Satellite Image Super-Resolution Using Multimodal Data

Lalit Kumar[1], Pawan Kumar[2], Nitin Kumar Das[3], and Amber Gupta[4]

Indian Institute of Technology Ropar, Punjab, India

Abstract. Satellite image resolution plays a crucial role in a wide range of applications, including agriculture, environmental monitoring, and urban planning. However, the low spatial resolution of freely available satellite data, such as Sentinel-2, limits its effectiveness for high-precision tasks. Enhancing satellite image resolution is essential to enable better decision-making, particularly in agriculture, where detailed spatial and spectral information is critical for crop health monitoring and yield estimation.

Solving this problem is challenging due to the need to simultaneously preserve spatial detail and spectral consistency across multiple image bands. Traditional upscaling techniques often fail to meet these dual requirements, especially when dealing with multispectral imagery.

In this study, we address these challenges using deep learning-based super-resolution models, specifically SwinIR, ESRGAN, and Real-ESRGAN. These models are applied to a dataset of 3,000 image pairs consisting of low-resolution Sentinel-2 images and high-resolution PlanetScope images, focused on agricultural fields and land areas. The goal is to generate visually sharper and spectrally accurate high-resolution outputs suitable for downstream geospatial analysis.

Additionally, we propose a novel caption-based image generation approach, where enriched captions, generated from models based on OpenAI's CLIP and Llama, are used to guide a GAN architecture for enhancing low-resolution satellite images. To explore alternative enhancement techniques, we also implement ARISGAN[2], an artistic super-resolution model, and train ESRGAN specifically for our dataset. These models are evaluated based on key image quality metrics, including PSNR, SSIM, and SAM, to assess their effectiveness in improving both spatial and spectral qualities of satellite imagery.

Keywords: Super-resolution · Satellite imagery · Deep learning · Agriculture · SwinIR[3] · ESRGAN[5]

1 Introduction

Satellite imagery serves as a fundamental data source for various domains, including agriculture, climate monitoring, urban development, and disaster management. However, a major limitation of widely accessible satellite platforms like Sentinel-2 is their relatively low spatial resolution, which restricts their applicability in tasks that demand fine-grained detail, such as precision farming or crop health analysis. To address this, we explore the use of deep learning-based super-resolution (SR) techniques for enhancing satellite images while maintaining their spectral fidelity.

In this work, we have implemented and evaluated super-resolution models, specifically SwinIR[3] and ESRGAN[5], on paired low-resolution Sentinel-2 and high-resolution PlanetScope satellite imagery. Our goal is to reconstruct high-resolution images that are visually sharp and spectrally consistent, enabling better performance in agricultural and environmental analysis tasks.

This problem is important because high-resolution satellite data is expensive and not always available, particularly for developing regions. Enhancing freely available low-resolution imagery can bridge this gap and empower data-driven decision-making in critical sectors like agriculture. However, applying SR techniques to satellite data introduces multiple challenges: handling multi-spectral inputs, ensuring spectral preservation, managing noise and atmospheric interference, and dealing with the domain gap between natural images and satellite imagery.

To overcome these issues, we tried to adopted different pipelines that integrates multimodal datasets and employs deep learning models originally designed for natural image SR. We evaluate the models on both spatial and spectral metrics such as PSNR, SSIM, and SAM to determine their effectiveness and identify which methods perform best for real-world satellite applications.

2 Related Works

Super-Resolution Techniques

Image super-resolution (SR) aims to reconstruct high-resolution (HR) images from low-resolution (LR) counterparts, and has become a vital area of research in computer vision, particularly in domains such as medical imaging, surveillance, and remote sensing. Traditional interpolation methods such as bicubic and Lanczos filtering often result in oversmoothed outputs with limited detail. In contrast, learning-based approaches—especially those leveraging deep convolutional neural networks (CNNs)—have demonstrated superior performance in capturing complex patterns and textures for more realistic and accurate image reconstruction.

