Babylonian Journal of Machine Learning
Vol.2026, **pp**. 235–257
DOI: https://doi.org/10.58496/BJML/2026/015; ISSN: 3006-5429
https://mesopotamian.press/journals/index.php/BJML

Research Article

Single-Image Super-Resolution of Sentinel-2 Imagery for Filling MisrSat-2 Acquisition Gaps Using Deep Learning Techniques

Bassant Raafat[1,*], Belal Elleithy[2], Mohammed Safy[1]

[1] *Arab Academy for Science, Technology and Maritime Transport, Faculty of Computing and Information Technology, Cairo, Egypt*

[2] *Egyptian Space Agency, Cairo, Egypt*

ARTICLEINFO

Article History
Received 20 Jul 2026
Revised 25 Aug 2026
Accepted 20 Sep 2026
Published 31 Oct 2026

Keywords
Super-Resolution,
MisrSat-2,
Sentinel-2,
Deep Learning,
Image Co-Registration.

ABSTRACT

Satellite imagery is of great importance in Earth observation, particularly in highly populated countries like Egypt, where environmental monitoring and urban planning are essential. The country has recently launched MisrSat-2, a high-resolution satellite capable of providing images with a spatial resolution of 2 m; yet its limited temporal coverage leads to considerable gaps in observation and therefore hampers its usefulness for monitoring rapid changes on the ground. This represents a major technical limitation since, for any given date, all the high-resolution spatial data is lost if there is no image acquired by MisrSat-2. To address this shortcoming, the current study employs deep learning-based super-resolution (SR) techniques to generate 2.5 m proxy images that resemble MisrSat-2, using freely available Sentinel-2 imagery (at 10 m), thereby helping to fill the temporal coverage gaps without the need for additional satellite acquisitions. It must be emphasized that the outputs are model-generated proxies and not real 2.5 m observations. Four different SR architectures were systematically evaluated: RCAN, SRGAN, ESRGAN, and SwinIR. The RCAN model obtained the best performance (PSNR = 26.629 dB, SSIM = 0.750), beating SRGAN (PSNR = 23.35, SSIM = 0.716), ESRGAN (PSNR = 24.38, SSIM = 0.735), and SwinIR (PSNR = 22.68, SSIM = 0.744), although the differences in SSIM values between the models are minor and the order of performance reflects the conditions under which the evaluation was carried out. Since the evaluation is conducted on histogram-matched imagery, the reported values relate to the structural and perceptual quality of the reconstruction rather than to absolute radiometric accuracy. A key aspect of this study is the development of a fully automated co-registration framework between Sentinel-2 and MisrSat-2 images, which does not require any manual input and which corrects spatial misalignments while considerably cutting down the preprocessing time. As MisrSat-2 is a satellite that has only recently been launched, this study is the first to apply and test deep learning super-resolution methods using MisrSat-2 as the reference, and it offers a practical and scalable approach to obtaining higher-resolution Earth observation over Egypt.

1. INTRODUCTION

Earth observation satellites are very important for a wide range of applications, for example environmental monitoring, the management of agricultural operations, the assessment of disasters, the analysis of urban expansion, and research into climate change [1]–[3]. The effectiveness of these applications is mainly determined by the availability of images that have both a high spatial resolution and a high temporal frequency. High spatial resolution enables detailed mapping and the accurate identification of objects, while frequent temporal coverage allows changes to the surface to be detected [4], [5]. There is, however, a well-known inverse relationship between these two characteristics in most satellite missions—those which offer high spatial resolution usually have long revisit times, whereas those that make frequent acquisitions do so at a coarser spatial resolution [6]. The compromise between spatial and temporal resolution is a basic limitation in operational Earth observation and means that no single sensor can at the same time satisfy the requirements of applications that demand both fine detail and continuous monitoring.

The rapid expansion of Egypt's urban areas and its infrastructure development shows this problem in a particularly severe way. The Egyptian MisrSat-2 satellite, which was launched in December 2023, offers high-resolution 2 m pansharpened multispectral imagery in which the panchromatic band has been combined with the multispectral bands [7], [8], making it suitable for detailed monitoring of urban and infrastructure features. Nevertheless, its 55-day nadir revisit time leads to large

*Corresponding author. Email: B.Elgetan63952@student.aast.edu

temporal gaps that greatly limit its usefulness when monitoring rapidly changing urban areas where considerable changes can take place within a few days. In contrast, the Sentinel-2 constellation supplies images with high frequency once every five days [9], but with a much coarser spatial resolution of 10 m for its visible and near-infrared bands, which is not sufficient for capturing the fine structural details required for urban mapping and for evaluating infrastructure. The main scientific question that this study sets out to answer is therefore whether deep learning-based super-resolution can reliably overcome the spatial resolution difference between Sentinel-2 and MisrSat-2, thus allowing the production of 2.5 m proxy images on those dates when there is no acquisition from MisrSat-2. Super-resolution (SR) methods based on deep learning have shown a remarkable ability to improve the spatial resolution of satellite imagery [10]–[12]. Unlike ordinary interpolation techniques that lead to blurred outputs with only limited recovery of detail, deep neural networks learn the complex spatial structures necessary for reconstructing sharp textures and edges with high fidelity.

Attention-based architectures such as the Residual Channel Attention Network (RCAN) [13], [14] and generative models such as the Super-resolution Generative Adversarial Network (SRGAN) [15], [16] have demonstrated substantial improvements over conventional interpolation methods. More recently, Enhanced SRGAN (ESRGAN) [17] introduced residual-in-residual dense blocks and relativistic adversarial training for sharper texture reconstruction, while SwinIR [18], [19] leverages Swin Transformer-based self-attention to capture both local and global spatial dependencies — capabilities particularly beneficial for large-scale satellite scene reconstruction. Although there have been these advances, there is still a significant gap in the existing literature in that no previous study has systematically benchmarked deep learning super-resolution architectures by using MisrSat-2 — a newly launched and operationally important satellite — as a high-resolution reference, and no study has dealt with the particular problem of filling in the temporal gaps of MisrSat-2 by means of super-resolved Sentinel-2 imagery.

Another issue which has not received much attention in the existing multi-sensor super-resolution literature is the spatial misalignment between images obtained by different satellite platforms that have different orbital geometries, sensor designs, and acquisition attitudes. If this misalignment is not corrected, it directly reduces the quality of training for super-resolution models by causing spatial inconsistencies between the low-resolution input images and the high-resolution reference images. In our previous study [20], we set up a basic pipeline for Sentinel-2/MisrSat-2 super-resolution using RCAN and SRGAN along with a manual co-registration process based on GIS. This study extends this foundation in the following ways: (1) by introducing a fully automated co-registration framework which calls for no manual intervention; (2) by increasing the size of the dataset to involve further urban areas and additional acquisition dates; and (3) by incorporating two more state-of-the-art architectures, namely ESRGAN and SwinIR, in order to establish a comprehensive comparative benchmark.

The main contributions of this study are therefore the following. First, we provide the first systematic benchmark of four deep learning super-resolution architectures — RCAN, SRGAN, ESRGAN, and SwinIR — which has been carried out under consistent experimental conditions with real MisrSat-2 imagery used as the high-resolution reference, obtaining a best PSNR of 26.629 dB and an SSIM of 0.750 using the RCAN model. Second, we develop a fully automated co-registration framework that eliminates the need for manual intervention and reduces the misalignment between Sentinel-2 and MisrSat-2, dropping the misalignment from tens of metres (with an image-level RMSE of 26–84 m) to an RMSE of 2.3–4.8 m in four out of the five evaluated images (10.7 m in the fifth), and to 2.83 m at independent check points not employed in the estimation of the correction, thus greatly enhancing the efficiency and scalability of data preparation. Third, we show the intended operational application of the pipeline by means of a temporal gap-filling case study carried out over 6th of October City, Egypt, presenting this as a feasibility demonstration and not as a validation of the reconstruction accuracy. These contributions help to promote the practical application of deep learning-based super-resolution for higher-resolution Earth observation over Egypt and offer a reproducible framework that can be used in other multi-sensor super-resolution situations involving newly launched satellites which have limited temporal coverage. This study focuses on single-image super-resolution: each Sentinel-2 image is improved individually, and there is no spatio-temporal fusion of multiple MisrSat-2 observations. The incorporation of explicit spatio-temporal fusion models has been identified as future work.

2. LITERATURE REVIEW

2.1 Evolution of Deep Learning-Based Super-Resolution

The area of image SR, especially when it comes to satellite images, has seen very rapid development since the introduction of deep learning. Early methods employed interpolation techniques such as bicubic or Lanczos interpolation [21], which in turn usually resulted in excessive smoothing and the loss of fine details.

A significant turning point occurred in 2014 with the introduction of Convolutional Neural Networks (CNNs), more specifically the Super-resolution Convolutional Neural Network (SRCNN) [22]. The SRCNN used a simple three-layer CNN architecture in order to learn end-to-end mappings from low-resolution (LR) to high-resolution (HR) images and, as a result, attained considerable improvements when compared to traditional techniques.

Based on this foundation, the VDSR (Very Deep Super-resolution) method [23] used a deeper CNN that was inspired by VGG-net, enabling the model to learn complex hierarchical features and residual mappings in order to speed up convergence. FSRCNN (Fast Super-resolution CNN) [24] came next, aiming at computational efficiency by incorporating deconvolution layers to achieve faster upsampling, whereas ESPCN (Efficient Sub-Pixel CNN) [25] suggested the use of sub-pixel convolutional layers for real-time super-resolution, a feature particularly advantageous for real-time video applications. Further development was achieved with the introduction of EDSR (Enhanced Deep Super-resolution) [26], a version of ResNet obtained by removing the batch normalization layers in order to improve both accuracy and training stability. In the same study [26], MDSR (Multi-Scale Deep SR) was also introduced, enabling the handling of multiple upscaling factors within a single model, and multi-scale feature extraction was later enhanced in MDCN [27]. This stage marked a shift towards more robust residual-based learning frameworks capable of extracting fine-grained image details.

A major advancement was achieved with the introduction of attention-based mechanisms through the development of the Residual Channel Attention Network (RCAN) [13], the network combining residual-in-residual structures with channel attention in order to enhance the informative features while at the same time suppressing the less useful ones, which allowed it to perform well in preserving fine spatial details. More recently, the Deep Recurrent Residual Channel Attention Networks [14] continued along this line by adding recurrent connections in order to improve feature reuse. Similarly, a study from 2025 produced an improved super-resolution framework for Sentinel-2 that incorporates high-pass frequency enhancement together with channel attention mechanisms and finds that it has lower reconstruction errors than bicubic interpolation and DSen2 [28].

2.2 Rise of Perceptual and Generative Models

A significant advancement was achieved with the introduction of Generative Adversarial Networks (GANs) into super-resolution applications. The SRGAN [15] model used both an adversarial loss and a VGG-based content loss and was the first approach to generate photo-realistic images at a 4× upscaling factor, producing results that looked visually realistic and aligned with human visual perception. Although SRGAN achieved considerable improvements in perceptual quality, it sometimes generated unwanted artifacts. ESRGAN [17] addressed these issues by introducing the Residual-in-Residual Dense Block (RRDB) and a relativistic discriminator, which enhanced texture recovery and reduced instability during training. Later, Real-ESRGAN [29] was developed with an emphasis on robustness regarding real-world degradations by employing a high-order degradation model to generate training data and using a U-Net discriminator. SPSR (Structure-Preserving Super-resolution) [30] included a gradient branch in order to better preserve geometric structures when increasing the image size—a feature that is important for remote sensing tasks in which spatial accuracy matters.

2.3 Transformer-Based Super-Resolution Models

The development of Transformer architectures introduced a new approach to super-resolution. Models like IPT (Image Processing Transformer) [31] and SwinIR [18] employed self-attention mechanisms in order to capture long-range dependencies in images, a capability that CNNs found difficult to achieve. SwinIR, which is based on the Swin Transformer architecture, combined attention based on local windows with hierarchical feature extraction and thus attained state-of-the-art results on tests using natural images. More recently, a transformer-based super-resolution model that was designed specifically for Sentinel-2 imagery used multi-head attention together with integrated spatial and channel attention mechanisms and achieved a PSNR of 33.52 dB and an SSIM of 0.862 on the Sentinel-2 RGB bands, beating the ResNet and ViT-based baseline models [32].

