Face Sketch-to-Photo Generation with Pix2Pix

Project overview:

Our baseline follows Pix2Pix which is a conditional GAN trained with paired sketch–photo data. The generator we used is a U-Net with skip connections and the discriminator is a 94x94 PatchGAN that scores local realism over image patches rather than one global score.
Input : A 2D grayscale face sketch image.
Output: A realistic grayscale face photo corresponding to the sketch, generated by the Pix2Pix model.

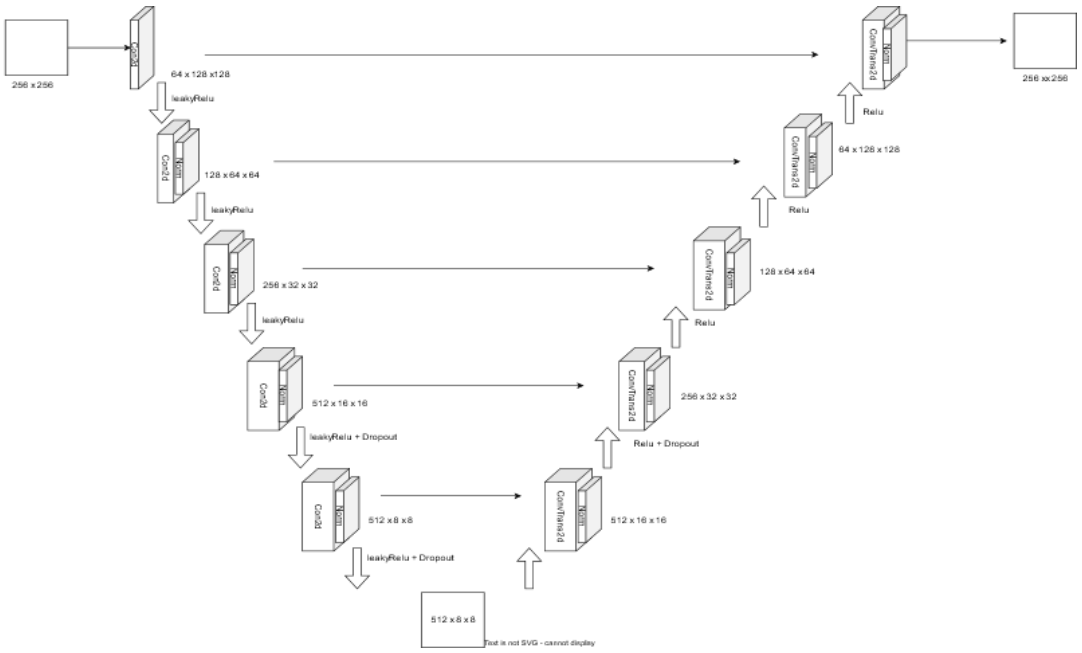
Dataset Details:

- Used CUHK Face Sketch Dataset with 606 sketch- image pairs
- Out of the total 606 available image pairs only 188 images were selected due to quality and suitability for our task.
- All the selected images are converted into grayscale. Several augmentation techniques were applied to the training data like horizontal flip, rotation, brightness adjustment and contrast adjustment.
- After augmentation the total number of samples increased to 1068 which was the used for training.The test dataset was split and kept aside before the data augmentation was done so that the model doesn't learn from the test set

Model Architecture:

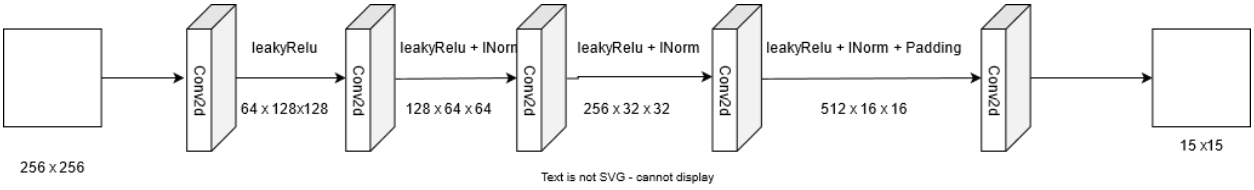
Generator(U-Net):

- Encoder–decoder with skip connections: 64→128→256→512→512 (down), symmetric upsampling with skip concatenations.
- Activation Functions used: LeakyReLU in encoder, ReLU in decoder; final Tanh.



