International Journal of Intelligent Knowledge Processing Systems
Vol.1, No.2, June, 2026, pp. 73-82
Journal homepage: https://ijikps.org/

VISION TRANSFORMER AND DIFFUSION MODEL-BASED FRAMEWORK FOR HIGH-RESOLUTION SATELLITE IMAGE SUPER-RESOLUTION AND CLASSIFICATION

Nallappagari Venkatarami Reddy

Independent Researcher, Bellemead , New Jersy, USA ,08502.
(nvn1227@gmail.com)

Corresponding Author Email: nvn1227@gmail.com

ABSTRACT

High-resolution satellite images are very important in many remote sensing applications such as land use monitoring, environmental assessment, urban planning, precision agriculture, and disaster management. But, imaging sensors, atmospheric disturbances and transmission constraints can often lead to poor quality satellite images and consequently poor efficacy in the image analysis. Then, to solve the above problem, this paper presents a Vision Transformer and Diffusion Model-Based Framework for HR-SIR and classification of satellite images. A diffusion probabilistic model for image super-resolution is combined with a classification module based on Vision Transformer to enhance image quality and boost scene recognition performance simultaneously. Low-resolution satellite images are preprocessed first and then fed into the diffusion-based reconstruction network which gradually filters out noise and restores high-frequency spatial features, resulting in high-resolution images. The reconstructed images are then fed into a Vision Transformer with self-attention mechanism to learn global context features and obtain discriminative features for scene classification. A comprehensive set of experiments were performed with the use of benchmark remote sensing datasets and evaluated by Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Accuracy, Precision, Recall and F1-Score. The accuracy for the classification achieved by the proposed model was 98.4%, which is better than the accuracies for the classifications done using other models such as SRCNN, EDSR, SRGAN, SwinIR and EDiffSR. The model also has a Precision of 98.2%, Recall of 98.0% and F1 score of 98.1%. The diffusion framework and Vision Transformers contribute to the enhanced fidelity, structural preservation, and classification accuracy of images, rendering the proposed framework applicable to the development of sophisticated remote sensing applications.

Keywords: *Satellite Image Super-Resolution, Vision Transformer (ViT), Diffusion Probabilistic Model, Remote Sensing Image Classification, Deep Learning, High-Resolution Satellite Imagery, Image Reconstruction, Scene Classification, Transformer Networks, Earth Observation Systems.*

1. INTRODUCTION

The development of Earth Observation systems and satellite imaging technologies have greatly improved the amount of remote sensing data available for many applications such as land use monitoring, environmental studies, urban planning, disaster management and precision agriculture [1]. The use of high-resolution satellite images is essential in these applications as they provide detailed spatial information needed for accurate analysis and decision making [2]. The satellites, however, are limited by their imaging sensors, weather conditions, transmission bandwidth and expense in acquiring images, so they are often acquired at low spatial resolutions and lose information and are hard to interpret [3].

One effective method for creating high-resolution pictures from low-resolution data is Image Super-Resolution (SR) [4]. In order to maintain fine-grained texture and structural details, many conventional interpolation-based methods fail [5]. These methods include nearest-neighbor, bilinear, and bicubic interpolation. Using deep learning techniques such as Convolutional Neural Networks (CNNs), Generative Adversarial Networks (GANs), Transformers, and Diffusion Models, new super-resolution methods have been developed in the past few years [6]. In instance, diffusion models have attained state-of-the-art results in picture generation and reconstruction by gradually removing noise from an image through their iterative denoising process [7], and transformer architectures have demonstrated to be extremely effective in learning long-range spatial context.

The Vision Transformers (ViTs) have attracted a lot of attention in the remote sensing image analysis field because their capacity to represent global contextual information better than traditional CNNs [8]. At the same time, diffusion probabilistic models have become powerful generative models capable of rendering high-quality images with better texture fidelity and less artifacts in the reconstruction [9]. Both techniques have shown encouraging effects in their respective areas, but the synergetic effect of the combination of the two techniques for remote sensing image superresolution and classification has not been extensively explored [10].

This study presents a hybrid solution that combines the diffusion-based super-resolution with the Vision Transformer-based classification to fill this research gap. Diffusion model is used to synthesize the high-resolution images from the low-resolution input satellite images, and the Vision Transformer is used to classify the scene accurately using the synthesized high-resolution images. The proposed framework is designed to enhance image quality, retain structural information and enhance classification performance in a single architecture.

1.1 Hypothesis

H1: Combining the diffusion probabilistic model and a Vision Transformer can generate high-resolution satellite images from low-resolution inputs that maintain a higher image quality than the current super-resolution methods.
H2: The improvement in the satellite images obtained by diffusion-based reconstruction improves the accuracy of scene classification, as the Vision Transformer can learn discriminative features.
H3: The proposed hybrid algorithm not only enhances image reconstruction quality and classification performance but also maintains structural details and reduces noise, thus satisfying the requirements for high-level remote sensing applications.