SwinIR

SwinIR [3] is a transformer-based super-resolution framework built upon the Swin Transformer architecture. It introduces a hierarchical feature representation with shifted window mechanisms, which significantly enhance the model's capacity to capture long-range dependencies while maintaining computational efficiency. SwinIR has shown remarkable performance on various image restoration tasks, including super-resolution, denoising, and JPEG artifact removal. Its window-based attention mechanism makes it especially suitable for tasks requiring global context, such as satellite image enhancement.

ESRGAN

Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) [5] is an advancement over SRGAN, aiming to produce perceptually more convincing images. ESRGAN introduces several improvements, including a Residual-in-Residual Dense Block (RRDB) to enhance feature representation, and a relativistic discriminator that focuses on the relative realism between generated and ground-truth images. ESRGAN has become a widely adopted benchmark in perceptual SR, especially in scenarios where sharpness and visual fidelity are prioritized.

Real-ESRGAN

Real-ESRGAN [4] extends the ESRGAN architecture to handle real-world image degradation, which is often more complex and less uniform than synthetic downscaling. It adopts a more generalized degradation model during training and introduces several modifications to improve robustness, including a U-Net discriminator and stronger data augmentation strategies. Real-ESRGAN is particularly useful in practical applications where LR images may contain noise, compression artifacts, or non-ideal downscaling patterns, such as satellite or drone imagery.

These approaches have each contributed to the advancement of super-resolution, with distinct strengths. SwinIR excels in global feature modeling, ESRGAN in perceptual quality, and Real-ESRGAN in real-world degradation robustness. Our work builds on this foundation to evaluate their applicability in enhancing agricultural and land-monitoring satellite imagery.

ARISGAN

ARISGAN[2], an open-source deep learning model designed to create high-resolution satellite images from lower-resolution ones. Although ARISGAN was originally made to improve Arctic satellite images, we apply it to enhance Sentinel-2 RGB images (10 m resolution) to the finer 3 m resolution of PlanetScope imagery. ARISGAN uses a type of neural network that learns how to add realistic details, helping us generate sharper and more detailed images for our study.

Multi Modal Models

Recent advancements in super-resolution (SR) have largely focused on enhancing image quality using deep learning models like ESRGAN [5], which improve visual fidelity but often struggle with preserving both spatial detail and spectral consistency. In recent years, multimodal approaches have emerged, combining textual information with visual data to guide the SR process. Text-guided SR models and CLIP-based models leverage textual descriptions to provide semantic context, improving both image quality and relevance. Our approach builds on these methods by integrating CLIP-generated captions with Sentinel-2 images in a GAN-based framework, enabling more accurate and contextually enriched high-resolution outputs for satellite image enhancement

3 Dataset Description

To train and evaluate our super-resolution models, we utilized two satellite imagery datasets with different spatial and spectral characteristics: Sentinel-2 and PlanetScope. These datasets were selected for their complementary features — Sentinel-2 provides free access to multispectral imagery at moderate resolution, while PlanetScope offers high-resolution imagery with comparable spectral bands.

3.1 Sentinel-2

Sentinel-2 is a multispectral satellite mission developed by the European Space Agency (ESA) as part of the Copernicus Program. It consists of two satellites (Sentinel-2A and 2B) that capture Earth's surface at high revisit frequencies (5 days at the equator) and in 13 spectral bands ranging from the visible to shortwave infrared (SWIR). For this work, we focused on the 30-meter resolution bands:

- Band 2 (Blue) – 490 nm
- Band 3 (Green) – 560 nm
- Band 4 (Red) – 665 nm

3.2 PlanetScope

PlanetScope is a constellation of small satellites operated by Planet Labs. It provides daily imagery at a spatial resolution of approximately 3 meters per pixel, with four spectral bands:

- Band 2 (Blue) – 490 nm
- Band 3 (Green) – 560 nm
- Band 4 (Red) – 665 nm

Fig. 1 is an example of Sentinel low-resolution imagery compared to PlanetScope high-resolution ground truth:

3.3 Data Preparation

To prepare the dataset for model training:

To ensure alignment, we collected paired Sentinel-2 and PlanetScope images captured over the same geographic regions and periods.

Images were georeferenced and resampled to a common grid for pixel-wise correspondence.