Discriminator(PatchGAN) :

- A convolutional classifier that outputs an N × N map of patch-level realism scores.
- Each 94×94 patch is judged as real or fake and encourages high-frequency detail generation.



Loss Functions:

We have used a combination of Least Squares GAN (LSGAN) loss and L1 reconstruction loss to achieve both realism and structural accuracy.

Discriminator Loss(LSGAN): This loss encourages discriminator to output values close to 1 for real images and 0 for generated ones.

$$L_D = \frac{1}{2}E[(D(x) - 1)^2] + \frac{1}{2}E[(D(G(z)))^2]$$

Generator Adversarial Loss(LSGAN): It encourages the generator to fool the discriminator by producing realistic outputs.

$$L_G^{adv} = \frac{1}{2}E[(D(G(z)) - 1)^2]$$

L1 Reconstruction Loss: L1 loss preserves structural similarity by minimizing the average absolute pixel difference between generated and real images.

$$L_{L1} = E[\|G(s) - y\|_1]$$

Final generator loss:

$$L_G = L_G^{adv} + \lambda \cdot L_{L1}$$

Here, λ controls the tradeoff between the realism and accuracy of the model.

Training Details:

Image Size	-	256×256 (grayscale)
Batch Size	-	10
Optimizer	-	Adam
Learning Rate	-	0.0002
Epochs	-	250
Framework	-	Pytorch

Design Choice:

This project draws inspiration from the research paper "Image-to-Image Translation with Conditional Adversarial Networks". Most of our design choices were a combination of architectural guidelines from the paper and our own experimental trial-and-error adjustments.

Our generator follows a U-Net architecture with skip connections, where each convolutional layer uses Instance Normalization. Even though our batch size is 8,

Instance Normalization normalizes features per individual image rather than across the batch. In the original Pix2Pix paper, a batch size of 1 was used (partly to avoid cross-sample feature leakage and to stabilize GAN training on smaller datasets), but our setup retained Instance Normalization for consistency in normalization behavior across different batch sizes.

For the discriminator, we implemented a PatchGAN with a receptive field of 94×94. The paper notes that receptive fields above 70×70 effectively capture local style and texture while reducing global overfitting. Larger patches can better enforce high-frequency details but increase computational overhead, which we accepted to gain improved local realism.

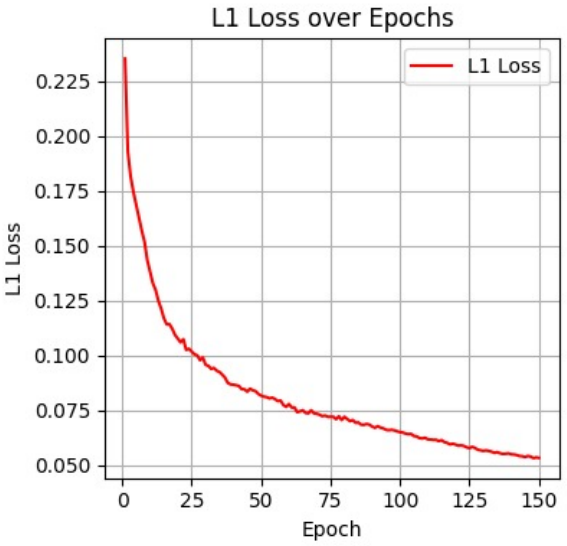
Initially, we trained the model using Vanilla GAN loss (BCEWithLogits) combined with L1 loss. However, we observed that the discriminator quickly became too accurate at distinguishing real from fake samples. This led to vanishing gradients for the generator, reducing its ability to improve. To address this, we switched to Least Squares GAN loss (LSGAN) combined with L1 loss. LSGAN replaces the binary cross-entropy term with a mean squared error formulation, which penalizes samples based on how far they are from the decision boundary rather than only their classification correctness. This provides smoother, non-saturating gradients, enabling the generator to receive stronger and more stable training signals even when the discriminator is strong.

