Enhancing Transparency of Neural Networks for Super-Resolution in Remote Sensing Using Local Attribution Maps

Martina Marani[1, 2], Xiaoyuan Wei[2, 3], Fabrizio Lamberti[1], Haopeng Zhang[2, 3, *]

[1] Dept. of Computer and Control Engineering, Politecnico di Torino, 10129 Torino, Italy - s309098@studenti.polito.it; fabrizio.lamberti@polito.it
[2] School of Astronautics, Beihang University, 102206 Beijing, China - (shiaoyuan, zhanghaopeng)@buaa.edu.cn
[3] Tianmushan Laboratory, Beihang University, 311115 Hangzhou, China

Keywords: Super Resolution (SR), Remote Sensing, Image Enhancement, Attribution Analysis, Explainable AI (XAI)

Abstract

As deep learning models advance, their use in Super Resolution (SR) tasks has become pivotal for enhancing remote sensing low-resolution (LR) satellite images. However, the decision making processes within these models remain opaque, especially in remote sensing applications where transparency is critical. This paper focuses on applying Explainable Artificial Intelligence (XAI) techniques, particularly Local Attribution Maps (LAMs), to analyze and interpret the internal behavior of both general purpose and remote sensing specific SR neural networks. General purpose models like Generative Adversarial Network for Super Resolution (SRGAN), Enhanced Deep Super-Resolution Network (EDSR), Efficient Super-Resolution Transformer (ESRT), and Hybrid-Attention Transformer (HAT), although highly effective in SR tasks, were originally designed for broader image enhancement challenges; in contrast, Hybrid-Scale Self-Similarity Exploitation Network (HSENet) and Multi-scale Enhanced Network (MEN) are tailored for the unique complexities of remote sensing, such as varied textures and intricate scene features. By leveraging LAMs, we highlight how different networks prioritize and process features such as edges, textures, and high frequency details to generate super resolved outputs. The comparative study between general purpose and remote sensing specific networks outlines each model's strengths and weaknesses in managing the data present in remote sensing imagery. This approach addresses the previously unexplored application of LAMs in remote sensing for SR, contributing to the research in this field and providing deeper insights into the interpretability and transparency of widely used SR models. Furthermore, by drawing parallels between feature attribution in SR and classification tasks, we suggest new pathways for integrating semantic information to refine model transparency and performance in remote sensing applications.

1. Introduction

Remote sensing is a crucial tool in various fields like agriculture, meteorology, urban planning, environmental monitoring, and military reconnaissance (Sishodia et al., 2020, Surendran et al., 2024, Tmušić et al., 2020, Smith and Doe, 2023). It enables the collection of extensive imagery from satellites and airborne platforms, providing essential data for resource management, environmental assessment, and strategic planning. However, the effectiveness of remote sensing imagery is often limited by its resolution (Johnson and Thompson, 2022). High-resolution images are vital for detailed analysis, yet their acquisition is restricted by current imaging technologies and the high costs of high-resolution sensors (Williams and Lee, 2021).

Super-Resolution (SR) techniques aim to mitigate these limitations by reconstructing high-resolution (HR) images from low-resolution (LR) inputs (Nasrollahi and Moeslund, 2014) using specialized neural networks. Recent advances in deep learning have significantly enhanced SR performance, resulting in the development of cutting-edge, highly effective neural networks like Generative Adversarial Network for Super Resolution (SRGAN) (Ledig et al., 2017), Enhanced Deep Super-Resolution Network (EDSR) (Lim et al., 2017), Efficient Super-Resolution Transformer (ESRT) (Lu et al., 2022), Hybrid-Attention Transformer (HAT) (Chen et al., 2023), Hybrid-Scale Self-Similarity Exploitation Network (HSENet) (Lei and Shi, 2021) and Multi-scale Enhanced Network (MEN) (Wang et al., 2023).

* Corresponding author

Despite the successes of SR techniques, their application to remote sensing imagery poses unique challenges, such as the complexity of scene features, texture diversity, varying illumination, and the presence of small, intricate objects (Wang et al., 2022a). Additionally, the scarcity of high-resolution training data further complicates the training of deep learning models. Our study evaluates the above state-of-the-art SR networks on remote sensing datasets, with a focus on using Explainable AI (XAI) techniques, specifically Local Attribution Maps (LAMs) (Gu and Dong, 2021). LAMs uses integrated gradients to analyze how different parts of the input image influence the output, emphasizing features such as edges and textures over mere pixel intensities. Such a detailed attribution analysis can provide deep insights into the internal mechanisms and decision-making processes of SR networks, elucidating how these models enhance image resolution. By applying LAMs, we aim to bridge the gap in understanding the interpretability of SR networks in remote sensing. This approach not only can advance comprehension of SR algorithms but also ensures their reliable and transparent application in critical remote sensing tasks, guiding the development of more effective and interpretable approaches tailored for remote sensing applications.

2. Related Work

In this section, we review the evolution in Single-Image Super Resolution (SISR) from early methods to advanced deep learning models, particularly in the context of remote sensing. Additionally, we examine the evolution of XAI, focusing on how

different techniques have emerged to provide transparency and interpretability in deep learning models, with a specific look at their application to remote sensing.