The dataset used in this project consists of paired satellite imagery, designed to support agricultural analysis and super-resolution tasks:

- **Low-Resolution Source:** Sentinel-2 imagery with 3 spectral bands (RGB) at resolutions varying from 10m to 60m.

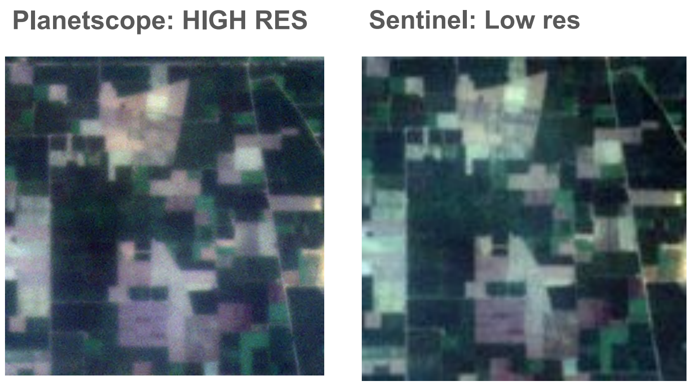


Fig. 1. Sentinel low-resolution image vs PlanetScope high-resolution ground truth image.

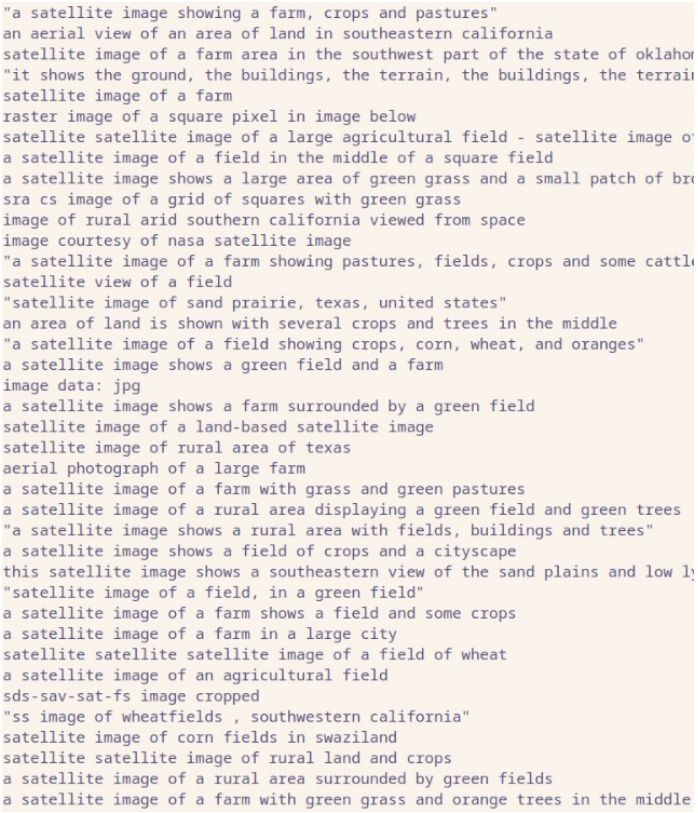
- **High-Resolution Ground Truth:** PlanetScope imagery with 3 corresponding spectral bands at a resolution of 3m.
- **Key Features:**
  - Geographically aligned low- and high-resolution image pairs.
  - Focused on agricultural regions to support crop monitoring and yield prediction.
  - Spectral bands include RGB data critical for vegetation analysis.

This dataset provides a robust foundation for training super-resolution models tailored to agricultural applications, addressing the unique challenges posed by multi-spectral satellite imagery.

4 Objectives

The objectives of this research are as follows:

- To test deep learning-based super-resolution models tailored for satellite images using spectral data.
- To utilize a dataset of 30,000 paired low- and high-resolution agriculture-related images for training and evaluation.
- To preserve spectral accuracy while improving spatial resolution for agricultural applications.
- To identify the most effective super-resolution techniques for satellite imagery enhancement.
- To train the existing architecture using our dataset and evaluating them.
- To propose a new GAN architecture that will combine two modalities; image caption (Low resolution image) and Low resolution image itself to generate the High resolution image.



```
"a satellite image showing a farm, crops and pastures"
an aerial view of an area of land in southeastern california
satellite image of a farm area in the southwest part of the state of oklaho
"it shows the ground, the buildings, the terrain, the buildings, the terrai
satellite image of a farm
raster image of a square pixel in image below
satellite satellite image of a large agricultural field - satellite image o
a satellite image of a field in the middle of a square field
a satellite image shows a large area of green grass and a small patch of br
sra cs image of a grid of squares with green grass
image of rural arid southern california viewed from space
image courtesy of nasa satellite image
"a satellite image of a farm showing pastures, fields, crops and some cattl
satellite view of a field
"satellite image of sand prairie, texas, united states"
an area of land is shown with several crops and trees in the middle
"a satellite image of a field showing crops, corn, wheat, and oranges"
a satellite image shows a green field and a farm
image data: jpg
a satellite image shows a farm surrounded by a green field
satellite image of a land-based satellite image
satellite image of rural area of texas
aerial photograph of a large farm
a satellite image of a farm with grass and green pastures
a satellite image of a rural area displaying a green field and green trees
"a satellite image shows a rural area with fields, buildings and trees"
a satellite image shows a field of crops and a cityscape
this satellite image shows a southeastern view of the sand plains and low l
"satellite image of a field, in a green field"
a satellite image of a farm shows a field and some crops
a satellite image of a farm in a large city
satellite satellite satellite image of a field of wheat
a satellite image of an agricultural field
sds-sav-sat-fs image cropped
"ss image of wheatfields , southwestern california"
satellite image of corn fields in swaziland
satellite satellite image of rural land and crops
a satellite image of a rural area surrounded by green fields
a satellite image of a farm with green grass and orange trees in the middle
```

Fig. 2. Caption generated by Blip model

5 Proposed Methodology

The objective of our study is to evaluate and compare the performance of three state-of-the-art super-resolution models—SwinIR, ESRGAN, and Real-ESRGAN—on satellite imagery enhancement tasks focused on land monitoring and agricultural analysis. Additionally, we aim to develop a new GAN-based architecture for the same purpose.

The methodology adopted for this research is summarized below.

We utilized a dataset comprising 3,000 pairs of corresponding low-resolution and high-resolution satellite images. The low-resolution inputs were derived from Sentinel satellite images, each resized or cropped to a spatial resolution of 128×128 pixels. The high-resolution ground truth images were obtained from PlanetScope imagery, known for its superior spatial and spectral fidelity. All image pairs were carefully selected to cover regions of agricultural fields and land surfaces, ensuring relevance for real-world applications in precision agriculture, crop monitoring, and land use analysis.

For each of the three models, we generated super-resolved (enhanced) images from the low-resolution Sentinel inputs. The outputs were then quantitatively evaluated against their corresponding high-resolution PlanetScope counterparts using the following image quality assessment metrics:

- **Peak Signal-to-Noise Ratio (PSNR)** – to measure reconstruction fidelity.
- **Structural Similarity Index Measure (SSIM)** – to assess perceptual image quality.
- **Spectral Angle Mapper (SAM)** – to evaluate spectral integrity.

All metrics were computed on a per-image basis and then averaged across the 3,000 samples to ensure robust and generalizable results. This evaluation strategy enables a comprehensive performance comparison across models, taking into account both spatial and spectral aspects of image quality.

This methodological pipeline ensures a standardized and fair evaluation for all models and reflects practical considerations relevant to satellite-based agricultural monitoring and land observation applications.

