Super-Resolution for Ahmedabad Satellite Imagery using Custom Fusion based Generator

A report submitted by

Bhargav Pipaliya
Enrollment No: 221040011003

Department of Computer Engineering
Institute of Infrastructure Technology Research and Management
Ahmedabad, Gujarat-380026

December 2025

Declaration

I, the undersigned, hereby declare that this Report is my own original work. All information and ideas from other sources have been properly cited and acknowledged. The work presented is free from plagiarism and fabrication. I understand that any violation of academic integrity may result in disciplinary action by the Institute.

Place: Ahmedabad

Date:

Signature:

Bhargav Pipaliya (221040011003)

Approval Sheet

This report, titled “Super-Resolution for Ahmedabad Satellite Imagery using Custom Fusion based Generator” is submitted by Bhargav Pipaliya (Enrollment No: 221040011003) in partial fulfillment of the requirements for the Bachelor of Technology degree.

Place:

Date:

Name of the supervisor:

Signature:

The committee reviewing this report confirms that the report titled “Super-Resolution for Ahmedabad Satellite Imagery using Custom Fusion based Generator” authored by Bhargav Pipaliya, meets the academic and research standards required for submission to the Department of Computer Engineering at the Institute of Infrastructure Technology Research and Management, Ahmedabad.

Name of the examiner-1: Professor Ashish Soni

Affiliation: Dept. of Computer, IITRAM, Ahmedabad

Signature with date:

Acknowledgements

I would like to express my sincere gratitude to my report supervisor, Professor Ashish Soni, for his invaluable guidance, insightful feedback, and constant encouragement throughout this research. His mentorship was instrumental in shaping this project and my approach to problem-solving. I am also thankful to the Department of Computer Engineering at IITRAM, Ahmedabad, for providing the necessary computational resources and a supportive academic environment. My appreciation also goes to my friends and colleagues. Their stimulating discussions and support provided fresh perspectives on the challenges encountered during this work. Finally, I owe a deep sense of gratitude to my family for their unwavering belief and un-conditional support throughout my academic journey. This work would not have been possible without them.

Dedication

This work is dedicated to the Almighty, whose blessings provided the strength to complete this research. I also dedicate this report to my supervisor, Professor Ashish Soni, for his exceptional guidance and unwavering support, which were key to the success of this project. Finally, this work is dedicated to the researchers and scientists in the field of satellite image enhancement. Their contributions have inspired and motivated this exploration.

Abstract

High-quality satellite imagery is essential for applications ranging from regional infrastructure development to urban planning. However, publicly available data often suffers from low resolution, limiting its practical utility. This thesis explores the use of a novel deep learning technique to enhance these images. Specifically, it focuses on implementing and evaluating a Custom Fusion Based Generator on a custom-curated Ahmedabad dataset derived from the ISRO Bhoonidhi Portal. The project involved rigorous dataset preparation using QGIS to create precise high-resolution pairings and designing a unique architecture that fuses Residual-in-Residual Dense Blocks (RRDB) with Multi-Head SEAttention mechanisms. The quality of the enhanced images was measured against standard benchmarks, including SRGAN and Real-ESRGAN. The results show that the proposed fusion model significantly improves upon state-of-the-art methods. It achieved a remarkable Average PSNR of 37.41 dB and an SSIM of 0.9477. This indicates a superior degree of accuracy and structural fidelity in the generated images compared to existing models. This study demonstrates that custom fusion-based architectures are powerful tools for maximizing the quality of regional satellite data for precise scientific analysis.

Keywords: Super-resolution, Fusion GAN, SEAttention, Ahmedabad dataset, ISRO Bhoonidhi, Google Earth pro, deep learning, Image Enhancement, PSNR, SSIM.

