CSCI 1430 Final Project Report: Super-Resolution for Medical Images

Team name: Super-Resolution (Yen Chu, Michelle Liu, Jessica Wan, Zilai Zeng).
TA name: Melvin He. Brown University

Abstract

Advancing medical diagnosis by using AI to analyze medical images is a critical topic in computer vision. While different types of artifacts and blurs are exhibited in medical images nowadays, effectively removing these noises and enhancing the image quality is essential to improve the robustness and accuracy of the diagnosis process. In this project, we develop a pipeline that takes in low-resolution medical images and restores their quality through super-resolution (SR). This pipeline is built upon ResShift, a state-of-the-art diffusion-based SR method, in which we first downsample medical images through various techniques, and upsample them with ResShift. We adopted CNNs to extract features from the original images and the upsampled images, comparing the features using (cosine) similarity scores to further evaluate restoration quality. Additionally, a proprietary CNN trained on the original images was used to classify the disease labels for both downsampled and upsampled images. We demonstrate that the classifier trained on images that are upsampled by our pipeline achieves significant improvement on diagnosis accuracy, highlighting the efficacy of our super-resolution pipeline.

1. Introduction

High-resolution medical images are necessary for reliable diagnosis. However, due to hardware limitations or other factors, images may often be noisy, low-resolution, or blurry. This can pose a risk for medical professionals and deep learning models to accurately identify and diagnose underlying medical conditions and diseases from low-resolution data. Currently, many disease classification models perform the best [5] on higher resolution images, due to the risk of losing important feature attributes as images decrease in resolution. In the meantime, it is crucial to prevent the classifier from overfitting trivial details provided by high-resolution images.

Super-resolution (SR) is a technique that attempts to restore higher-resolution images from low-resolution or corrupted image inputs. In this project, we investigate a super-resolution pipeline built upon ResShift [10] for compressed and low-resolution medical images. We focus on evaluating how well these super-resolved results 1) perform under standalone image quality metrics, such as PNSR, SSIM, CLIPIQA, MUSIQ 2) compare to their ground-truth counterparts by computing the similarity score between their CNN-extracted features and 3) perform in a custom CNN trained on disease classifications compared to their downsampled counterparts.

2. Related Work

Super-resolution (SR) pipelines for medical imaging currently exist. Prior studies show the potential for using Super-Resolution Generative Adversarial Networks (SRGANs) for enhancing imaging data and improving diagnostic accuracy. Kamashita and Markov (2020) used an enhanced SRGAN to denoise speckle noise in optic nerve head tissue images [9]. Ahman et al. (2022) use a combination of SRGANs and ResNet34 to upsample noisy medical images to achieve high classification accuracy results [8]. In our project, we attempt to use ResShift, a pre-trained diffusion-based SR model, to upsample the low-resolution data on NIH Chest X-Ray imaging data and use a pretrained ResNet54 to extract image features for similarity score computation. We also design our own CNN architecture to perform disease classification tasks on both low-resolution and high-resolution images, which demonstrates the significance of high resolution for accurate diagnosis.

3. Method

We apply a three step process in our methodology: (1) creating a dataset of downsampled images, (2) utilizing a diffusion-based SR model to reconstruct them, and (3) comparing the performance SR images to the original with a CNN model.

3.1. Downsampling and Preprocessing

To evaluate the performance of our SR pipeline, we first construct a dataset of downsampled, noisy images using `/images_001` of the NIH Chest X-Ray dataset [4]. We aim to replicate common image degradations found in real-world scenarios, such as artifacts. Each original, high-resolution image is processed through `downsample.py`. Within that pipeline, the following perturbations are applied:

- **Downsampling**: A 4x downsampling factor is applied on the 1024×1024 x-ray images, resulting in a 256×256 output.
- **Gaussian noise**: With a given probability (default 100%), we apply Gaussian noise with mean 0 and standard deviation 5.
- **Gaussian blur**: The noisy image can then be blurred with a Gaussian kernel of size 25×25.
- **Compression artifacts**: Finally, we encode and decode the image with JPEG compression at just a 10% quality level. This mimics heavy block artifacts that may arise in normal medical workflows.

All image processing is done through OpenCV [1]. The downsampled and noisy dataset is then fed into our diffusion-based SR model to restore quality.

3.2. Diffusion Super Resolution

To perform diffusion-based SR on downsampled medical images, we utilize ResShift [10], a state-of-the-art diffusion SR model pre-trained on ImageNet, to restore the quality of downsampled low-resolution (LR) images. In contrast to previous diffusion-based SR methods which usually take hundreds of sampling steps to obtain a high-resolution (HR) image, ResShift reduces the sampling steps to 15 steps by shifting the residual between LR images and HR images through a constructed Markov chain, without sacrificing the SR performance. In addition, it introduces a noise schedule to control the sampling speed and noise strength for the SR process flexibly.