A new proposed pipeline tries to integrates text-guided conditioning into a GAN-based super-resolution framework to enhance Sentinel-2 (10m) imagery to PlanetScope (3m) resolution. First, Sentinel-2 and PlanetScope image pairs are preprocessed into tensor format, accompanied by enriched textual captions generated using a hybrid approach combining OpenAI's CLIP [6] and Llama [1] models. These captions (e.g., "agricultural fields with radial irrigation patterns and adjacent rural settlements") provide semantic context for guiding the enhancement process. The architecture modifies SRGAN with a dual-branch generator: one branch processes CLIP-based text embeddings (512D projected to 256D) that are concatenated with visual features, while the other employs 16 residual blocks and PixelShuffle layers for 4× spatial upsampling. Training leverages multi-GPU optimization (NCCL backend) with a hybrid loss balancing L1 reconstruction (λ_1=1.0), VGG-based perceptual quality (λ_2=0.1), and adversarial training (λ_3=0.01). During inference, the system generates high-resolution outputs conditioned on either auto-generated enriched captions (CLIP + Llama) or user-defined text, validated through pixel-level metrics (PSNR/SSIM), perceptual quality (LPIPS/FID), and text-image alignment (CLIPScore). This end-to-end workflow—preprocessing → caption enrichment → training → inference—enables precise, semantics-driven super-resolution tailored to geospatial applications, outperforming conventional methods by prioritizing text-specified features like urban grids or ecological patterns.

- **Goal**: Improve low-resolution Sentinel-2 (10m) images to PlanetScope (3m) quality using text guidance
- **Key Components**:
  1. **Data Preparation**:
     - Paired Sentinel-2 (blurry) and PlanetScope (clear) images
     - Auto-generated descriptive text captions using CLIP [6] + Llama [1]
     - Example caption: "farm fields with irrigation channels"
  2. **AI Model**:
     - Modified SRGAN architecture with two parallel branches:
       - Text processing branch: Converts captions to numerical vectors
       - Image upscaling branch: 16-layer network with 4× resolution boost
  3. **Training**:
     - Uses three simultaneous objectives:
       - Match pixel colors ($\lambda_1 = 1.0$)
       - Maintain realistic textures ($\lambda_2 = 0.1$)
       - Align with text descriptions ($\lambda_3 = 0.01$)
     - Multi-GPU accelerated training
  4. **Usage**:
     - Input: Blurry image + text (auto-generated or custom)
     - Output: High-quality image focusing on text-specified features
     - Quality checks: Accuracy (PSNR/SSIM), Realism (LPIPS/FID), Text match (CLIPScore)
- **Advantage**: Focuses enhancement on text-prioritized features (roads, crops, etc.) better than traditional methods

Workflow: Preprocess images → Generate captions → Train model → Produce enhanced images

Furthermore, we implemented ARISGAN, an artistic super-resolution model, to assess its potential application in enhancing satellite imagery with an artistic touch. The ARISGAN framework was evaluated on our dataset to explore its ability to generate high-resolution images with artistic nuances that could be useful in specific remote sensing tasks.

Finally, we also trained ESRGAN specifically for our dataset, allowing us to fine-tune the model on satellite images to achieve the best possible performance for land monitoring and agricultural analysis tasks.

6 Results

The performance of different super-resolution models was evaluated using PSNR, SSIM, and SAM metrics. Among the models, our fine-tuned ESRGAN ("Trained - ESRGAN") achieved the highest PSNR value of 21.0684 dB and the highest SSIM score of 0.5649, indicating better reconstruction quality and structural similarity compared to other models. Although the SAM value of the fine-tuned ESRGAN (0.2535 radians) is slightly higher than that of the standard ESRGAN (0.2545 radians), it remains competitive and shows improved visual fidelity. Overall, these results demonstrate that fine-tuning ESRGAN on our satellite image dataset significantly enhances its performance compared to both the original ESRGAN and other pre-trained models like SwinIR and Real-ESRGAN. The results are summarized in Table 1.