2.1 Remote Sensing Single-Image Super Resolution

In recent years, various methods have been proposed for SISR in remote sensing (Wang et al., 2022b). Initially, methods such as interpolation and reconstruction were used to enhance image resolution (Cherifi et al., 2020). However, these approaches had limitations, including loss of high-frequency information and high computational costs (Wang et al., 2022a). The advent of deep learning has revolutionized SISR, allowing for more accurate and efficient high-resolution image reconstruction (Yang et al., 2019). Early deep learning based approaches, such as Super Resolution Convolutional Neural Network (SRCNN) (Dong et al., 2015), introduced a lightweight three-layer CNN that demonstrated significant improvements over traditional methods. However, SRCNN's limited depth constrained its ability to capture complex image structures. To address this, Deeply Recursive Convolutional Network (DRCN) (Kim et al., 2016b), introduced deep recursive layers, allowing for a more refined image reconstruction but at the cost of increased memory consumption and training time. The Very Deep Super Resolution (VDSR) model (Kim et al., 2016a), further optimized performance by incorporating residual learning and gradient clipping, making it more effective for handling multi-scale images. An example of notable deep learning-based models is SRGAN (Ledig et al., 2017), which leveraged a generative adversarial network (GAN) framework to generate high-resolution images with realistic textures. Despite its ability to produce visually appealing results, SRGAN's limitations include the potential for artifact and limited generalization to diverse real-world images. To enhance both efficiency and performance EDSR (Lim et al., 2017) removed unnecessary batch normalization layers, reducing memory overhead while improving image quality. More recent models incorporate transformer based architectures to enhance feature extraction. ESRT (Lu et al., 2022) combines CNNs and transformers, leveraging their strength to improve performance while addressing high computational costs. However, transformer-based models often tend to excessively smooth out fine details. To mitigate this, HAT (Chen et al., 2023) overcomes limitations of traditional transformer models by combining channel attention and window-based self-attention mechanisms, effectively utilizing a larger input region while reducing redundancy, reaching a better image reconstruction. Specialized models have also been developed for remote sensing. HSENet (Lei and Shi, 2021), introduces spatial and channel-wise attention mechanisms to enhance feature extraction, particularly suited for the complex textures and varying resolutions of remote sensing images. MEN (Wang et al., 2023), employs a multi-scale ensemble network, effectively handling diverse image characteristics while balancing computational complexity and performance.

2.2 Explainable AI in Deep Learning

XAI has emerged as an important area of research, driven by the need to make complex models interpretable and transparent. Various methods have been introduced to generate human understandable explanations for AI decision making processes, broadly categorized into local or global, back-propagation-based or perturbation-based approach and intrinsic or post-hoc methods. Intrinsic methods integrate explainability directly into the model architecture, whereas post-hoc methods, applied to already trained models, offering flexibility across different architectures. Several methods have been developed to enhance the explainability, such as saliency maps, Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), and SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) interpret the model's decisions by highlighting the input features most influencing the output. These methods help to understand the inner workings of deep learning models, building trust, as well as facilitating model debugging and improvement.

In remote sensing, XAI methods address specific challenges posed by Earth Observation (EO) tasks such as classification, segmentation, and scene understanding. Techniques like explainable CNNs for land use classification, physically explainable CNNs for Synthetic Aperture Radar (SAR) image classification, and attention-based mechanisms for land cover mapping provide clear and interpretable insights into how remote sensing models process and analyze data. These enhancements improve their reliability and applicability in real-world scenarios. Differently than other areas of remote sensing, the field of SR remains underexplored in terms of explainability. The complexity of generating and explaining new image details, the continuous nature of the output, and the sophistication of model architectures make providing clear and interpretable explanations more challenging than in classification and segmentation tasks.

LAMs (Gu and Dong, 2021) is a local back-propagation post-hoc explainable algorithm that represent a promising approach in this area. LAMs leverage path integral gradients for attribution analysis, using a blurred image baseline to represent missing high-frequency components. This method focuses on specific patches, challenging areas, and local features such as edges and textures to understand individual pixel contributions. LAMs can provide a detailed analysis of how SR networks process and enhance image resolution, revealing the network's learned knowledge and capabilities, which is particularly important for identifying the network's performance in difficult-to-reconstruct areas.

3. Methodology

First, the models under study, i.e., SRGAN, EDSR, ESRT, HAT, HSENet, and MEN have been trained on a large remote sensing dataset. To conduct the training, the dataset adopted is divided between the training set and the validation set. Once trained, these models are tested on a separate remote sensing dataset to assess their ability to generate HR images from LR inputs in a different dataset context. To conduct both the training and the testing the original images from the datasets are taken and considered as the HR images and the LR version of them were obtained using bicubic interpolation downsampling method, computed by:

$$I_{\text{LR}} = (I_{\text{HR}} \otimes k) \downarrow_s + \sigma \tag{1}$$

where $(I_{\text{HR}} \otimes k)$ is the convolution operation between the HR image I_{HR} and the degenerate blur kernel k (e.g., Gaussian blur kernel), $\downarrow_s$ is the down sampling operation with scale factor s (with values such as 2, 3 and 4) and σ is an additive noise term. Subsequently, local attribution analysis has been conducted to gain insights into how these networks utilize input information to enhance resolution. Specifically, we used the LAMs method which adopt a progressive path function defined as:

$$\gamma_{pb}(k/L) = \omega(\sigma(1 - k/L)) \otimes I \tag{2}$$

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-575-2025 | © Author(s) 2025. CC BY 4.0 License.

The Gaussian kernel ω parameters include σ, representing the standard deviation of the Gaussian blur applied to the input image. Higher σ values lead to more blurred baselines, resulting in broader but less detailed attributions. Another key component is the use of the path function for integrated gradients, which involves the definition of the number of steps L that control the smoothness of the gradient approximation. Higher L values provide more accurate attributions but increase computation time. Additionally, the $Fold$ parameter determines the number of subdivisions in the input space. A higher $Fold$ results in finer granularity in attribution analysis but increases computational complexity.