In 2023, SRFormer [33] proposed permuted self-attention, which increased the size of the attention window at a relatively low computational cost. Likewise, the efficient frequency feature aggregation transformer (EFATSR) [34] improved the representation of information in both the spatial and frequency domains, resulting in better visual quality and more accurate reconstruction.

2.4 Super-Resolution for Satellite Imagery

The use of deep learning SR in relation to satellite imagery has now become a key area of research, this being due to the increasing demand for high-resolution remote sensing data in the fields of environmental monitoring, agriculture and urban development [1]–[3]. A recent review of deep-learning-based SR methods and their applications [35] points out that the majority of the existing approaches are assessed using synthetically degraded data, which limits their practical usefulness in real-world situations. More recently, Vu-Duc et al. [11] looked at over 50 SR remote sensing datasets that were publicly available and found that the widespread use of unrealistic (synthetic) degradation models and the geometric misalignment between the image pairs represented key limitations, stressing the importance of using real, co-registered image pairs for supervised training. A number of studies have applied SR specifically to Sentinel-2 imagery. Lanaras et al. [12] proposed a deep neural network that is suitable for global use in order to super-resolve the lower-resolution Sentinel-2 bands to 10 m, thus showing that it is possible to apply learning-based SR to multispectral satellite data. Malczewska and Wielgosz [36] applied SRCNN to Sentinel-2 imagery and obtained a PSNR of 61.8 and an SSIM of 0.99 when enhancing images from 20

m to 10 m, markedly surpassing traditional interpolation methods. Research that has combined Sentinel-2 with other high-resolution satellite sensors as reference high-resolution images has also yielded promising results. Wang et al. [37] increased the resolution of Sentinel-2 from 10 m to 4 m by using GAN models trained on Gaofen-2 data, while Müller et al. [38] trained CNNs on pairs of multispectral and pan-sharpened images, with the residual encoder–decoder network (RedNet) [39] proving to be the best performing among the architectures they evaluated. A major drawback in many of the studies currently available is their use of synthetically degraded data for training, since this does not adequately represent actual imaging conditions in the real world. In order to overcome this limitation, Kowaleczko et al. developed the MuS2 benchmark, a real-world dataset that enables super-resolution of Sentinel-2 images with WorldView-2 serving as the actual high-resolution reference [40], and thus underlined the need for training data that involves real sensor pairs. In a similar way, more recent research into multi-sensor super-resolution has looked at using one satellite as a high-resolution reference to super-resolve images from another, in order to bridge the resolution gap between different types of sensors [41]. Nevertheless, none of these studies have used MisrSat-2, Egypt's newly launched high-resolution satellite, as a high-resolution reference, which constitutes a significant gap in the context of Earth observation applications in Egypt.

2.5 Co-Registration in Multi-Sensor Super-Resolution

A problem which is often neglected but important in multi-sensor super-resolution is the spatial misalignment between images obtained by different satellites having different viewing geometries, acquisition angles, and temporal baselines. If the co-registration is not accurate, it directly reduces the performance of the super-resolution model by causing inconsistencies between the low-resolution inputs and the high-resolution references during training. More recent studies have shown the close interdependence between super-resolution and image registration, demonstrating that improving image resolution by means of super-resolution can enhance the accuracy of downstream registration by making the edges and corner features sharper [42]. In our previous work [20] we developed a co-registration procedure for use with Sentinel-2 and MisrSat-2 by means of manual registration using GIS tools. The current study extends this by presenting a fully automated framework requiring no manual intervention. This contribution directly addresses this challenge in the context of a newly launched satellite with limited prior co-registration literature.

3. METHODS

This study extends our prior published work [20], in which a deep learning pipeline was first established for super-resolving Sentinel-2 imagery using MisrSat-2 as a high-resolution reference. In that study, the RCAN and SRGAN architectures were trained and evaluated on a dataset covering four urban areas of interest in Egypt — 6th of October City, 10th of Ramadan City, El Shorouk City and Madinaty — using a co-registration procedure based on manual GIS georeferencing to align image pairs before training. Fig. 1 illustrates the pipeline used in the current study, an extension of the earlier study in three important respects: first, the spatial coverage is increased from the original four areas to nine by incorporating Kayan City, the New Capital, Sheikh Zayed City, El Sadat City and New Cairo; second, the time frame is widened by adding further acquisition dates (June–August 2024); and third, the manual co-registration process is replaced with a completely automated approach that does not require any manual input, thereby greatly reducing the preprocessing time and enhancing scalability when applied to various areas of interest.

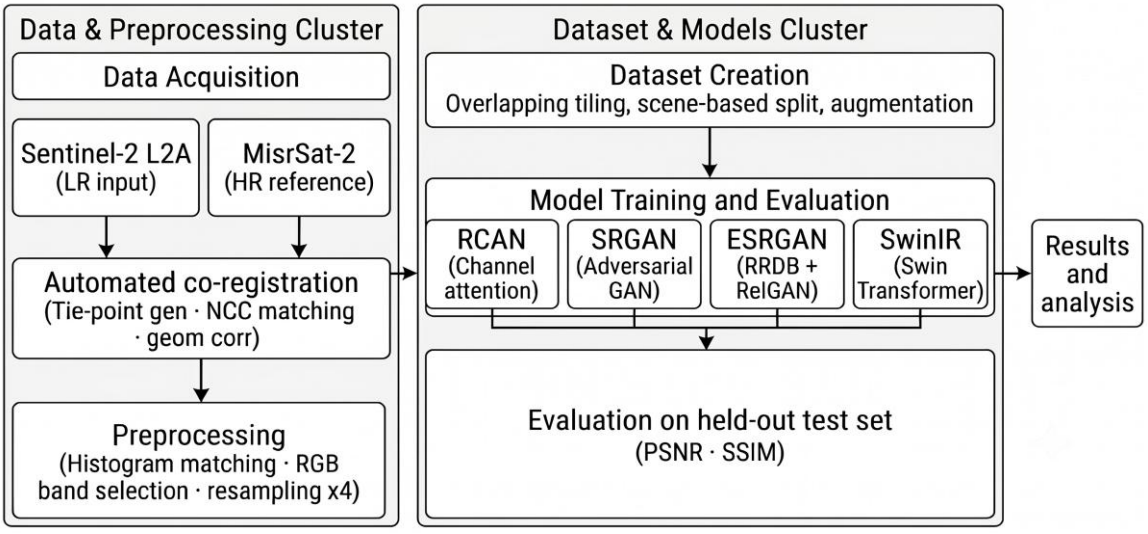


Fig. 1. Experimental pipeline for super-resolution of Sentinel-2 imagery using MisrSat-2 as a high-resolution reference, covering data acquisition, automated co-registration, preprocessing, dataset creation, model training, and evaluation.

3.1 Dataset Preparation

a. Data Sources

The research utilizes Sentinel-2 Level-2A (L2A) imagery as the LR input and MisrSat-2 data as the HR reference. Sentinel-2, part of the European Space Agency's Copernicus Programme, provides freely available optical imagery with 13 spectral bands at spatial resolutions of 10 m, 20 m, and 60 m, and a revisit frequency of approximately five days at the equator and two to three days at higher latitudes. Despite its high temporal frequency, Sentinel-2's coarser spatial resolution limits its ability to capture fine urban details.

MisrSat-2, operated by the Egyptian Space Agency, complements Sentinel-2 by providing higher spatial resolution imagery through a 2 m panchromatic band and 8 m multispectral bands covering the Blue, Green, Red, and NIR spectral regions. These characteristics make MisrSat-2 well suited for detailed urban mapping and infrastructure monitoring. However, its temporal revisit interval — 55 days at nadir and approximately 4 days at ±35° off-nadir — restricts its temporal coverage and creates observational gaps. By integrating both datasets, this study leverages Sentinel-2's temporal frequency and MisrSat-2's spatial detail, applying super-resolution methods to enhance Sentinel-2 imagery during periods when MisrSat-2 coverage is unavailable.

The MisrSat-2 2 m pansharpened product was used as the high-resolution reference. Because the Sentinel-2 input has a 10 m ground sampling distance and the super-resolution scale factor is ×4, the MisrSat-2 imagery was resampled from its native 2 m ground sampling distance to 2.5 m using nearest-neighbour interpolation and aligned to the Sentinel-2 grid (the grid alignment procedure is detailed in the Preparing the Dataset for the Model subsection), so that each LR–HR tile pair corresponds to an exact ×4 relationship. The target resolution of the super-resolved output is therefore 2.5 m rather than the 2 m native resolution of the MisrSat-2 product. The same resampling procedure was applied in our earlier work [20].

The 2.5 m target grid follows from the ×4 scale factor: it corresponds to an integer factor relative to the 10 m Sentinel-2 grid, so that every Sentinel-2 pixel maps exactly onto a 4 × 4 block of target pixels, which keeps each LR–HR pair geometrically consistent and matches the integer up-sampling of the evaluated architectures and the setting of [20]. Reaching the 2 m native resolution of MisrSat-2 would require a ×5 factor, making the already ill-posed reconstruction more difficult, while 2.5 m remains close to the native resolution. Because the ratio between 2 m and 2.5 m is not an integer, nearest-neighbour resampling preserves the original MisrSat-2 pixel values but skips one in every five source pixels along each axis; this introduces positional offsets of less than 1 m and slight aliasing in the HR targets, which form part of the residual differences between the super-resolved outputs and their references.

TABLE I. COMPARISON BETWEEN THE EQUIVALENT BANDS OF MISRSAT-2 AND SENTINEL-2.

MisrSat-2 Band	Wavelength Range (µm)	Closest Sentinel-2 Band	Sentinel-2 Wavelength Range (µm)	Center Wavelength (nm)	Bandwidth (nm)
Blue	0.45–0.52	B2 (Blue)	0.458–0.523	490	66
Green	0.52–0.59	B3 (Green)	0.543–0.577	560	36
Red	0.63–0.69	B4 (Red)	0.650–0.680	665	31
NIR	0.77–0.89	B8 (NIR)	0.785–0.899	842	106
Panchromatic	0.50–0.90	Covers B2–B8A	0.458–0.865	490–865	66–106

b. Study Areas

The expanded dataset covers nine urban growth areas across Egypt: 6th of October City, Sheikh Zayed City, 10th of Ramadan City, El Shorouk City, Madinaty, Kayan City, part of the New Capital, El Sadat City and New Cairo. Compared with [20], which covered 6th of October, 10th of Ramadan, El Shorouk and Madinaty, the present study adds Kayan City, the New Capital, Sheikh Zayed City, El Sadat City and New Cairo. For dataset construction these areas were clipped from four MisrSat-2 acquisitions into seven scenes (Table II): El Shorouk and Madinaty form a single scene, as do Kayan City and the clipped part of the New Capital. These areas represent government-prioritized development zones designed to accommodate Egypt's growing population and support national urban expansion plans. Their inclusion ensures that the dataset captures a diverse range of urban morphologies, land-cover types, and spatial patterns, supporting more reliable model training, validation, and evaluation for monitoring Egypt's rapidly evolving urban landscape.