1 Introduction

Satellite imagery is a fundamental asset for urban planning and environmental analysis, yet the scarcity of high-resolution data often restricts its effectiveness for precise regional monitoring. This project addresses this limitation by developing a specialized super-resolution pipeline for the Ahmedabad region. A critical component of this work was the construction of a custom dataset, beginning with the acquisition of 10m resolution imagery from the ISRO Bhoonidhi Portal. Due to the massive scale of the original data (approximately 50,000 * 50,000 pixels), the block was segmented into 200 smaller units using QGIS. These low-resolution (LR) blocks served as reference coordinates to manually acquire corresponding 5m High-Resolution (HR) imagery from Google Earth Pro. To address the geometric mismatch between the rectangular Google Earth exports and the square ISRO blocks, the QGIS Georeferencer tool was employed to align and crop the data, resulting in a precise dataset of 400 paired LR and HR images. To further optimize the pipeline and resolve Out-Of-Memory (OOM) errors during training, the HR images was converted into the Cloud Optimized GeoTIFF (COG) format. While initial experiments with standard models like SRGAN and ESRGAN and Real-ESRGAN yielded suboptimal results for this specific terrain, the proposed Custom Fusion Based Generator integrating Residual-in-Residual Dense Blocks with Multi-Head Attention mechanisms successfully overcame these challenges, delivering superior structural fidelity and significantly higher PSNR and SSIM scores.

2 Dataset Preparation

The creation of a high-quality, region-specific dataset was a foundational element of this research. Unlike standard benchmark datasets, real-world satellite imagery often faces challenges related to massive file sizes, varying resolutions, and geometric misalignments. This section details the rigorous pipeline established to curate the custom Ahmedabad dataset, ensuring precise pixel-to-pixel correspondence between Low-Resolution (LR) and High-Resolution (HR) pairs.

2.1 Data Acquisition and Initial Partitioning

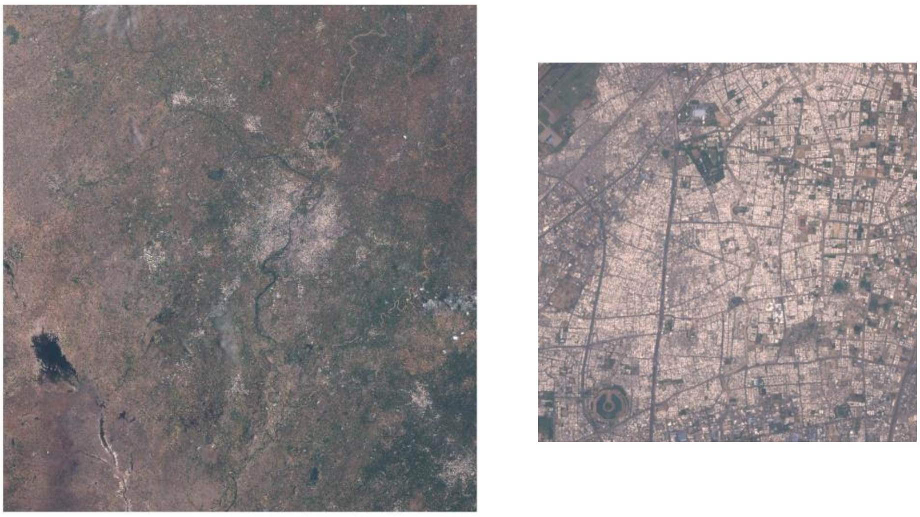


Figure 1: (Left) Original 50,000×50,000-pixel Sentinel-2A/2B(10m) sensor block of the Ahmedabad region Downloaded from the ISRO Bhoonidhi Portal. (Right) the small tile cut-

down from (Left) Ahmedabad block by QGIS tool.

The primary Low-Resolution (LR) data was sourced from the ISRO Bhoonidhi Portal, specifically targeting the Ahmedabad region. The raw data consisted of Sentinel-2A/2B imagery with a spatial resolution of 10 meters. The initial download presented a significant computational challenge as it arrived as a massive single block, approximately 50,000 * 50,000 pixels in size (Figure 1: Left side). Processing such a large file directly was infeasible, necessitating a partitioning strategy. Using QGIS (Figure 1: Right side). This large raster was spatially segmented into 200 smaller, manageable square blocks. This step was essential not only to make the file sizes tractable but also to isolate specific regions of interest, which facilitated the precise acquisition of corresponding ground truth data in subsequent steps.

2.2 High-Resolution Ground Truth Generation

To train the Super-Resolution model effectively, "Ground Truth" or High-Resolution (HR) target images were required for each of the 200 LR blocks. We utilized Google Earth Pro as the primary source for this data. A key advantage of this approach was accessibility; while commercial high-resolution satellite datasets are often prohibitively expensive or unavailable for academic research, Google Earth Pro provides free access to high-quality, time-stamped historical imagery. This allowed us to bypass the financial barrier of purchasing proprietary datasets while still securing the necessary High-Resolution targets.