Specifically, we leverage the pre-trained diffusion model checkpoints released by ResShift authors. We adopt a single or a combination of downsampling techniques introduced in the previous section to corrupt the sample images and perform SR operation on the downsampled images. For different downsampling processes, we also utilize different SR options provided by ResShift. In all, we use two different upscaling options for different downsampling processes:

- **Real-World Super Resolution (realsr)**: Vanilla downsampling and the combination of downsampling and compression artifacts.
- **Image Deblurring (deblur)**: Deblurs a Gaussian blur.

3.3. CNN Feature Similarity

After we complete the upsampling process, we hope that upsampled images maintain the key features for correct diagnosis. To this end, we compare features from original images and upsampled images extracted by a pretrained ResNet50.

In `similarity.py`, we use the TensorFlow library to load in a pre-trained CNN (ResNet50) model that extracts features from images in two folders: one folder with the original images and the other with upsampled images from the ResShift pipeline. A cosine similarity matrix is then constructed from these features (by taking the dot product of normalized feature vectors), and images from the upsampled images folder are matched to images from the original images folder based on calculated similarity scores from the matrix. After this process completes, the number of correct matches is counted based on the file names (if image A is matched to image B based on similarity score and both have the same file name, then the images were correctly matched).

3.4. CNN Disease Classification

In addition to feature similarity comparison, we design a custom CNN and train it on the original dataset and their true labels which can then be used to classify our upsampled and downsampled datasets. The motivation behind this process is to determine the significance of super-resolution in classification performance, supplementing the importance of super-resolution in real medical diagnosis settings. The CNN architecture consists of three convolutional layers with ReLU activations, each followed by 2x2 max-pooling layers for feature extraction and downsampling, culminating in a fully connected layer with 64 neurons and an output softmax layer for classification. It also supports feature extraction by returning flattened features before the dense layers, enabling reuse for tasks like clustering or similarity measurements.

3.5. Additional Evaluation Metrics

To evaluate the visual quality of the images, we also utilize the following four metrics:

- **Structural Similarity Index (SSIM)**: SSIM measures the perceived image change using visual indicators like luminance, constrast, and structural information [7].
- **Peak Signal-to-Noise Ratio PSNR**: PSNR calculates the peak signal-to-noise ratio between two images, and the higher the PSNR the better the quality of the reconstructed image [3].
- **CLIP Image Quality Assessment (CLIPIQA)**: CLIP-IQA is a metric that measures the perceived quality of images based on the CLIP neural network model and semantic similarity [6].
- **Multi-Scale Image Quality Transformer (MUSIQ)**: MUSIQ is a similarity model using multi-scale image representations to capture image quality at different scales [2].

4. Results

4.1. Diffusion-based Super Resolution on Medical Images

To understand how well existing diffusion-based SR methods work on medical images, we first perform different downsampling processes (e.g. downsampling, blurring, and compression) on a set of medical images through our preprocessing pipeline. We then adopt a ResShift model, which is pre-trained on general images (e.g. ImageNet), to upsample or deblur the downsampled medical images.

To evaluate the performance of restoration, we start with some qualitative examples in 1, Figure 2, and Figure 3. We observe that compression operation has a relatively large impact on the quality of downsampled images. In Figure 3, the compressed image (left) loses a large amount of fine-grained information, which is hard to store even after performing super-resolution operation (right). Conversely, vanilla downsampling 1 appears to have the least information loss, with the reconstructed image (right) being visually and showing a reasonable amount of details.

To further understand the quantitative performance, we include evaluations in three folds. We first evaluate the visual image quality of 1,250 upsampled images from NIH Chest X-Ray using four different metrics as introduced in Section 3.5. The results are shown in Table 1. Specifically, we run 2 metrics that require reference images (SSIM and PSNR) as well as 2 reference-free metrics (CLIPQA and MUSIQ) to comprehensively assess the image quality. Similarly, we observe that the compressed images in general are harder to achieve high image quality after super-resolution operation. The vanilla $4\times$ downsampling process has the least information loss and the image quality can be well preserved after upscaling.

We observe that compression operation has a relatively large impact on the quality of downsampled images. Therefore, we further conducted experiments on feature similarity with two quality levels: 10% and 80%, and the downsampled images are stored in "reconstructed.zip" and "reconstructed_light_80.zip" respectively.

Based on the feature extraction and similarity matrix from "similarity.py", 1226/1250 images from "reconstructed_light_80" were matched correctly with an average similarity score of 0.956. In regard to the images from "reconstructed", 575/1250 images were matched correctly with an average similarity score of 0.875. Conceptually, these results make sense because the downsampling process was not as extreme for the images in "reconstructed_light_80" compared to "reconstructed".