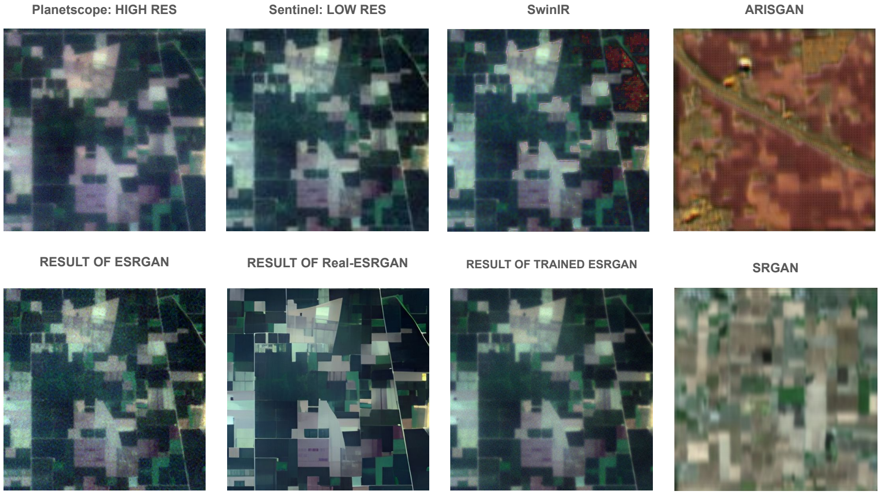


Fig. 3. Comparison of original and output of multiple models we tried.

Table 1. Quantitative comparison of super-resolution models.

Model	**PSNR (dB)**	**SSIM**	**SAM (radian)**
SwinIR	17.2769	0.1821	0.1668
ESRGAN	20.7759	0.5145	0.2545
Real-ESRGAN	19.4473	0.5082	0.2501
Trained - ESRGAN	21.0684	0.5649	0.2535

6.1 SwinIR Results

Fig.4 is a comparison of low-resolution Sentinel imagery vs the enhanced image generated by using SwinIR, and in this visualisation we can see that SwinIR creates some artifacts in the top right region these is some reddish spots and this is correlated by the not optimal metrics results we got in Table 1.

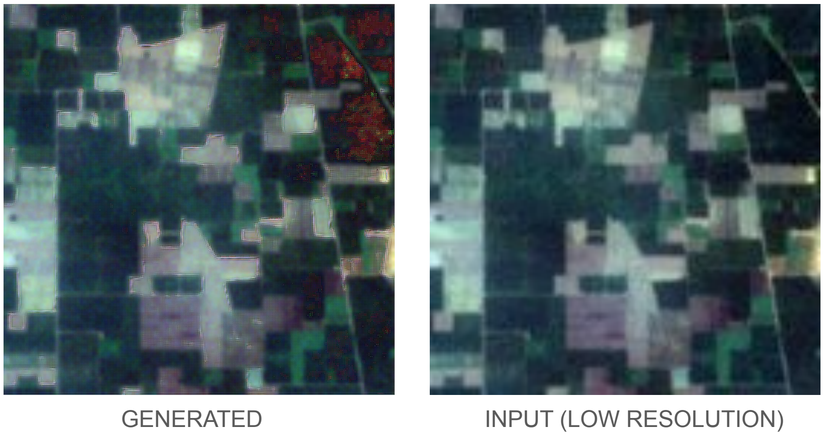


Fig. 4. Comparison of Sentinel low-resolution image (right) vs SwinIR-enhanced image (left).

6.2 ESRGAN Results

In the Figure 5 the quality of the generated image has increased by many folds to the naked eye but the underlying distribution of pattern has been disturbed leading to a low SSIM value.

6.3 Real-ESRGAN Results

In Figure 6 the generated image is even sharper than the generated image of ESRGAN but this sharpness comes at the cost of disturbing the structure of the underlying pattern even more resulting in even poorer SSIM value than ESRGAN.

6.4 Caption-Based Diffusion Imaging Results

For Figure 7 we first used the BLIP model to create captions from low-resolution images. Then, we improved these captions with the "Falcon-7b" language model.

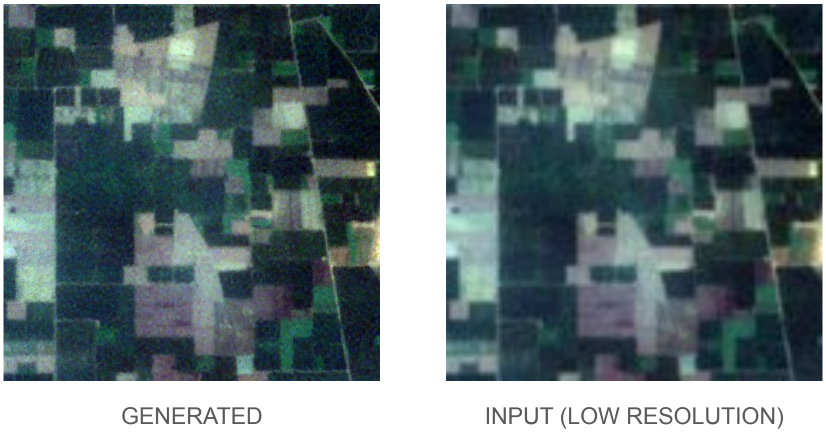


Fig. 5. Comparison of Sentinel low-resolution image (right) vs ESRGAN-enhanced image (left).