The target resolution for these HR images was set at 5 meters, providing a 2x super-resolution factor over the source 10m data. For the acquisition process, each of the 200 LR tiles served as a geospatial reference. The corresponding areas were manually located in Google Earth Pro and exported at maximum clarity to ensure the highest possible fidelity for model training.

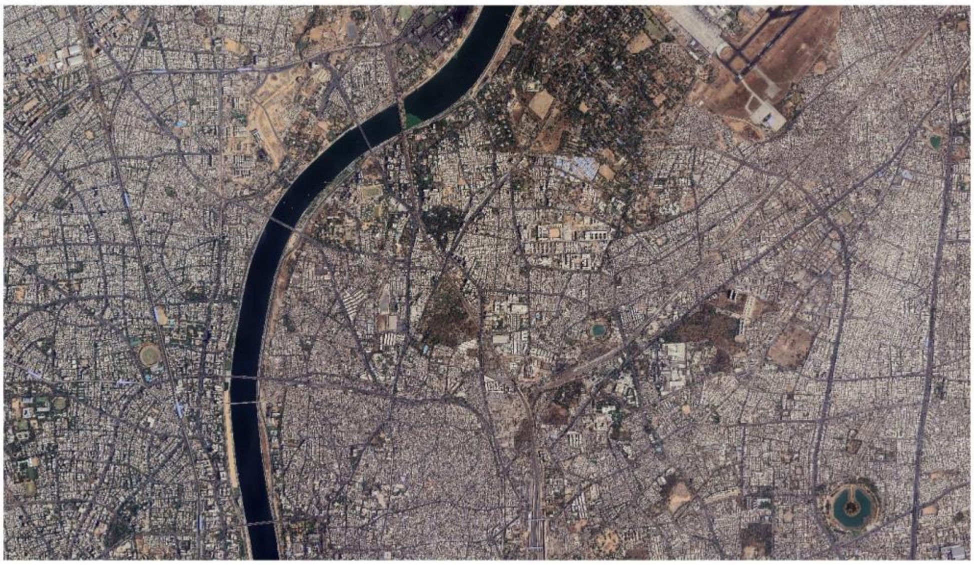


Figure 2: Captured Image from Google Earth Pro based on LR tile (Figure 1).

2.3 Geometric Alignment and Georeferencing

A critical technical challenge arose during the HR acquisition phase due to formatting discrepancies. Google Earth Pro exports images in a rectangular viewport format(Figure 2), whereas the original ISRO 10m blocks were perfectly square. This mismatch in aspect ratio and geospatial alignment would have prevented the neural network from learning accurate spatial features. To resolve this, the QGIS Georeferencer Tool was employed to align the datasets. Key distinct features, such as road intersections and building corners, were identified as control points in both the 10m ISRO tile and the 5m Google Earth export. A transformation algorithm was then applied to warp and crop the rectangular Google Earth image, ensuring it perfectly matched the geometry and spatial extent of the square 10m tile. This process yielded 200 perfectly aligned pairs where the LR input and HR target shared identical geospatial boundaries.

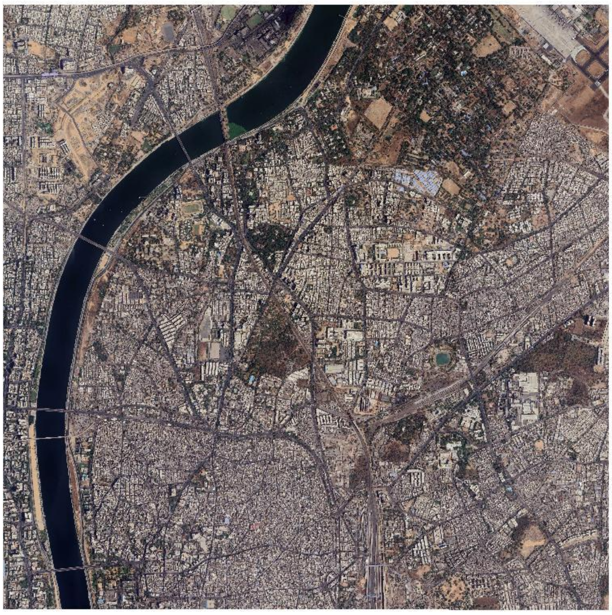


Figure 3: Geometrically corrected and cropped 5m HR tile aligned to the 10m LR input.