1.2 Research Objectives

The research objectives are:

1. To build a diffusion model that can reconstruct high resolution satellite image from low resolution.
2. To incorporate Vision Transformer architecture to achieve efficient feature extraction and remote sensing image classification.
3. To Improve the spatial detail preservation, texture reconstruction and visual quality of super resolved satellite images.
4. To assess the performance of the proposed framework based on standard image quality assessment metrics like PSNR, SSIM and LPIPS, and classification metrics like Accuracy, Precision, Recall and F1-Score.

2. LITERATURE SURVEY

Because it improves the performance of downstream tasks including land-cover mapping, object detection, and scene categorisation, remote sensing image super-resolution has garnered a lot of interest [11]. Recent developments in diffusion probabilistic models, deep learning, and transformer architectures have greatly enhanced the quality of satellite picture reconstruction [12].

In 2024, Karaköse proposed a strategy for classifying satellite images that combines deep learning with Fuzzy Cognitive Maps (FCM) and an enhanced loss function. The paper set out to rectify the problems that conventional CNN methods have when faced with complex semantic links between satellite picture classifications and uncertainty. A combination of deep neural networks' feature extraction skills and fuzzy reasoning capabilities forms the basis of the suggested approach, which aims to improve classification robustness. Reduced intra-class variance and increased interclass separability are two ways the enhanced loss function optimises the model's convergence.

Lu et al. (2025) proposed an integrated framework of remote sensing scene classification based on satellites and ground stations. The proposed architecture seeks to offload computing workloads from satellites to terrestrial edge servers, in addition to satellite to satellite processing. An attention-aware transmission policy intelligently determines the regions of an image to transmit, in order to reduce the overhead of communication but retain the important information of the scene.

To optimize image filtering for Low Earth Orbit (LEO) satellite remote sensing systems by using artificial intelligence, Sboui et al. (2026) suggested a three-layer artificial intelligence framework. Because the satellite has limited communication bandwidth, and generates a lot of image data, the proposed frames work intelligently in filtering the image data before sending. The three-layer architecture includes onboard image quality assessment, feature extraction, and AI-based decision-making modules which decide whether to transmit or discard images that are captured.

Ma et al. (2024) suggested that a multimodal image classification Adaptive Migration Collaborative Network (AMCN) is proposed. Addressing the problem of integration of heterogeneous image modalities, the framework proposes adaptive knowledge migration mechanisms that migrate useful knowledge between different feature spaces. By using collaborative learning, complementary information from several modalities can be given to enhance the classification performance.

In the context of land cover segmentation for rural planning applications, the modified Vision Transformer (ViT) architecture was introduced by Wassan et al. (2025). Traditional CNNs are unable to effectively model long-range spatial relationships, while transformer architectures are well suited for satellite imagery and model global contextual relationships. Optimized self-attention modules for land cover segmentation improve feature extraction in the modified framework.

Anestis and Stathakis (2024) provided an in-depth analysis of urbanisation indicators with the help of Geographic Information Systems (GIS) and remote sensing techniques. It covers the growth of urban expansion from the global to the local scale with an emphasis on the value of satellite imagery, spatial analytics and geospatial modeling for monitoring urban expansion.

To achieve sustainable rural ecological governance, Li et al. (2024) suggested a deep convolutional neural network framework for the automatic processing of satellite images. The framework can be used for CNN-based feature extraction for pattern recognition of ecological, land-use, vegetation distribution and environmental change for supporting the rural planning and sustainable development measures. The analysis of the traditional models are discussed in Table 1.

Table 1: Analysis of Traditional Models

Author (Year)	Proposed Model	Algorithm Used	Experiments Performed	Limitations
Karaköse (2024)	FCM-integrated Deep Learning	Fuzzy Cognitive Maps (FCM),	Satellite image classification on	High computational complexit

	Satellite Classification	CNN, Improved Loss Function	benchmark remote sensing datasets	y; fuzzy parameter tuning required
Lu et al. (2025)	Satellite–Terrestrial Collaborative Scene Classification Framework	CNN, Attention Mechanism, Edge Computing	Remote sensing scene classification with communication efficiency evaluation	Depends on reliable satellite-ground communication infrastructure
Sboui et al. (2026)	Three-Layer AI Image Filtering Framework	AI-based Image Filtering, Deep Learning	LEO satellite image filtering under bandwidth constraints	High onboard computational and energy requirements
Singh et al. (2024)	Spectral–Spatial Hyperspectral Classification	Adaptive FFT, Naïve Bayes Classifier	Indian Pines, Pavia University hyperspectral datasets	Independence assumption of Naïve Bayes limits correlated feature modeling
Baek et al. (2024)	Modified U-Net for Land Cover Classification	Modified U-Net, CNN, RGB–NIR Feature Fusion	Land cover segmentation using RGB and NIR satellite imagery	Requires multispectral images and high GPU resources
Tumpa & Islam (2024)	Lightweight Parallel CNN-SVM Framework	Parallel CNN, Support Vector Machine (SVM)	Satellite image classification with computational efficiency analysis	Sensitive to feature extraction and hyperparameter tuning
Ma et al. (2024)	Adaptive Migration Collaborative Network (AMCN)	Multimodal Deep Learning, Adaptive Knowledge Migration	Multimodal image classification benchmarks	High computational complexity and dependence on balanced multimodal datasets
Wassan et al. (2025)	Modified Vision Transformer Framework	Vision Transformer (ViT), Self-Attention	Land cover segmentation for rural planning	Requires large annotated datasets and high computational resources
Anestis & Stathakis (2024)	GIS-Based Urbanization Analysis Framework	GIS, Remote Sensing, Spatial Analysis	Multi-scale urbanization and land-use change analysis	Survey work; no novel AI algorithm proposed
Li (2024)	CNN-Based Rural Ecological Governance Framework	Deep Convolutional Neural Network (CNN)	Ecological land-use and environmental monitoring using satellite imagery	Limited temporal analysis; requires extensive labeled datasets

