Multispectral Satellite Image Super-Resolution with Progressive Focused Transformers

Luigi Pio Castaldo, Marco Mario Ciamarra, Pietro Conte

Università degli Studi di Napoli Federico II
MSc students in Computer Engineering
castaldoluigipio@gmail.com, marco.ciamarra0@gmail.com, pietroconte20033@gmail.com

Abstract. This report presents our approach to the Enel Innothon 2025 Challenge #4: NextGen Satellite Super-Resolution. We developed a Transformer-based method for 4× super-resolution of multispectral Sentinel-2 imagery, focusing on the four highest-resolution bands (R, G, B, NIR at 10m ground sampling distance). Our solution adapts the Progressive Focused Transformer (PFT) architecture to handle multispectral data while maintaining spectral integrity and structural fidelity. Despite limited computational resources as a student team, we achieved promising results that demonstrate the feasibility of Transformer-based approaches for operational satellite image enhancement.

1 Problem Statement and Motivation

Image Super-Resolution aims to reconstruct high-resolution details from low-resolution inputs, enhancing both visual quality and extraction of meaningful information. This task is particularly relevant for remote sensing applications, where spatial resolution is often constrained by sensor physics, orbital parameters, and cost considerations. Enhanced spatial resolution enables more precise identification of objects, structures, and phenomena that would otherwise remain blurred or ambiguous in the original imagery.

Traditional SR methods relied on interpolation techniques or early learning-based approaches using sparse coding and dictionary learning. The advent of deep learning revolutionized the field, with Convolutional Neural Networks achieving substantial improvements. However, CNN-based methods are fundamentally limited by their local receptive fields, which constrains their ability to model long-range dependencies and global context factors for reconstructing complex spatial structures in satellite imagery.

Recent advances in Transformer architectures have opened new possibilities for SR tasks. Self-attention mechanisms enable modeling of global dependencies across the entire image, potentially capturing structural patterns and correlations that span large spatial extents. The Progressive Focused Transformer (PFT) [1] represents a significant advancement by introducing focused attention mechanisms that balance computational efficiency with the ability to capture both local details and global context through progressive refinement.

For satellite imagery, particularly multispectral data from Sentinel-2, super-resolution presents additional challenges: (1) maintaining spectral consistency across multiple bands, (2) preserving radiometric accuracy for quantitative analysis, (3) handling the unique noise characteristics of satellite sensors, and (4) ensuring computational tractability for large-scale operational deployment.

2 Dataset and Preprocessing

2.1 Dataset Composition

Our training and validation data consist of Sentinel-2 multispectral imagery focusing on the four bands with native 10m ground sampling distance: Band 2 (Blue), Band 3 (Green), Band 4 (Red), and Band 8 (Near-Infrared). This selection maximizes the available spatial information while covering the spectral range most relevant for land cover analysis and infrastructure detection.

The dataset is composed of:

- **Training set:** 5,000 image pairs (low-quality and ground-truth)
- **Validation set:** 100 image pairs

Each image pair consists of:

- **Ground truth (GT):** 256×256×4 crops from Sentinel-2 imagery at full 10m resolution
- **Low-quality (LQ) input:** 64×64×4 downscaled versions, simulating effective 40m resolution

This configuration implements a 4× super-resolution task, where the model learns to reconstruct spatial details lost during the downscaling process. The dataset alternates between natural and urban environments, providing diversity in texture patterns, spectral signatures, and structural complexity.

2.2 Data Preparation and Normalization

The preprocessing pipeline consists of the following steps:

Spatial cropping: Original Sentinel-2 tiles are cropped into 256×256 patches to create manageable training samples while preserving local spatial coherence. Random cropping ensures diversity in the training set.

Downscaling: Ground truth patches are downscaled by a factor of 4 using bicubic interpolation to generate the corresponding 64×64 low-quality inputs. This simulates the degradation process the model must learn to invert.

Normalization: All images are pre-normalized to the [0,1] range. During training, we apply mean subtraction to center the data distribution, computed separately for each spectral band from the training set. This normalization strategy improves convergence stability and gradient flow through the network.

Data augmentation: Augmentation is handled by the BasicSR framework and includes random horizontal and vertical flips, 90° rotations, and random crops. These transformations ensure rotational and reflection invariance while expanding the effective training set size.

3 Methodology

3.1 Architecture: Progressive Focused Transformer

Our approach is based on the Progressive Focused Transformer (PFT) [1], a state-of-the-art architecture specifically designed for single image super-resolution. The PFT addresses the computational challenges of applying Transformers to high-resolution image reconstruction through three key innovations:

1. Progressive refinement: The architecture operates through multiple stages, each progressively increasing spatial resolution while refining details. This hierarchical approach allows

the model to first establish coarse structures and then focus computational resources on fine-grained detail recovery.