Finally, we trained our CNN on 800 original NIH images and their true labels, and tested our model on 1250 reconstructed images ("reconstructed_light_80") and 816 downsampled images (blurred by a Gaussian kernel of size 25, downsampled by a factor of 2, and introduced with variable noise). We saw that "reconstructed_light_80" achieved 0.635 accuracy, while the downsampled images achieved only 0.269 accuracy on disease classification. Based on the disease classification results from our own CNN, we observed that upsampled medical images perform significantly better than low-resolution medical images in accurate diagnosis. Our model was also used for feature extraction and similarity calculation, and performed relatively similar to ResNet feature extraction with a average similarity score of 0.931 on "reconstructed_light_80", reaffirming the effectiveness of our super-resolution pipeline.

4.2. Technical Discussion

In Table 1, we can observe that SSIM scores remain very high across all downsampling techniques. One possible explanation is SSIM focuses more on coarse-level structural patterns rather than exact pixel-wise matches. As medical images have high contrast-to-noise ratios and clear contours, these fundamental structures are preserved after our downsampling processes, which eventually cause high SSIM scores.

Comparing Table 1 and the similarity matching scores, we observe that when the compression quality is higher, both visual quality metrics and similarity score will be improved. With heavy compression artifacts, the restoration of HR images becomes much more difficult, causing lower image quality scores, similarity scores against original images, as well as classification accuracy. However, we also discover some inconsistencies across different visual quality metrics in Table 1. For example, even though the restoration quality of Gaussian Blur (the second row) is higher than $4\times$ Downsample (the first row) as shown by PSNR scores, the reference-free metrics (CLIPIQA and MUSIQ) indicate the opposite findings. As we are evaluating the quality of medical images, their structures are quite different from those used to pre-train the image encoders of these metrics. This can further affect the faithfulness of the scores in Table 2. To identify whether the upsampled images can be employed for unbiased diagnosis, rigorous evaluation metrics, such as our similarity scores, and classification accuracies on the upsampled images would be more robust choices of metrics to utilize in safety-sensitive scenarios.

5. Conclusion

We have designed and implemented a preprocessing pipeline to degrade NIH Chest X-Ray images [4] and utilized a state-of-the-art diffusion-based SR model, ResShift [10], to restore them. Our findings show that while the SR model can reconstruct many details, some corruptions still pose challenges. In medical settings where the quality of images is important for diagnostic purposes, having hundred-percent accuracy is critical because the tiniest of mistakes

could result in extremely unfavorable outcomes. We demonstrated that our pipeline was able to effectively upsample medical images, however, it could be argued that further regulation is necessary in regard to the medical use-cases of this pipeline because our model does not produce a one-to-one or hundred-percent accurate image conversion, especially with severe image corruptions. One potential future direction is to perform fine-tuning on the SR diffusion model with medical images, which could enhance the domain-specific understanding of medical image structures and further improve the restoration quality. This work demonstrates the possibility of diffusion-based SR models in enhancing medical images, providing the foundation for future work in AI-assisted medical diagnosis.

References

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 2

[2] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: multi-scale image quality transformer. *CoRR*, abs/2108.05997, 2021. 2

[3] MathWorks. Psnr - peak signal-to-noise ratio. https://www.mathworks.com/help/vision/ref/psnr.html, 2024. Accessed: 2024-12-14. 2

[4] National Institutes of Health. Nih chest x-ray dataset. https://www.kaggle.com/datasets/nih-chest-xrays/data/data, 2017. Accessed: 2024-12-14. 1, 3

[5] Carl F. Sabottke and Bradley M. Spieler. The effect of image resolution on deep learning in radiography. In *Radiology: Artificial Intelligence*, 2020. 1

[6] PyTorch Lightning Team. Torchmetrics - clip-iqa metric. https://lightning.ai/docs/torchmetrics/stable/multimodal/clip_iqa.html, 2024. Accessed: 2024-12-14. 2

[7] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 2

[8] Zubair Shah Waqar Ahmad, Hazrat Ali and Shoaib Azmat. A new generative adversarial network for medical images super resolution. In *Scientific Reports*, 2022. 1

[9] Koki Yamashita and Konstantin Markov. Medical image enhancement using super resolution methods. In *Computational Science – ICCS 2020*, 2020. 1

[10] Zongsheng Yue, Jianyi Wang, and Chen Change Loy. ResShift: efficient diffusion model for image super-resolution by residual shifting. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2, 3

Appendix

Team contributions

Yen Chu "similarity.py", excluding the command-line parsing and the proprietary CNN. Worked on the Abstract, Section 3.3, 4.1, parts of 3.4 and the conclusion.

Michelle Liu Developed the downsampling pipeline and did analysis and preprocessing of the dataset ("downsampling.py"). Worked on parts of the report including sections 3.1, 3.4, and 5.