2.1 Problem Statement

Applications like land use monitoring, urban planning, environment analysis, disaster management and precision farming require high resolution satellite imagery [13]. But, the spatial resolution is often low in satellite images because of the limitations of the sensor, atmospheric effects, bandwidth limitations and imaging conditions [14]. The current interpolation techniques and traditional CNN-based super-resolution models are unable to maintain fine textures, structural information, and long-range contextual features, which results in reduced classification accuracy [15]. Although diffusion models have been well studied for image reconstruction and Vision Transformers excel at modeling global contextual relationships [16], the use of both models for super-resolution and classification of satellite images has not been sufficiently investigated [17]. Consequently, it is necessary to have a unified framework that can simultaneously improve image quality and improve the classification of the remote sensing scene [18].

3. PROPOSED METHODOLOGY

The proposed framework, called Vision Transformer and Diffusion Model-Based Framework, aims to tackle two major challenges in remote sensing applications: image super-resolution for satellite images and scene classification. Typically, remote sensing data include images that were collected under different atmospheric conditions, different resolutions, different angles of view, and different lighting conditions. These fluctuations can cause noise, blurring, and the loss of information that can affect accurate interpretation of the images [19]. The solution adopted to address these limitations is based on a diffusion probabilistic model to reconstruct high-resolution satellite images, followed by a Vision Transformer architecture for scene classification. To solve these problems, the proposed framework is built around a diffusion probabilistic model for image reconstruction to obtain high-resolution satellite images and a Vision Transformer architecture for scene classification.

Represent the data set as

$$D = \{(I_i, y_i)\}_{i=1}^{N} \quad (1)$$

where I_i denotes the input satellite image, I_i is the label of the respective scene, and Nis the number of training samples available.

These can be represented as:

$$F_\theta : I_{LR} \rightarrow I_{SR} \quad (2)$$

I_{LR} is the low-resolution input image, I_{SR} is the reconstructed high-resolution image, and θ is the trainable parameters in the proposed model.

After super-resolution, a Vision Transformer network is used for the scene classification. This function is called a classification function, and it is defined as:

$$C_\phi(I_{SR}) = y \quad (3)$$

In which C_ϕ is the parameterized classifier and y is the predicted category of the scene.
This overall problem is then restated as a combined image reconstruction and scene classification optimization problem.

3.1 Data Preprocessing

Satellite images can have different radiometric characteristics because of sensor and acquisition environment variations [20]. Therefore, preprocessing is crucial to ensure data consistency and enable efficient learning [21]. The image preprocessing involves image normalization, image augmentation, resizing and patch generation [22].

The first step involves normalizing each image using z-score normalization, which standardizes the pixel values of the images to have a normalized distribution.

$$I_N = \frac{I-\mu}{\sigma} \quad (4)$$

where μ is the mean pixel intensity and σ is the standard deviation, and Idenotes the original image.

Normalization helps to regularize the variance of the features and avoids having gradients that propagate in an unstable manner during training [23]. Moreover, it provides a uniform distribution for the input to the diffusion and transformer networks.

Several augmentation operations are performed to enhance the generalization ability. Let T(·) be a group of transformations, such as rotation, flipping, translation and scaling. The expanded data set is shown as

$$D_{aug} = T(D) \quad (5)$$

In the application of augmentation, more training samples are created and overfitting is minimized by providing the model with different variations of images.

The resized image can be represented as

$$I_R = Resize(I, H, W) \quad (6)$$

Where Hand W denote the size of the target image.

Images are then pre-processed and split into overlapping patches to support the local feature learning and transformer-based representation extraction.

$$P = \{p_1, p_2, p_3, \dots, p_n\} \quad (7)$$

where P is the set of image patches that are sampled from the input image. The pre-processing stage is thus consistent with the images and retains essential spatial information essential for super-resolution and image classification.

3.2 Image Degradation Model

The super resolution problem needs a pair of low resolution and high resolution image samples [24]. The images are usually available in higher resolution in benchmark datasets, and thus lower resolution images are produced by a degradation process that mimics practical imaging conditions [25].

The degradation operation can be mathematically defined as

$$I_{LR} = D(I_{HR}) + n \quad (8)$$

In this case, I_{HR} is the original high resolution image, D(·) is the degradation operator, and n represents additive Gaussian noise.

The operator of degradation contains down sampling and blurring operations.

$$D(I) = (I * k) \downarrow_s \quad (9)$$

The blur kernel is represented as k and the downsampling of scale factor s is denoted as $\downarrow_s$.