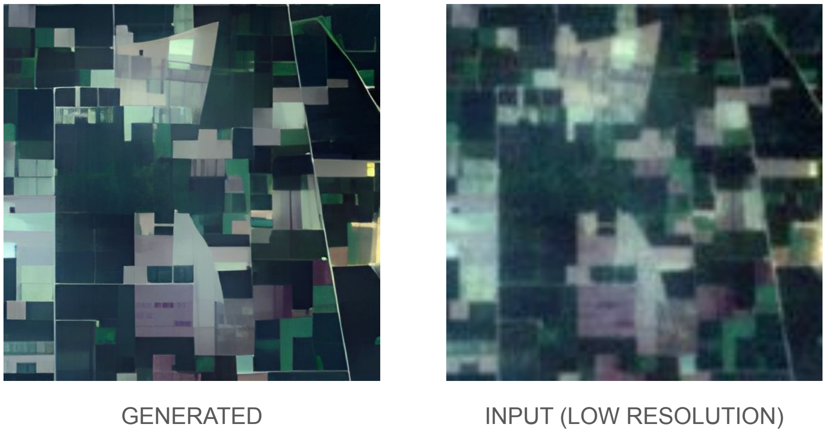


Fig. 6. Comparison of Sentinel low-resolution image (right) vs Real-ESRGAN enhanced image (left).

After that, we used the better captions and the low-res images together in a pre-trained "Stable Diffusion v1-5" model to generate high-resolution images. Finally, generated image is compared with high resolution image.

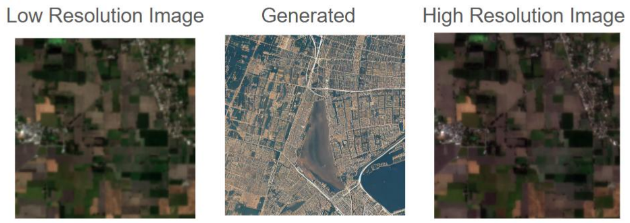


Fig. 7. Caption and Diffusion model based High Resolution Image Generation(center)

6.5 ARISGAN Results

For Figure 9 the ARISGAN-generated image exhibits significantly enhanced resolution and clarity compared to the low-resolution image. It showcases finer details and textures that are absent in the pixelated and blurry low-resolution counterpart. Overall, the generated image demonstrates a substantial improvement in visual quality, appearing more refined and realistic.

6.6 Trained ESRGAN

We fine-tuned the ESRGAN model on our satellite image dataset to better adapt it to the specific characteristics of our data. After training, the fine-tuned model demonstrated superior performance compared to the pre-trained versions, producing higher-quality super-resolved images with improved detail preservation and reduced artifacts. This highlights the importance of domain-specific fine-tuning for achieving optimal results in satellite image enhancement tasks.

7 Conclusion

The experimental results demonstrate that the Trained-ESRGAN model achieves the best overall performance among the evaluated methods. It recorded the highest Peak Signal-to-Noise Ratio (PSNR) of 21.0684 dB and the highest Structural

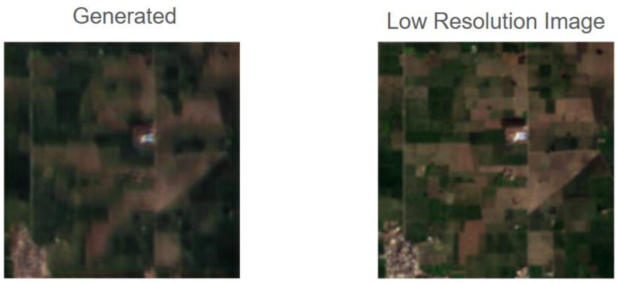


Fig. 8. Comparison of Sentinel low-resolution image (right) vs ARISGAN-enhanced image (left)

