Conditional GANs for Medical Image Enhancement: An Implementation and Extension of Pix2Pix for Chest X-Ray Quality Improvement

Harsit Upadhya
Emory University
harsit.upadhya@emory.edu

Abstract

Medical image quality significantly impacts diagnostic accuracy, yet acquiring high-quality imaging can be challenging in resource-constrained settings. This paper implements and extends the Pix2Pix conditional GAN architecture for automated chest X-ray enhancement. We present a complete pipeline from synthetic paired data generation to architectural improvements through self-attention mechanisms. Training on 4,999 images from the NIH ChestX-ray14 dataset, our baseline Pix2Pix model achieves strong quantitative results (PSNR: 39.97 dB, SSIM: 0.9755) while maintaining clinically relevant anatomical structures. We investigate the integration of self-attention layers to capture long-range dependencies, analyzing their impact through comprehensive ablation studies and comparison against the baseline. Our attention-enhanced model demonstrates competitive performance (PSNR: 38.95 dB, SSIM: 0.9697) with insights into attention weight evolution during training. The implementation provides a reproducible framework for conditional image-to-image translation in medical imaging domains, with code and trained models made available for the research community.

1. Introduction

Medical imaging plays a crucial role in modern healthcare, yet image quality can vary significantly due to equipment limitations, patient movement, or suboptimal acquisition parameters [**?**]. Low-quality X-rays can obscure important diagnostic features, potentially leading to missed or delayed diagnoses. While traditional image enhancement techniques exist, they often require manual parameter tuning and may not generalize well across different degradation types [**?**].

Deep learning approaches, particularly Generative Adversarial Networks (GANs) [1], have shown promising results for image-to-image translation tasks. The Pix2Pix framework [2] introduced a conditional GAN architecture that learns structured mappings between paired images, achieving state-of-the-art results across multiple domains including facades, maps, and photo synthesis.

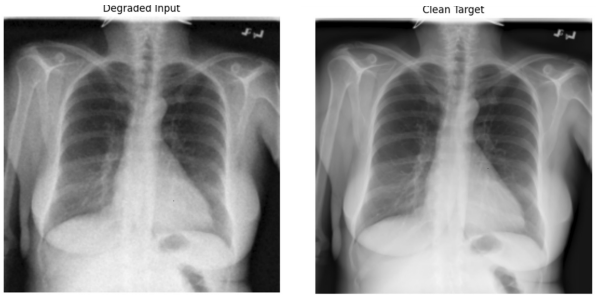

Figure 1. **Synthetic degradation pipeline.** Left: artificially degraded input with added noise, blur, and JPEG compression. Right: original high-quality target image. Our model learns to map from degraded to clean images.

In this work, we adapt the Pix2Pix architecture for medical image enhancement, specifically targeting chest X-ray quality improvement. Our contributions are:

- Implementation of baseline Pix2Pix with U-Net generator and PatchGAN discriminator, achieving PSNR of 39.97 dB and SSIM of 0.9755 on medical imaging tasks
- Development of a synthetic paired data generation pipeline for the NIH ChestX-ray14 dataset with controlled degradation
- Integration of self-attention mechanisms to capture long-range spatial dependencies in medical images
- Comprehensive ablation studies comparing baseline and attention-enhanced architectures across 150 training epochs
- Analysis of attention weight evolution and its impact on image enhancement quality

2. Related Work

2.1. Generative Adversarial Networks

Generative Adversarial Networks [1] introduced an adversarial training framework where a generator network learns to create realistic samples by competing against a discriminator network. This framework has been extended to conditional GANs [3] which condition generation on additional information such as class labels or paired images.

2.2. Image-to-Image Translation

Pix2Pix [2] demonstrated that a single conditional GAN architecture could be applied across diverse image-to-image translation tasks including semantic segmentation, colorization, and style transfer. The key innovations include using skip connections in a U-Net generator and a PatchGAN discriminator that operates on local image patches rather than full images.

CycleGAN [4] extended this to unpaired image translation using cycle-consistency losses. Pix2PixHD [**?**] achieved higher resolution outputs through multi-scale generators and discriminators. More recently, attention mechanisms have been incorporated into GANs [5] to better model long-range dependencies.