The choice of *window size* for LAMs is also critical. A larger *window size* can capture more contextual information from the image, leading to broader attribution but potentially diluting specific details. Conversely, a smaller *window size* focuses on finer details but may miss broader context, affecting the overall interpretability. The *position of the window* is also important, placing it over regions with relevant elements of the image (e.g., edges, textures) ensures that the most informative features are highlighted in the attribution analysis. Incorrect placement can lead to misleading attributions, thereby affecting the reliability of the interpretation.

4. Experimentns and Results

In this section, we present the experimental setup and results of our study. We start by describing the hardware and datasets used for training and evaluation. Then, we present the implementation specifics, including training parameters and attribution analysis settings. Afterwards, we analyze the performance of the considered super-resolution models across different image classes and scale factors. Finally, we provide insights into their effectiveness based on Peak Signal-to-Noise Ratio (PSNR) and Structure Similarity Index Measure (SSIM), and discuss the results of attribution analysis using LAMs.

4.1 Experimental Data and Settings

4.1.1 Hardware Specifications We employed a 13th Gen Intel(R) Core(TM) i9-13900k with two NVIDIA GeForce GTX 4090 GPUs to train the models and a 12th Gen Intel(R) Core(TM) i7-12700H with an NVIDIA GeForce RTX 3070 GPU to test and perform attribution analysis.

4.1.2 Datasets and Metrics In this work, we use the Aerial Image Dataset (AID) (Xia et al., 2017) for training, that contains 10,000 aerial images and each image has a resolution of 600×600 pixels and the UCMerced dataset for testing containing 2,100 images and each image has a resolution of 256×256 pixels. PSNR and SSIM are widely-used metrics for SR evaluation, which reflect the pixel-level difference between the SR result and the groundtruth. PSNR is defined as follows:

$$10 \cdot \log_{10}\left(\frac{\text{MAX}_I^2}{\text{MSE}}\right) \tag{3}$$

where MAX_I is the maximum possible pixel value of the image (for 8-bit RGB images, $MAX_I = 255$). MSE is the Mean Squared Error (MSE) between the original and the reconstructed image and it is defined as follows:

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}\left(I(i,j) - K(i,j)\right)^2 \tag{4}$$

where m and n are the dimensions of the image, $I(i,j)$ is the pixel values of the original picture at position (i,j) and $K(i,j)$ is the pixel values of the reconstructed image at position (i,j). The higher the PSNR values, the better the reconstruction quality. Whereas, SSIM metric is computed as:

$$\frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{5}$$

where μ_x and μ_y are the local means, σ_x^2 and σ_y^2 are the variances, $\sigma_x y$ is the covariance of x and y, C_1 and C_2 are two constants. Higher SSIM value indicates greater structural coherence and thus better SR capability. To evaluate the attribution analysis result the Diffusion Index (Gu and Dong, 2021), which measure how many input pixels contribute to the final SR output and it is defined as:

$$DI = (1 - G)100 \tag{6}$$

where and G is calculated by:

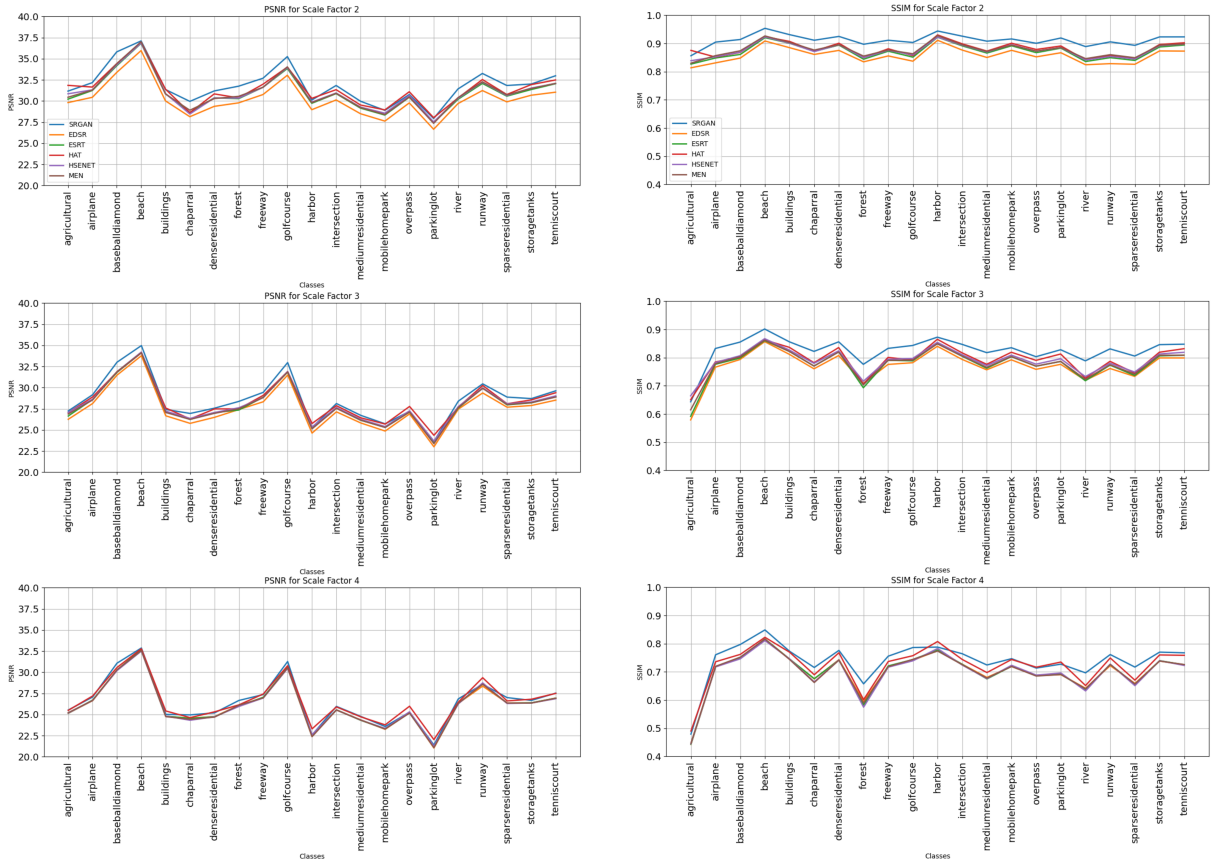
$$G = \frac{\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}|g_i - g_j|}{2n^2\bar{g}} \tag{7}$$