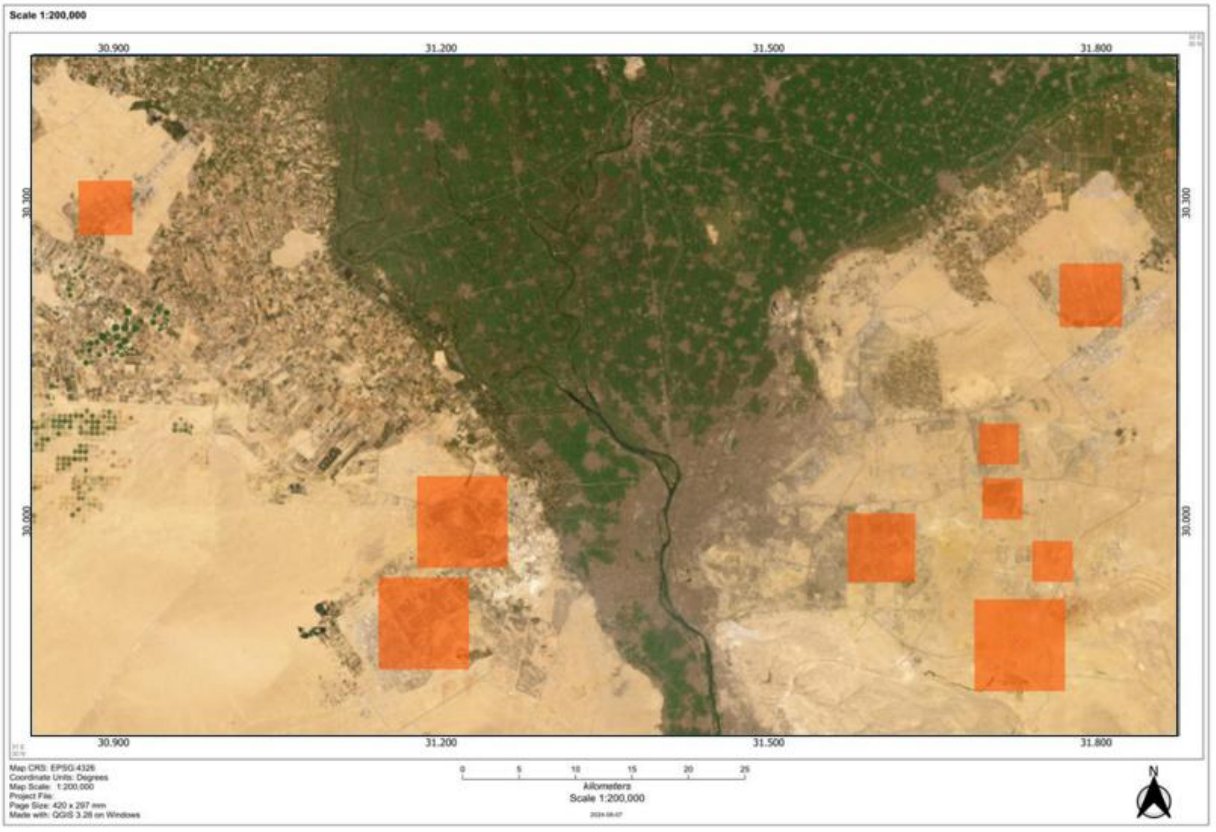


Fig. 2. Geographic distribution of the selected areas of interest (AOIs) across Egypt used for dataset creation and model evaluation. Nine areas are shown; El Shorouk and Madinaty, and Kayan City and part of the New Capital, were each clipped as a single scene, giving the seven scenes of Table II.

c. Dataset Creation

In this study, imagery from MisrSat-2 and its equivalent from Sentinel-2 were selected for the same geographic location and the closest possible acquisition date for several Areas of Interest (AOIs) across Egypt, as illustrated in Fig. 2. The selected locations include 6th of October City, 10th of Ramadan City, El Shorouk City and Madinaty, Kayan City and part of the New Capital, El Sadat City, Sheikh Zayed City, and New Cairo, clipped into seven scenes. The corresponding acquisition dates and the subset to which each scene was assigned are summarized in Table II.

The interval between the two acquisitions of each pair is reported in Table II. It ranges from 3 days (El Sadat City) to 66 days (10th of Ramadan), with three scenes at or below 11 days and four between 38 and 66 days. All acquisitions fall between June and August 2024, within the same dry summer season, which limits seasonal and phenological differences between the paired images; the AOIs are predominantly urban and desert land cover with limited vegetation. Summer dates were also selected to obtain cloud-free imagery, and both images of every scene pair were confirmed by visual inspection to be free of clouds and cloud shadows.

To screen for genuine land-cover change, each co-registered image pair was inspected visually by the first author in ArcGIS Pro, with the Sentinel-2 and MisrSat-2 images displayed side by side. Areas showing new urban development or construction activity, crop rotation, or major vegetation change between the two dates were flagged as changed. No significant changes were found in the three scenes with intervals of 11 days or less (6th of October, Sheikh Zayed City and El Sadat City). In each of the four scenes with longer intervals (El Shorouk, Kayan City, New Cairo and 10th of Ramadan), between one and three localized changed areas were identified; these were small relative to the AOI and were excluded before tiling so that the retained LR–HR pairs depict the same ground conditions. Minor differences below the threshold of visual detection, together with differences in building shadows and relief displacement arising from the different acquisition times and viewing geometries of the two sensors, cannot be fully excluded. Such residual differences act as label noise during training and, in the test scenes, would be expected to lower rather than inflate the reported scores, since no model can reconstruct a feature that is absent from its Sentinel-2 input. After data selection, several preprocessing steps were performed to prepare the dataset for the super-resolution model training. These steps are described in the following subsections.

TABLE II. SELECTED AOIS WITH MISRSAT-2 AND SENTINEL-2 ACQUISITION DATES AND THE INTERVAL BETWEEN THEM, WITH THE SUBSET TO WHICH EACH SCENE WAS ASSIGNED AND THE NUMBER OF TILES IN EACH SUBSET (BEFORE AUGMENTATION). SCENES WITH THE SAME MISRSAT-2 DATE WERE CLIPPED FROM THE SAME ACQUISITION.

Area of Interest	MisrSat-2 Date	Sentinel-2 Date	Interval (days)	Subset	Tiles
6th of October	13-06-2024	24-06-2024	11	Training	1,238
El Shorouk & Madinaty	29-07-2024	21-06-2024	38		
El Sadat City	27-06-2024	24-06-2024	3		
Sheikh Zayed City	13-06-2024	24-06-2024	11	Validation	1,008
New Cairo	29-07-2024	21-06-2024	38		
10th of Ramadan	26-08-2024	21-06-2024	66	Test	1,008
Kayan City & New Capital (part)	29-07-2024	21-06-2024	38		

d. Satellite Image Registration

For super-resolution, both the LR images (Sentinel-2) and the HR images (MisrSat-2) must be precisely co-registered. Any slight co-registration deviation between the two datasets can lead to inaccurate model training and misleading outputs.
Because MisrSat-2 and Sentinel-2 were acquired by different satellite missions, they exhibit different geometric distortions due to differences in orbit, sensor design, and acquisition viewing angles. To address this spatial misalignment, the ArcGIS Georeferencing Tool was previously used to manually register both HR and LR images.
The co-registration process is time-consuming and not feasible for multiple AOIs. In this study, an automated co-registration method has been developed. This method facilitates precise co-registration of satellite images through automated tie-point identification and geometric correction, aligning uncorrected MisrSat-2 images to the Sentinel-2 reference images. MisrSat-2 images were aligned using Sentinel-2 as the reference dataset. This ensured that all LR-HR image pairs were spatially consistent and well prepared for training the deep learning super-resolution model.
Tie points are generated directly from the Sentinel-2 reference image. Sentinel images were clipped using the extension of MisrSat-2 equivalent images, but with a buffer to ensure that the generated tie points would cover the entire area of the MisrSat-2 images. Each tie point was located within the corresponding Sentinel-2 tile. This tile was subsequently pattern-matched with the larger MisrSat-2 tile, including the same tie point. Through this matching process, the equivalent tile that includes the tie point is identified within the larger MisrSat-2 tile. Based on the offsets between the two sets of tie points, a geometric transformation is applied to the MisrSat-2 images to achieve accurate co-registration with the Sentinel-2 reference image.

e. Tie Points Generation

An automated identification code was developed to generate tie points from Sentinel-2, which serves as the reference image for correcting registration between Sentinel-2 and MisrSat-2 images. The process involved defining a buffer around each MisrSat-2 image, clipping the corresponding Sentinel-2 image within this buffer to ensure the generated tie points covered the entire MisrSat-2 image, and then using the clipped Sentinel-2 image as a reference to extract Tie points (TPs). Three approaches were implemented for generating tie-point chips from Sentinel-2 reference images:

- Shapefile-based extraction: Chips cropped around polygon centroids defined by the AOI shapefile.
- Random sampling: Chips randomly extracted from the image while avoiding no-data areas.
- Grid-based sampling: the image was divided into a grid, and chips were generated at the grid centers.

In this technique, the chip size was adaptively selected based on the dimensions of the reference image, large images (>7000 px) had a tile size of 512 × 512 pixels, medium images (5000–7000 px) had a tile size of 256 × 256 pixels, and small images (<5000 px) had a tile size of 128 × 128 pixels. This adaptive sizing strategy ensured that tie points were distributed uniformly across the entire reference image.
The GeoTIFF chips were obtained by using a moving window with its center at the selected point (whether this was random, grid-based or at the centroid of a shapefile). The dimension of the moving window corresponded to the tile size which had been defined in accordance with the size of the reference image as already described, and only those chips containing valid (non-zero) data were kept, thus eliminating the blank or noisy samples.
For each TP chip, relevant metadata was recorded in a central CSV database, the information comprising the Chip ID, the UTM and geographic coordinates, the Zone and CRS, the Image bounds, the Ground Sampling Distance (GSD), and the Acquisition and processing details. This arrangement ensured that all the generated TPs could be traced and reproduced, which in turn allowed for the exact determination of the tie-point reference tile needed to correct each MisrSat-2 image. The

automated registration workflow produced approximately 48,000 tie points, as shown in Fig. 3, covering the entirety of Egypt, ensuring a dense, spatially comprehensive distribution of control points across all regions.

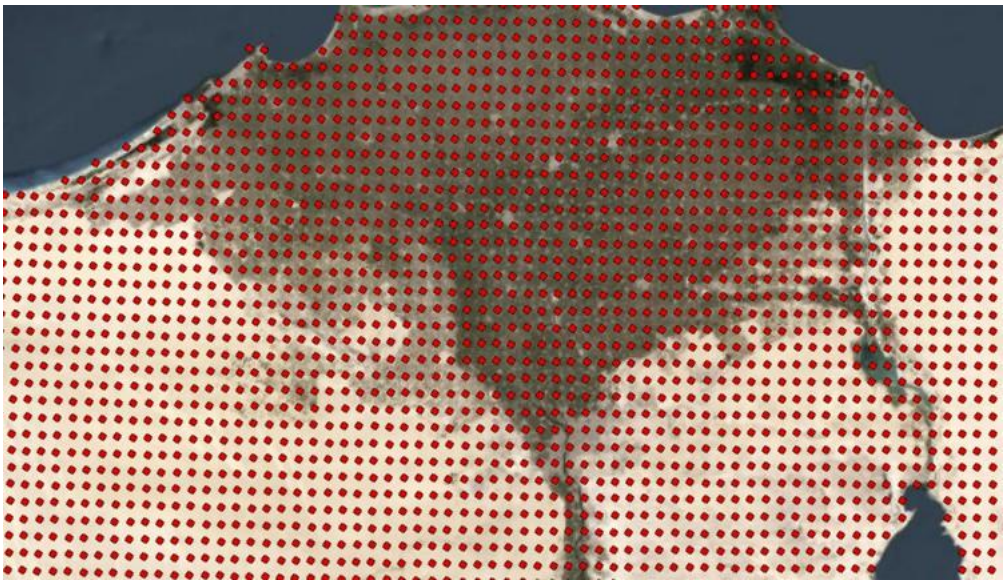


Fig. 3. Generated Tie points from Sentinel-2 Images cover all Egypt.

f. MisrSat-2 Correction Using Sentinel Tie Points

When the tie points (TP) images and the related metadata dataframe have been generated, any MisrSat-2 image that needs correcting will make use of the TPs that lie within the image's spatial boundaries. The filtering process keeps only those tie points for which the latitude and longitude are within the extent of the uncorrected image. Once the dataframe is filtered, additional information is appended, including the row and column location of each TP within the uncorrected image, calculated by transforming the TP's center coordinates (x, y) into the uncorrected image's coordinate system, and the x and y coordinates of the TP in the reference image.



Fig. 4. Representative tie-point chip images generated from Sentinel-2 reference imagery.



(a) Tie Points in MisrSat-2 Image 1 (b) Tie Points in MisrSat-2 Image 2

Fig. 5. Distribution of generated tie points within MisrSat-2 images used for geometric correction.

Because the Sentinel-2 tie-point chips and the MisrSat-2 images may differ in coordinate reference system and pixel size, each chip is brought to the geometry of the MisrSat-2 image before matching. If the CRS of a chip differs from that of the MisrSat-2 image, the chip is reprojected to the image CRS using nearest-neighbour resampling; if its pixel size differs from that of the image, it is resampled to the image pixel size (2 m) using bilinear interpolation. Template matching, the 20-pixel threshold and all displacements reported in Tables III and IV are therefore computed on the 2 m MisrSat-2 pixel grid.
Once the dataframe containing the filtered TPs was finalized, pattern matching is applied between each TP image and its corresponding portion of the uncorrected image. The geometric correction is carried out in the following way:

- Go through the filtered dataframe: Deal with each TP one at a time in sequence.
- Set up a search window: Taking the center of the approximate TP, extract a search window from the uncorrected image; the window has to be twice the size of the TP image in order to allow for pattern matching.
- Check that the clipped image size is valid: Make sure the clipped area taken from the uncorrected image has the required dimensions.
- Load the TP image: Obtain and read the TP image from the path given in the dataframe.
- Carry out pattern matching: Use the template matching method [43] to find the top-left corner (row, column) of the TP within the uncorrected image by matching the TP image against the search window.
- Calculate the TP center: Derive the TP center by adding half the width and height of the TP image to the matched top-left coordinates.