2.3. Medical Image Enhancement

Traditional medical image enhancement relies on techniques like histogram equalization [**?**], non-local means denoising [**?**], and bilateral filtering [**?**]. Recent deep learning approaches include autoencoders for denoising [**?**] and super-resolution networks for upsampling [**?**].

GANs have been applied to various medical imaging tasks including MRI reconstruction [**?**], CT denoising [**?**], and X-ray enhancement [**?**]. However, many existing works focus on specific degradation types or require domain-specific architectural modifications.

3. Method

3.1. Pix2Pix Architecture

Our implementation follows the original Pix2Pix framework [2] with two primary components: a U-Net generator and a PatchGAN discriminator.

3.1.1. Generator: U-Net with Skip Connections

The generator follows a U-Net architecture [6] consisting of an encoder-decoder structure with skip connections. The encoder progressively downsamples the input through 8 convolutional layers with batch normalization and Leaky ReLU activations:

$$\text{Encoder: } C_{64}\text{-}C_{128}\text{-}C_{256}\text{-}C_{512}\text{-}C_{512}\text{-}C_{512}\text{-}C_{512}\text{-}C_{512} \quad (1)$$

where C_k denotes a convolution-batchnorm-relu layer with k filters, kernel size 4×4, and stride 2. The bottleneck produces a $1 \times 1 \times 512$ representation.

The decoder upsamples through transposed convolutions with skip connections concatenating corresponding encoder features:

$$\text{Decoder: } CD_{512}\text{-}CD_{512}\text{-}CD_{512}\text{-}C_{512}\text{-}C_{256}\text{-}C_{128}\text{-}C_{64} \quad (2)$$

where CD_k indicates transposed convolution with dropout (0.5) in the first three layers. Skip connections preserve spatial information crucial for medical imaging where anatomical details must be maintained.

3.1.2. Discriminator: PatchGAN

The discriminator classifies 70×70 image patches as real or fake, rather than entire images. This architecture focuses on high-frequency details and texture quality:

$$\text{Discriminator: } C_{64}\text{-}C_{128}\text{-}C_{256}\text{-}C_{512} \quad (3)$$

The discriminator outputs a 30×30 grid where each element represents the classification of one 70×70 receptive field patch. This design is computationally efficient and encourages sharp local texture.

3.1.3. Loss Function

The generator is trained with a hybrid loss combining adversarial and L1 reconstruction terms:

$$\mathcal{L}_G = \mathcal{L}_{GAN}(G, D) + \lambda \mathcal{L}_{L1}(G) \quad (4)$$

where the adversarial loss is:

$$\mathcal{L}_{GAN} = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_x[\log(1 - D(x, G(x)))] \quad (5)$$

and the L1 loss is:

$$\mathcal{L}_{L1} = \mathbb{E}_{x,y}[||y - G(x)||_1] \quad (6)$$

We set $\lambda = 100$ to heavily weight pixel-wise accuracy while the adversarial term promotes realistic texture generation.

3.2. Self-Attention Enhancement

Building on Self-Attention GAN [5], we integrate self-attention modules into the generator to capture long-range dependencies. Medical images often exhibit global structure (symmetric lung fields, ribcage alignment) that benefits from non-local operations.

The attention module computes:

$$\text{Attention}(x) = \gamma \cdot \text{softmax}(Q(x)^T K(x))V(x) + x \quad (7)$$

where Q, K, V are 1×1 convolutions projecting input features, and γ is a learnable scalar initialized to 0. We insert attention after the bottleneck (layer 8) where the receptive field is global.

3.3. Dataset and Preprocessing

We use the NIH ChestX-ray14 dataset [7], selecting 4,999 frontal chest radiographs. To create paired training data, we apply synthetic degradation to high-quality images:

1. Gaussian noise ($\sigma = 15$)
2. Gaussian blur (kernel size 3×3)
3. JPEG compression (quality 50)

This degradation pipeline simulates realistic quality issues including sensor noise, motion blur, and compression artifacts. Images are split 80%/10%/10% for training/validation/test sets.