where g_i means the absolute value of the ith dimension of the attribution map and $\bar{g}$ is the average value of g_i. Higher DI values indicate that the network utilizes information from a broader range of pixels.

4.1.3 Implementation Details During training, all methods utilize Adam optimizer with a learning rate of 10^{-4} except for ESRT with a value of $2 * 10^{-4}$ and HAT adopting a value of 10^{-5}. Patch sizes of 96, 144, and 192 were considered to achieve a good trade-off between performance and computational cost for the adopted scale factors of 2, 3, and 4, respectively. Additionally, in this work attribution analysis was performed using two different window sizes of 32 and 64, extending the experiments did in the official paper where a window size of 16 was adopted. These windows were positioned in the center of the images, where relevant features were most prominent. For the integrated gradient with gaussian blur, the following parameters were employed $L = 100$, $Sigma = 2$ and $Fold = 9$.

4.2 Networks Investigation and Analysis

4.2.1 SR models results In our study, we utilized HR images from the UCMerced dataset and generated LR counterparts through bicubic downsampling with scale factors of 2, 3, and 4. These LR images, along with their corresponding HR images, were used to produce the corresponding SR images. Figure 1 illustrates the performance comparison of the considered models across the 21 image classes of the UCMerced dataset and various scale factors. In particular, Figure 1a shows the PSNR values, whereas Figure 1b shows the SSIM values across the image classes of UCMerced dataset. Based on these results, classes such as “beach”, “baseballdiamond” and “golfcourse” show higher PSNR and SSIM values across all networks, possibly because these classes contain more distinct and repetitive patterns that are easier for the networks to learn and reconstruct. On the other hand, classes like “parkinglot”, “harbour”, and “mobilehomepark” have more complex and less structured features, making them more challenging for the networks to super-resolve effectively. Additionally, we analyzed the average PSNR and SSIM values on the whole UCMerced dataset,



(a) PSNR across image classes

(b) SSIM across image classes

Figure 1. Performance comparison of the considered models across the image classes of the UCMerced dataset and different scale factors.

Model	Method	Scale	Param	PSNR↑	SSIM↑
SRGAN	Generative Adversarial Network	2x	5,801,931	31.88	0.91
		3x	5,986,571	28.56	0.83
		4x	5,949,644	26.53	0.74
EDSR	Enhanced Deep Super-Resolution Network	2x	40,729,603	30.23	0.86
		3x	43,680,003	27.57	0.77
		4x	43,089,923	26.11	0.70
ESRT	Efficient Super-Resolution Transformer	2x	677,783	30.96	0.87
		3x	770,263	27.96	0.78
		4x	751,767	26.15	0.70
HAT	High-Attention Transformer	2x	20,624,795	31.40	0.88
		3x	20,809,435	28.27	0.79
		4x	20,772,507	26.58	0.72
HSENet	High-Resolution Swin Efficient Network	2x	5,285,659	31.07	0.88
		3x	5,470,299	28.07	0.79
		4x	5,433,371	26.13	0.70
MEN	Multi-scale Enhanced Network	2x	2,592,899	31.07	0.88
		3x	2,777,539	27.96	0.78
		4x	2,740,611	26.12	0.70

Table 1. Comparison of models with PSNR and SSIM metrics for different scale factors.

as shown in Table 1. The table reflects the expected trend where SR images obtained with a scale factor of 2 exhibit higher PSNR and SSIM values compared to those with scale factors of 3 and 4, which generally have lower values. Among the models, SRGAN demonstrates slightly superior results for both PSNR and SSIM, attributed to its enhanced ability to capture and reconstruct fine details in SR images. HAT, HSENet, and MEN also offer competitive performance, with HAT achieving the highest PSNR value for scale factor of 4 among all models. Although EDSR and ESRT perform well, they are generally outperformed by the other models. Additionally, the number of parameters plays a significant role in its performance, capacity, and computational requirements. On the one hand, models with more parameters, such as EDSR and HAT, show an increased memory usage and computational demands and require more time and resources during training and inference. On the other hand, models with fewer parameters, such as ESRT and MEN, are generally lighter and faster, although these models might struggle to capture intricate details possibly affecting the SR quality. Instead, with its moderate parameter count, SRGAN and HSENET is designed to balance quality and efficiency.