Template matching makes use of normalized cross-correlation (NCC), and the location at which the NCC response is a maximum is taken as the candidate match since it is known that the tie-point (TP) tile lies within the clipped area of the uncorrected image. The highest correlation value therefore indicates the most likely and physically reasonable match. The dataframe is then updated with information obtained from pattern matching, including the row coordinate of the TP image within the uncorrected image, the column coordinate of the TP image within the uncorrected image, transformed x and y coordinates corresponding to the matched row and column, pixel-level shift between the TP in the reference and uncorrected images, and equivalent shift in meters, computed from the pixel-level distance.
Tie points whose displacement between the reference and uncorrected images exceeded 20 pixels (40 m on the 2 m MisrSat-2 grid) were excluded from estimation of the transformation; tie points within this threshold were retained as control points. Images with fewer than three retained tie points were not corrected. Otherwise, the transformation model was selected according to the number of retained tie points: a first-order polynomial for fewer than 10, a second-order polynomial for 10–20, and a spline transformation for more than 20. A first-order transformation was also used when the retained tie points did not adequately cover the image. The transformation is then applied to register the MisrSat-2 image with the Sentinel-2 reference image, and template matching is re-applied on the corrected image to re-measure the displacement of every tie point, including those excluded from estimation of the transformation. Tables III and IV report the displacement before and after correction. The number of retained and excluded tie points differs from image to image, depending on the distribution of tie-point displacements within each image; the number retained for each evaluated image is given in Table IV.
The automated registration workflow may encounter failures under certain conditions. Dense cloud cover can obscure ground features, preventing reliable template matching and resulting in inaccurate control point detection. Similarly, low-texture

regions, such as deserts or calm water bodies, provide minimal distinct features for matching, which can lead to ambiguous or incorrect tie-point placements. Additionally, areas experiencing significant seasonal changes, such as vegetation growth or snow cover, may exhibit temporal variations that disrupt feature correspondence between images.

TABLE III. DISPLACEMENT OF A SUBSET OF TIE POINTS PER MISRSAT-2 IMAGE BEFORE AND AFTER AUTOMATED CORRECTION. EACH SAMPLE IS ONE TIE POINT.

Image No.	Dist Before Corr (Pixels)	Dist Before Corr (Meters)	Dist After Corr (Pixels)	Dist After Corr (Meters)
Image 1 — independent check points (not used in fit)				
Image1-sample1	26.2488095	52.49762	1	2
Image1-sample2	29.73213749	59.46427	1.414213562	2.828427125
Image1-sample3	25.61249695	51.22499	2	4
Image1-sample4	24.8394847	49.67897	1.414213562	2.828427125
Image1-sample5	28.8444102	57.68882	1	2
Image1-sample6	25	50	1.414213562	2.828427125
Image 2 — control points (used in fit)				
Image2-sample1	16.12452	32.24903	1	2
Image2-sample2	13.60147	27.20294	2.236068	4.472136
Image2-sample3	17	34	1	2
Image2-sample4	14.21267	28.42534	0	0
Image2-sample5	17.20465	34.4093	0	0
Image2-sample6	17	34	1	2
Image 3 — control points (used in fit)				
Image3-sample1	9.433981	18.86796	2.236068	4.472136
Image3-sample2	13.34166	26.68333	1.414214	2.828427
Image3-sample3	13.89244	27.78489	1.414214	2.828427
Image3-sample4	8.062258	16.12452	0	0
Image3-sample5	12.64911	25.29822	1	2
Image3-sample6	18.68154	37.36308	0	0

TABLE IV. EVALUATING THE AUTOMATED CORRECTION USING RMSE TO MEASURE THE DIFFERENCE BETWEEN PIXELS BEFORE AND AFTER THE AUTOMATED REGISTRATION. RMSE IS COMPUTED OVER THE EVALUATED TIE POINTS OF EACH IMAGE.

Image No.	Number of Tie-Points used for correction	RMSE Before Corr (Meters)	RMSE After Corr (Meters)
Image1	29	53.56	2.828
Image2	32	31.843	2.309
Image3	23	26.255	2.582
Image4	14	74.79	10.69
Image5	26	83.63	4.80

Tables III and IV assess the correction at a number of tie points chosen from each image. In the case of Image 1, the six tie points that were evaluated were selected from among those which exceeded the 20-pixel threshold; since they were not used in estimating the transformation, they act as independent check points. With regard to Images 2 and 3, the six evaluated tie points lie within the threshold and were used as control points; as the spline transformation goes through its control points, the displacements of these points after correction show that they are consistent with the control points rather than indicating independent accuracy. For Images 4 and 5, the tie points that have been evaluated include both those that were retained and those that were excluded. In Table IV, the RMSE of each image is the square root of the mean squared displacement over the evaluated tie points of that image, where the displacement of a tie point is the distance in metres between its position in the reference image and in the uncorrected or corrected MisrSat-2 image.

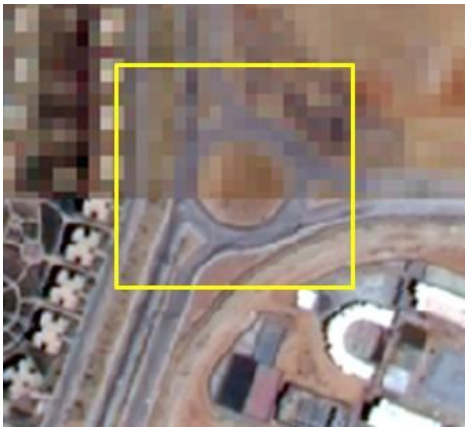
At the six independent check points of Image 1, the displacement fell from an RMSE of 53.56 m (median 51.86 m; range 49.7–59.5 m) before correction to an RMSE of 2.83 m after correction (median 2.83 m, mean 2.75 m, standard deviation 0.74 m, range 2–4 m). Because displacement magnitudes were recorded, the systematic component (bias) cannot be resolved by direction, but its magnitude is bounded by the mean displacement and therefore does not exceed 2.75 m. This error is comparable to that at the control points of Images 2 and 3 (RMSE 2.45 m, median 2.0 m; n = 12), indicating that the transformation generalizes to locations not used in its estimation. These check points are the tie points with the largest initial

displacements in Image 1 rather than a random selection, which makes them a demanding test. Displacements are measured by integer-pixel template matching and are therefore quantized in steps of one 2 m pixel.

The automated registration workflow significantly reduces preprocessing time compared to manual georeferencing using ArcGIS. For each area of interest (AOI), manual registration typically requires approximately 30 minutes to one hour, depending on the size and complexity of the AOI. By contrast, the automated method completes co-registration in approximately 3–5 minutes per AOI, depending on the number of tie points and AOI dimensions. This substantial reduction in processing time demonstrates the efficiency and scalability of the proposed workflow for multiple AOIs.



(a) Sample 1 Before Correction



(b) Sample 1 After Correction

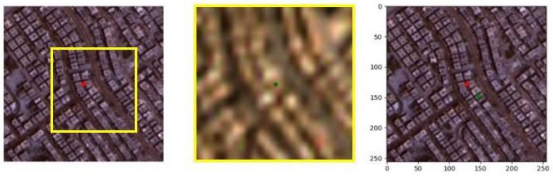


(c) Sample 2 Before Correction

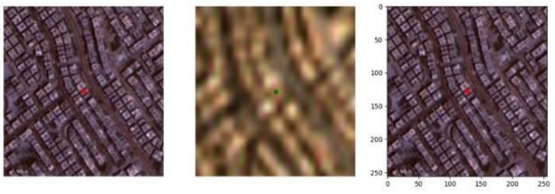


(d) Sample 2 After Correction

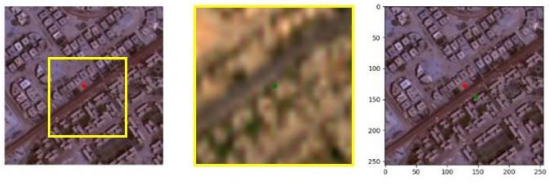
Fig. 6. Representative MisrSat-2 image samples before and after automated co-registration correction.



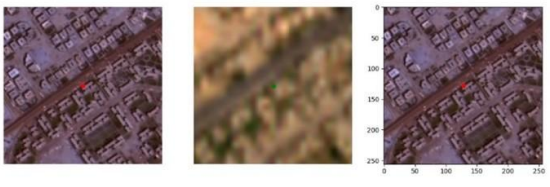
(a) Sample 1 Before Correction



(b) Sample 1 After Correction



(c) Sample 2 Before Correction



(d) Sample 2 After Correction



(e) Sample 3 Before Correction



(f) Sample 3 After Correction

Fig. 7. MisrSat-2 Images Before and After Correction. The green dot indicates the center in the reference image, and the red dot indicates the center of the uncorrected image. Images on the right showing only the red dot are corrected (the green dot at the bottom of the red dot indicates no or a very small shift).

3.2 Preparing the Dataset for the Model

All images lie within UTM zone 36N (WGS 84; EPSG:32636), the common projection of the Sentinel-2 L2A tiles and the MisrSat-2 products over the study area, so no reprojection of either dataset was required. The Sentinel-2 bands B2, B3 and B4 were used at their original 10 m pixel size for the super-resolution dataset (resampling to 2 m was applied only to the tie-point chips during co-registration), and their pixel grid served as the reference grid. After co-registration, which was performed at the native 2 m MisrSat-2 resolution, each MisrSat-2 image was resampled to 2.5 m using nearest-neighbour interpolation and aligned to the Sentinel-2 grid in ArcGIS Pro, so that the 2.5 m grid shares its origin with a Sentinel-2 pixel corner and every 10 m Sentinel-2 pixel covers exactly 4 × 4 MisrSat-2 pixels. Both images of each pair were then clipped to the same extent before tiling. Consequently, each 40 × 40 LR tile and its 160 × 160 HR counterpart cover the same 400 m × 400 m ground area, and the ×4 correspondence holds pixel by pixel.

The red, green, and blue bands from both the MisrSat-2 and Sentinel-2 satellites were chosen for training after the MisrSat-2 images had been registered with the Sentinel-2 images. Since there are differences in the radiometric distributions between MisrSat-2 and Sentinel-2, histogram matching was carried out on the MisrSat-2 images, taking Sentinel-2 as a reference, in order to reduce the radiometric discrepancies between the two sensors and to enable stable model training. The main aim of this preprocessing step was to enhance perceptual consistency and promote convergence during learning for the purpose of super-resolution reconstruction, not to maintain the actual physical reflectance values. Baseline experiments that did not include histogram matching resulted in lower similarity metrics because of the inherent radiometric differences between the two sensors, although the general performance trends did not change. For this reason, histogram matching was adopted as the normalization method in the experiments reported here. Applications that demand strict radiometric accuracy, such as the calculation of vegetation indices or change detection, were therefore not included in this study.

Histogram matching was applied separately to each scene after it had been clipped to its AOI polygon, and before tiling, using the co-registered Sentinel-2 image clipped to the same polygon as reference; only the MisrSat-2 (HR) images were transformed, and the Sentinel-2 (LR) inputs were left unchanged. Each scene is therefore matched using only its own pair of images, with no statistics pooled across scenes; in particular, each test reference was transformed using statistics from its own scene only, and no information from the training or validation scenes enters the test data. Because matching is performed independently for each scene, its result does not depend on the order of matching and subset assignment. All images were represented as 8-bit RGB (0–255), each scaled linearly between its own minimum and maximum values, so no clipping was required.

The upscaling factor was set to 4, following previous work [20]. To prevent spatial data leakage, each scene (the area clipped for one AOI, Table II) was assigned a unique numerical scene identifier, which forms the first field of the name of every tile derived from that scene. Subset assignment was performed on these scene identifiers before tiling, so that every tile of a given scene belongs to exactly one subset. The overlapping tiling described below was then carried out separately within each scene. Overlapping tiles can therefore occur only within a single scene, and hence within a single subset: scene-level assignment guarantees by construction that no overlapping tile is shared between the training, validation and test subsets. This strategy ensured that the test data originated from scenes unseen during training and validation.