2. Focused attention mechanism: Rather than computing full self-attention over all spatial positions (which scales quadratically), PFT employs focused attention that restricts computation to local windows while maintaining global receptive fields through cross-stage connections.

3. Multi-scale feature fusion: Features from different refinement stages are progressively fused, allowing the network to leverage both low-frequency structural information and high-frequency texture details in the reconstruction process.

3.2 Adaptation for Multispectral Data

The original PFT architecture was designed for 3-channel RGB images. We made the following modifications to handle 4-channel multispectral Sentinel-2 data:

Input/output adaptation: The first convolutional layer (`conv_first`) was modified to accept 4 input channels (R, G, B, NIR) instead of 3. Similarly, the final reconstruction layer was adapted to output 4 channels. This modification ensures the network processes all spectral bands simultaneously, enabling it to learn and preserve inter-band correlations.

Weight initialization: Model weights were initialized from scratch rather than using RGB-pretrained weights, as the NIR band has fundamentally different spectral characteristics that would not benefit from RGB-based initialization. Random initialization allows the network to learn appropriate feature extractors for all four bands without bias toward visible-spectrum patterns.

Training from scratch: Given the architectural modifications and the different input domain, we trained the complete network from random initialization. This approach avoids potential negative transfer from RGB-centric features while allowing the model to fully adapt to multispectral data characteristics.

3.3 Loss Function and Training Objective

We employ a pixel-wise L1 loss (Mean Absolute Error) as our primary training objective:

$$\mathcal{L} = \frac{1}{C \cdot H \cdot W} \sum_{c=1}^{C} \sum_{h=1}^{H} \sum_{w=1}^{W} |I_{c,h,w}^{SR} - I_{c,h,w}^{GT}| \tag{1}$$

where I^{SR} is the super-resolved output, I^{GT} is the ground truth, and $C = 4$, $H = 256$, $W = 256$ represent channels, height, and width respectively. The L1 loss is applied uniformly across all four spectral bands with equal weighting, reflecting our commitment to maintaining consistent spectral fidelity across visible and near-infrared ranges.

We chose L1 loss over L2 (MSE) for its robustness to outliers and its tendency to produce sharper reconstructions. While perceptual losses (based on VGG features) and adversarial training can improve visual quality, they are less appropriate for multispectral data where radiometric accuracy and spectral integrity are paramount for downstream quantitative analysis.

4 Training Configuration and Implementation

4.1 Optimization Strategy

The model was trained using the Adam optimizer with the following configuration:

– **Learning rate schedule:** MultiStepLR with milestones at iterations [150k, 225k, 270k, 290k], reducing the learning rate by a factor of $\gamma = 0.5$ at each milestone
– **Total iterations:** 300,000
– **Warmup:** 10,000 iterations with gradual learning rate increase from 0 to the base learning rate (2e-4), improving training stability in early stages
– **Batch size:** 32
– **Input patch size:** 64×64×4 (low-resolution)
– **Target patch size:** 256×256×4 (high-resolution)

The learning rate schedule was designed to provide aggressive initial learning followed by fine-tuning phases. The warmup period prevents early instability that can occur when training deep networks from scratch with large learning rates.

4.2 Regularization and Stability

Exponential Moving Average (EMA): We employ EMA of model weights with a decay factor of 0.999 to maintain a temporally smoothed version of the model. This technique improves generalization and provides more stable validation performance. The EMA model is used for evaluation and inference.

Data augmentation: As mentioned in Section 2.2, we use geometric augmentations (flips, rotations, crops) to improve robustness and reduce overfitting.

4.3 Implementation Framework

Our implementation is built on **BasicSR** [2], a comprehensive PyTorch-based framework for super-resolution research. BasicSR provides modular components for datasets, models, losses, and training pipelines, enabling rapid experimentation while maintaining reproducibility. The framework handles data loading, augmentation, distributed training, logging, and checkpointing automatically.

4.4 Computational Resources and Constraints

As a student team participating in this challenge, we face significant computational constraints. Training deep Transformer models on multispectral imagery is computationally intensive, requiring substantial GPU memory and processing time. Our limited access to high-performance computing resources has constrained both the model scale and the extent of hyperparameter exploration we could conduct.

Despite these limitations, we have focused on efficient implementation practices and careful experimental design to maximize the utility of available resources. We believe that with access to professional-grade infrastructure (multi-GPU clusters, longer training durations, larger datasets), the performance of our approach could be substantially improved.

