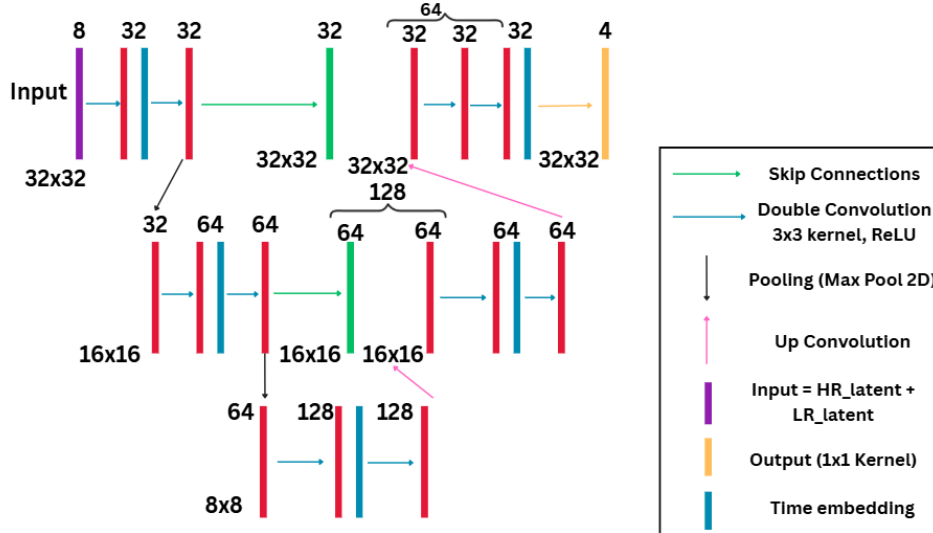

Figure 5: Structure of the U-Net.

4.3 INFERENCE AND REVERSE DIFFUSION

During inference, the process starts from pure Gaussian noise in latent space ($4 \times 32 \times 32$). Reverse diffusion is performed from $t = 299$ to 0, iteratively denoising with the trained U-Net. At each step,

the LR latent is concatenated with the noisy HR latent, and the predicted $\hat{\epsilon}_t$ is used to reverse the noise. After all steps, the final latent is decoded via the VAE to produce a $3 \times 256 \times 256$ super-resolved image. PSNR, SSIM and LPIPS are used to measure performance against ground-truth HR images, with bicubic interpolation as the baseline.

4.4 EDGE DEPLOYMENT FOR PERFORMANCE BOOST

To accelerate inference and prepare the model for deployment on resource-constrained edge devices, we optimized the trained U-Net using NVIDIA TensorRT. After training, the PyTorch checkpoint (`.pt`) containing only the U-Net weights was exported to the ONNX format (`.onnx`) using `torch.onnx.export`, preserving the same input-output interface (x: 8×32×32, t: scalar timestep). The ONNX model was then compiled into a TensorRT engine file (`.trt`) using `trtexec`, TensorRT's command-line compiler.

During inference, the TensorRT runtime loads the optimized engine, allocates GPU and CPU buffers, and executes inference using an asynchronous CUDA stream. The LR and noisy HR latents are flattened into contiguous NumPy arrays and copied to GPU memory, where the U-Net performs a forward pass to predict noise. The resulting output is transferred back to CPU memory and reshaped into the original latent format.

5 BASELINE MODEL

A bicubic interpolation algorithm (Keys, 1981) is used as the baseline model. The input LR image is resized to an HR image via PyTorch's `F.interpolate` function with `mode='bicubic'` and `align_corners=False` parameters. Bicubic interpolation is widely used in image processing due to its computational efficiency and smooth results, though it cannot recover high-frequency details lost during downsampling.

6 QUANTITATIVE RESULTS

The dataset was split into 70% train, 20% validation, and 10% test. Training ran for 50 epochs, and all twelve available training batches were loaded. Both training and validation losses dropped smoothly to about 0.18 (Figure 6). Evaluation was performed on the NIH ChestX-ray14 patient-held-out test split, computing PSNR, SSIM, and LPIPS on images and reporting the means over all samples in the test set. Bicubic interpolation serves as a non-learned baseline. Higher is better for PSNR/SSIM and lower is better for LPIPS. The latent-diffusion model achieved PSNR/SSIM values that are slightly lower than bicubic but still comparable, while achieving substantially lower LPIPS (Table 1). These results indicate greater perceptual similarity but slightly degraded pixel-wise similarity.

On the test set, the TensorRT-optimized U-Net reduced inference time, achieving a $4.7\times$ speedup compared to the standard PyTorch model. PSNR, SSIM, and LPIPS remained within negligible differences, indicating that the performance gains came without meaningful loss in image quality.