2.4 Optimization: Cloud Optimized GeoTIFF (COG)

During the initial training phases, the pipeline encountered significant Out-Of-Memory (OOM) errors. Standard image formats, such as standard TIFF or PNG, typically require the entire file to be loaded into memory, which overwhelmed the system RAM when processing batches of satellite data. To mitigate this, all 400 images (200 LR and 200 HR) were converted into the Cloud Optimized GeoTIFF (COG) format. This conversion was crucial because COG allows for efficient "tiled" reading of data, enabling the data loader to access specific chunks of the image file without loading the entire raster into RAM. This optimization completely resolved the OOM issues and stabilized the training pipeline.

2.5 Final Dataset Specification:

The final curated dataset consists of 400 images forming 200 aligned LR-HR pairs, specifically optimized for the unique terrain of Ahmedabad. This custom dataset serves as the benchmark for evaluating the proposed Fusion Based Generator against standard models. Here after macking complete dataset we will train all models which includes (SRGAN, ESRGAN, Real-ESRGAN, Real-ESRGAN with Self-Attention and Our custom fusion-based generator model).

3. Architecture Design and Methodology

3.1 What is a GAN?

A Generative Adversarial Network (GAN) is a deep learning framework composed of two neural networks that contest with each other in a game-theoretic scenario. The first network, the Generator, creates synthetic data (in this case, images) from random noise or low-resolution inputs. The second network, the Discriminator, acts as a binary classifier that attempts to distinguish between "real" data from the training set and "fake" data produced by the Generator. As training progresses, the Generator learns to produce increasingly realistic images to fool the Discriminator, while the Discriminator becomes better at spotting fakes. This adversarial process results in the generation of highly detailed and realistic imagery.

3.2 The Foundation: ESRGAN

For this project, we utilize the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN). Originally designed for super-resolving natural images (such as photographs of people or objects), ESRGAN improves upon standard GANs (like SRGAN) by removing artifacts and generating more realistic textures. However, satellite imagery contains complex terrain information that standard natural-image models often miss. Therefore, we have adopted the core ESRGAN framework but significantly modified the Generator architecture to better handle the unique spatial and spectral characteristics of remote sensing data.

3.3 How the Custom ESRGAN is Built?

The proposed model consists of three critical components: the Discriminator, the modified Fusion-Based Generator, and the fundamental building block known as the RRDB.

3.3.1 The Discriminator Network

The Discriminator acts as the "judge" in this system. Instead of a simple black box, it follows a specific VGG-style structure designed to process inputs.

- Input Layer: Accepts either the Real HR image or the Generated SR image.
- Feature Extraction: Utilizes SN-Conv2d (Spectral Normalization) layers to prevent gradient explosion and stabilize training.
- Activation Function: Uses Leaky ReLU to allow a small gradient when the unit is not active, preventing "dead neurons".
- Dimensionality Reduction: Applies a Flatten layer to convert the 2D feature maps into a 1D vector.
- Classification Head:
    1. Linear Dense (1024): A fully connected layer with 1,024 neurons to interpret high-level features.
    2. Linear Dense (1): The final output neuron that produces the probability score

(Real vs. Fake).

3.3.2 The Fusion-Based Generator Network

The Generator is the core of our innovation. It is designed with a dual-branch architecture to separately process semantic and structural features before fusing them.

A. Input Stage

- Input Data: Receives the Low-Resolution (LR) satellite tile.
- Multi-Scale Splitting: The input is simultaneously converted into three parallel Resolutions for: the 30m branch, the 20m branch, and the 10m branch.

B. The Upper Branch (30m Context - Coarse Features)

- Convolution: Initial feature extraction.
- Deep Feature Extraction: Passes through a block of 16 RRDBs to capture semantic context.
- Upsample: Increases the spatial resolution of the feature maps.
- SE Attention: Applies Self Attention to weigh the importance of different feature channels, helping the model focus on relevant spectral bands to learn image structure.

C. The Middle Branch (20m Context - Medium Features)

- Convolution: Initial feature extraction.
- Deep Feature Extraction: Passes through a block of 16 RRDBs.
- Convolution: Applies a standard convolution layer to refine features without altering the spatial resolution.

