SDP Research Report

Project Name	A Unified Deep Learning Benchmark for Satellite Image Restoration and Generation
Faculty Supervisor Name	Ismayil Shahaliyev
Team Lead Name & ID	Royana Huseynova 11270
Team member #1 Name & ID ● What sections did this person contribute to?	Royana Huseynova 11270 Introduction
Team member #2 Name & ID ● What sections did this person contribute to?	Nijat Alisoy 17366 Proposed Solution
Team member #3 Name & ID ● What sections did this person contribute to?	Pasha Zulfugarli 13731 Review
Team member #4 Name & ID ● What sections did this person contribute to?	Huseyn Sadatkhanov 13770 Expected Outcomes

1.Introduction

Background and Context

Our project is linked to synthetic satellite data, which is critical for a variety of applications such as agriculture, development planning, disaster response, and many more.There are three main limitations when it comes to satellite imagery:

(i) cloud cover obstruction,

(ii) low spatial resolution

(iii) the lack of high-definition imagery

- For instance, *Optical Remote Sensing* (ORS) images are often obstructed by cloud cover, leading to the distortion or loss of Earth's surface information. As a result, cloud removal has become an essential preprocessing step in ORS image analysis [1]. Insufficient image spatial resolution is also a serious limitation in many scenarios, especially when acquiring images at a bigger scale is impractical or brings higher costs. This is inherent to remote sensing, including Sentinel-2 satellite images that are available free of charge at a high revisit frequency, but whose spatial resolution is limited to 10 meters on the ground, making it difficult to capture details. [2] Furthermore, the limited availability of naturally high-definition satellite images restricts the level of detail possible in applications like urban mapping and environmental monitoring. To overcome these limitations, various image enhancement techniques including color adjustment, noise reduction, and super-resolution have been developed to improve the quality of images from both current and older satellites. [3]

Problem Statement / Research Questions and Hypotheses

It should be noted that the present benchmarks and assessment frameworks for satellite imagery enhancement methods are largely isolated and task-specific, focusing on a single task such as **cloud removal (CR)** or **super resolution (SR)**. For example, the AllClear benchmark [4] offers a comprehensive dataset and tools tailored for evaluating cloud removal methods. Likewise, OpenSR-test [5] provides a benchmark and metric suite for assessing super-resolution models. Therefore, there is currently no unified evaluation framework that supports and compares algorithms across CR , SR, and HD image generation within a single, integrated pipeline.

Proposed Solution Objectives

This gap shows the importance of our project: **to create an evaluation pipeline** that lets researchers systematically benchmark and compare models across all three tasks, using consistent metrics and a wide range of datasets.

We plan to have in our pipeline data loaders that can work with a number of publicly available datasets, including **AllClear [4], SEN12MS-CR-TS [6], WorldStrat [7], PROBA-V-SR [8], and SAT25K [9]**. The evaluation framework will include code for testing across metrics, such as **Peak Signal-to-Noise Ratio (PSNR)**, **Structural Similarity Index (SSIM)**, **Spectral Angle Mapper (SAM)**, and others.

Impact of the Proposed Solution

By standardizing the evaluation process, our pipeline will support fair, consistent, and repeatable comparisons of satellite imagery methods which researchers and students will be able to use in the future. Additionally, it will help identify the most effective techniques for different types of satellite imagery, guiding future research and practical applications. Over time, this can accelerate advancements in areas such as environmental monitoring, urban planning, and disaster management.

2. Review

Overview of Existing Market

Synthetic satellite image enhancement methods have accelerated alongside specialized datasets and evaluation metrics. Publicly available datasets support different satellite image enhancement tasks, each with a specific focus. SEN12MS-CR-TS contains images captured at different times and from multiple sensors, designed for **cloud removal**, helping models reconstruct clear images under varying conditions. WorldStrat provides diverse landscapes and atmospheric conditions for **general image reconstruction and benchmarking**, enabling fair comparison between models. The PROBA-V Super-Resolution Challenge dataset is aimed at **multi-image super-resolution**, where several low-resolution images of the same area are combined to produce higher-resolution outputs. SAT25K generates **synthetic images from semantic maps**, supporting experiments in generative modeling, while AllClear, focuses on **large-scale cloud removal**, facilitating the development and evaluation of restoration pipelines at scale. [11]