Jessica Wan Built disease classification pipeline with training data and labels preparation portion of similarity.py, with custom CNN stored in model.py and cnn.py. Added command-line arguments for similarity comparison. Worked on Introduction, Related Work, Section 3.4, and part of Section 3.5.

Zilai Zeng Built upsampling pipeline from downsampled images using diffusion-based SR ("corrupt_to_restore.py"), and set up visual quality metrics calculation ("calc_metrics.py"). Worked on parts of Abstract, Section 3.2, Section 4 and Conclusion.

Powerpoint Link

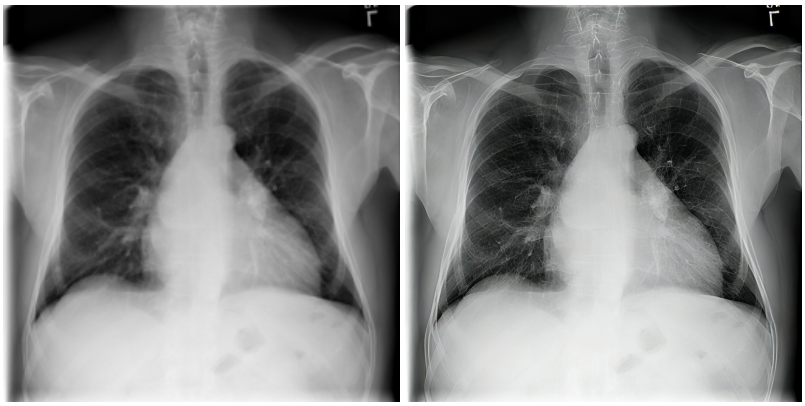

Figure 1. *Left:* 4× Downsampled Image (256 × 256) *Right:* Reconstructed Image (1024 × 1024).

Table 1. Quantitative Results on Reconstructed Image Quality

Corruption	SSIM (↑)	PSNR (↑)	CLIPIQA (↑)	MUSIQ (↑)
4× Downsample	0.99	78.6	0.61	58.6
Gaussian Blur (w/ Kernel Size 25)	0.99	79.5	0.19	22.5
4× Downsample + Compression (w/ 10% Quality)	0.99	75.2	0.526	33.6
4× Downsample + Compression (w/ 80% Quality)	0.99	77.1	0.54	44.7

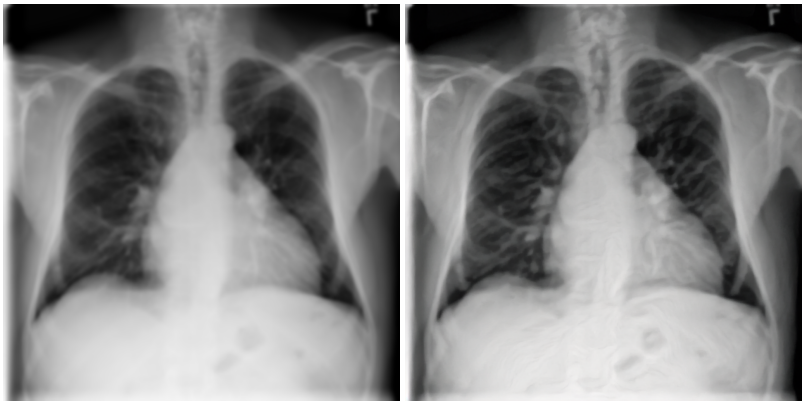

Figure 2. *Left:* Blurred Image with Gaussian Kernel Size 25 (1024 × 1024) *Right:* Reconstructed Image (1024 × 1024).

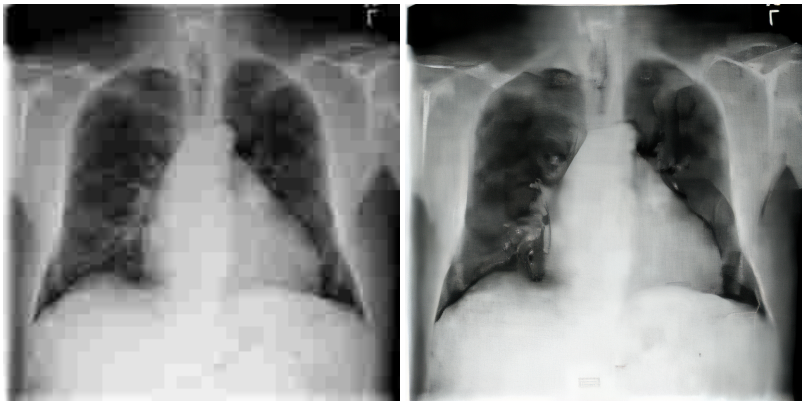

Figure 3. *Left:* 4× Downsample + Compression with 10% Quality (256 × 256) *Right:* Reconstructed Image (1024 × 1024).