D. The Lower Branch (10m Context - Fine Features)

- Convolution: Initial feature extraction.
- Deep Feature Extraction: Passes through a block of 16 RRDBs to capture high-frequency details.
- Downsample: Reduces the spatial resolution, forcing the network to compress information and focus on dense structural features.
- Spatial Attention: Applies attention maps to emphasize specific spatial locations, such as edges and boundaries.

E. Fusion and Reconstruction Stage

- Concatenate: Merges the outputs from all three branches (30m, 20m, and 10m) into a single, rich feature set It generally convert 3D Input into 1D Output.
- Conv 1x1: Reduces the channel dimensionality of the concatenated features to make processing efficient.
- Spatial Multihead SA: A Self-Attention mechanism that captures long-range dependencies across the entire image, ensuring texture consistency.
- Conv 3x3: Further smooth and integrates the features.
- Upsample x2 (Stage 1): The first step of image enlargement.

- Upsample x2 (Stage 2): The second step of image enlargement to reach the final target resolution.
- Conv + PReLU: Final feature adjustment layer using Parametric ReLU activation for flexibility.
- Tanh: The final activation function that normalizes the output pixel values to the range [-1, 1] for the generated Super-Resolution (SR) image.

3.3.2 The RRDB Block

The fundamental building unit of our Generator is the Residual-in-Residual Dense Block (RRDB).

- Structure: A complex block containing three dense blocks linked by a "residual" connection.
- Internal Layers: Composed of tightly connected Convolution layers and Leaky ReLU activations.
- Key Feature: Removes Batch Normalization (BN) to preserve the natural contrast of satellite imagery and reduce artifacts.