For assessing image quality, remote sensing studies typically report **PSNR**, **SSIM**, and **SAM**. Fusion and SR literature often includes **Error Relative Global Dimensionless Synthesis (ERGAS)** and **Quality with No Reference (QNR)**, while generative model evaluation increasingly uses **Fréchet Inception Distance (FID)**. [12]

In Azerbaijan, documented activities mainly focus on analyzing satellite imagery through Azercosmos, indicating limited publicly described efforts in synthetic image generation. Utilizing **Earth Observation (EO)** data and **Geographic Information Systems (GIS)**, Azerbaijan has national efforts in monitoring and mitigating climate change. Capacity-building initiatives are underway, such as the **Azercosmos–Bayanat partnership**, which provides GIS/remote sensing cooperation, AI and SAR training, and temporary access to Bayanat's geospatial platform [13]. Contrary to earlier assumptions, there is now public evidence of Azerbaijan's involvement in synthetic satellite image generation using generative models. The SATDM framework, which is developed by researchers including contributors from Azercosmos. It employs diffusion models to generate realistic satellite images conditioned on semantic layouts, representing one of the first known national-scale contributions to generative satellite data. Similarly, the SAT25K dataset provides high-resolution satellite image patches curated for use in training and evaluating generative models, further supporting Azerbaijan's emerging role in AI-driven Earth observation [11].

Gaps in Current Market (Azerbaijan)

Azerbaijan does not have its own projects for creating synthetic satellite data. This is a good opportunity to start small local projects using open international datasets like WorldStrat, PROBA-V SR, and AllClear, so the results are more useful for local needs. At the moment, there are also very few public reports that use standard ways to measure image quality, such as PSNR, SSIM, SAM, ERGAS, QNR, or FID. Using these common international tools would make future research clearer and easier to compare. In terms of technology and skills, Azerbaijan still depends a lot on international platforms such as **Baynat** for training and support. This shows that local knowledge and infrastructure are still developing. To improve, the country needs better access to **High Performance Computing (HPC)** and more experts in data engineering and AI for satellite images (EO-AI). Building local skills and teamwork between universities, government, and research centers will be very important for long-term progress in these fields.

3.Proposed Solution

The objective of this project is to design and implement a standardized evaluation pipeline that allows fair and reproducible comparison of image restoration and enhancement methods applied to satellite imagery. The project focuses on three major tasks: Cloud Removal, Super-Resolution, and High-Definition Image Generation. The pipeline aims not only to assess visual quality but also to ensure the spectral and physical correctness of the generated or restored images, which is critical for reliable downstream analysis in remote sensing applications.

System Architecture

The core of the system is an API designed to visualize major datasets and assess the accuracy of models. The architecture is composed of three main components: *Data Loader, Metrics, Evaluation Engine* codebase.

Data Loader is responsible for dataset management and **input standardization**. It supports several benchmark datasets, including *AllClear*, *SEN12MS-CR-TS*, *WorldStrat*, *PROBA-V SR*, and *SAT25K*. Each dataset can differ in pixel value, scales, number of bands, and formats. Through **normalization** (e. g. adjusting pixel values so they all follow the same scale) and **tensor conversion** (turning raw images into numerical arrays that machine learning frameworks can process)**,** datasets are automatically processed into a unified format. Standardization is done to ensure that models can operate with all datasets. This layer also manages dataset splitting into **training, validation, and testing** subsets to maintain consistent experimental protocols.