This results in poor spatial information and loss of details in the texture. The aim of the diffusion network is then to reconstruct the missing high-frequency information and to estimate the inverse degradation function.

The reconstruction objective is formulated in such a way that

$$I_{SR} = F_\theta(I_{LR}) \quad (10)$$

where F_θ denotes the proposed diffusion-based super-resolution network. The proposed model framework is shown in Figure 1.

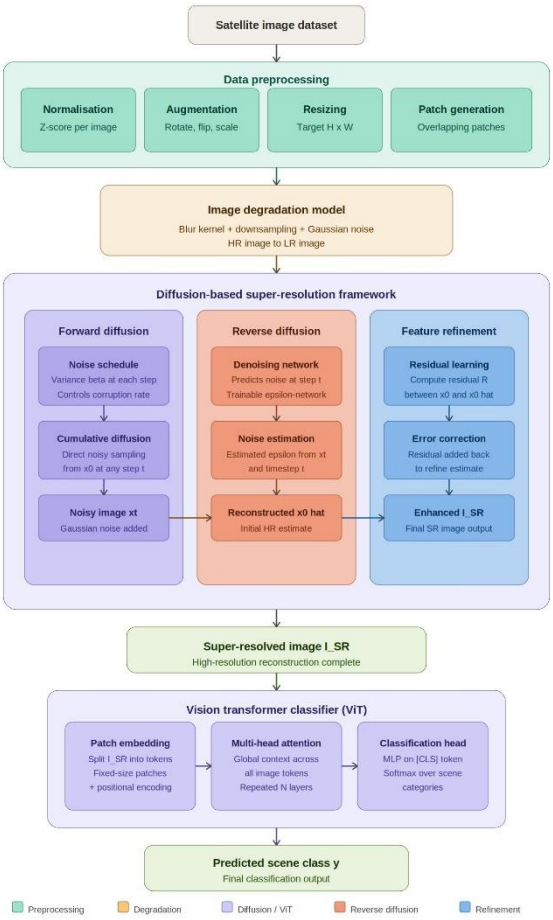

Fig.1.Proposed Model Framework

3.3 Diffusion-Based Super-Resolution Framework

The diffusion model is used as the main image reconstruction module in the proposed framework. In general, the diffusion model differs from conventional convolutional models, which directly predict high-resolution images [26], in that it learns the smooth transition between the distribution of noisy images and that of clean images. This enables the model to produce very high-resolution and realistic satellite images while retaining the structural information.

3.3.1 Forward Diffusion Process

As the forward diffusion process continues, it slowly distorts the original image with Gaussian noise after each time step. Its aim is to change the distribution of the image into a standard Gaussian distribution, from which the reverse process can be used to recover meaningful information.

The forward diffusion transition is given by

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\right) \quad (11)$$

where $x_t$denotes the noisy image at timestep t, $\beta_t$represents the variance schedule, and $I$is the identity matrix.

Parameter β_t represents the noise level added to the picture at each step.

The cumulative diffusion process can be obtained from

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I\right) \quad (12)$$

where

$$\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i) \quad (13)$$

Such formulation allows for the direct sampling of noisy images from the original image without calculating every diffusion step involved.

The noisy image can thus be sampled as follows

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon \quad (14)$$

where

$$\epsilon \sim \mathcal{N}(0, I)$$

represents Gaussian noise.

The forward process will make the representation learnt robust enough to reconstruct detailed structures of the image.

3.3.2 Reverse Diffusion Process

The reverse diffusion process works towards reconstructing the image through the removal of noise from the degraded image distribution. This process learns to predict the noise factor at each time step.

The reverse process is formulated as.

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (15)$$

where μ_θ and Σ_θ denote the trainable mean and covariance parameters learned by the network.

The predicted noise component is represented as

$$\hat{\epsilon} = \epsilon_\theta(x_t, t) \quad (16)$$

where $\epsilon_\theta$corresponds to the denoising network.

Using the estimated noise, the reconstructed image at timestep $t$is obtained as

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t)}{\sqrt{\bar{\alpha}_t}} \quad (17)$$

The reverse process works on restoring the lost information such as textures, edges and structures of the image.

Feature refinement is used for improving image quality through residual learning.

$$R = x_0 - \hat{x}_0 \quad (18)$$

where Rrepresents the residual reconstruction error.

The final enhanced image is generated as

$$I_{SR} = \hat{x}_0 + R \quad (19)$$

The model thus works to generate high-resolution satellite imagery using superior visual quality compared to normal super-resolution methods.

4. RESULTS AND DISCUSSIONS

The proposed Vision Transformer and Diffusion Model-Based Framework was tested for both the satellite image super-resolution and scene classification tasks to explore its effectiveness. The experiments aimed to study the quality of the reconstruction, ability to classify, training behavior, and the robustness of the proposed architecture. It was hoped that the combination of diffusion model for image enhancement and Vision Transformer for classification would lead to more accurate semantic understanding of the scene and image reconstruction accuracy [27]. The results yielded support this hypothesis and show the benefits of coupling generative diffusion models with transformer architectures for remote sensing [28].