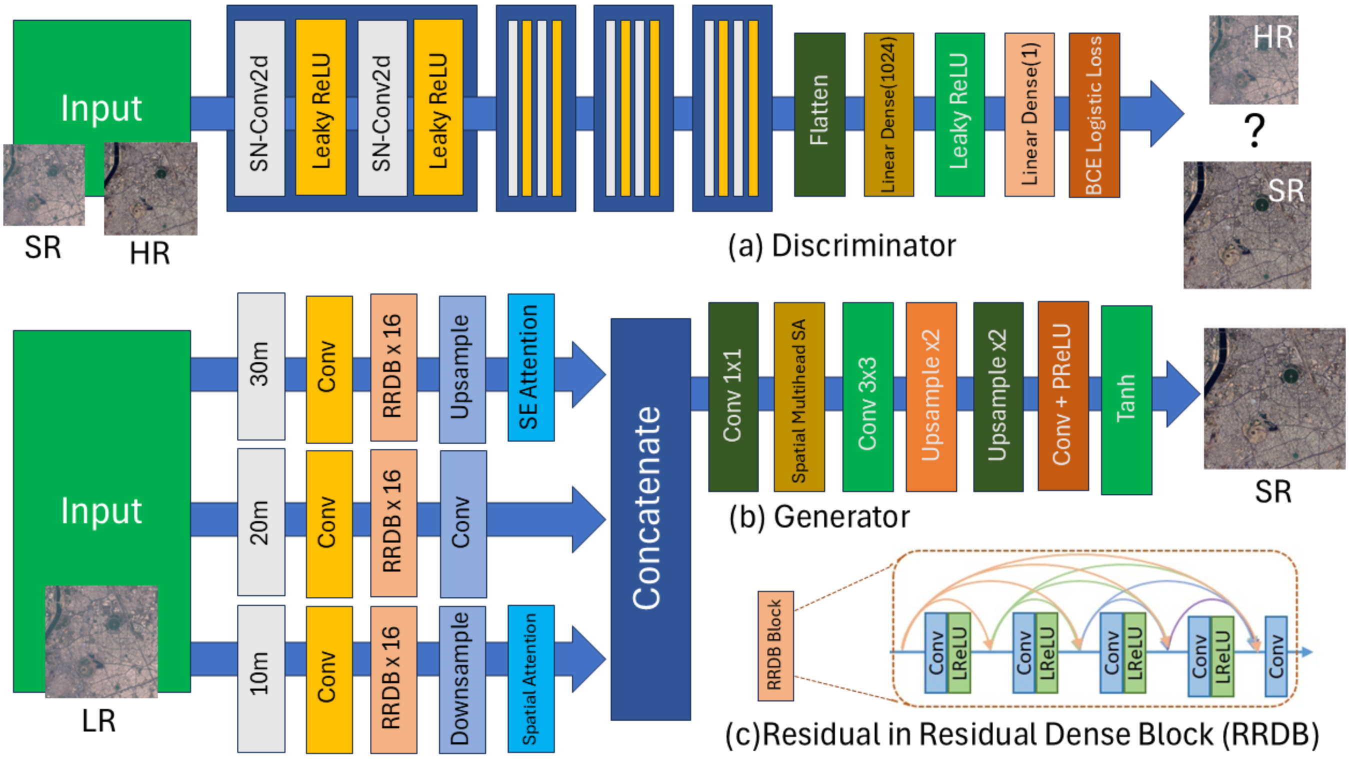


Figure 3: Complete Architecture of the Proposed Fusion Based GAN. (a) The Discriminator network utilizing Spectral Normalization (SN-Conv2d) for stable adversarial training. (b) The proposed Generator featuring a triple-branch multi-scale architecture (30m, 20m, 10m) utilizing RRDBs, Attention mechanisms, and Multi-Head Self-Attention. (c) The internal structure of the Residual-in-Residual Dense Block (RRDB)

3.4 How ESRGAN Learns (Loss Functions)

To make images sharper and more real, ESRGAN uses a smart mix of “loss functions.” These are like scores that tell the model how well it is doing and what it needs to improve.

3.4.1 Perceptual Loss

This score makes sure the generated image has similar high-level patterns and details as the original clear image. It uses a pre-trained VGG network to check features before the activation layers, which gives a richer comparison. This is very important for making textures look real.

3.4.2 Adversarial Loss

This score pushes the Generator to create images that are so real-looking that the Discriminator cannot tell them apart from actual high-resolution images. This helps bring back small details and textures. The Relativistic Average GAN (RaGAN) loss is defined as:

$$L^{Ra}_{D} = -E_{x_r}[\log(D_{Ra}(x_r, x_f))] - E_{x_f}[\log(1 - D_{Ra}(x_f, x_r))]$$

Where x_r is a real image, x_f is a fake (generated) image, and D_{Ra} is the relativistic discriminator’s output.

3.4.3 Content Loss

This score makes sure that the generated image is very close to the original clear image, pixel by pixel. ESRGAN often uses L1 loss, which tends to make images less blurry compared to the L2 loss used in older models. It provides the basic structure for the clearer image.

$$L_{\text{content}} = |G(I_{LR}) - I_{HR}|_1$$

Where $G(I_{LR})$ is the generated super-resolved image, I_{LR} is the low-resolution input, I_{HR} is the ground-truth high-resolution image, and $|| \cdot ||_1$ denotes the L1 norm.

All these scores are combined to guide the Generator. The total loss for the Generator is a weighted sum of the perceptual loss, adversarial loss, and content loss. This smart combination helps ESRGAN avoid making images too smooth, makes them sharper, and improves their textures.

4. Experimental Results and Comparative Analysis

4.1 Quantitative Evaluation

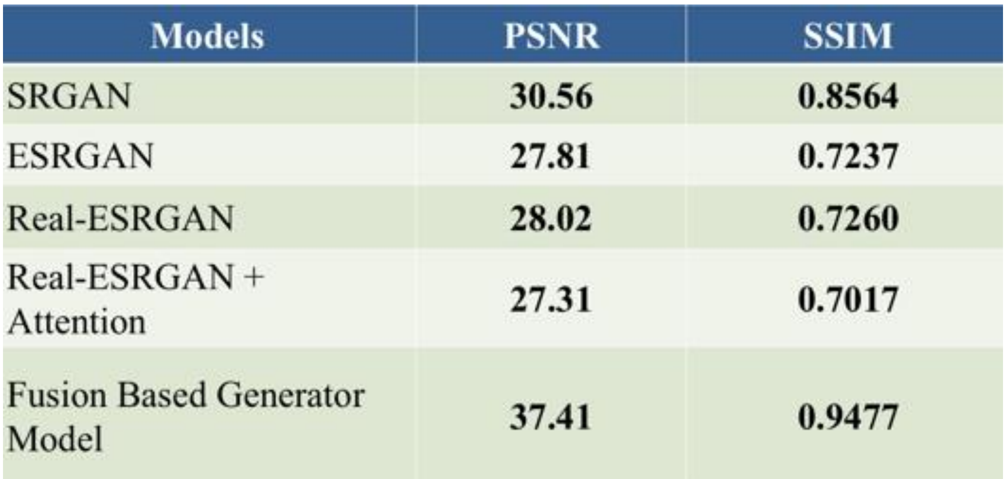
In the Evaluation we calculate two standard quantitative metrics: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM).