Metrics Library is a collection of mathematical tools that measure how good the model's outputs are compared to the reference, or when there is no reference, how physically consistent they look. It provides a wide range of evaluation tools, including both traditional image quality metrics and remote-sensing-specific measures. Metrics such as **PSNR** and **SSIM** are used to evaluate the **visual fidelity** of reconstructed images. Visual fidelity means how perceptually close the reconstructed image is to the original. It focuses on the sharpness of the edges, preservation of structures such as roads and buildings, and the absence of distortions. **PSNR** measures pixel-level accuracy between two images by comparing how much **error** exists between the **generated image** and the **reference**. **SSIM** measures structural similarity, looking at local patterns of texture and geometry, luminance, and contrast [15]. On the other hand, **SAM** and **ERGAS** measure **spectral accuracy** and **global radiometric consistency**. These are unique to remote sensing because satellite images capture spectral bands, not just colors. Reflectance at multiple wavelengths represents real physical properties of surfaces, so errors have scientific consequences. **SAM** treats each pixel's band values as a vector in multi-dimensional space and measures the angle between the reference and reconstructed pixel vectors [17]. A smaller angle (close to 0°) means **better spectral similarity**. **ERGAS** measures how much overall error exists across all bands. Lower **ERGAS** means less numerical error, and this is important because **ERGAS** ensures that no major brightness or spectral shifts occur during enhancement (e.g., the image does not get overall darker or brighter) [18]. Sometimes, especially in HD image generation, there is no exact ground truth image. Additional metrics like **QNR** and **FID** allow evaluation of methods even in the absence of a reference image, making the pipeline applicable to **generative models** as well. QNR computes spectral distortion (color/reflectance changes) and spatial distortion (detail/edge changes) and combines these two indices to evaluate image fusion [16]. When QNR is 1, it means the quality is perfect (no distortion), however, a lower QNR means worse fusion. QNR lets you judge how natural the result is without needing a true high resolution multispectral image. FID on the other hand, compares how realistic and natural generated images look compared to real ones. It takes a set of generated images and a set of real images, extracting high-level features from both sets using a pre-trained deep network, then comparing their feature distributions (means and covariances). Lower FID equates to images that look more like real images statistically. This comprehensive suite of metrics will ensure that the evaluation captures not only visual appearance but also **physical correctness** [12, 15-19].

Evaluation Engine will allow models to be tested across all supported datasets, eliminating inconsistencies between experiments. This component automates metric calculation, result storage, and report generation, producing summaries and visual plots that compare model performance. Furthermore, the engine will manage reproducibility to ensure that every evaluation can be replicated precisely.

Technologies and Tools

The implementation of the system will rely on modern open-source technologies. The primary development environment is **Python**, supported by deep learning framework **PyTorch** for model integration. Data processing and geospatial analysis will be handled using libraries like **Rasterio** and **GDAL** (Geospatial Data Abstraction Library), which facilitate the manipulation of multispectral satellite imagery.

Data Collection and Analysis

This project relies on established, open satellite datasets that are widely used in remote sensing research and benchmarking. Because our goal is a **unified evaluation pipeline**, the datasets were selected to cover three target tasks—**Cloud Removal (CR), Super-Resolution (SR), and High-Definition (HD) / generative image synthesis**—while also representing diverse sensors, resolutions, and data formats. Across all datasets, inputs will be standardized through a common preprocessing interface (normalization, band alignment, and resolution handling) before model execution and metric computation.

AllClear (Cloud Removal)

AllClear is a large-scale benchmark designed specifically for cloud removal in satellite imagery [4]. It is structured around **23,742 globally distributed Regions of Interest (ROIs)** and contains **approximately 4 million images** in total [4]. Each ROI includes temporal coverage (for the year 2022) and provides multiple complementary data sources: **multi-spectral optical imagery from Sentinel-2 and Landsat 8/9**, **SAR imagery from Sentinel-1**, and supporting auxiliary products such as **cloud masks and land-cover maps** [4]. These multimodal components make AllClear suitable for evaluating cloud removal methods that leverage optical-only restoration, SAR-guided reconstruction, or multi-temporal fusion, while also enabling controlled benchmarking using consistent ROI definitions and metadata [4].