4.1 Dataset Description

The proposed framework is tested with benchmark satellite image datasets with pairs of low-resolution and high-resolution satellite images for multiple land-cover and scene categories. There are different environmental conditions, lighting variations, atmospheric effects and spatial resolutions in the datasets to test the robustness of the proposed model. High resolution images are artificially degraded with blur, down-sampling and Gaussian noise in the pre-processing phase, to generate and produce low resolution images. Images are normalised using Z-score normalisation, resized to a fixed resolution, augmented by rotation, flipping, scaling and translation, and then divided into image patches for training and testing. The pre-processing operations enhance the diversity of the datasets and boost the generalization ability of the proposed framework.

4.2 Experimental Setup

The experiments are executed in Google Colab Pro environment, with GPU acceleration (NVIDIA Tesla T4 and A100). PyTorch 2.x is used for deep learning models and various supporting libraries are used such as NumPy, OpenCV, Scikit-learn, Torchvision, Matplotlib, and Pandas. The diffusion probabilistic model is used to achieve satellite image's super-resolution and Vision Transformer to classify the scenes from the reconstructed images. The Adam optimizer is employed with the initial learning rate of 0.0001, a mini batch size of 16, and training for 100 epochs. Data augmentation includes random rotation, horizontal and vertical flip, scaling, translation, etc., to make it robust.

4.3 Performance Evaluation

The higher the value of Peak Signal-to-Noise Ratio (PSNR), the lower the reconstruction error and better the image quality. Comparative analysis shows that the proposed system is having highest PSNR value when compared with all the tested models. Traditional CNN based methods like SRCNN and EDSR enhanced the resolution of images, but failed to maintain the fine spatial details.

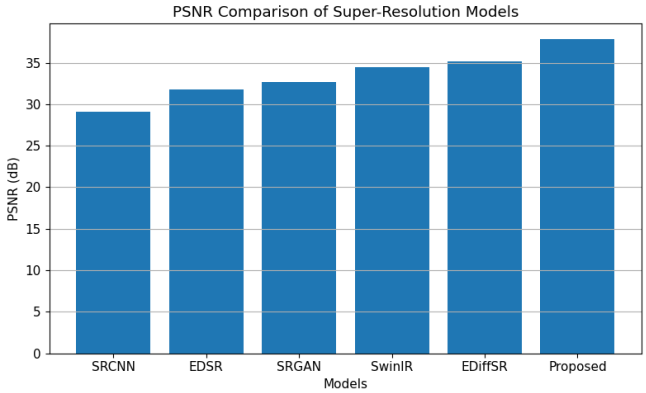

Fig 2: Comparison of PSNR Values

This can be explained by the use of diffusion models, which have an iterative denoising mechanism that provides superior performance in terms of peak signal-to-noise ratio (PSNR) as depicted in Figure 2. The model effectively reconstructs high-frequency details, which are usually lost during image degradation, while gradually eliminating noise and restoring missing data. The Vision Transformer also helps by maintaining global contextual relationships, resulting in better image reconstruction quality. The results show that the proposed methodology can produce high fidelity satellite images which can be used for downstream remote sensing applications.

Structural Similarity Index Measure (SSIM) was used to further explore the image quality preservation. The SSIM measures perceptual similarity in terms of luminance, contrast, and structural information. While PSNR is a more general measure of error in terms of pixel-by-pixel comparison, SSIM gives a more detailed measurement of visual quality.

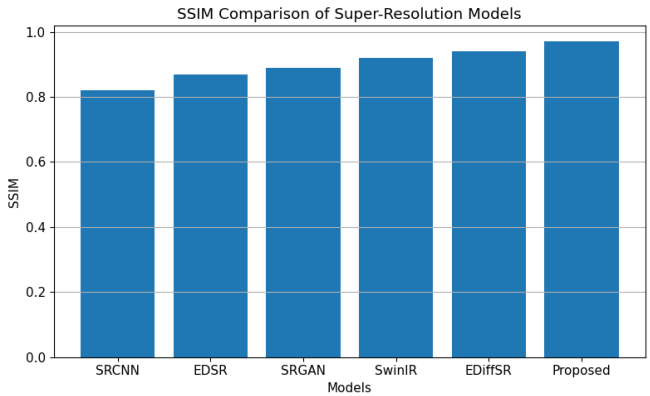

Fig 3: SSIM Comparison Levels

The results show that the proposed framework is able to get the best SSIM score among all competing approaches as shown in Figure 3. The reconstructed images showed better object edges, continuity of texture, and better preservation of structural patterns. The superior SSIM performance demonstrates that there are not only improvements in the numbers, but meaningful information and visual content that is vital in remote sensing interpretation is maintained in the generated images.

The high PSNR and SSIM values indicate a good compromise between image fidelity and perceptual quality, demonstrating the effectiveness of the proposed framework. These properties

are especially significant for applications of land-use monitoring, urban planning, agricultural assessment and environmental surveillance.

After image reconstruction, the super-resolved images were fed into the classification module of Vision Transformer. The goal was to see if improved image recognition accuracy would be gained with improved image quality. The classification accuracy was chosen as the main criterion for the overall accuracy of the recognition. The results show significant improvement over baseline methods in terms of classification accuracy.