PSNR (Peak Signal-to-Noise Ratio): This metric measures the ratio between the maximum possible power of a signal (image) and the power of corrupting noise. A higher PSNR value typically indicates better image reconstruction quality.

SSIM (Structural Similarity Index**):** Unlike PSNR, which measures absolute pixel errors, SSIM measures the perceived change in structural information (luminance, contrast, and structure). It ranges from -1 to 1, where a value of 1 indicates perfect structural identity with the ground truth.

4.2 Performance Comparison

We benchmarked our Fusion-Based model against several state-of-the-art super-resolution models: SRGAN, standard ESRGAN, Real-ESRGAN, and Real-ESRGAN augmented with Attention mechanisms. The results on the custom Ahmedabad dataset are summarized below:



Models	PSNR	SSIM
SRGAN	**30.56**	**0.8564**
ESRGAN	**27.81**	**0.7237**
Real-ESRGAN	**28.02**	**0.7260**
Real-ESRGAN + Attention	**27.31**	**0.7017**
Fusion Based Generator Model	**37.41**	**0.9477**

Table 1: Quantitative training comparison of standard models vs. the proposed Fusion-Based Generator on the custom Ahmedabad dataset.

4.2 Analysis of Results

As evidenced by Table 1, the proposed **Fusion-Based Generator Model** significantly outperforms all baseline models.

PSNR Improvement: Our model achieved a PSNR of **37.41 dB**, which is a substantial improvement over the standard ESRGAN (27.81 dB) and SRGAN (30.56 dB). This indicates that our multi-scale architecture effectively minimizes reconstruction errors.

SSIM Improvement: The most striking improvement is in the SSIM score, where our model achieved **0.9477**. This is nearly a perfect score compared to the 0.72–0.85 range of competing models. This demonstrates that the triple-branch (30m/20m/10m) approach successfully preserves the complex geometric patterns of the Ahmedabad terrain, which standard models often distort.

4.2 Qualitative (Visual) Evaluation

We compared the visual output of the models on a specific tile containing mixed agricultural and urban features.