Similarity Index Measure (SSIM) of 0.5649, indicating superior reconstruction quality and structural preservation. Although its Spectral Angle Mapper (SAM) value (0.2535 radians) is slightly higher compared to SwinIR, it remains competitive relative to the other GAN-based models. Among the pretrained models, ESRGAN outperformed Real-ESRGAN, achieving better PSNR and SSIM values, while Real-ESRGAN exhibited slightly better spectral consistency based on the SAM metric. In contrast, SwinIR showed the weakest performance across all evaluated metrics, with notably lower PSNR, SSIM, and SAM scores. Furthermore, the ArisGAN model exhibited poor performance during evaluation, and the caption-based model failed to generate coherent images, often producing random and inconsistent outputs.

Overall, fine-tuning ESRGAN significantly enhanced both the visual and spectral quality of the reconstructed images, making it the most effective approach among those studied.

Acknowledgments. This project was conducted as part of the CS546: Introduction to Agriculture Cyber Physical Systems course at IIT Ropar.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

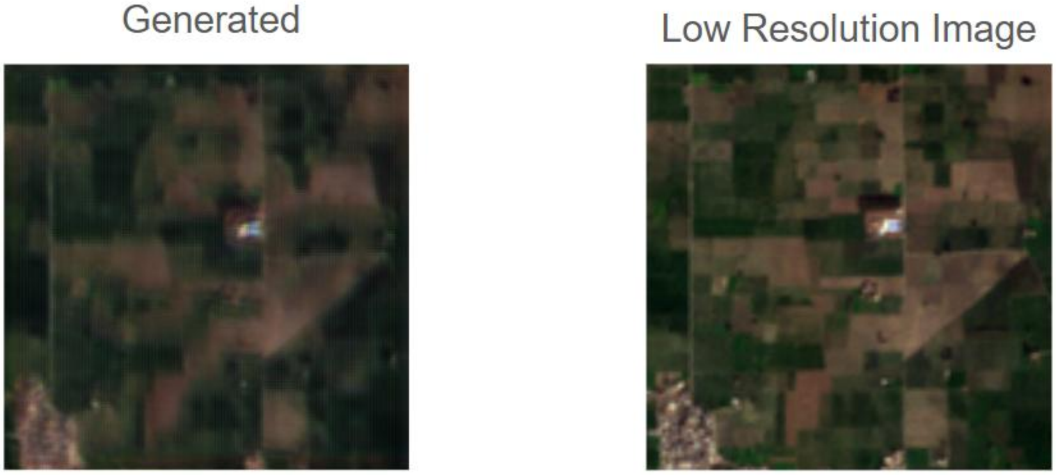


Fig. 9. Comparison of Sentinel low-resolution image (right) vs ARISGAN-enhanced image (left)

2. Au, C., Tsamados, M., Manescu, P., Takao, S.: ARISGAN: Extreme super-resolution of arctic surface imagery using generative adversarial networks **5**. https://doi.org/10.3389/frsen.2024.1417417, https://www.frontiersin.orghttps://www.frontiersin.org/journals/remote-sensing/articles/10.3389/frsen.2024.1417417/full
3. Liang, J., Cao, J., Sun, G., Zhang, K., Gool, L.V., Timofte, R.: SwinIR: Image restoration using swin transformer. https://doi.org/10.48550/arXiv.2108.10257, http://arxiv.org/abs/2108.10257
4. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. https://doi.org/10.48550/arXiv.2107.10833, http://arxiv.org/abs/2107.10833
5. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C.C., Qiao, Y., Tang, X.: ESRGAN: Enhanced super-resolution generative adversarial networks. https://doi.org/10.48550/arXiv.1809.00219, http://arxiv.org/abs/1809.00219
6. Zeng, Y., Jiang, C., Mao, J., Han, J., Ye, C., Huang, Q., Yeung, D.Y., Yang, Z., Liang, X., Xu, H.: Clip^2: Contrastive language-image-point pretraining from real-world point cloud data (2023), https://arxiv.org/abs/2303.12417