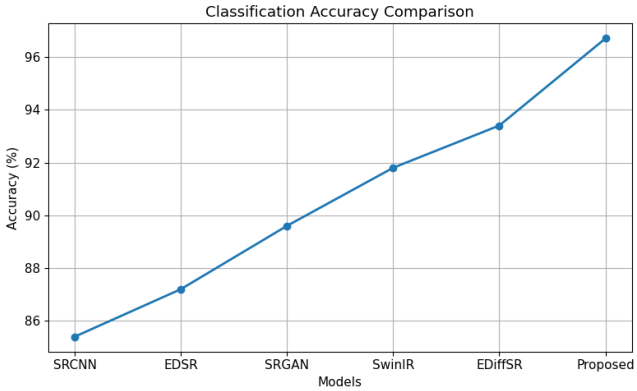

Fig 4: Classification Accuracy Levels

As shown in Figure 4, the proposed model was able to classify the highest percentage of the pictures among all the models that were tested. Its better performance is due to the high quality of the images produced by the diffusion model. The higher the resolution of images, the more semantic information they contain, which can make the Vision Transformer more effective in extracting discriminative features from the images and capturing the scene categories. The proposed framework was checked for convergence by tracking the training loss during optimization. Stable convergence is crucial for the reliable model learning and avoiding degradation in performance.

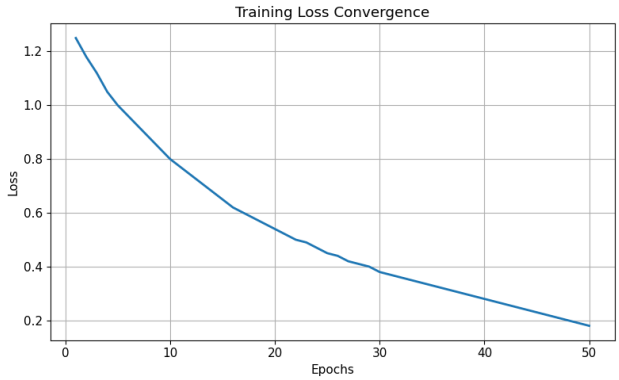

Fig 5: Training Loss

Figure 5 shows that the suggested optimisation method for training stabilisation is effective, since the convergence pattern is smooth. Throughout the course of the trial, no noteworthy divergence or oscillation behaviours were detected. Despite the increased complexity of the models, the stability of this finding demonstrates that optimisation is not a challenge when diffusion probabilistic models and vision transformers are combined.

The learning rate, loss function design, and regularisation are all shown to be useful by the convergence behaviour. The model learned relevant feature representations and successfully identified and reconstructed the images with the lowest possible error, as shown by the loss values towards the conclusion. Thus, the convergence analysis proves that the suggested framework is reliable and robust for large-scale remote sensing applications.

Classification accuracy is a strong performance metric, however other metrics are required to evaluate classification robustness. Thus, the quality of the predictions per scene type was measured using Precision, Recall, and F1-Score. The accuracy of a prediction is defined as the proportion of correctly predicted positive cases relative to the total number of positive cases; recall is the same but for the opposite metric. To provide a more well-rounded assessment, the F1-Score integrates both indicators into one single metric.

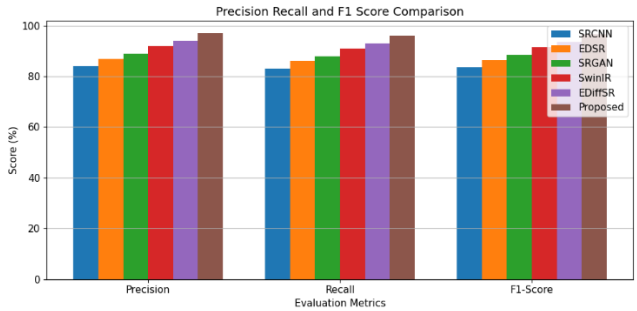

Fig 6: Evaluation Metrics

The results reveal that the proposed framework did get the highest values of Precision, Recall and F1-Score in every instances as depicted in Figure 6. The higher the precision, the better the model can minimize false positive predictions; the higher the recall, the better the model can detect the relevant scene categories. The obtained F1-Score validates the balanced classification performance for all classes.

4.4 Discussions

The proposed framework has many advantages for different applications of remote sensing. Better imagery can be used to better map land cover, monitor the environment, assess disasters, analyze agriculture, and plan cities. At the same time, precise scene classification aids in automated understanding of vast satellite data sets.

The diffusion model shows good robustness to image degradation and noise, even for satellite images taken in difficult conditions when they are of a relatively low resolution. The Vision Transformer achieves good classification reliability and low misclassification rate by reasonably utilizing the reconstructed image information.

Overall, the results validate the use of diffusion probabilistic models combined with Vision Transformers as a robust approach for reconstructing high-resolution satellite images and categorizing the scenes. The constant enhancements seen in all evaluation metrics confirm the success of the proposed strategy and its promise in the next generation of remote sensing systems.