SEN12MS-CR-TS (Cloud Removal, Multi-Temporal & Multi-Modal)

SEN12MS-CR-TS is a global dataset built for **multi-modal, multi-temporal cloud removal**, combining **Sentinel-1 SAR** and **Sentinel-2 multispectral** observations [6], [10]. The dataset consists of **53 globally distributed ROIs**, each covering roughly **40×40 km^2** (over **4000×4000 pixels** per ROI) and spanning more than **80,000 km^2** total surface area [10]. For each ROI, the dataset provides **30 paired time points** across 2018, where Sentinel-1 and Sentinel-2 acquisitions are matched within a short temporal window [10]. Importantly for standardized evaluation, preprocessing in the dataset includes **upsampling all bands to 10 m resolution** and slicing full scenes into **non-overlapping 256×256 patches** [10]. Sentinel-1 is provided through **VV and VH polarization channels**, while Sentinel-2 data is provided as **13-band multispectral observations** (with band handling unified at 10 m in the dataset's prepared format) [10]. These properties make SEN12MS-CR-TS ideal for benchmarking both single-frame and temporal-sequence cloud removal methods under realistic cloud coverage conditions [6], [10].

WorldStrat (Super-Resolution / Multi-Resolution Benchmarking)

WorldStrat is designed to support benchmarking of reconstruction and super-resolution by providing **paired multi-resolution imagery** across geographically diverse locations [7]. The dataset links **high-resolution commercial satellite imagery** (SPOT 6/7 products) with **lower-resolution Sentinel-2** observations over the same areas, enabling evaluation of resolution enhancement and cross-sensor reconstruction [7]. It is organized around a large number of curated areas of interest (AOIs) and includes consistent geospatial alignment between sources [7]. This structure is especially valuable for our benchmark because it supports: (i) training/testing on paired low–high resolution data, and (ii) evaluating whether super-resolution models preserve both visual structure and remote-sensing consistency when moving across resolutions and sensors [7].

PROBA-V Super-Resolution (Multi-Image SR Benchmark)

The PROBA-V Super-Resolution dataset (Kelvins / ESA challenge data) targets **multi-image super-resolution**, where multiple low-resolution observations from the same scene are fused into a higher-resolution output [8], [14]. The dataset is built from **74 hand-selected regions** globally and provides **Top-Of-Atmosphere (TOA) reflectance** data for the **RED and NIR bands** at two resolutions: **300 m (low-resolution)** and **100 m (high-resolution)** [8], [14]. Data is distributed as grayscale images where **300 m inputs are 128×128 pixels** and **100 m targets are 384×384 pixels**, alongside **quality maps** indicating unusable pixels (clouds, shadows, missing data, etc.) [14]. Each datapoint contains **one high-resolution image** and **multiple low-resolution images** for the same scene [14]. In total, the dataset contains **1450 scenes**, split into **1160 training scenes** and **290 test scenes**, with an average of **~19 low-resolution images per scene** (and at least 9) [14]. This dataset is therefore well-suited for consistent SR benchmarking where temporal/frame fusion is part of the model design [8], [14].

SAT25K (HD / Generative Satellite Image Synthesis)

SAT25K is a curated dataset created to support **semantic-conditioned satellite image generation**, particularly for built-environment structures [11]. It is derived from **Google Earth imagery** associated with the **Pleiades** satellite, processed as an **orthophotomosaic**, and geographically focused on **Mardakan, Baku, Azerbaijan** [11]. The dataset is **RGB (Red, Green, Blue)** and includes paired **pixel-wise binary building footprint annotations** (semantic layouts), enabling conditional generation and evaluation workflows [11]. SAT25K includes both “organic” (real) and augmented tiles; reported splits include **5,000 real test images**, and training sets on the order of **~24–25k tiles** depending on resolution configuration (e.g., 64×64 and 128×128 experimental settings) [11]. These characteristics make SAT25K appropriate for HD/generative benchmarking where ground-truth “future” imagery may not exist, but realism and structure consistency can still be evaluated through no-reference metrics and controlled semantic conditioning [11].