5 Experimental Results

5.1 Quantitative Evaluation

We evaluate super-resolution quality using two standard metrics. We consider the Peak Signal-to-Noise Ratio (PSNR), which measures the pixel-wise reconstruction accuracy (higher values indicate better fidelity to the ground truth). Then we compute the Structural Similarity Index

(SSIM), that assesses perceptual similarity by comparing local patterns of luminance, contrast, and structure. SSIM values range from 0 to 1, with 1 indicating perfect similarity. Both metrics are computed on the full 4-channel outputs, providing an aggregate measure of reconstruction quality across all spectral bands. For detailed analysis, we also report per-channel metrics on selected validation examples to assess spectral consistency.

Validation Set Performance On our validation set of 100 image pairs, our model achieves the results shown in Figure 1 and in the Table below. These results are promising considering the challenging real-world conditions, and the relatively small training dataset of 5k images.

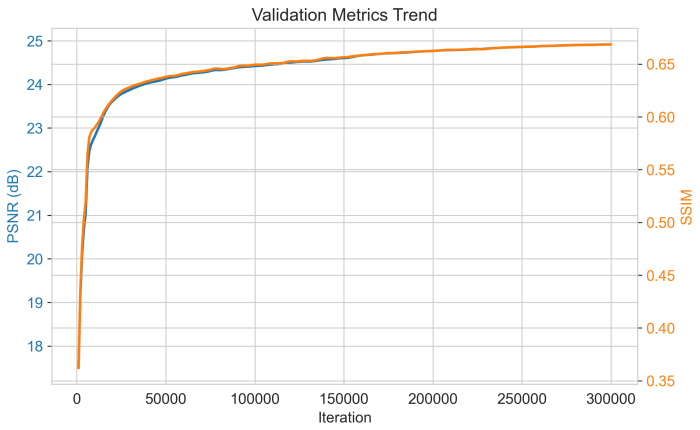

Fig. 1. Validation metrics: PSNR in blue, SSIM in yellow

Metric	**Value**
PSNR (dB)	24.92
SSIM	0.669
Final Training Loss (L1)	8.45e-02

Training. Fig. 2 shows the training loss evolution over 300k iterations. The curve exhibits stable convergence without significant overfitting, suggesting that our regularization strategy (EMA, data augmentation) is effective.

5.2 Qualitative Analysis

The visual analysis (shown in Fig. 3) reveals several key characteristics:

Edge preservation: The PFT architecture effectively reconstructs sharp boundaries between different land cover types, such as field edges, building outlines, and road networks. This is critical for infrastructure detection and urban planning applications.

Texture restoration: Fine-scale textures in natural environments (vegetation patterns, terrain variation) and urban areas (building rooftops, parking lots) show significant enhancement compared to bicubic interpolation, demonstrating the model's ability to hallucinate plausible high-frequency details.

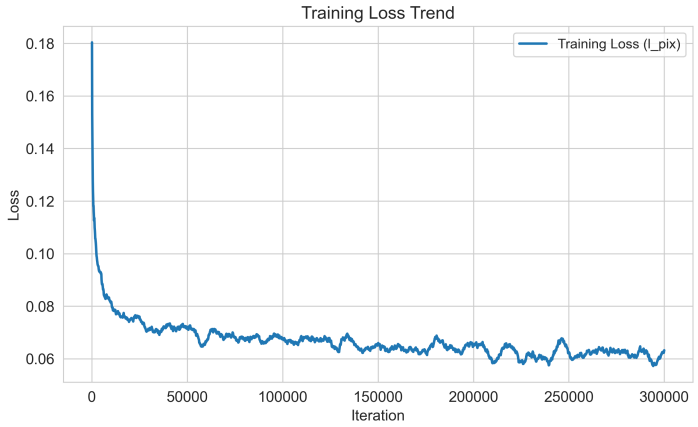

Fig. 2. Training L1 loss vs. iterations

Spectral consistency: Preliminary observations indicate that the model maintains consistent spectral relationships across the four bands. Vegetation appears with appropriate characteristics, and artificial structures retain their expected spectral signatures. This is crucial for ensuring that super-resolved imagery remains suitable for multispectral analysis and classification tasks.

5.3 Application Scenarios

As the challenge requires, we test our super-resolution approach on images coming from different scenarios. For each scenario, we provide:

- Original low-resolution input (64×64×4);
- Ground truth high-resolution reference (256×256×4);
- Super-resolved output (256×256×4);
- Per-channel PSNR and SSIM metrics.