Subset assignment was performed at the level of complete scenes, where a scene is the area clipped for one AOI from a MisrSat-2 acquisition. The seven scenes listed in Table II were clipped from four MisrSat-2 acquisitions, so some scenes share an acquisition, and every tile derived from a given scene was assigned exclusively to a single subset. Consequently no area of interest is shared between the training, validation, and test subsets, and no geographic location seen during training reappears at test time. The reported test metrics therefore reflect generalization to geographic locations that are entirely unseen during training, rather than generalization to new acquisition dates over locations already observed.

The training subset comprises three scenes (6th of October; El Shorouk with Madinaty; El Sadat City), the validation subset two scenes (Sheikh Zayed City and New Cairo) and the test subset two scenes (10th of Ramadan; Kayan City with part of the New Capital), as listed in Table II. The AOIs of different subsets do not overlap, so no ground area is shared between

subsets. Some scenes in different subsets were, however, clipped from the same MisrSat-2 acquisition: 6th of October (training) and the neighbouring Sheikh Zayed City (validation) come from one acquisition, and El Shorouk with Madinaty (training), New Cairo (validation) and Kayan City with the New Capital (test) from another. Performance on Kayan City therefore reflects generalization to a new location under acquisition conditions represented in training, whereas 10th of Ramadan, taken from an acquisition not used for training, provides the more independent test of the two.

The HR images were then split into 160 × 160-pixel tiles, while the corresponding LR images were split into 40 × 40-pixel tiles. Initially, tiling was performed without overlap; however, the resulting number of samples was insufficient for robust training. Consequently, an overlapping tiling strategy was adopted, producing 3,254 tiles. Subset assignment was carried out at scene level before tiling and before any augmentation, yielding 3,254 tiles in total: 1,238 training, 1,008 validation and 1,008 test tiles. The subset to which each scene was assigned and the number of tiles in each subset are reported in Table II; tile counts were recorded per subset rather than per scene. Data augmentation was subsequently applied exclusively to the training subset, expanding it from 1,238 to 3,025 tiles and giving 5,041 samples in total. These augmentations included mirroring, flipping, rotation, and adjustments to contrast and brightness. The validation and test subsets were left unaugmented, so all reported performance metrics are computed on unmodified imagery.

TABLE V. DATASET CONFIGURATION AND PARTITIONING SUMMARY.

Parameter	Value
Tiles before augmentation	3,254
Training tiles (before augmentation)	1,238
Training tiles (after augmentation)	3,025
Validation tiles (unaugmented)	1,008
Test tiles (unaugmented)	1,008
Total samples (after augmentation)	5,041
Augmentation applied to	Training subset only
HR patch size	160 × 160 px
LR patch size	40 × 40 px
Upscaling factor	×4

3.3 Implemented Models

This study builds upon the previously published framework [20], which introduced and described in detail a deep learning pipeline for super-resolving Sentinel-2 imagery using MisrSat-2 data through the Residual Channel Attention Network (RCAN) and Super-Resolution Generative Adversarial Network (SRGAN) models. In the current study, these models served as baseline comparators against the newly evaluated ESRGAN and SwinIR architectures. Expanding upon that foundation, two additional state-of-the-art architectures—Enhanced Super-Resolution GAN (ESRGAN) and SwinIR, a Transformer-based image restoration model—were introduced and evaluated to further improve perceptual and structural image quality. All models were trained from scratch under identical preprocessing and postprocessing pipelines so that the comparison is made under a common protocol.

a. Enhanced Super-Resolution Generative Adversarial Network (ESRGAN)

ESRGAN builds upon the original SRGAN framework by introducing architectural and training improvements that enable the generation of sharper and more realistic images. ESRGAN replaces the standard residual blocks with Residual-in-Residual Dense Blocks (RRDB), which combine residual learning and dense connections to enable deeper networks without introducing training instability.

The generator consists of an initial convolutional layer followed by a sequence of 23 RRDB blocks, each composed of three 3×3 convolutional layers with LeakyReLU activations [44] and a local residual connection [45]. After deep feature extraction, the network applies two successive upsampling stages implemented through convolutional layers followed by PixelShuffle operations and ReLU activations. A final 3×3 convolution layer reconstructs the super-resolved output.

The discriminator is composed of a series of convolutional blocks with increasing channel depths (64 → 128 → 256 → 512), each followed by Batch Normalization and LeakyReLU activations. The final convolution outputs a single-channel probability map. ESRGAN employs a relativistic discriminator (RaGAN) that estimates the probability that real images are more realistic than generated images.

The loss function combines three components: pixel-wise L1 loss to maintain fidelity with the ground truth; perceptual loss computed from the conv5_4 feature maps (before activation) of a VGG19 network pretrained on ImageNet and kept frozen, to encourage perceptual similarity; and adversarial loss to push the generator towards producing photo-realistic textures.

Training was performed consistently with the RCAN and SRGAN experiments, with a scale factor of ×4 and image patches of size 40×40 (LR) and 160×160 (HR). This configuration was shared by all models.
For ESRGAN, separate Adam optimizers [46] were employed for the generator and discriminator with a fixed learning rate of 2×10^{-4}. The optimizer parameters were set to $\beta1 = 0.9$ and $\beta2 = 0.999$, with zero weight decay. The generator loss consisted of a weighted combination of pixel-wise $\ell1$ loss, perceptual feature loss based on VGG-19, and adversarial loss. The adversarial component followed the relativistic GAN formulation. The discriminator was updated once per generator update (D update ratio = 1), and no discriminator warm-up iterations were applied.

b. **SwinIR**

SwinIR is a method for image restoration and super-resolution based on the transformer architecture, using the Swin Transformer to efficiently capture both local and global dependencies in a hierarchical way.
The network starts with a shallow feature extraction stage which makes use of a convolutional layer and an embedding dimension. This is then followed by a number of Residual Swin Transformer Blocks (RSTBs), each consisting of several stacked Swin Transformer Blocks connected by residual connections. A convolutional layer is used afterwards for feature fusion and image reconstruction.
To allow for the exchange of both local and cross-window information, Window-based Multi-Head Self-Attention (W-MSA) together with a shifted window strategy (SW-MSA) [47] is applied, and relative position bias is included in order to maintain the spatial relationships within each window.
SwinIR has an embedding dimension of 60, three RSTB layers, and a window size of 8. After the feature extraction and transformation phases, the model applies a PixelShuffle upsampling module and then a reconstruction convolutional layer to generate the final super-resolved image.
The model, following the official network architecture [18], was trained using a combination of pixel-wise L1 loss, a perceptual loss derived from VGG19 features and TV loss. The Adam optimizer was also employed, using the same dataset, patch sizes, and ×4 upscaling factor as in the other models, so that the evaluation could be kept consistent. Dynamic padding is used to deal with images whose dimensions are not divisible by the window size. SwinIR was trained with a fixed learning rate of 2×10^{-4}, $\beta1 = 0.9$, $\beta2 = 0.999$, and zero weight decay.

TABLE VI. SUMMARY OF SUPER-RESOLUTION MODEL MECHANISMS.

Model	Core Mechanism	Key Feature	Configuration	Upsampling	Loss
RCAN	Channel Attention (RCAB)	Deep residual learning	6 RG × 16 RCAB, 64 ch, r = 16 [20]	PixelShuffle	L1
SRGAN	ResNet-based GAN	Perceptual adversarial learning	5 residual blocks, 64 ch [20]	PixelShuffle	MSE + Perceptual + Adv + TV
ESRGAN	RRDB + Relativistic GAN	Dense residuals, sharper textures	23 RRDB blocks	PixelShuffle	L1 + Perceptual + Adv
SwinIR	Swin Transformer	Window-based Self-Attention	3 RSTB, embed. dim. 60, window 8	PixelShuffle	L1 + Perceptual + TV

3.4. Performance Evaluation Metrics

In order to maintain consistency and enable a fair comparison with previously published results [20], the same image quality evaluation metrics were used. The Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) were employed as quantitative measures of the performance of each super-resolution model. As this study focuses exclusively on RGB super-resolution rather than multispectral reflectance reconstruction, PSNR and SSIM — widely adopted in the image super-resolution literature — are considered appropriate for evaluating perceptual reconstruction quality within the defined scope of this work.

TABLE VII. TRAINING HYPERPARAMETERS APPLIED UNIFORMLY ACROSS ALL EVALUATED MODELS. ONLY THE GPU DIFFERS BETWEEN MODELS.

Hyperparameter	Value
Epochs	100
Batch size	32
Optimizer	Adam
Learning rate	2×10^{-4}
β1	0.9
β2	0.999
Weight decay	0

Hyperparameter	Value
Weight initialization	Random
Perceptual-loss network (SRGAN, ESRGAN, SwinIR)	VGG19, ImageNet-pretrained, frozen
Learning-rate schedule	Fixed (no decay)
Stopping criterion	Fixed 100 epochs (no early stopping)
Model selection	Checkpoint retained only when validation PSNR and SSIM both improved
GPU (RCAN, SRGAN, ESRGAN)	NVIDIA P100
GPU (SwinIR)	Dual NVIDIA T4

All four architectures were trained from randomly initialised weights under the identical configuration listed in Table VII. The learning rate was held fixed at 2×10^{-4} throughout training, with no decay schedule; training ran for a fixed budget of 100 epochs without early stopping, and a checkpoint was retained only when both validation PSNR and SSIM improved on the best values obtained so far, the most recent such checkpoint being used for testing. No model-specific hyperparameter search was carried out: the same settings were applied uniformly to all four models so that the comparison reflects architectural differences rather than differences in tuning effort. This uniformity is a deliberate choice for fairness, but it also means that the reported figures represent performance under a common configuration rather than each architecture's individually optimised performance.

The settings held constant across the four models are the training, validation and test data, the augmentation, the patch sizes and ×4 scale factor, the optimizer and its parameters, the learning rate and schedule, the number of epochs, the batch size and the checkpoint-selection rule. The architecture and the training objective differ by design, since they define each method (Table VI); the perceptual losses of SRGAN, ESRGAN and SwinIR use a VGG19 network pretrained on ImageNet and kept frozen, while the super-resolution networks themselves are randomly initialised. Loss weights were not tuned: for SRGAN, ESRGAN and SwinIR they were kept at the default values of the publicly available implementations on which our code is based, the ESRGAN code being adapted from the BasicSR toolbox. RCAN, SRGAN and ESRGAN were trained on an NVIDIA P100 GPU and SwinIR on dual NVIDIA T4 GPUs; the hardware affects training time but not the optimisation. Equal settings do not by themselves guarantee equal suitability for every architecture, so the comparison should be read as a comparison under a common, untuned protocol. The training and evaluation procedure shared by all four models is summarized in Algorithms 1 and 2.

Algorithm 1. Training and checkpoint selection, applied to each model m ∈ {RCAN, SRGAN, ESRGAN, SwinIR}.

```
Input:  augmented training pairs (40×40 LR, 160×160 HR); validation pairs;
        epochs E = 100; batch size B = 32; learning rate η = 2e-4;
        loss L_m as listed in Table VI
Output: selected generator G*
 1: initialise G_m (and discriminator D_m for SRGAN and ESRGAN) randomly
 2: if L_m has a perceptual term: load VGG19 with ImageNet weights, freeze it
 3: best_psnr ← −∞;  best_ssim ← −∞
 4: for epoch = 1, …, E do
 5:     for each mini-batch (x, y) of size B do
 6:         ŷ ← G_m(x)
 7:         compute L_m(ŷ, y); update G_m with Adam(η, β1 = 0.9, β2 = 0.999)
 8:         if m is a GAN: update D_m with its adversarial loss on (y, ŷ)
 9:             (relativistic for ESRGAN; G_m and D_m updated alternately)
10:     end for
11:     compute mean PSNR_val and SSIM_val of G_m on the validation tiles
12:     if PSNR_val > best_psnr and SSIM_val > best_ssim then
13:         best_psnr ← PSNR_val;  best_ssim ← SSIM_val;  G* ← G_m
14:     end if
15: end for
16: return G*
```

Algorithm 2. Evaluation on the held-out test tiles.

```
Input:  selected generator G*; unaugmented test tiles
Output: mean PSNR and SSIM
 1: for each test tile (x, y) do
 2:     ŷ ← G*(x), converted to 8-bit RGB
 3:     compute PSNR_t and SSIM_t of (ŷ, y) on valid (non-no-data) pixels
 4:         (scikit-image, data_range = 255)
 5: end for
 6: return the mean of PSNR_t and SSIM_t over all test tiles
```

SwinIR's lower PSNR is consistent with the data requirements of transformer-based architectures. Self-attention models learn spatial relationships with fewer built-in inductive biases than convolutional networks, and therefore typically require substantially larger training sets to reach their potential. With 1,238 training tiles before augmentation, the dataset used here is small by the standards on which SwinIR was originally benchmarked. The convolutional models — and RCAN in particular, whose channel-attention mechanism operates on top of a convolutional backbone — are better suited to this data regime. This result should therefore be read as reflecting SwinIR's behaviour under a limited-data, fixed-budget training configuration rather than an inherent limitation of transformer-based super-resolution.