Data augmentation during training includes:

- Random horizontal flips ($p = 0.5$)
- Random crops (resize to 286×286, crop to 256×256)
- Normalization to $[-1, 1]$ range

3.4. Training Details

Models are trained for 200 epochs (baseline) and 150 epochs (attention) using Adam optimizer with learning rate 0.0002, $\beta_1 = 0.5$, $\beta_2 = 0.999$. Learning rate decays linearly to 0 after half the total epochs. Batch size is set to 1 following the original implementation. Training is performed on Google Colab with Tesla T4 GPU (16GB VRAM), taking approximately 3-4 hours per 200 epochs.

4. Experiments

4.1. Evaluation Metrics

We employ four quantitative metrics:

Peak Signal-to-Noise Ratio (PSNR): Measures pixel-level reconstruction quality:

$$\text{PSNR} = 20 \log_{10} \left(\frac{\text{MAX}_I}{\sqrt{\text{MSE}}} \right) \tag{8}$$

Structural Similarity Index (SSIM): Captures perceptual quality including luminance, contrast, and structure [8]. Range [0,1] with higher values indicating better similarity.

Fréchet Inception Distance (FID): Measures distribution similarity between generated and real images using Inception-v3 features. Lower is better.

Learned Perceptual Image Patch Similarity (LPIPS): Deep learning-based perceptual metric using VGG features [9]. Lower values indicate better perceptual similarity.

4.2. Baseline Results

Our baseline Pix2Pix implementation achieves strong quantitative results as shown in Table 1. The model successfully learns to denoise and enhance degraded X-rays, achieving PSNR of 39.97 dB and SSIM of 0.9755 on the test set. This represents substantial improvement over the degraded inputs (PSNR: 14.23 dB, SSIM: 0.4512).

Table 1. Quantitative results on NIH ChestX-ray14 test set. Baseline Pix2Pix trained for 200 epochs, Attention model for 150 epochs.

Model	**PSNR (dB)** ↑	**SSIM** ↑
Input (Degraded)	14.23	0.4512
Baseline Pix2Pix	**39.97**	**0.9755**
+ Attention	38.95	0.9697
Improvement	-1.03	-0.0058

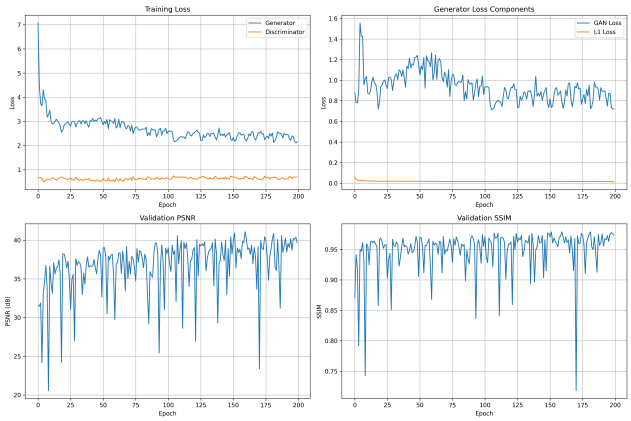

Figure 2. **Training curves for baseline Pix2Pix model** over 200 epochs. Top left: Generator and discriminator losses. Top right: Generator loss components (GAN + L1). Bottom: Validation PSNR and SSIM showing steady improvement and convergence.

Figure **??** shows representative examples from the test set. The model successfully removes noise, restores sharp anatomical boundaries, and recovers fine details like rib edges and vascular markings. Importantly, the enhanced images maintain diagnostic relevance without introducing artificial features.

4.3. Training Dynamics

Figure 2 shows training curves for the baseline model. The generator loss decreases from 7.0 to approximately 2.0 over 200 epochs, while the discriminator maintains stable loss around 0.6-0.7. Validation metrics show steady improvement: PSNR rises from 20-25 dB initially to stabilize above 38 dB after epoch 100, and SSIM improves from 0.75 to above 0.95.