4.2.2 Attribution Analysis As it can be seen in Figure 2, the LAMs analysis reveals significant differences in how various neural network models focus on image features during super-resolution tasks at different scales. Figure 2a, 2b and 2c shows examples of LAMs analysis using window size of 32 for scale factor of 2,3 and 4 respectively, whereas in Figure 2d, 2e and 2f a window size of 64 is used. In each figure, on the left it is shown the original images with the window size considered, while the subsequent images display LAMs, highlighting which parts of the images are used by each model to achieve super resolution. To better understand these examples, Table 2 suggests that models like HAT, HSENet show concentrated and well-defined focus areas, indicating their superior ability to identify and enhance critical image features. More in details, HAT's

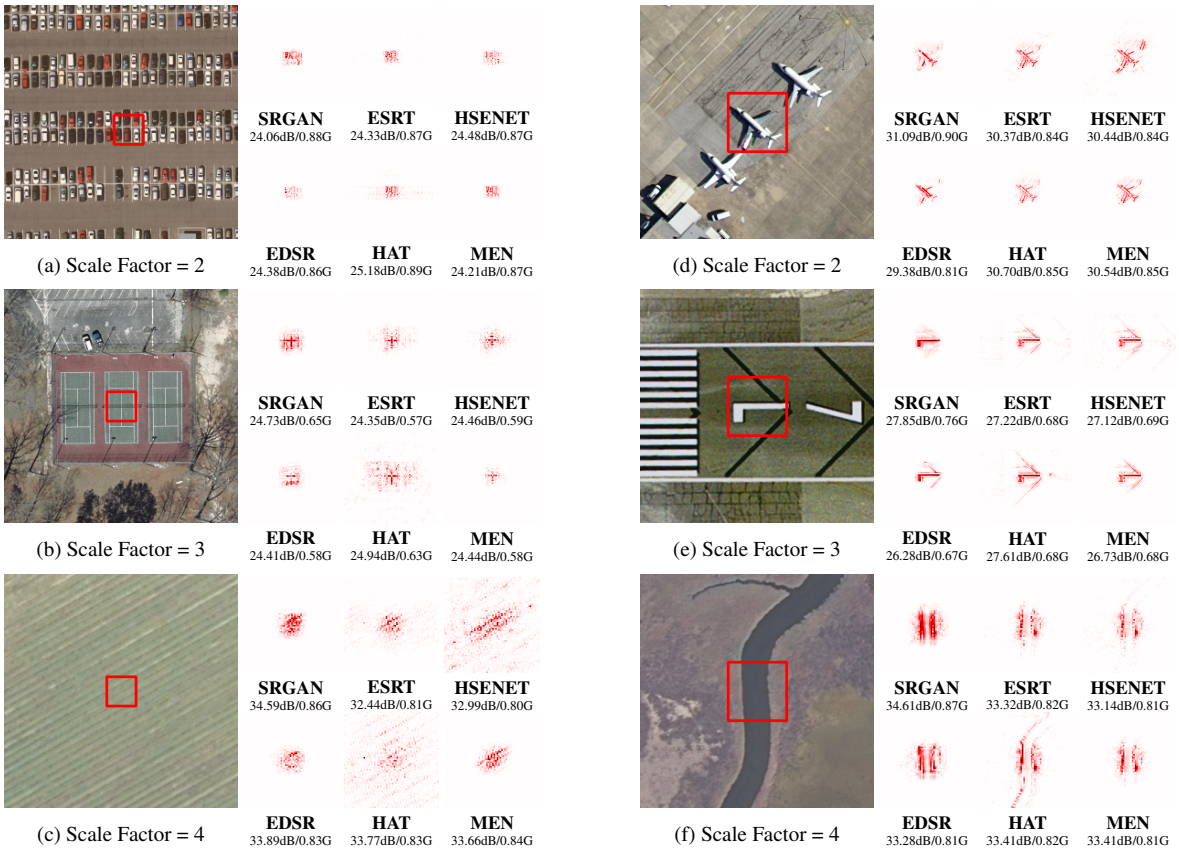


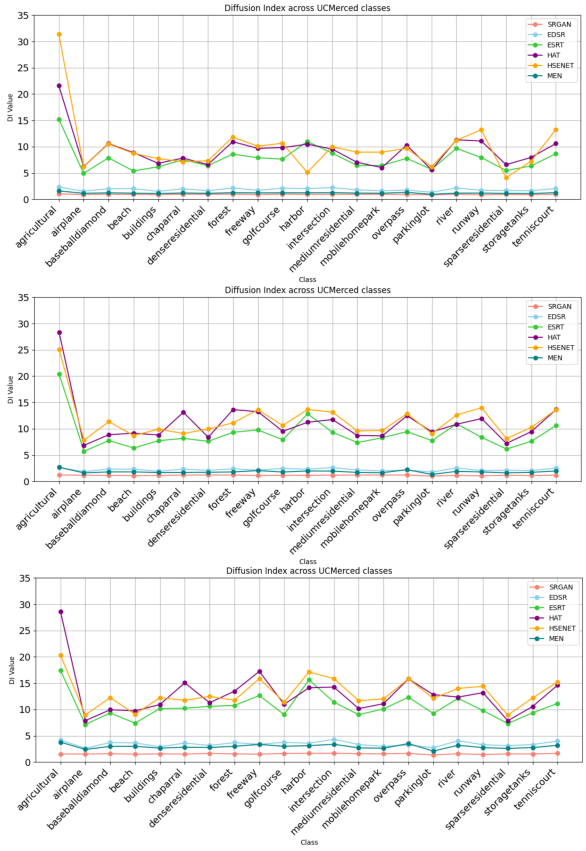
Figure 2. Examples of LAMs analysis for the considered models on UCMerced images from different classes. On the left, window size = 32 and different scale factors (a) scale factor = 2, (b) scale factor = 3, (c) scale factor = 4. On the right, window size = 64 and different scale factors (d) scale factor = 2, (e) scale factor = 3, (f) scale factor = 4.