Table 1: Inference time and quality metrics for bicubic, diffusion, and TensorRT models. Lower LPIPS is better, while higher PSNR/SSIM are better.

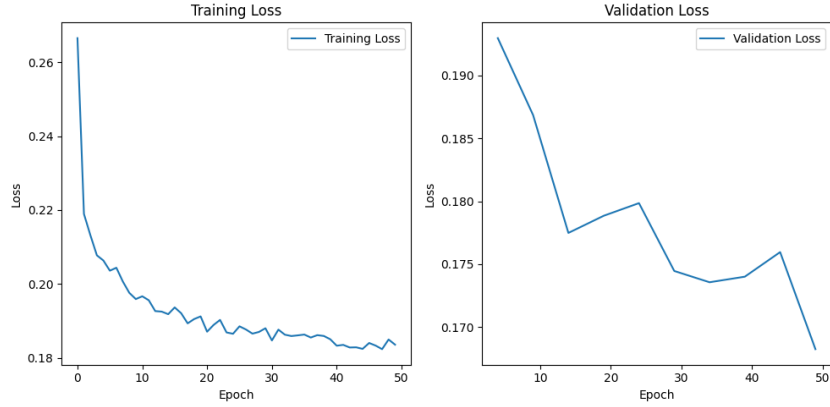| Inference Type	Time	Model	LPIPS	PSNR	SSIM
N/A	N/A	Bicubic	0.3472	29.59	0.9176
Standard	7.02s	Diffusion	0.1232	27.34	0.8647
TensorRT	1.49s	Diffusion	0.1223	27.11	0.8501

Figure 6: Training and validation curves.

7   QUALITATIVE RESULTS

Figure 7 compares two test samples across LR (64×64), bicubic (256×256), diffusion SR (256×256), and HR (256×256). Diffusion SR yields sharper bone edges, clearer diaphragm contours, and finer lung textures than bicubic, aligning with its lower LPIPS despite slightly lower PSNR/SSIM. The model struggles with distortions, artifacts, or text in LR inputs, sometimes smoothing or hallucinating details. For example, the white 'L' marker above the right shoulder in the bottom-right HR image is blurred in the diffusion output. Ribs also appear slightly less pronounced than in HR, though still more accurate than bicubic.
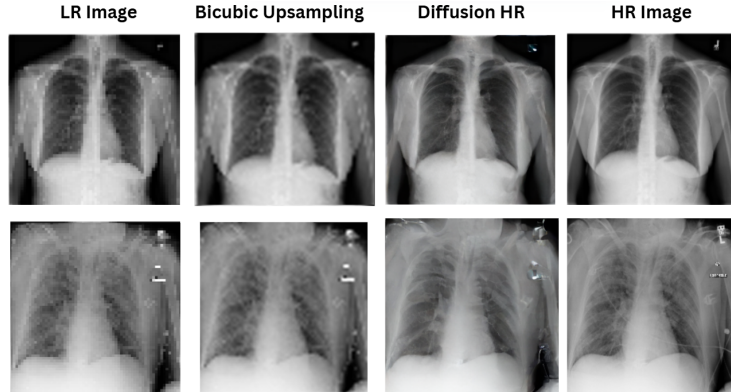

Figure 7: Sample results using test data.

TensorRT outputs were nearly identical to PyTorch, with only slight contrast/sharpness differences in fine structures, thus matching the small PSNR changes (Figure 8). A subtle tendency was observed for TensorRT outputs to produce slightly darker regions in already dark areas of the image.

8   EVALUATION OF MODEL ON NEW DATA

The model was trained exclusively on the NIH ChestX-ray14 dataset. To assess true generalization, the model was tested on the Indiana University Chest X-Ray Collection which is a fully independent frontal chest x-ray dataset sourced from a different institution with no shared images. The Indiana Collection introduces imaging conditions not present during training, thus challenging the model's adaptability. Its independent origin ensures that these test samples were not used in any training or tuning of hyperparameters, providing a clean, unbiased evaluation of performance on new data.
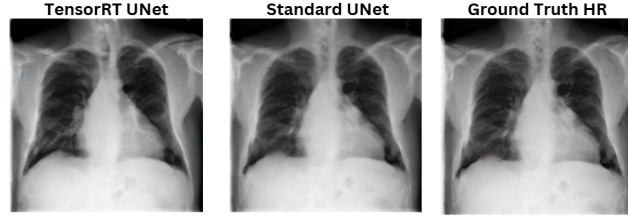
Figure 8: Comparison of model with TensorRT on test set.

On the unseen data, the latent diffusion model performed similarly to the NIH dataset, clearly outperforming bicubic by recovering fine anatomical details and diagnostic features lost with traditional methods (Figure 9), though it still hallucinated fine text, noise, and artifacts.