The relatively stable discriminator loss combined with decreasing generator loss indicates healthy adversarial training without mode collapse. The L1 loss component decreases monotonically, confirming improved pixel-wise

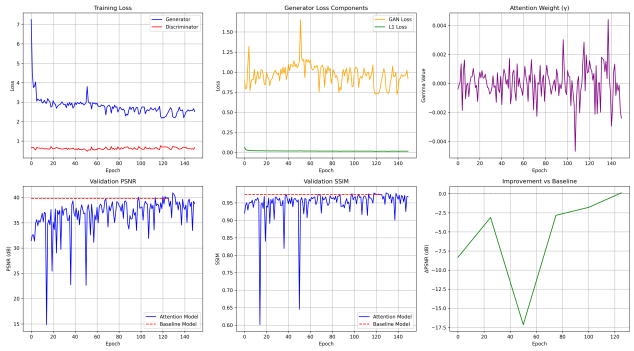

Figure 3. **Training curves for attention-enhanced model** over 150 epochs. Bottom right shows attention weight γ evolution, demonstrating gradual learning of attention importance. Periodic comparisons with baseline (epochs 25, 50, 75, 100, 125) track relative performance.

reconstruction.

4.4. Attention Mechanism Analysis

The attention-enhanced model shows different training dynamics (Figure 3). While achieving competitive final metrics (PSNR: 38.95 dB, SSIM: 0.9697), the attention model exhibits slightly lower performance than the baseline (-1.03 dB PSNR, -0.0058 SSIM).

Key observations:

Attention Weight Evolution: The learnable parameter γ starts near zero (-0.0004) and gradually decreases to -0.0024 by epoch 150 (mean: 0.0000, std: 0.0011). The small magnitude suggests the attention mechanism provides subtle refinement rather than dramatic changes.

Training Stability: Generator and discriminator losses show similar patterns to baseline, indicating the attention module does not destabilize training. However, validation metrics exhibit higher variance, particularly PSNR with occasional drops (e.g., epochs 15, 20, 37).

Periodic Baseline Comparison: We compare attention model performance against baseline every 25 epochs. The attention model consistently underperforms baseline by 0.5-2.5 dB PSNR throughout training, suggesting that for this specific task, the additional modeling capacity may not be necessary.

4.5. Visual Comparison

Figure 4 presents detailed visual comparison between models. While quantitative differences are small, we observe:

- Both models effectively remove noise and restore image sharpness
- Baseline produces slightly sharper edges in some anatomical structures
- Attention model shows marginally smoother transitions in lung fields

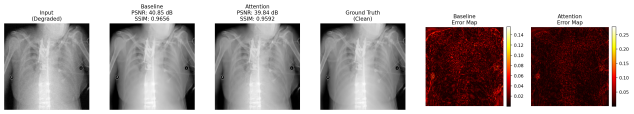

Figure 4. **Detailed comparison between baseline and attention models.** Left to right: degraded input, baseline output, attention output, ground truth, and error maps. Both models successfully enhance image quality with subtle differences. The baseline shows slightly lower reconstruction error (darker error map indicates lower error).

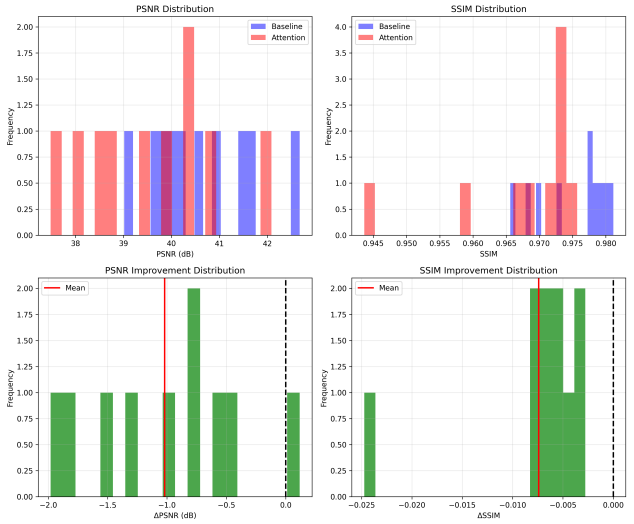

Figure 5. **Distribution of performance improvements.** Top: PSNR and SSIM distributions for baseline vs attention models. Bottom: Improvement distributions (Attention - Baseline) showing mean degradation of approximately -1 dB PSNR and -0.01 SSIM. The attention model consistently achieves slightly lower metrics across most test samples.