We also experimented with Perceptual Loss by incorporating a pretrained VGG network. This approach compares high-level feature representations of real and generated images, improving texture fidelity and perceptual quality. With perceptual loss, the model achieved good results in fewer epochs, but at the cost of training speed, increasing from ~25 seconds per epoch to ~55 seconds per epoch. Given our aim for a lighter and faster model, we ultimately opted not to include perceptual loss in the final training pipeline.

Evaluation Metrics:

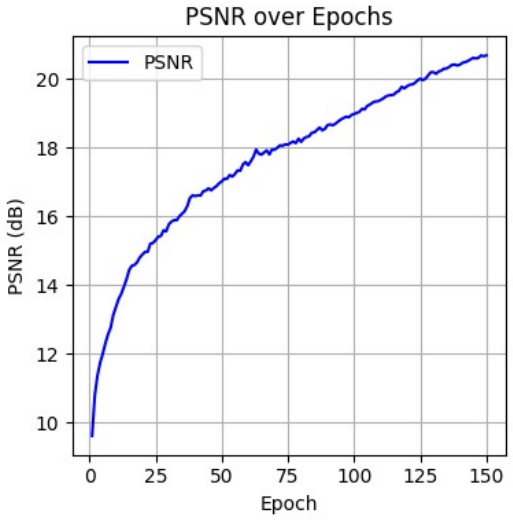
L1 Loss:

- The loss values steadily decreased from 0.23 to 0.045 across 250 epochs.
- Shows a consistent reduction in pixel-wise difference between generated and real images.



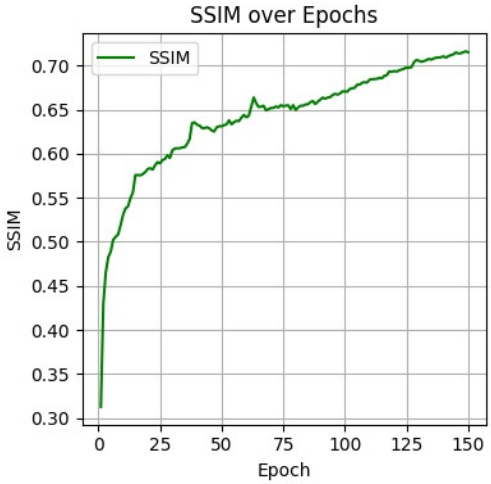
PSNR:

- PSNR has improved from around 10 dB to 22 dB.
- Higher PSNR values indicate better reconstruction quality with lower distortion.



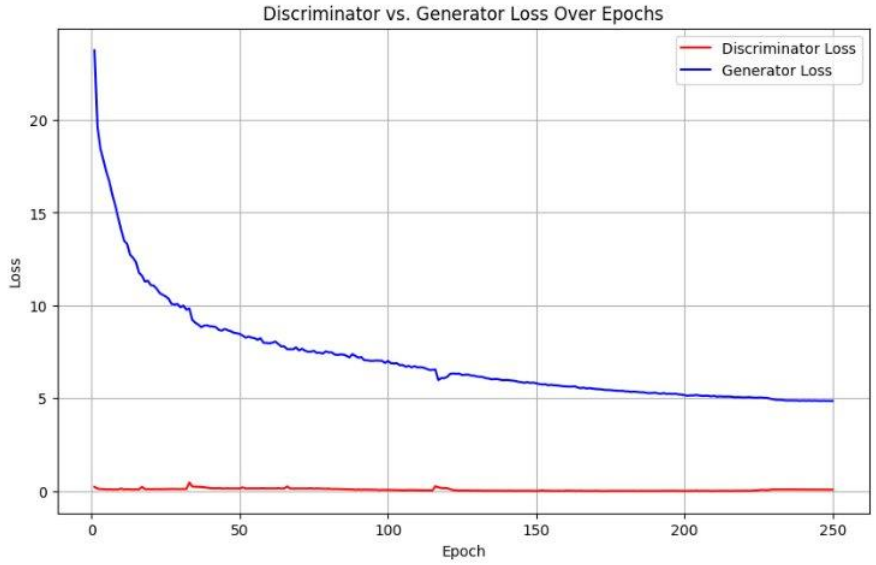
SSIM:

- The graph shows steady increase of SSIM values from 0.30 to 0.73 over the epochs.
- It indicates better preservation of structural and perceptual details.



Metric	Description	Value (Mean)
L1 / MAE	**Measures average pixel-level error**	**0.045**
SSIM	**Structural Similarity Index for perceptual quality**	**0.73**
PSNR	**Peak Signal-to-Noise Ratio for image clarity**	**22 dB**

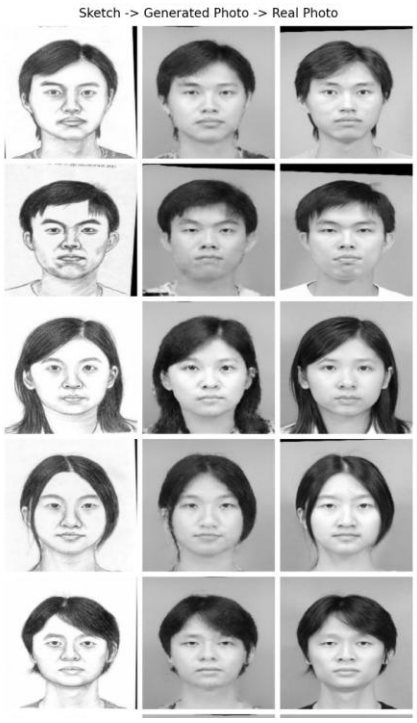
Loss curves over epochs:



Qualitative Results:

For 20 epochs:

For 250 epochs:



Analysis & Observations

- High detail preservation was achieved in high-contrast facial regions such as eyes, eyebrows, and lips.
- Lower accuracy in low-contrast regions (e.g., cheek areas, smooth skin) where the generator produced slightly blurred textures.
- Occasional artifacts/blemishes appeared in background and skin regions, especially in earlier training epochs.
- Switching from BCE to LSGAN loss improved gradient stability, reduced mode collapse, and led to better fine details.
- Data augmentation (flips, rotation, brightness/contrast changes) helped the model generalize better despite the small dataset.
- The PatchGAN discriminator with 94×94 receptive field successfully captured local texture details but increased computational load.
- SSIM improvement from 0.30 to 0.73 shows structural detail preservation, while PSNR improvement from ~10 dB to 22 dB indicates significant reduction in reconstruction distortion.

Conclusion:

The Pix2Pix-based conditional GAN, combined with LSGAN loss and L1 reconstruction, effectively translated grayscale face sketches into realistic grayscale photos. The model successfully preserved overall facial structure while improving perceptual similarity over the course of training. Despite the small dataset size, augmentation strategies and architectural choices allowed for stable convergence and good visual quality. However, limitations remain in fine texture realism and robustness to low-contrast regions.

Future Work

- Colored Image Generation: Extend the model to produce photorealistic colored outputs from sketches.
- Multi-Modal Conditioning: Incorporate textual descriptions or additional modalities (e.g., depth maps) to guide generation.
- High-Resolution Outputs: Train on higher-resolution images for sharper detail reproduction.
- Enhanced Training Stability: Explore advanced GAN stabilization techniques such as spectral normalization, TTUR, or progressive growing.
- Style and Attribute Control: Add controls for facial attributes (e.g., hair style, expression) to make generation more flexible.

References

[1]Image-to-Image Translation with Conditional Adversarial Networks
Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros-
https://doi.org/10.48550/arXiv.1611.07004