Means for metrics on the Indiana dataset closely matched the means for metrics on the NIH dataset, with PSNR/SSIM slightly lower and LPIPS showing a significant improvement (Table 2).
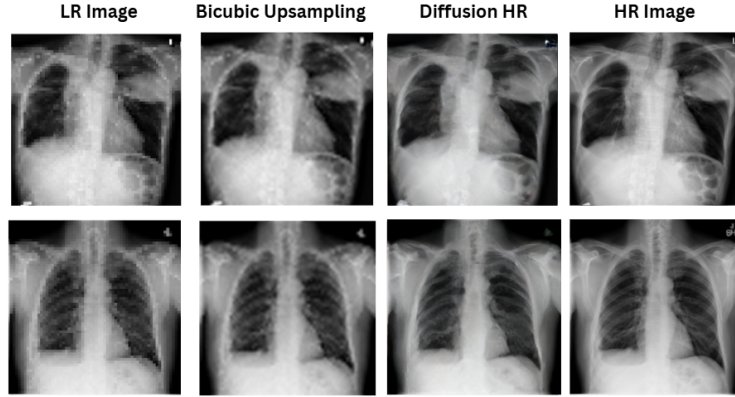

Figure 9: Sample results using unseen new data.

Table 2: Inference time and quality metrics on the unseen new data for bicubic, diffusion, and TensorRT models. Lower LPIPS is better, while higher PSNR/SSIM are better.

Inference Type	Time	Model	LPIPS	PSNR	SSIM
N/A	N/A	Bicubic	0.3444	28.36	0.9019
Standard	7.11s	Diffusion	0.1285	26.74	0.8621
TensorRT	1.70s	Diffusion	0.1290	27.01	0.8280

The TensorRT-optimized U-Net reduced inference time, achieving a similar $4.2\times$ speedup compared to the standard PyTorch model. This improvement is similar to that of the NIH results and it was expected, as runtime depends on model architecture and TensorRT's optimizations rather than the specific dataset. Output quality remained comparable, with only minor contrast differences as noted in the validation results (Figure 10).

9 DISCUSSION

The latent-diffusion SR produces chest X-rays that are perceptually closer to the HR targets than bicubic, as seen by richer rib edges, diaphragm contour, and textures. This aligns with lower LPIPS even when PSNR/SSIM are similar or slightly lower. This is expected since after downsampling, the LR image has lost high-frequency content. A linear interpolator like bicubic can only smooth
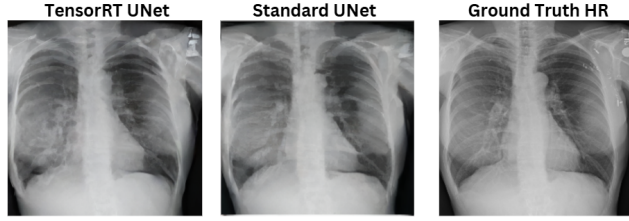
Figure 10: Comparison of model with TensorRT on unseen new data.

between existing samples and so it could not recover lost high-frequency information. Diffusion, by contrast, used a learned prior over HR chest X-rays and estimates samples from the conditional distribution $p(\text{HR} \mid \text{LR})$, thus synthesizing a brand new image from noise. That prior injected high-frequency details (edges/texture) that are consistent with the LR evidence, as opposed to simply smoothing from existing samples. Because many HR images are plausible for the same LR input, a pixel-wise metric could penalize these valid choices, explaining why perceptual quality improves (LPIPS) while PSNR/SSIM may not.

Looking at the qualitative results, the shortcomings arise when non-anatomical artifacts (compression blockiness, clipped borders, markers) are present. In these cases, the model may smooth or invent details, as visible around borders and in regions with distortions/unnatural noise. Additionally, thin, high-contrast anatomy such as the clavicles and ribs appeared slightly faint, perhaps due to the VAE bottlenecking very thin edges and/or pre-encoding upsample adds blur which the model does not always fully reverse. A somewhat surprising finding was how strongly LPIPS improved even when PSNR/SSIM slightly dipped, highlighting that pixel-aligned metrics can undervalue clinically reasonable reconstructions.

For deployment on resource-constrained hardware, the observed inference speedup from TensorRT stems from kernel fusion, reduced precision operations, and hardware-specific optimizations that lower computational overhead. The slight quality degradation likely results from quantization and approximation during these optimizations, which can slightly affect fine-scale contrast and edge definition.

Overall, the model's strengths are perceptual realism and anatomical accuracy under typical conditions. Conditioning in latent space keeps global structure anchored while the denoiser invents plausible high-frequency detail. The TensorRT export also delivered a meaningful speedup with only minor, mostly contrast-level differences, suggesting practicality for edge deployment.