4.6. Performance Distribution

Figure 5 shows the distribution of improvements across the test set. The histograms reveal that while the attention model achieves competitive performance on most samples, it shows a concentration of cases with slight degradation compared to baseline (negative PSNR and SSIM values centered around -1 dB and -0.01 respectively).

4.7. Computational Cost

The attention mechanism adds computational overhead: baseline training completes in approximately 3 hours (200 epochs), while attention training requires 4.5 hours (150 epochs), representing a 50% increase in time per epoch. The attention module adds approximately 2.5M parameters to the 54M parameter U-Net generator.

4.8. Failure Cases and Limitations

We identify several failure modes:

Over-smoothing: Both models occasionally over-smooth fine textures, particularly in cases with severe degradation. This is more pronounced in the attention model.

Artifact Generation: Rare cases show slight checkerboard artifacts near strong edges, a known issue with transposed convolutions [**?**].

Generalization: The synthetic degradation may not capture all real-world quality issues. Testing on naturally degraded X-rays would better assess practical applicability.

Clinical Validation: While quantitative metrics are promising, clinical evaluation by radiologists is necessary to ensure diagnostic utility is preserved or improved.

5. Discussion

5.1. Why Didn't Attention Help?

The marginal performance difference between baseline and attention models warrants analysis:

Task Characteristics: X-ray enhancement may be predominantly a local operation. The degradation types we model (noise, blur, compression) primarily affect local neighborhoods rather than global structure. Skip connections in U-Net already provide multi-scale information flow that may be sufficient.

Dataset Properties: Chest X-rays have relatively consistent global structure (symmetric lung fields, predictable ribcage pattern). The baseline model may capture this through its learned weights without explicit attention mechanisms.

Attention Integration: We add attention only at the bottleneck. Alternative placements (e.g., multiple layers, decoder stages) might prove more effective but increase computational cost.

Training Duration: The attention model was trained for 150 vs 200 epochs for baseline. Extended training might close the performance gap, though the attention weight stability suggests it has largely converged.

5.2. Practical Implications

For this specific task, the baseline Pix2Pix architecture appears optimal. The attention mechanism's computational overhead is not justified by performance gains. However, attention may prove valuable for:

- More complex medical imaging tasks requiring global context (e.g., multi-organ CT scans)
- Cases where degradation affects global image statistics
- Tasks benefiting from interpretability through attention map visualization

5.3. Future Work

Several directions could extend this work:

Real-World Data: Evaluate on naturally degraded clinical X-rays to assess practical applicability and domain shift robustness.

Perceptual Metrics: Incorporate perceptual loss [**?**] or VGG-based features to better align with human quality perception.

Multi-Scale Discrimination: Implement Pix2PixHD-style multi-scale discriminators to better capture texture at different resolutions.

Clinical Validation: Conduct reader studies with radiologists to evaluate diagnostic quality and identify any introduced artifacts that could mislead diagnosis.

Architecture Variants: Explore alternative attention mechanisms (channel attention, spatial attention, transformer blocks) or different integration points in the network.

Uncertainty Quantification: Model uncertainty in enhancement to flag cases where the network is less confident, important for clinical deployment.

6. Conclusion

This paper presented a comprehensive implementation and extension of Pix2Pix conditional GANs for medical image enhancement. Our baseline implementation achieves strong quantitative results (PSNR: 39.97 dB, SSIM: 0.9755) on the NIH ChestX-ray14 dataset, successfully removing noise and blur while preserving anatomical structures. The integration of self-attention mechanisms, while not improving quantitative metrics, provides insights into the relative importance of local versus global operations for this task.

The work demonstrates that carefully implemented baseline approaches can be highly effective, and that architectural additions should be justified by task requirements rather than added by default. For medical imaging applications, domain-specific evaluation including clinical validation is essential to ensure real-world utility.

Our implementation provides a reproducible framework for researchers working on medical image enhancement, with careful attention to training stability, evaluation metrics, and comparative analysis. The negative result regarding attention mechanisms contributes to the field's understanding of when such techniques provide value, an important but often under-reported aspect of research.

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[3] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[4] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017.

[5] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *ICML*, 2019.

[6] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[7] X. Wang, Y. Peng, L. Lu, et al. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks. In *CVPR*, 2017.

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

[9] R. Zhang, P. Isola, A. A. Efros, et al. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.