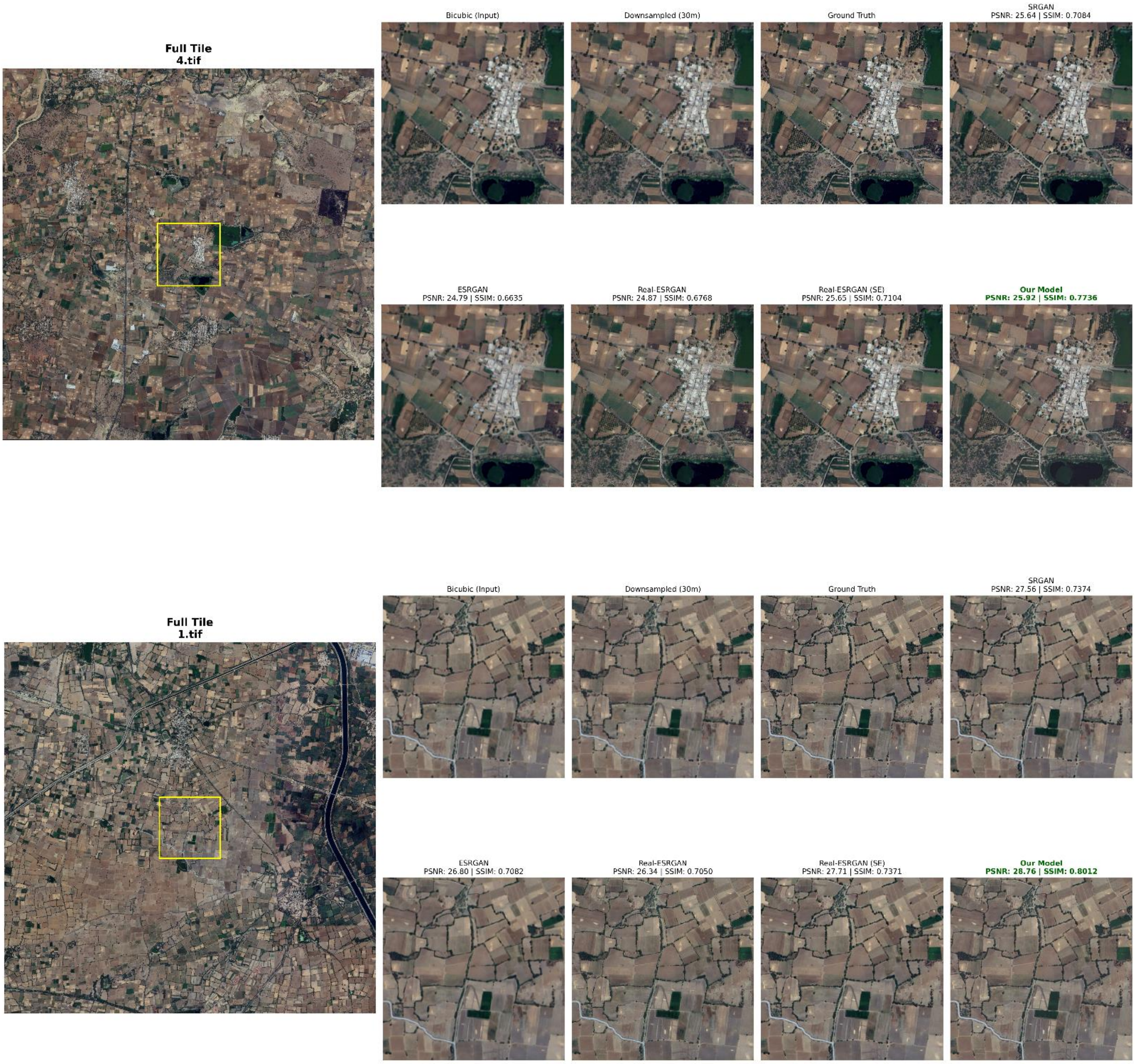

Full Tile
4.tif
Bicubic (Input)
Downsampled (30m)
Ground Truth
SRGAN
PSNR: 25.64 | SSIM: 0.7084
ESRGAN
PSNR: 24.79 | SSIM: 0.6635
Real-ESRGAN
PSNR: 24.87 | SSIM: 0.6768
Real-ESRGAN (SE)
PSNR: 25.65 | SSIM: 0.7104
Our Model
PSNR: 25.92 | SSIM: 0.7736
Full Tile
1.tif
Bicubic (Input)
Downsampled (30m)
Ground Truth
SRGAN
PSNR: 27.56 | SSIM: 0.7374
ESRGAN
PSNR: 26.80 | SSIM: 0.7082
Real-ESRGAN
PSNR: 26.34 | SSIM: 0.7050
Real-ESRGAN (SE)
PSNR: 27.71 | SSIM: 0.7371
Our Model
PSNR: 28.76 | SSIM: 0.8012

Figure 5: Visual comparison between Ground Truth, Bicubic interpolation, and various Super-Resolution models on a test tile.

4 Future Work

While this project shows the strong performance of the Fusion-Based Generator for satellite image enhancement, several areas remain for future improvement. Hyperparameters like learning rate and the weights λ and η can be further tuned to boost PSNR and SSIM. Using more realistic degradation models—such as atmospheric effects, sensor noise, or motion blur—would make the system more robust to real-world data. Exploring Transformer-based designs like SwinIR may also capture longer-range details better than the current architecture. Expanding the dataset with sources such as Sentinel-2, PlanetScope, or WorldView would improve generalization. For real-time use, techniques like quantization and pruning could enable efficient edge deployment. Finally, creating a simple GUI or QGIS plugin would make the tool accessible to non-technical users.

References

[1] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, C. C. Loy, Y. Qiao, and X. Tang, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, Munich, Germany, 2018.

[2] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4681-4690, 2017.

[3] Z. 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*, vol. 13, no. 4, pp. 600-612, 2004.

[4] N. C. Rakotonirina and A. Rasoanaivo, "ESRGAN+: Further Improving Enhanced Super-Resolution Generative Adversarial Network," *arXiv preprint arXiv:2001.08073*, Jan. 2020.

[5] Open Geospatial Consortium, "OGC Cloud Optimized GeoTIFF Standard," OGC Document 21-026, 2023.