Key challenges involved designing and integrating the diffusion scheduler with the VAE and U-Net. As a latent diffusion model, the system required a solid understanding of discrete-time stochastic processes and variational inference to generate images from pure noise. A significant portion of time was spent reading relevant literature and working through the underlying math to ensure correct implementation. Another key challenge was sourcing the Jetson Nano. To address the observed artifact sensitivity and faint bones, we will scale and tune the system: train on the full dataset for a longer duration with larger effective batch sizes (feasible via gradient accumulation), adopt a deeper U-Net to address bottlenecks, and implement cross-attention between the LR and HR latents prior to entering the U-Net, instead of the current naive concatenation of feature vectors. Together, these steps should further elevate the perceptual gains into more reliable, clinically usable outputs.

10 ETHICAL CONSIDERATIONS

While promising for improving diagnostic support by enhancing low-resolution chest X-ray images, this system should be used with caution and further validated for safe, fair clinical use. Public datasets, like the one being used for this project, usually lack demographic diversity, which risks bias and reduced performance on underrepresented populations (Galanty et al., 2024). The diffusion-based architecture relies on generative modeling in a latent space, which sometimes introduces hallucinated details not present in the original scan. These enhancements must not be mistaken for real anatomical features, as they could mislead diagnoses if not carefully validated (Shin et al., 2024).

REFERENCES

Chinmayee Athalye and Rima Arnaout. Domain-guided data augmentation for deep learning on medical imaging. *PloS one*, 18(3):e0282532, 2023. ISSN 1932-6203. doi: 10.1371/journal.pone. 0282532. URL https://europepmc.org/articles/PMC10035842.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012. ISBN 9781118585771. URL https://books.google.ca/books?id=VWq5GG6ycxMC.

Dina Demner-Fushman, Marc D. Kohli, Martin B. Rosenman, Sarah E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association: JAMIA*, 23(2):304–310, March 2016. doi: 10.1093/jamia/ocv080.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Image super-resolution using deep convolutional networks. *CoRR*, abs/1501.00092, 2015. URL http://arxiv.org/abs/1501.00092.

M. Galanty, D. Luitse, S.H. Noteboom, S.H. Noteboom, P.M.A. van Ooijen, M. Homan, M. Nagtegaal, J.A.W.M. van der Laak, and H. Huisman. Assessing the documentation of publicly available medical image and signal datasets and their impact on bias using the beamrad tool. *Scientific Reports*, 14:31846, 2024. doi: 10.1038/s41598-024-83218-5. URL https://doi.org/10.1038/s41598-024-83218-5.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801, 2008. doi: 10.1049/el:20080522. URL https://digital-library.theiet.org/doi/abs/10.1049/el%3A20080522.

R. Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(6):1153–1160, 1981. doi: 10.1109/TASSP.1981.1163711.

Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114, 2017. doi: 10.1109/CVPR.2017.19.

Mehdi Noroozi, Isma Hadji, Victor Escorcia, Anestis Zaganidis, Brais Martinez, and Georgios Tzimiropoulos. Edge-sd-sr: Low latency and parameter efficient on-device super-resolution with stable diffusion via bidirectional conditioning, 2025. URL https://arxiv.org/abs/2412.06978.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 9351 of *LNCS*, pp. 234–241. Springer, 2015. URL http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a. (available on arXiv:1505.04597 [cs.CV]).

Amir Sanaat, Iraj Shiri, Somayeh Ferdowsi, et al. Robust-deep: A method for increasing brain imaging datasets to improve deep learning models' performance and robustness. *Journal of Digital Imaging*, 35(2):469–481, 2022. doi: 10.1007/s10278-021-00536-0. URL https://doi.org/10.1007/s10278-021-00536-0.

Minji Shin, Mohan Seo, Kyung Lee, and Kyuri Yoon. Super-resolution techniques for biomedical applications and challenges. *Biomedical Engineering Letters*, 14(3):465–496, March 2024. doi: 10.1007/s13534-024-00365-4.

Sabina Umirzakova, Sevara Mardieva, Shakhnoza Muksimova, Shabir Ahmad, and Taegkeun Whangbo. Enhancing the super-resolution of medical images: Introducing the deep residual feature distillation channel attention network for optimized performance and efficiency. *Bioengineering*, 10(11):1332, 2023. doi: 10.3390/bioengineering10111332.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3462–3471, 2017. doi: 10.1109/CVPR.2017.369.

Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In Laura Leal-Taixé and Stefan Roth (eds.), *Computer Vision – ECCV 2018 Workshops*, pp. 63–79, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11021-5.

Zhou Wang, Alan Bovik, Hamid Sheikh, Student Member, and Eero Simoncelli. Image quality assessment: From error measurement to structural similarity. *IEEE Trans. Image Process.*, 13, 11 2003.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.