The PSNR is determined by calculating the logarithm of the ratio between the squared highest possible pixel value and the mean squared error (MSE) between the reconstructed and the reference images. The MSE between a reference image I and a super-resolved image K of size m×n is defined as:

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2$$

PSNR is subsequently computed as:

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right)$$

where MAXI denotes the maximum possible pixel value of the image (255 for 8-bit images). Higher PSNR values indicate greater reconstruction fidelity.

SSIM assesses structural similarity between the super-resolved image and the ground truth by jointly comparing luminance, contrast, and structural components, and is defined as:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}$$

where μx and μy are the local mean intensities, σx² and σy² are the local variances, σxy is the covariance of images x and y, and C1 = (k1L)², C2 = (k2L)² are stabilization constants, with L denoting the dynamic range of pixel values and k1 = 0.01, k2 = 0.03. SSIM values range from -1 to 1, where a value of 1 indicates perfect structural similarity. Higher PSNR and SSIM values indicate better reconstruction accuracy and perceptual similarity, respectively. The same evaluation procedure was applied to all four models, ensuring direct comparability between RCAN, SRGAN, ESRGAN, and SwinIR results.

PSNR and SSIM were computed with scikit-image (peak_signal_noise_ratio and structural_similarity) directly on the 8-bit RGB images, without color-space conversion, using data_range = 255 and default settings (a 7 × 7 window for SSIM, computed for each RGB channel and averaged over the three channels). No-data pixels were excluded from the computation. Both metrics were computed for each test tile and then averaged over the 1,008 test tiles.

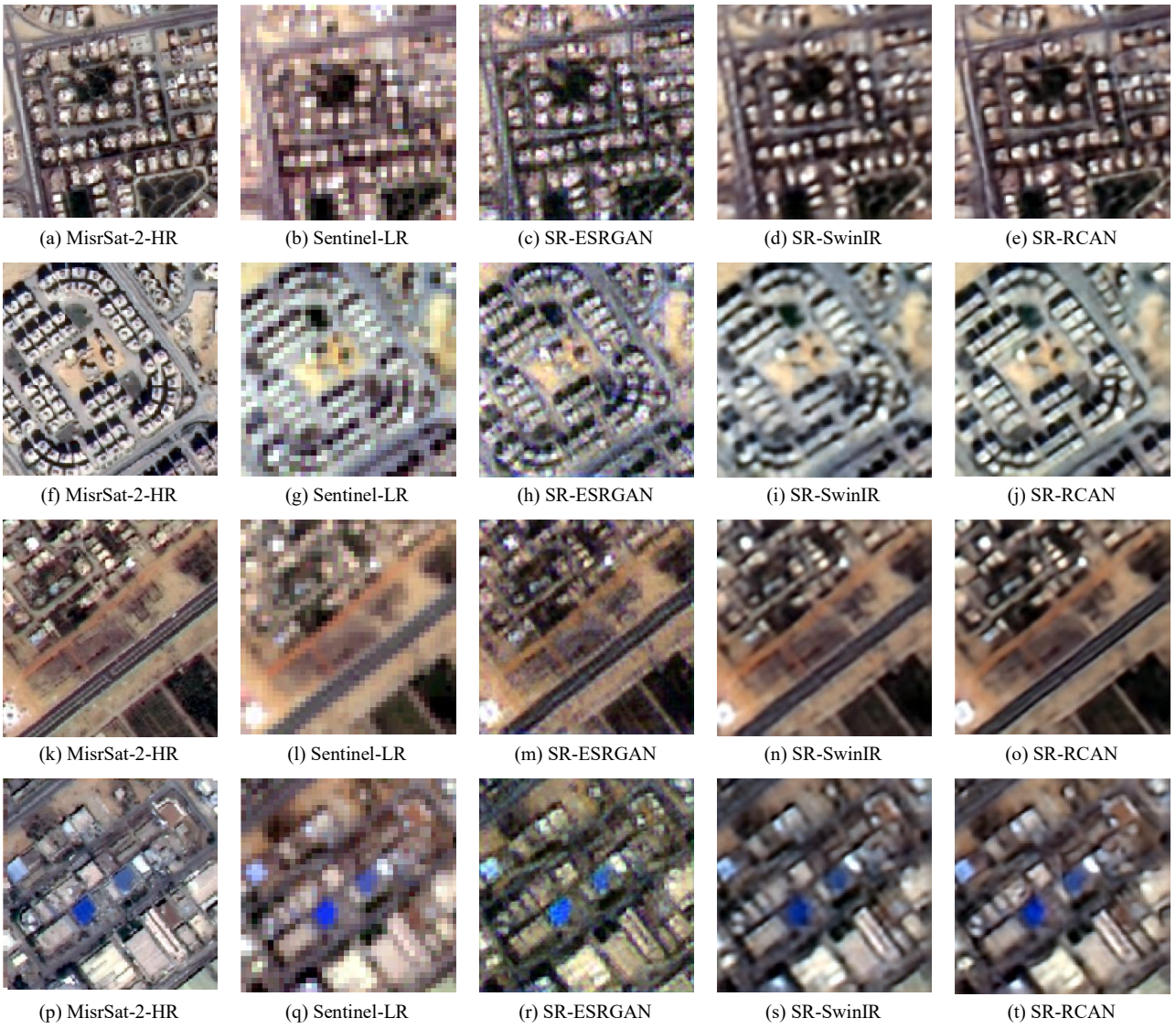
4. RESULTS

4.1 Algorithm Setup

RCAN, SRGAN and the Enhanced Super-Resolution Generative Adversarial Network (ESRGAN) were trained on an NVIDIA P100 GPU, while the Swin Transformer for Image Restoration (SwinIR) model was trained on dual NVIDIA T4 GPUs. All four models were trained for 100 epochs with a batch size of 32 under the shared configuration listed in Table VII. Full architectural descriptions of the evaluated models are provided in the Methods section.

4.2 Model Performance

Fig. 8 presents a visual comparison of the super-resolution results produced by different models alongside the original MisrSat-2 high-resolution reference and Sentinel-2 low-resolution input at four sample locations drawn from the two test scenes.



(a) MisrSat-2-HR (b) Sentinel-LR (c) SR-ESRGAN (d) SR-SwinIR (e) SR-RCAN

(f) MisrSat-2-HR (g) Sentinel-LR (h) SR-ESRGAN (i) SR-SwinIR (j) SR-RCAN

(k) MisrSat-2-HR (l) Sentinel-LR (m) SR-ESRGAN (n) SR-SwinIR (o) SR-RCAN

(p) MisrSat-2-HR (q) Sentinel-LR (r) SR-ESRGAN (s) SR-SwinIR (t) SR-RCAN

Fig. 8. Comparison of super-resolution results at four sample locations from the two test scenes. Each row shows a different location with the following columns: MisrSat-2 high-resolution reference, Sentinel-2 low-resolution input, and SR outputs from ESRGAN, SwinIR, and RCAN models.

Table VIII presents the quantitative evaluation of all tested super-resolution methods in comparison to traditional interpolation baselines on the held-out test set, comprising 1,008 tiles originating from scenes entirely unseen during training and validation. The test tiles were selected from the AOIs depicted in Fig. 2, as explained in the Dataset Preparation section, and the mean PSNR and SSIM values for the test tiles are given for both the histogram-matched (HM) and the non-histogram-matched (No-HM) training setups. Results for bilinear and bicubic interpolation are included only for the non-histogram-matched setting.

The results demonstrate that applying histogram matching as a preprocessing step yields modest improvements in PSNR and SSIM. However, performance without histogram matching remains comparable, indicating that the models maintain spatial structure and reconstruction quality even without radiometric normalization. The small performance gap suggests that histogram matching primarily enhances radiometric alignment rather than fundamentally improving the spatial reconstruction capability. The ranking of the four models is identical with and without histogram matching, for both PSNR and SSIM.

TABLE VIII. EVALUATION OF SUPER-RESOLUTION IMAGES WITH REFERENCE. METRICS ARE REPORTED FOR BOTH HISTOGRAM-MATCHED (HM) AND NON-HISTOGRAM-MATCHED (NO-HM) DATA. BILINEAR AND BICUBIC INTERPOLATION METHODS DO NOT INVOLVE HISTOGRAM MATCHING AND WERE APPLIED DIRECTLY TO THE LOW-RESOLUTION SENTINEL-2 IMAGES.

Method Name	PSNR (HM)	PSNR (No-HM)	SSIM (HM)	SSIM (No-HM)
Bilinear	-	18.90	-	0.34
Bicubic	-	18.80	-	0.31
RCAN model	26.629	25.80	0.750	0.740
SRGAN model	23.35	22.95	0.716	0.705
ESRGAN model	24.38	23.90	0.735	0.720
SwinIR model	22.68	22.15	0.744	0.730

The RCAN and SRGAN models were originally introduced in our earlier work [20] on a smaller dataset covering four areas of interest. For the present study both models were retrained from random initialisation on the training subset described in Section 3.2, using the configuration of Table VII, and evaluated on the same test set as ESRGAN and SwinIR, so that all four architectures are trained and scored on identical data under identical conditions. The values reported in Table VIII for RCAN and SRGAN therefore supersede those of [20], and differ slightly from them (26.76 dB / 0.756 and 23.40 dB / 0.720 respectively) as a consequence of the larger and more geographically diverse dataset and of the different training configuration ([20] used a learning rate of 10^{-3} on different hardware and a 60/20/20 percentage split of the tiles rather than the scene-level split used here).

After generating the super-resolved tiles using the RCAN model, a specialized merging pipeline was implemented to reconstruct the full HR image without visible seams or boundary artifacts. The large satellite images were first split into overlapping tiles, with extra buffer margins placed around the edges of the tiles in order to retain the contextual information needed for the RCAN prediction. The overlap together with the buffer meant that the pixels close to the edges of the tiles could be predicted with sufficient surrounding context, which in turn reduced the discontinuities between neighbouring tiles. The overlapping super-resolved tiles were then reassembled by means of a weighted averaging merging technique instead of the conventional "first-value overwrite" method which is normally used in raster mosaicking. Specifically, all the corresponding pixel values from the overlapping areas were added up and then divided by the number of tiles contributing to the average. This method of averaging successfully removed the gridline artifacts which usually occur when the tiles are simply stitched together, resulting in an output that was both visually smooth and spatially continuous. This last step guarantees the seamless combination of all the predicted tiles and maintains both radiometric continuity across the boundaries between tiles and spatial coherence over the whole reconstructed image.

4.3 Temporal Gap-Filling Case Study

On 26 July 2025, no MisrSat-2 acquisition was available over 6th of October City, whereas a Sentinel-2 acquisition was. The trained RCAN model was applied to this Sentinel-2 image to produce a 2.5 m super-resolved proxy (Fig. 9). The nearest available MisrSat-2 acquisition, from 16 July 2025, ten days earlier, is shown for visual comparison, and Table IX reports PSNR and SSIM against it. Because the two images are ten days apart in an area of active construction, and are also affected by residual registration error, these values combine reconstruction error with genuine ground change and misregistration; they are therefore not a measure of reconstruction accuracy for the missing date. In addition, 6th of October City is one of the training scenes (acquired in 2024), so this example concerns a new acquisition date over a location represented in training rather than an unseen location. The case study is therefore presented as a feasibility illustration of how the pipeline would be used between MisrSat-2 acquisitions; quantitative accuracy is assessed on the held-out test scenes (Table VIII).