4.Expected Outcomes

By the end of this project, we expect to have a fully operational evaluation pipeline that significantly simplifies the assessment of cloud removal, super-resolution, and high-definition image generation methods. Furthermore, the evaluation will assess **spectral fidelity** across all tasks to verify that the enhanced images are not only visually realistic but also retain their original **spectral characteristics**. For instance, cloud-removed images should be carefully checked to make sure that the **restored pixel values** match real clear-sky images, without introducing any unnatural color and brightness. Restored pixel values are the numerical values that represent light intensity or color at each pixel in an image. In satellite imagery, each pixel stores reflectance information (is the data showing how much light a surface reflects at different wavelengths) for different wavelengths (red, green, blue, or infrared). Similarly, high-resolution images should be compared with their low-resolution version to confirm that the enhancement did not cause any changes or distortions in the original spectral information. Ultimately, this project aims to develop a clear and user-friendly evaluation system for satellite image enhancement. The pipeline, together with open-source code, documentation, and usage examples, will be made publicly available to enable other researchers to test their models.

Anticipated Deliverables

To accomplish the abovementioned outcomes, the project will deliver:

A. **Unified Evaluation Pipeline** – The main outcome of our project is a Python-based unified evaluation pipeline for satellite imagery enhancement.
It will include all the necessary components to automatically load different datasets such as AllClear, SEN12MS-CR-TS, WorldStrat, PROBA-V SR, etc., and to calculate performance metrics such as PSNR, SSIM, ERGAS, etc. The system will feature an interface that allows researchers to easily connect their models, choose a dataset, and run evaluations with minimal setup. Along with the source code, we will provide detailed documentation that explains how to use the evaluation API. This will make the system easy to understand and ensure that even researchers with less code experience successfully run evaluations on different datasets.

B. **Reference Model** – To verify the performance and reliability of the evaluation, we will create a reference model to demonstrate initial results. We plan to create a final report of our results that show how different methods may perform on each dataset. These results will provide reference points for future work, making it easier to test and measure the performance of new models.

These deliverables ensure that the project remains useful even after its completion. The code and documentation will allow others to use or improve the pipeline, the initial result will offer useful information for future research, and the final report and presentation will highlight what was achieved and what was learned during the project.

Potential Impact and Applications

This project aims to make a meaningful contribution to both research and practical applications in satellite imagery and remote sensing. The planned evaluation pipeline will provide a unified, well-structured framework for testing and improving cloud removal, super-resolution, and high-definition image generation models across multiple datasets. By simplifying the evaluation process, it will allow researchers to compare models across different datasets and metrics, producing reliable and consistent results.

Improving the quality and reliability of satellite imagery has the potential to make a real difference in fields like agriculture, environmental monitoring, urban planning, and disaster response. More accurate imagery will help track vegetation, land use, and infrastructure, support better resource management and urban development, and improve planning and response during emergencies. These advances will help decision-makers, planners, and researchers make more informed choices and use satellite data more effectively across a range of applications.

Taken together, the outcomes of this project highlight its potential to advance both research and real-world applications. The planned pipeline will address core challenges such as cloud cover, low resolution, and limited availability of high-definition data. By integrating multiple datasets and using both visual and remote sensing–specific metrics, it aims to ensure that restored and generated images are not only clear and detailed but also scientifically accurate ,meaning their spectral values, brightness, and spatial structures closely reflect real-world conditions, making them dependable for analysis and decision-making.

By providing a standardized and reproducible framework, the planned pipeline will make it easier to benchmark models for cloud removal, super-resolution, and high-definition image generation. Together with reference models and detailed documentation, it is designed to give researchers and practitioners a practical and reliable platform for improving satellite imagery. Ultimately, the project aims to enhance the usefulness of satellite data, supporting more accurate analysis, informed planning, and better decision-making in real-world applications.

References:

[1] J. Ning, L. Xie, J. Yin, and Y. Liu, “Cloud Removal Advances: A Comprehensive Review and Analysis for Optical Remote Sensing Images,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. PP, no. 99, pp. 1–17, Jan. 2025, doi: https://doi.org/10.1109/JSTARS.2025.3580718

[2] P. Kowaleczko *et al.*, “A Real-World Benchmark for Sentinel-2 Multi-Image Super-Resolution,” *Scientific Data*, vol. 10, no. 1, Sep. 2023, doi: https://doi.org/10.1038/s41597-023-02538-9

[3] Sandeep Kumar Jangir and Reza Bahmanyar, “Enhancing Very High-Resolution Satellite Images At 15 cm: A Novel Pipeline With Synthetic Data and Single Image Superresolution,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 17, pp. 2670–2689, Dec. 2023, doi:https://doi.org/10.1109/jstars.2023.3346181