scattered red points and widespread focus show that its attention mechanism prioritizes different parts of the image, providing a more global view; this can be seen from the higher DI values. HSENet reveals distinctive patterns, frequently emphasizing edges and specific structures within the images. These patterns become more defined and extensive with a larger window size, indicating that HSENet effectively utilizes the larger window to refine its focus on significant structures. Furthermore, although ESRT also uses attention mechanisms similar to HAT, its attribution maps display denser red points and a focus on slightly finer features. This suggests that ESRT utilizes attention mechanisms both to capture long-range contextual information and enhance details resolution. In comparison, MEN demonstrates a moderate density of red points in its attribution maps and shows a balanced focus on features across different resolutions. MEN's approach involves a multi-scale strategy that aggregates information from various scales to enhance image resolution. Thanks to this strategy, MEN maintains robust performance and consistency across a range of scales and resolution variations. In contrast, SRGAN and EDSR show dense scattered red points, indicating their focus on various small, localized features for reconstruction and their attention to fine local details. SRGAN, focuses on generating realistic textures and high-frequency details, emphasizing perceptual quality and realism through adversarial training. EDSR maintains a concentration of red points that remain relatively focused but spread slightly with a larger window size. This suggests that EDSR continues to prioritize fine details through its deep residual learning framework, but it also balances its focus between local features and broader context when provided with more information.

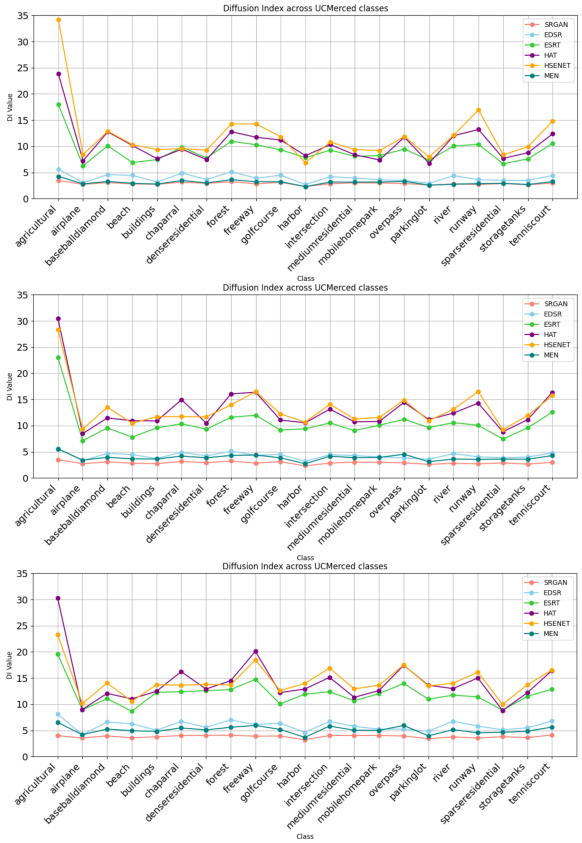
Models	WS32			WS64		
	SF=2	SF=3	SF=4	SF=2	SF=3	SF=4
SRGAN	0.9389	1.1627	1.5743	2.2985	2.8985	3.8363
EDSR	1.8648	2.2384	3.4693	3.9658	4.2671	5.8866
ESRT	7.7127	9.0338	10.5880	9.1631	10.4540	11.9729
HAT	9.3201	11.2121	12.9570	10.5320	13.0948	14.2539
HSENET	9.9865	11.6216	13.1241	12.0296	13.3035	14.4163
MEN	1.2053	1.8424	2.9482	3.0919	3.8955	5.1134

Table 2. Comparison of the average DI value on the overall UCMerced dataset with window sizes of 32 and 64 with different scale factors

Image context analysis: Images not containing defined objects (like cars, boats, houses or airplanes) exhibit clearer and more consistent LAM results across all models, leading to higher PSNR values, as shown in Figure 2. Images from classes like “parkinglots” or “harbors” with finer details and complex features are characterized by a lower PSNR than images like “agricultural fields” or “rivers”. This clarity can be attributed to the homogeneous nature of these images, where textures and patterns are more uniform, making it easier for the models to identify and enhance relevant features consistently. Conversely, images with defined objects present a greater challenge due to the complexity and variety of features that need to be accurately identified and enhanced. This complexity often results in more scattered focus areas, especially in models without attention mechanisms or a hierarchical structure like SRGAN, EDSR or MEN.



(a) DI values with window size of 32



(b) DI values with window size of 64

Figure 3. DI values comparison of the considered models across the image classes of the UCMerced dataset with different scale factors and window size of 32 and 64

Scale factor analysis: Using a lower scale factor can be more useful for complex images with more details. At lower scales, the models have less difficulties in preserving and enhancing intricate features, leading to higher-quality results. As the scale factor increases, the challenge of maintaining detailed feature integrity becomes more pronounced. Therefore, higher scale factors are better suited for images with fewer details, where the focus can be on enhancing broader, less complex features without losing critical information.

Window size analysis: The window size plays a crucial role in the quality of the super-resolution results. Smaller window sizes provide a more localized view, enabling the models to focus on finer details within a limited area. This can be beneficial for detailed images but may lead to less context awareness for broader patterns. Larger window sizes, instead, capture larger patterns and context but potentially dilute focus on finer details. The choice of window size thus depends on the specific characteristics of the images and the scale factor used, with larger windows being more advantageous for less detailed images and smaller windows better suited for highly detailed images. To gain a more nuanced understanding, we examined the relationship between the DI and window size across models for the entire testing dataset and for all scale factors, comparing window sizes of 32 and 64. The results are illustrated in Figure 3. This figure highlights that larger window size tend to conduct higher DI values. Notably, larger window sizes capture a broader pixel context, leading to an approximate increase of 2 points in DI across all cases. With a window size of 64, models such as SRGAN, EDSR, and MEN achieve higher DI values and generate more detailed maps, facilitating a more precise analysis. This observation opens the possibility for a discussion on the trade-offs between window size, model focus and computational cost. Larger windows generally require more memory and processing power, which could be a limiting factor. Conversely, smaller windows enable the model to focus on a smaller area, but may come with the benefit of lower computational cost.