(a) LR-Sentinel-2 L2A (26-07-2025) (b) SR-RCAN (26-07-2025) (c) MisrSat-2 (16-07-2025)

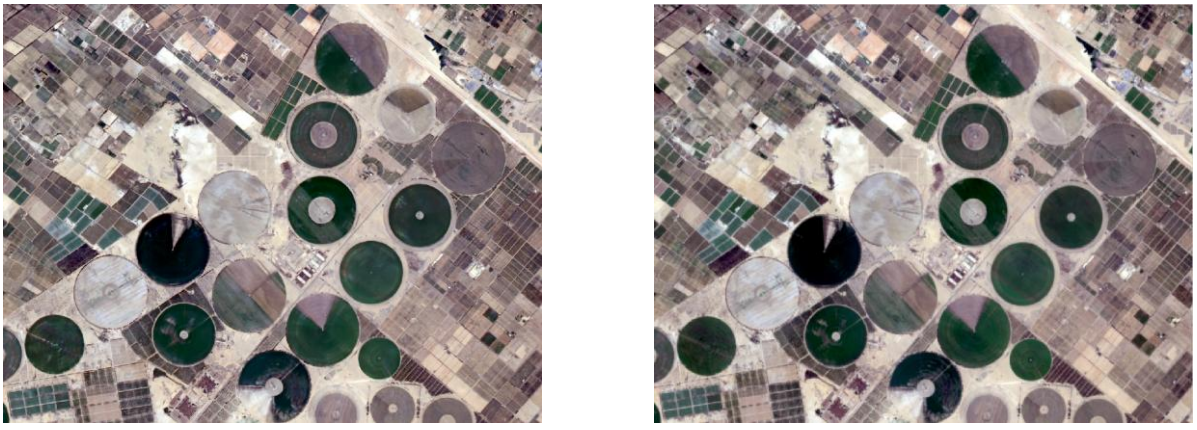
Fig. 9. Temporal gap-filling case study [48] over 6th of October City. (a) Low-resolution Sentinel-2 L2A input, (b) super-resolved RCAN output, and (c) nearest available MisrSat-2 reference image. Image (c) was acquired ten days before (a) and is not an exact reference.

TABLE IX. SIMILARITY BETWEEN THE SUPER-RESOLVED IMAGE OF 26 JULY 2025 AND THE NEAREST AVAILABLE MISRSAT-2 IMAGE (16 JULY 2025). VALUES ARE AFFECTED BY THE TEN-DAY INTERVAL AND ARE NOT A MEASURE OF RECONSTRUCTION ACCURACY.

Method Name	PSNR	SSIM
RCAN model	24.208	0.692

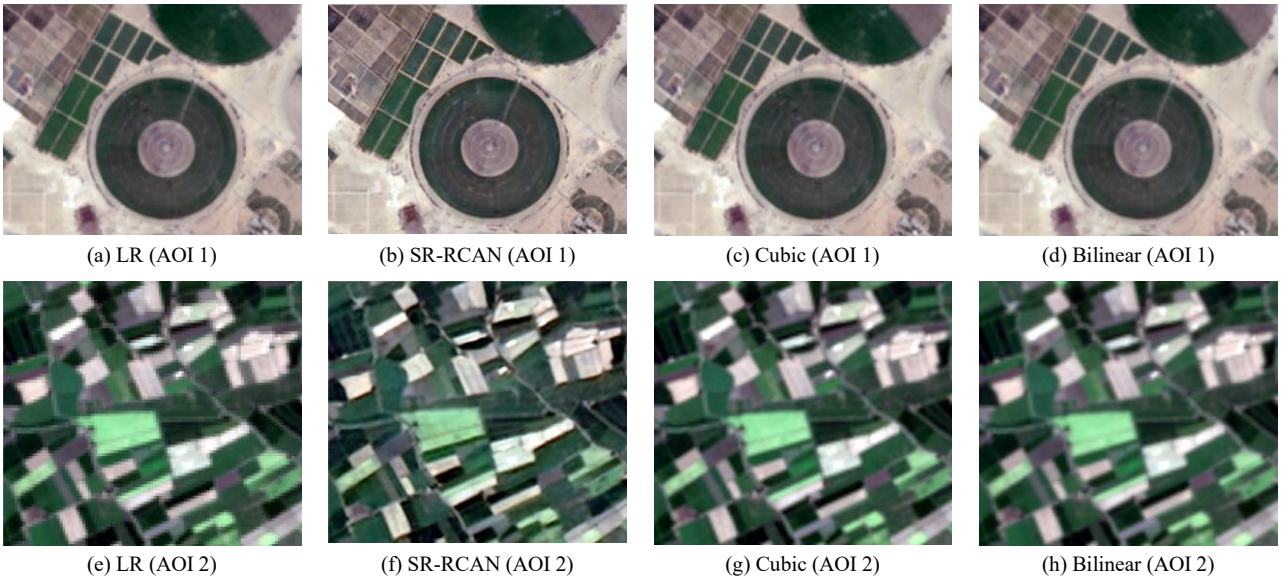
4.4 Illustrative Application to Agricultural Areas

As an illustrative example, the RCAN model was also applied to Sentinel-2 imagery over two agricultural areas, one with centre-pivot irrigation and one with small cultivated fields, a land-cover type not represented in the training scenes, which are predominantly urban and desert (Figs. 10 and 11). No MisrSat-2 or other high-resolution reference was available for these areas, which mirrors the operational setting the pipeline is designed for, in which no same-date high-resolution reference exists. The super-resolved images appear sharper than the Sentinel-2 input and than bicubic and bilinear interpolation, with crisper field boundaries. Because no reference is available, the output was assessed with a low-resolution consistency check (Table X): when downsampled to the 10 m Sentinel-2 grid, the super-resolved output agrees closely with the input in structure (SSIM of 0.95 and 0.85), indicating that the model does not distort the information present in the input. This check is a necessary but not a sufficient condition: it cannot establish that the high-frequency detail added by the model, such as sharper field boundaries, corresponds to real structures on the ground. The agricultural examples are therefore presented as visual illustrations supported by a consistency check, not as evidence of generalization to agricultural land cover.



(a) LR-Sentinel-2 L2A (b) SR-RCAN

Fig. 10. Pivot irrigation areas comparison: low-resolution Sentinel-2 L2A input and super-resolved RCAN output.



(a) LR (AOI 1) (b) SR-RCAN (AOI 1) (c) Cubic (AOI 1) (d) Bilinear (AOI 1)

(e) LR (AOI 2) (f) SR-RCAN (AOI 2) (g) Cubic (AOI 2) (h) Bilinear (AOI 2)

Fig. 11. Qualitative comparison of SR-RCAN with traditional interpolation methods across two unseen AOIs. Each row represents a different AOI.

a. **Low-Resolution Consistency Check**

When no high-resolution reference image is available, the SR output can be downsampled to the original LR grid and compared with the input Sentinel-2 image. It should be emphasised that this comparison measures low-resolution consistency — whether the SR output remains faithful to the information present in its own input — and does not validate the correctness of the high-frequency detail introduced by the model. Establishing the accuracy of newly generated detail requires a high-resolution reference and is therefore outside the scope of this check. PSNR and SSIM are accordingly reported here as consistency indicators rather than as measures of reconstruction accuracy.

Table X presents the PSNR and SSIM metrics for downscaled SR images from the two agricultural AOIs of Fig. 11. When the SR image is reduced to the original LR size, a moderate pixel-wise error is observed relative to the LR image. In the context of super-resolution evaluation, PSNR values in the mid-20s indicate that the SR output does not achieve perfect pixel-level consistency with the LR reference, which may result from minor spatial shifts, smoothing, or the generation of new textures. Conversely, the high SSIM score indicates that the structural layout, shapes, and overall scene geometry in the downscaled SR image closely correspond to those of the LR image. This combination of high SSIM and moderate PSNR suggests that, although exact pixel-level agreement is lacking, the SR output preserves the structural content of its input. This demonstrates internal consistency of the reconstruction; it does not establish that the high-frequency detail added by the model corresponds to features actually present on the ground.

TABLE X. LOW-RESOLUTION CONSISTENCY OF THE RCAN SUPER-RESOLVED OUTPUT, MEASURED BY DOWNSAMPLING TO THE SENTINEL-2 INPUT GRID.

AOI	PSNR	SSIM
Area 1	24.15	0.95
Area 2	23.20	0.85

5. DISCUSSION

This study presented and evaluated a scalable workflow for generating 2.5 m proxy imagery over Egypt for dates on which no MisrSat-2 acquisition is available, using deep learning-based super-resolved Sentinel-2 data (10 m GSD) against the newly launched MisrSat-2 satellite as reference. Four state-of-the-art SR architectures were systematically evaluated — RCAN, SRGAN, ESRGAN, and SwinIR — under identical experimental conditions using real MisrSat-2 imagery as the high-resolution reference, representing the first such benchmark conducted with this sensor.

The fully automated co-registration framework introduced in this study reduced the misalignment between Sentinel-2 and MisrSat-2 from an image-level RMSE of 26–84 m to 2.3–4.8 m in four of the five evaluated images (10.7 m in the fifth), and to 2.83 m at independent check points that were not used to estimate the correction, requiring no manual intervention and completing processing in approximately 3–5 minutes per area of interest compared to 30–60 minutes for manual GIS-based georeferencing — representing a substantial improvement in both accuracy and preprocessing efficiency.

Among the evaluated models, RCAN achieved the highest scores on both metrics, with a PSNR of 26.629 dB and SSIM of 0.750 under histogram-matched conditions, outperforming SRGAN (PSNR = 23.35 dB, SSIM = 0.716), ESRGAN (PSNR = 24.38 dB, SSIM = 0.735), and SwinIR (PSNR = 22.68 dB, SSIM = 0.744). These results indicate that, under the present experimental conditions, the attention-based RCAN architecture achieved the highest structural fidelity and the lowest pixel-wise error of the four models tested. The case study involving the temporal gap-filling for 6th of October City demonstrates the intended way in which the pipeline is to be used. The PSNR of 24.208 dB and SSIM of 0.692 as compared with a MisrSat-2 image taken ten days earlier include both the reconstruction error and actual ground changes together with residual misregistration, and thus show only feasibility rather than providing validated reconstruction accuracy. The agricultural examples are given as visual illustrations accompanied by a low-resolution consistency check; since there are no high-resolution references, they do not demonstrate that the method generalizes to non-urban land cover. Taken together, these results give Egyptian monitoring authorities a practical foundation for carrying out higher-resolution observations of dynamic urban and environmental change.

The 2.5 m outputs produced by the model must be regarded as proxies similar to MisrSat-2 imagery, not as new 2.5 m observations; the only information that they contain which has actually been measured is that from the 10 m Sentinel-2 input, while the detail at a scale below 10 m is inferred by the network from the patterns it acquired during training. The model can therefore produce plausible-looking structures, such as building edges or field boundaries, that are not supported by the input, and the low-resolution consistency check cannot rule this out. This matters most for downstream mapping and change detection: features smaller than a Sentinel-2 pixel should not be mapped from the super-resolved product alone, and an apparent change between two super-resolved images may reflect differences in generated detail rather than real change on the ground. We therefore recommend using the super-resolved products for visual interpretation and as a complement to,

rather than a substitute for, real high-resolution observations, with changes at sub-10 m scale confirmed against the Sentinel-2 input or high-resolution imagery.

a. **Limitations**

When interpreting these results several constraints should be taken into account. Firstly, the test dataset includes two full scenes (10th of Ramadan and Kayan City), a number which is too small to allow for inferential statistical testing; the comparative statements mentioned above therefore refer only to the ranking obtained under these experimental conditions and not to any claim of statistical significance. Secondly, the metrics reported are based on histogram-matched images and measure the structural and perceptual quality of the reconstruction rather than the absolute radiometric fidelity; hence applications that demand strict reflectance accuracy are not within the scope of this study. Thirdly, independent validation of the co-registration is restricted to six check points in one image, and for the other images the tie points evaluated include control points that were used to calculate the transformation; a more rigorous estimate of the residual registration error could be obtained with a larger, spatially distributed set of independent check points. Fourthly, the temporal gap-filling case study has no high-resolution image from the same date and relates to a location that was included in the training data; thus, the scores obtained against an image taken ten days earlier cannot be interpreted as the reconstruction accuracy for the missing date. Lastly, the agricultural examples were created without the use of high-resolution references; the low-resolution consistency check only confirms agreement with the Sentinel-2 input and does not verify the correctness of the added detail, so these examples do not show generalization to agricultural land cover.