[4] H. Zhou, C.-H. Kao, P. C. Perng, U. Mall, B. Hariharan, and K. Bala, “AllClear: A Comprehensive Dataset and Benchmark for Cloud Removal in Satellite Imagery,” *arXiv.org*, 2024. https://arxiv.org/abs/2410.23891

[5] C. Aybar, D. Montero, S. Donike, F. Kalaitzis, and L. Gómez-Chova, “A Comprehensive Benchmark for Optical Remote Sensing Image Super-Resolution,” *IEEE Geoscience and Remote Sensing Letters*, vol. 21, pp. 1–5, Jan. 2024, doi:https://doi.org/10.1109/lgrs.2024.3401394

[6] P. Ebel, “SEN12MS-CR-TS Dataset,” *Github.io*, 2018. https://patricktum.github.io/cloud_removal/sen12mscrts/

[7] WorldStrat, “WorldStrat,” *Github.io*, 2022. https://worldstrat.github.io

[8] “Welcome To Zscaler Directory Authentication,” *Esa.int*, 2025.

https://kelvins.esa.int/proba-v-super-resolution/data/

[9] A. Ray *et al.*, “SAT: Dynamic Spatial Aptitude Training for Multimodal Language Models,” *arXiv.org*, 2024.
https://arxiv.org/abs/2412.07755

[10] P. Ebel, Y. Xu, M. Schmitt, and X. Zhu, “SEN12MS-CR-TS: A Remote Sensing Data Set for Multi-modal Multi-temporal Cloud Removal,” *arXiv:2201.09613 [cs, eess]*, Jan. 2022, doi: https://doi.org/10.1109/TGRS.2022.3146246

[11] O. Baghirli, H. Askarov, I. Ibrahimli, I. Bakhishov, and N. Nabiyev, “SatDM: Synthesizing Realistic Satellite Image with Semantic Layout Conditioning using Diffusion Models,” *arXiv.org*, 2023. doi:https://doi.org/10.48550/arXiv.2309.16812

[12] J. Liu, J. Huang, S. Liu, H. Li, Q. Zhou, and J. Liu, “Human visual system consistent quality assessment for remote sensing image fusion,” vol. 105, pp. 79–90, Jul. 2015, doi: https://doi.org/10.1016/j.isprsjprs.2014.12.018

[13] “Azercosmos,” *Azercosmos.az*, 2023.
https://azercosmos.az/en/media-center/news/azercosmos-and-bayanat-signed-an-agreement-on-the-development-of-geographic-information-systems

[14] M. Märtens, D. Izzo, A. Krzic, and D. Cox, “Super-Resolution of PROBA-V Images Using Convolutional Neural Networks,” *arXiv.org*, Jul. 03, 2019. https://arxiv.org/abs/1907.01821

[15] 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, Apr. 2004, doi: https://doi.org/10.1109/tip.2003.819861

[16] T. Ranchin and L. Wald, “Fusion of high spatial and spectral resolution images : The ARSIS concept and its implementation,” *Photogrammetric Engineering and Remote Sensing*, 2017.
https://www.semanticscholar.org/paper/Fusion-of-high-spatial-and-spectral-resolution-%3A-Ranchin-Wald/ae3cf7f2e2a9afbcee463d9666ab344bc118f14d#extracted

[17] K. F. A *et al.*, “The Spectral Image Processing System (SIPS): Software for integrated analysis of AVIRIS data,” *JPL, Summaries of the Third Annual JPL Airborne Geoscience Workshop. Volume 1: AVIRIS Workshop*, Jun. 1992. https://ntrs.nasa.gov/citations/19940012201

[18] L. Alparone, Stefano Baronti, A. Garzelli, and F. Nencini, “A Global Quality Measurement of Pan-Sharpened Multispectral Imagery,” vol. 1, no. 4, pp. 313–317, Oct. 2004, doi: https://doi.org/10.1109/lgrs.2004.836784

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” *arXiv:1706.08500 [cs, stat]*, Jan. 2018, Available: https://arxiv.org/abs/1706.08500