These results are further validated by the analysis of the heat maps, which highlight the areas of interest for the various neural networks on example images (see Figure 4). Figures 4a, 4b and 4c consider a window size of 32 for scale factor of 2,3 and 4, respectively, whereas Figures 4d, 4e and 4f use a window size of 64. In each figure, the leftmost image shows the original image with the selected window size, while the subsequent images display the heat maps, illustrating the regions that different models prioritize for super-resolution. For instance, SRGAN and EDSR, which focus on localized, fine details, are better suited for images containing intricate textures or small, well-defined features, like individual cars in parking lots or specific textures in agricultural fields. Their attention to fine details is critical when HR enhancement of such small structures is required. In contrast, ESRT and HAT, which focus on broader, more generalized patterns, perform better in images with simpler, repetitive features, such as agricultural fields with striped or grid-like textures. HSENet’s heat maps, which emphasize larger regions and recurrent parts of the image, indicate that it is well-suited for scenes with broad, structural patterns. For surface features that require recognizing large, linear structures

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-X-G-2025-575-2025 | © Author(s) 2025. CC BY 4.0 License.

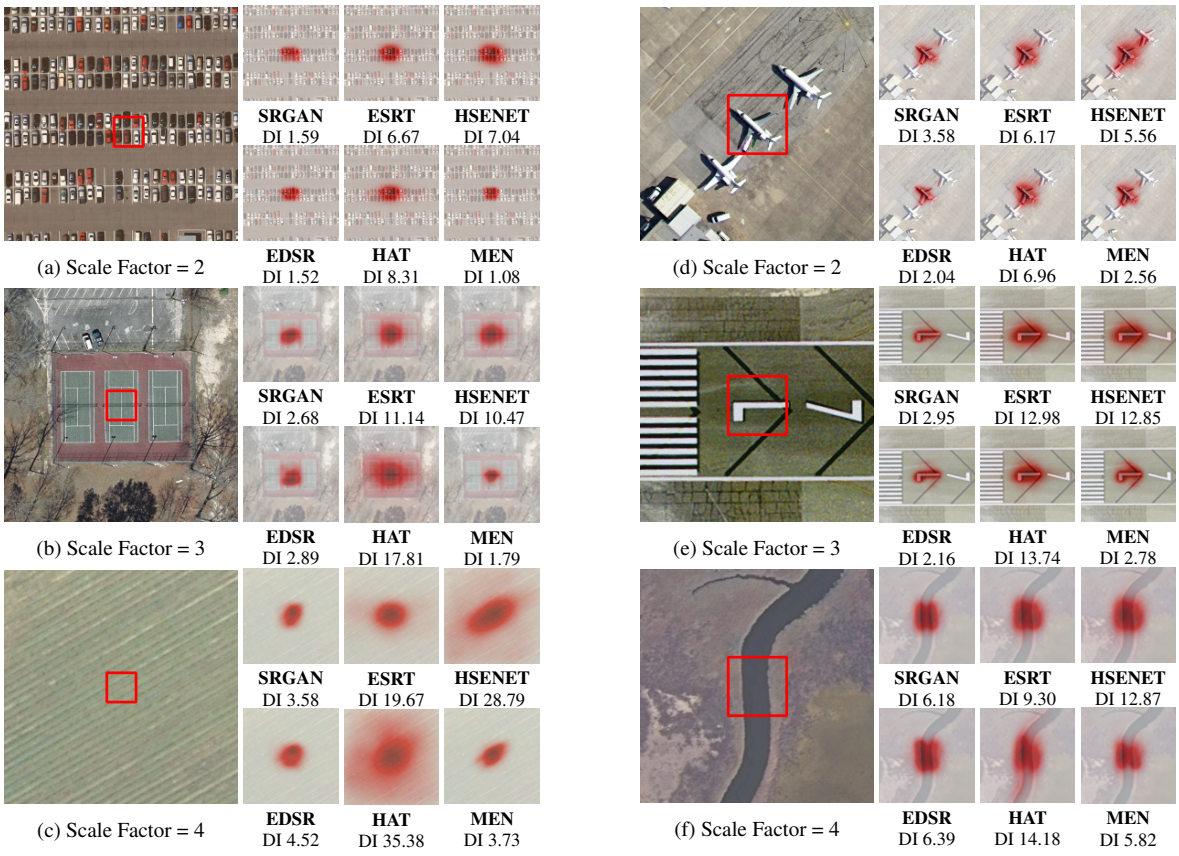


Figure 4. Examples of the area of contribution for the considered models on UCMerced images from different classes. On the left, window size = 32 and different scale factors (a) scale factor = 2, (b) scale factor = 3, (c) scale factor = 4. On the right, window size = 64 and different scale factors (d) scale factor = 2, (e) scale factor = 3, (f) scale factor = 4.

(e.g., roads, runways, or highways), ESRT, HAT, and HSENet are particularly effective, as they leverage the larger context to enhance the overall profile of these elements. HSENet and HAT's higher DI values across all scale factors confirm their strength in capturing the structural integrity of large surface areas, such as agricultural fields or urban landscapes with linear features. Meanwhile, MEN exhibit a balanced approach, highlighting structural and textural features effectively across various scales, making it a good choice for images where consistency across different scales is important, such as in the case of diverse environmental conditions with varying levels of detail. For more complex, detailed scenes, such as “airplane,” “beach,” or “sparseresidential” images, models like SRGAN and EDSR may struggle to capture the full range of detail. In these cases, it is essential to choose models that prioritize specific features over broad patterns, as the complex textures and diverse elements in these classes make it challenging for models to process the image globally. For simple, repetitive surface features (like agricultural patterns or urban roads), models like HAT, ESRT, and HSENet offer strong performance due to their ability to leverage broader context and structural information. For more detailed or complex surface features, SRGAN and EDSR should be preferred for their ability to enhance fine textures.