There are still several avenues open for future research. Applying the SR pipeline to the multispectral bands would allow the derivation of vegetation and environmental indices such as the Normalized Difference Vegetation Index (NDVI). Further improvements in visual sharpness while preserving quantitative accuracy could be achieved by optimizing hybrid loss functions that combine perceptual, adversarial, and fidelity terms. The integration of spatiotemporal fusion models would improve temporal consistency in the generated imagery. Finally, rigorous benchmarking of the automated co-registration workflow under a variety of geographic and seasonal conditions would establish its wider industrial applicability.

Conflicts of Interest

The authors declare no competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment

The authors would like to acknowledge the Egyptian Space Agency for providing access to MisrSat-2 imagery and the European Space Agency for the freely available Sentinel-2 data through the Copernicus programme. The authors acknowledge the use of Claude AI (Anthropic) to assist with text organization and improve the clarity and readability of this manuscript. All scientific content, analysis, and conclusions were independently developed and verified by the authors.

Author Contributions

B.R. conceived the experiments and conducted the data preprocessing, model training, and evaluation. B.E. provided access to MisrSat-2 data and contributed to data acquisition. M.S. supervised the research and contributed to methodology design. All authors reviewed the manuscript.

Data Availability

Sentinel-2 data used in this study are publicly available from the Copernicus Data Space Ecosystem. Access to MisrSat-2 imagery is subject to authorization by the Egyptian Space Agency due to data policy restrictions. Therefore, the raw MisrSat-2 datasets cannot be publicly redistributed. To support reproducibility, a representative sample of super-resolved outputs, along with their corresponding low-resolution inputs, as well as sample image pairs before and after the automated co-registration process, are publicly available online [48]. All preprocessing procedures, dataset partitioning protocols, model configurations, and evaluation methodologies are fully described in the manuscript to ensure reproducibility. Trained model weights and inference scripts may be provided for academic research purposes upon reasonable request and subject to approval.

References

[1] J. Yang, D. Ahn, J. Bahk, S. Park, N. Rizqihandari, and M. Cha, “Assessing climate risks from satellite imagery with machine learning: A case study of flood risks in Jakarta,” Climate Risk Management, vol. 46, art. no. 100651, 2024, doi: 10.1016/j.crm.2024.100651.

[2] T. Vaiyapuri and G. Julie, “AI and geospatial technologies for climate change mitigation: Opportunities, challenges, and pathways to sustainability,” Procedia Computer Science, vol. 259, pp. 1346–1355, 2025, doi: 10.1016/j.procs.2025.04.089.

[3] Y. Xi et al., “A satellite imagery dataset for long-term sustainable development in United States cities,” Scientific Data, vol. 10, no. 1, art. no. 866, 2023, doi: 10.1038/s41597-023-02576-3.

[4] C. Liu et al., “An efficient approach to capture continuous impervious surface dynamics using spatial-temporal rules and dense Landsat time series stacks,” Remote Sensing of Environment, vol. 229, pp. 114–132, 2019, doi: 10.1016/j.rse.2019.04.025.

[5] Z. Gu and M. Zeng, “The use of artificial intelligence and satellite remote sensing in land cover change detection: Review and perspectives,” Sustainability, vol. 16, no. 1, art. no. 274, 2024, doi: 10.3390/su16010274.

[6] H. Song, Q. Liu, G. Wang, R. Hang, and B. Huang, “Spatiotemporal satellite image fusion using deep convolutional neural networks,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 11, no. 3, pp. 821–829, 2018, doi: 10.1109/JSTARS.2018.2797894.

[7] S. Fazilov, O. Yusupov, S. Radjabov, E. Eshonqulov, and K. Abdieva, “Analysis of pansharpening methods for multispectral satellite imagery,” in AIP Conference Proceedings, vol. 3244, art. no. 030080, 2024, doi: 10.1063/5.0241414.

[8] B. Kamala, V. Ramana, S. Giridharan, and J. Ranjani, “Advanced pansharpening techniques for satellite image fusion,” in Proc. 2024 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), 2024, pp. 1–6, doi: 10.1109/ICPECTS62210.2024.10780411.

[9] A. Gatti and A. Bertolini, “Sentinel-2 products specification document,” Thales Alenia Space, Tech. Rep. S2-PDGS-TAS-DI-PSD, Issue 13.1, Nov. 2015. [Online]. Available: https://sentinels.copernicus.eu/documents/247904/349490/S2_MSI_Product_Specification.pdf

[10] X. Wang et al., “A review of image super-resolution approaches based on deep learning and applications in remote sensing,” Remote Sensing, vol. 14, no. 21, art. no. 5423, 2022, doi: 10.3390/rs14215423.

[11] A. Vu-Duc, K. Nguyen-Vi, B. Bui-Quoc, and N. Kamel, “A comprehensive survey of super-resolution remote sensing image datasets: Evolution, challenges, and future directions,” IEEE Access, vol. 13, pp. 145350–145372, 2025, doi: 10.1109/ACCESS.2025.3599535.

[12] C. Lanaras, J. Bioucas-Dias, S. Galliani, E. Baltsavias, and K. Schindler, “Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network,” ISPRS Journal of Photogrammetry and Remote Sensing, vol. 146, pp. 305–319, 2018, doi: 10.1016/j.isprsjprs.2018.09.018.

[13] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proc. European Conference on Computer Vision (ECCV), 2018, pp. 294–310, doi: 10.1007/978-3-030-01234-2_18.

[14] Y. Liu, D. Yang, F. Zhang, Q. Xie, and C. Zhang, “Deep recurrent residual channel attention network for single image super-resolution,” The Visual Computer, vol. 40, no. 5, pp. 3441–3456, 2024, doi: 10.1007/s00371-023-03044-0.

[15] C. Ledig et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 105–114, doi: 10.1109/CVPR.2017.19.

[16] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” IEEE Signal Processing Magazine, vol. 35, no. 1, pp. 53–65, 2018, doi: 10.1109/MSP.2017.2765202.

[17] X. Wang et al., “ESRGAN: Enhanced super-resolution generative adversarial networks,” in Proc. European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 63–79, doi: 10.1007/978-3-030-11021-5_5.

[18] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, “SwinIR: Image restoration using Swin Transformer,” in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 1833–1844, doi: 10.1109/ICCVW54120.2021.00210.

[19] M. Puttagunta, R. Subban, and C. N. K. Babu, “SwinIR Transformer applied for medical image super-resolution,” Procedia Computer Science, vol. 204, pp. 907–913, 2022, doi: 10.1016/j.procs.2022.08.110.

[20] B. Raafat, M. Safy, B. Elleithy, and A. M. EL-Shal, “Sentinel-2 images’ super-resolution through MisrSat-2 and deep learning methods,” Journal of Al-Azhar University Engineering Sector, vol. 21, pp. 239–250, 2026, doi: 10.21608/auej.2026.482882.

[21] Z. Wang, J. Chen, and S. C. H. Hoi, “Deep learning for image super-resolution: A survey,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 10, pp. 3365–3387, 2021, doi: 10.1109/TPAMI.2020.2982166.

[22] C. Dong, C. C. Loy, K. He, and X. Tang, “Learning a deep convolutional network for image super-resolution,” in Proc. European Conference on Computer Vision (ECCV), 2014, pp. 184–199, doi: 10.1007/978-3-319-10593-2_13.

[23] J. Kim, J. K. Lee, and K. M. Lee, “Accurate image super-resolution using very deep convolutional networks,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1646–1654, doi: 10.1109/CVPR.2016.182.

[24] C. Dong, C. C. Loy, and X. Tang, “Accelerating the super-resolution convolutional neural network,” in Proc. European Conference on Computer Vision (ECCV), 2016, pp. 391–407, doi: 10.1007/978-3-319-46475-6_25.

[25] W. Shi et al., “Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1874–1883, doi: 10.1109/CVPR.2016.207.

[26] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017, pp. 1132–1140, doi: 10.1109/CVPRW.2017.151.

[27] J. Li, F. Fang, J. Li, K. Mei, and G. Zhang, “MDCN: Multi-scale dense cross network for image super-resolution,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 7, pp. 2547–2561, 2021, doi: 10.1109/TCSVT.2020.3027732.

[28] K. Nguyen-Vi, B. Bui-Quoc, and N. Kamel, “Deep learning-based Sentinel-2 super-resolution via channel attention and high-frequency feature enhancement,” Frontiers in Remote Sensing, vol. 6, art. no. 1644460, 2025, doi: 10.3389/frsen.2025.1644460.

[29] X. Wang, L. Xie, C. Dong, and Y. Shan, “Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data,” in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 1905–1914, doi: 10.1109/ICCVW54120.2021.00217.

[30] C. Ma, Y. Rao, J. Lu, and J. Zhou, “Structure-preserving image super-resolution,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 11, pp. 7898–7911, 2022, doi: 10.1109/TPAMI.2021.3114428.

[31] H. Chen et al., “Pre-trained image processing transformer,” in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 12294–12305, doi: 10.1109/CVPR46437.2021.01212.

[32] A. Sharifi and M. M. Safari, “Enhancing the spatial resolution of Sentinel-2 images through super-resolution using transformer-based deep-learning models,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 18, pp. 4805–4820, 2025, doi: 10.1109/JSTARS.2025.3526260.

[33] Y. Zhou, Z. Li, C.-L. Guo, S. Bai, M.-M. Cheng, and Q. Hou, “SRFormer: Permuted self-attention for single image super-resolution,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12734–12745, doi: 10.1109/ICCV51070.2023.01174.

[34] J. Song, A. Sowmya, and C. Sun, “Efficient frequency feature aggregation transformer for image super-resolution,” Pattern Recognition, vol. 167, art. no. 111735, 2025, doi: 10.1016/j.patcog.2025.111735.

[35] H. Su, Y. Li, Y. Xu, X. Fu, and S. Liu, “A review of deep-learning-based super-resolution: From methods to applications,” Pattern Recognition, vol. 157, art. no. 110935, 2025, doi: 10.1016/j.patcog.2024.110935.

[36] A. Malczewska and M. Wielgosz, “How does super-resolution for satellite imagery affect different types of land cover? Sentinel-2 case,” IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 340–363, 2024, doi: 10.1109/JSTARS.2023.3328997.

[37] X. Wang, Z. Ao, R. Li, Y. Fu, Y. Xue, and Y. Ge, “Super-resolution image reconstruction method between Sentinel-2 and Gaofen-2 based on cascaded generative adversarial networks,” Applied Sciences, vol. 14, no. 12, art. no. 5013, 2024, doi: 10.3390/app14125013.

[38] M. U. Müller, N. Ekhtiari, R. M. Almeida, and C. Rieke, “Super-resolution of multispectral satellite images using convolutional neural networks,” ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, vol. V-1-2020, pp. 33–40, 2020, doi: 10.5194/isprs-annals-V-1-2020-33-2020.

[39] X.-J. Mao, C. Shen, and Y.-B. Yang, “Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 29, 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/hash/0ed9422357395a0d4879191c66f4faa2-Abstract.html

[40] P. Kowaleczko et al., “A real-world benchmark for Sentinel-2 multi-image super-resolution,” Scientific Data, vol. 10, no. 1, art. no. 644, 2023, doi: 10.1038/s41597-023-02538-9.

[41] P. W. Shin et al., “Towards high-resolution alignment and super-resolution of multi-sensor satellite imagery,” in Proc. IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2025, pp. 2799–2807, doi: 10.1109/ICCVW69036.2025.00292.

[42] Y. Im and Y. Lee, “Robust multi-resolution satellite image registration using deep feature matching and super resolution techniques,” Applied Sciences, vol. 16, no. 2, art. no. 1113, 2026, doi: 10.3390/app16021113.

[43] N. S. Hashemi, R. B. Aghdam, A. S. B. Ghiasi, and P. Fatemi, “Template matching advances and applications in image analysis,” arXiv preprint arXiv:1610.07231, 2016, doi: 10.48550/arXiv.1610.07231.

[44] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in Proc. ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013. [Online]. Available: https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf

[45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[46] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. International Conference on Learning Representations (ICLR), 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[47] Z. Liu et al., “Swin Transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 9992–10002, doi: 10.1109/ICCV48922.2021.00986.

[48] B. Raafat, M. Safy, and B. Elleithy, “Super-resolution dataset sample,” GitHub repository, 2026. [Online]. Available: https://github.com/Bassantraafat98/Super-Resolution-Dataset-Sample (accessed Oct. 25, 2026).