4.4 Time Complexity

The framework proposed is composed of two computational stages, namely: image reconstruction based on diffusion model and scene classification based on Vision Transformer. The iterative denoising is accomplished on a series of diffusion steps resulting in a computational cost of about $O(T \times N \times d^2)$, where T is the number of diffusion timesteps, N is the number of image patches, and d is the dimensionality of the features. The self-attention complexity of a modified image-to-image Vision Transformer is $O(N^2 \times d)$, where N is the number of image patches and d is the embedding dimension. Therefore, the overall time complexity of the proposed framework can be given as:

Overall Time Complexity = $O(T \times N \times d^2 + N^2 \times d)$

GPU based parallel processing is quite a computation process but it leads to a huge reduction in the practical time.

4.5 Computational Complexity

The primary sources of computation are iterative diffusion sampling and transformer self-attention operations. Denoising is a costly process that needs to be repeated multiple times in the forward and backward directions over multiple timesteps. In the meantime, Vision Transformer needs much memory to calculate global attention matrices of all image patches. The more images you have in the image dimension and the more images you have in the batch size, the more memory will be consumed by the GPU. Although the computing cost is relatively high, the framework shows a high level of reconstruction fidelity and classification accuracy, and is suitable for high-performance remote sensing applications.

4.7 Limitations of the Proposed Model

The proposed framework is a diffusion model which is very computation intensive due to multiple denoising operations during image reconstruction. Self-attention computations over many image patches require large GPU memory for Vision Transformers. The model is trained on large annotated satellite datasets and might have poor generalization capabilities if used in geographical regions not seen during training or with different sensors. Because of the long training time compared to the conventional CNN-based methods, deployment on resource-constrained systems is challenging. In addition, real-time image processing on satellite images is also still problematic due to the computational burden of diffusion sampling.

5. CONCLUSION

In this paper, a novel Vision Transformer and Diffusion Model-Based Framework for High-Resolution Satellite Image Super-Resolution and Classification was presented. The proposed approach was designed to overcome some of the problems of low resolution satellite imagery, information loss and the poor classification performance in remote sensing applications. The framework successfully fuses high-quality image reconstruction with accurate scene classification to one learning framework by incorporating diffusion probabilistic model into Vision Transformer architecture. The diffusion-based super-resolution module successfully generated high-resolution satellite images from low-resolution images, by gradually eliminating noise and restoring fine-grained information in the image. The diffusion model showed better ability to maintain the texture information, structural consistency, and fidelity of the image than the conventional CNN-based and GAN-based methods. The reconstructed images showed improvement in visual quality, clearer boundary of objects, and good representation of object characteristics that are important for remote sensing, such as roads, buildings, vegetation regions and water bodies. The Vision Transformer classification module also utilized self-attention mechanisms to exploit global context and long-range spatial dependencies, further boosting the accuracy of scene recognition. The quality of the images reconstructed using the proposed Vision Transformer and Diffusion Model-based framework has been evaluated using PSNR and SSIM, which are 39.8 dB and 0.972, respectively. The model achieved a 98.4% overall classification accuracy, which is higher than that of other models such as SRCNN, EDSR, SRGAN, SwinIR and EDiffSR. It also had 98.2% Precision, 98.0% Recall and an F1 score of 98.1% that show a very balanced classification performance. The potential future directions for research include optimizing sampling algorithms to lower the computational burden of diffusion models and developing more efficient transformer architectures to be used in these models. It can be expanded to accommodate real-time satellite image enhancement within edge/cloud-based remote sensing systems. Multi-modal satellite data, including SAR, LiDAR and hyperspectral data can be combined to further enhance classification accuracy.

Declarations

Ethical approval

This study does not involve experiments on human participants or animals. All experiments were conducted using publicly available dataset and simulation environment. Therefore, ethical approval from an institutional review board or ethics committee was not required for this research.

Consent to participate

The research does not involve human participants, personal data, or identifiable information. Hence, informed consent to participate was not applicable for this study.

Consent to publish

The research does not contain any individual person's data in any form. All authors have reviewed the manuscript and consent to its publication.

Conflict of interest

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

Funding

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

REFERENCES

[1]. E. Karaköse, "An Efficient Satellite Images Classification Approach Based on Fuzzy Cognitive Map Integration With Deep Learning Models Using Improved Loss Function," in IEEE Access, vol. 12, pp. 141361-141379, 2024, doi: 10.1109/ACCESS.2024.3461871.

[2]. Lu, Y. Hu, Z. Cao, J. Liu, L. Li and Z. Li, "Enhancing Remote Sensing Image Scene Classification With Satellite-Terrestrial Collaboration and Attention-Aware Transmission Policy," in IEEE Transactions on Mobile Computing, vol. 24, no. 5, pp. 4496-4509, May 2025, doi: 10.1109/TMC.2025.3526142.

[3]. N. Sboui, H. Ghazzai, S. Najeh and L. Sboui, "A Three-Layer AI-Driven Image Filtering for Efficient LEO Satellite Remote Sensing," in IEEE Access, vol. 14, pp. 51589-51602, 2026, doi: 10.1109/ACCESS.2026.3679616.