5. Conclusion

In this study, we evaluated the performance of state-of-the-art neural networks on remote sensing datasets, addressing the critical challenge of enhancing image resolution. By integrating XAI techniques, particularly LAMs, we provided a novel perspective on the internal mechanisms of these networks, offering detailed insights into how they process and enhance low-resolution images. The usage of LAMs also demonstrated the potential to bridge the gap in understanding the interpretability of SR networks in remote sensing, an area that has remained under explored. For future work, incorporating semantic information into SR models, as demonstrated in (Mc Cutchan et al., 2021), could improve the interpretability of SR models in remote sensing. Integrating high-level semantic metadata, such as land cover types or object categories, with LAMs could enhance transparency and align with key features in SR models. Additionally, examining feature prioritization, as seen in classification tasks, could offer further insights into how SR models process edges, textures, and high-frequency details. Overall, this research contributes to the advancement of SR techniques in remote sensing by providing deeper insights into model behavior, guiding the development of more effective and transparent SR algorithms tailored for remote sensing applications.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62271017), and the Fundamental Research Funds for the Central Universities.

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear

classifier decisions by layer-wise relevance propagation. *PLOS One*, 10(7), e0130140.

Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C., 2023. Activating more pixels in image super-resolution transformer. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 22367–22377.

Cherifi, T., Hamami-Metiche, L., Kerrouchi, S., 2020. Comparative study between super-resolution based on polynomial interpolations and whittaker filtering interpolations. *2020 1st International Conference on Communications, Control Systems and Signal Processing*, IEEE, 235–241.

Dong, C., Loy, C. C., He, K., Tang, X., 2015. Image super-resolution using deep convolutional networks. *IEEE transactions on pattern analysis and machine intelligence*, 38(2), 295–307.

Gu, J., Dong, C., 2021. Interpreting super-resolution networks with local attribution maps. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9199–9208.

Johnson, E. R., Thompson, M. L., 2022. Limitations of remote sensing imagery due to resolution constraints. *International Journal of Remote Sensing*, 43(7), 1456–1478.

Kim, J., Kwon Lee, J., Mu Lee, K., 2016a. Accurate image super-resolution using very deep convolutional networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1646–1654.

Kim, J., Lee, J. K., Lee, K. M., 2016b. Deeply-recursive convolutional network for image super-resolution. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1637–1645.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4681–4690.

Lei, S., Shi, Z., 2021. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–10.

Lim, B., Son, S., Kim, H., Nah, S., Lee, K. M., 2017. Enhanced deep residual networks for single image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 136–144.

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., Zeng, T., 2022. Transformer for single image super-resolution. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 457–466.

Lundberg, S. M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Mc Cutchan, M., Comber, A. J., Giannopoulos, I., Canestrini, M., 2021. Semantic Boosting: Enhancing Deep Learning Based LULC Classification. *Remote Sensing*, 13(16). https://www.mdpi.com/2072-4292/13/16/3197.

Nasrollahi, K., Moeslund, T. B., 2014. Super-resolution: a comprehensive survey. *Machine Vision and Applications*, 25, 1423–1468.

Sishodia, R. P., Ray, R. L., Singh, S. K., 2020. Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sensing*, 12(19).

Smith, J. A., Doe, R. B., 2023. Remote sensing technologies in military reconnaissance: Advances and applications. *Journal of Defense Technology*, 15(2), 123–145.

Surendran, U., Nagakumar, K. C. V., Samuel, M. P., 2024. *Remote Sensing in Precision Agriculture*. Springer International Publishing, Cham, 201–223.

Tmušić, G., Manfreda, S., Aasen, H., James, M. R., Gonçalves, G., Ben-Dor, E., Brook, A., Polinova, M., Arranz, J. J., Mészáros, J., Zhuang, R., Johansen, K., Malbeteau, Y., de Lima, I. P., Davids, C., Herban, S., McCabe, M. F., 2020. Current Practices in UAS-based Environmental Monitoring. *Remote Sensing*, 12(6).

Wang, P., Bayram, B., Sertel, E., 2022a. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Science Reviews*, 232, 104110.

Wang, X., Yi, J., Guo, J., Song, Y., Lyu, J., Xu, J., Yan, W., Zhao, J., Cai, Q., Min, H., 2022b. A review of image super-resolution approaches based on deep learning and applications in remote sensing. *Remote Sensing*, 14, 5423.

Wang, Y., Shao, Z., Lu, T., Wu, C., Wang, J., 2023. Remote sensing image super-resolution via multiscale enhancement network. *IEEE Geoscience and Remote Sensing Letters*, 20, 1–5.

Williams, S. J., Lee, C. K., 2021. Challenges in acquiring high-resolution remote sensing images: Imaging technology and cost considerations. *Remote Sensing of Environment*, 258, 112383.

Xia, G.-S., Hu, J., Hu, F., Shi, B., Bai, X., Zhong, Y., Zhang, L., Lu, X., 2017. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7), 3965–3981.

Yang, W., Zhang, X., Tian, Y., Wang, W., Xue, J.-H., Liao, Q., 2019. Deep learning for single image super-resolution: A brief review. *IEEE Transactions on Multimedia*, 21(12), 3106–3121.