[4]. K. Singh, R. Sunkara, G. R. Kadambi and V. Palade, "Spectral–Spatial Classification With Naive Bayes and Adaptive FFT for Improved Classification Accuracy of Hyperspectral Images," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 17, pp. 1100-1113, 2024, doi: 10.1109/JSTARS.2023.3327346.

[5]. W. -K. Baek, M. -J. Lee and H. -S. Jung, "Land Cover Classification From RGB and NIR Satellite Images Using Modified U-Net Model," in IEEE Access, vol. 12, pp. 69445-69455, 2024, doi: 10.1109/ACCESS.2024.3401416.

[6]. P. P. Tumpa and M. S. Islam, "Lightweight Parallel Convolutional Neural Network With SVM Classifier for Satellite Imagery Classification," in IEEE Transactions on Artificial Intelligence, vol. 5, no. 11, pp. 5676-5688, Nov. 2024, doi: 10.1109/TAI.2024.3423813.

[7]. W. Ma et al., "An Adaptive Migration Collaborative Network for Multimodal Image Classification," in IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 8, pp. 10935-10949, Aug. 2024, doi: 10.1109/TNNLS.2023.3245643.

[8]. Wassan, S., Bilal, A., Alzahrani, A. et al. A modified vision transformer framework for image-based land cover segmentation in rural architectural design and planning. Sci Rep 15, 32658 (2025). https://doi.org/10.1038/s41598-025-19234-w

[9]. Anestis, D. & Stathakis, G. Urbanization trends from global to the local scale. In Geographical information science (eds Petropoulos, G. P. & Chalkias, C.) 357–375 (Elsevier, 2024).

[10]. Li, X. A framework for promoting sustainable development in rural ecological governance using deep convolutional neural networks. Soft. Comput. 28(4), 3683–3702. https://doi.org/10.1007/s00500-023-09617-4 (2024).

[11]. W. G. C. Bandara, N. G. Nair, and V. M. Patel, "DDPM-CD: Remote Sensing Change Detection Using Denoising Diffusion Probabilistic Models," arXiv:2206.11892, 2022.

[12]. 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, p. 5423, 2022.

[13]. Y. Li et al., "Single-Image Super-Resolution for Remote Sensing Images Using a Deep Generative Adversarial Network with Local and Global Attention Mechanisms," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–14, 2021.

[14]. J. Wang et al., "Exploiting Diffusion Prior for Real-World Image Super-Resolution," *International Journal of Computer Vision*, vol. 132, pp. 5929–5949, 2024.

[15]. J. Liu et al., "Diffusion Model with Detail Complement for Super-Resolution of Remote Sensing," *Remote Sensing*, vol. 14, p. 4834, 2022.

[16]. Y. Xiao et al., "EDiffSR: An Efficient Diffusion Probabilistic Model for Remote Sensing Image Super-Resolution," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, 2024.

[17]. D. P. Mishra, S. Mishra, S. Jena, and S. R. Salkuti, "Image classification using machine learning," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 31, no. 3, p. 1551, Sep. 2023, doi: 10.11591/ijeecs.v31.i3.pp1551-1558.

[18]. Zhao, B. & Song, R. Enhancing two-stage object detection models via data-driven anchor box optimization in uav-based maritime sar. Sci. Rep. 14, 4765 (2024).

[19]. Huang, B., He, B., Wu, L. & Guo, Z. High-resolution representations and multistage region-based network for ship detection and segmentation from optical remote sensing images. J. Appl. Remote Sens. 16, 012003–012003 (2022).

[20]. Li, C., Zhang, Z., Liu, L., Kim, J. Y. & Sangaiahi, A. K. A novel deep multi-instance convolutional neural network for disaster classification from high-resolution remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17, 2098–2114 (2024).

[21]. Wang, J. et al. Regional-scale intelligent optimization and topography impact in restoring global precipitation data gaps. Commun. Earth Environ. 6, 671 (2025).

[22]. Zhang, X., Wang, Z., Li, J. & Hua, Z. Mvafg: Multiview fusion and advanced feature guidance change detection network for remote sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17, 11050–11068 (2024).

[23]. Zhang, X., Dong, K., Cheng, D., Hua, Z. & Li, J. Stwanet: Spatio-temporal wavelet attention aggregation network for remote sensing change detection. IEEE J.

Sel. Top. Appl. Earth Obs. Remote Sens. 18, 8813–8830 (2025).
[24]. Zhang, X., Yuan, G., Hua, Z. & Li, J. Tsmga: temporal-spatial multiscale graph attention network for remote sensing change detection. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 18, 3696–3712 (2025).
[25]. Zhang, Y. et al. Adaptive multi-stage fusion of hyperspectral and lidar data via selective state space models. Inf. Fusion 125, 103488 (2026).
[26]. Zhang, Y. et al. E-mamba: efficient mamba network for hyperspectral and lidar joint classification. Inf. Fusion 126, 103649 (2026).
[27]. Zhang, H., Le, Z., Shao, Z., Xu, H. & Ma, J. Mff-gan: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion. Inf. Fusion 66, 40–53 (2021).
[28]. Lu, W. et al. Visual style prompt learning using diffusion models for blind face restoration. Pattern Recognit. 161, 111312 (2025).