Scenario 1: Centurion, USA — Please note: the .kmz provided leads to some coordinates in Portugal. In Fig. 3 we show the selected crop:

Metric	Value	R-G-B-NIR
PSNR (dB)	29.86	[31.9, 34.4, 35.2, 25.47]
SSIM	0.864	[0.87, 0.89, 0.91, 0.77]

Scenario 2: Bologna, Italy — September 2024

Metric	Value	R-G-B-NIR
PSNR (dB)	27.47	[28.5, 30.2, 27.8, 24.9]
SSIM	0.849	[0.85, 0.86, 0.87, 0.79]

Scenario 3: Bitonto, Italy — 12/04/2022

Metric	Value	R-G-B-NIR
PSNR (dB)	25.14	[24.9, 26.6, 26.0, 23.5]
SSIM	0.839	[0.83, 0.85, 0.87, 0.79]

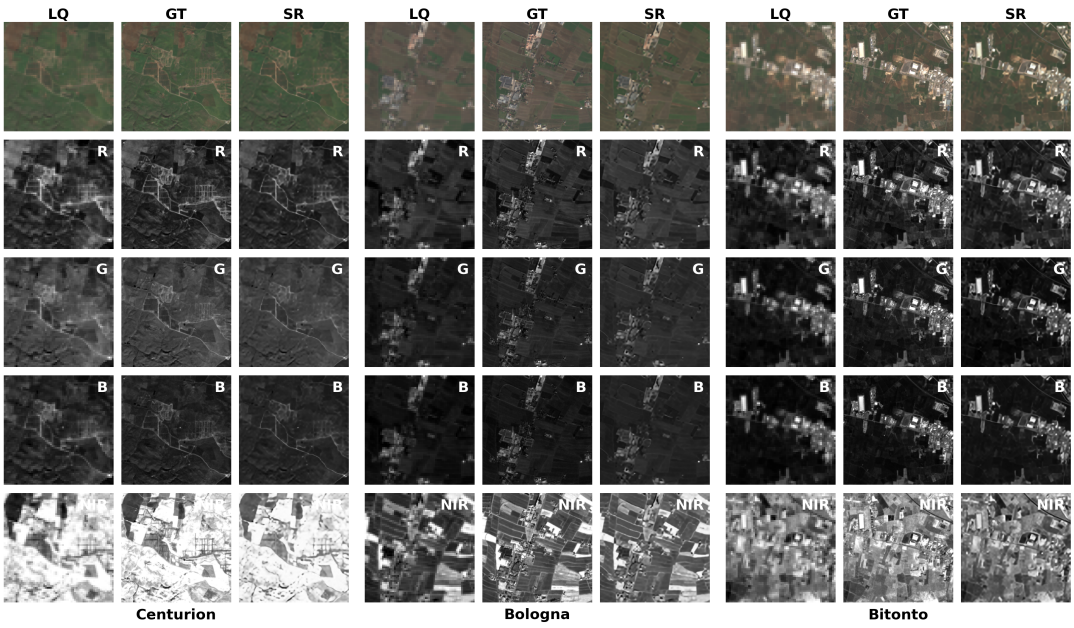

Fig. 3. Our selection of crops from the scenes. The horizontal layout is respectively LQ, GT and SR, while the vertical one is RGB, R, G, B and NIR

6 Discussion

6.1 Innovation and Contributions

Our work advances the state of the art in satellite image super-resolution through the following contributions:

1. Multispectral Transformer adaptation: We demonstrate that the Progressive Focused Transformer, originally designed for RGB imagery, can be successfully adapted to multispectral satellite data. The focused attention mechanism proves particularly effective for capturing long-range spatial dependencies in remote sensing contexts, where structural patterns (field boundaries, road networks, urban layouts) span large spatial extents.

2. Spectral integrity preservation: By training with equal-weighted L1 loss across all four bands and mean-centered normalization, our approach maintains spectral fidelity essential for downstream quantitative analysis. Unlike RGB-focused methods that prioritize visual quality, our solution respects the radiometric nature of satellite data.

3. Computational efficiency: The progressive focused attention mechanism significantly reduces computational complexity compared to standard Transformer architectures, making the approach more tractable for operational deployment.

6.2 Limitations and Future Work

As a research prototype developed by a student team with limited computational resources, our current implementation has several limitations:

Training scale: Given more extensive GPU resources, we could train on larger datasets (10k+ samples), explore deeper architectures, and conduct comprehensive hyperparameter optimization. Extended training (500k-1M iterations) would likely yield additional performance gains.

Model capacity: Computational constraints limited our ability to experiment with larger model variants. Increasing the number of Transformer blocks and attention heads could improve reconstruction quality, particularly for complex urban scenes.

Multi-scale evaluation: While we have focused on 4× super-resolution as required by the challenge, exploring 10× super-resolution would require cascaded architectures or progressive upsampling strategies that we have not yet investigated due to time and resource constraints.

6.3 Potential Value for Enel

For Enel's operational needs in renewable energy infrastructure monitoring, our super-resolution approach offers several benefits:

Enhanced detection capability: By improving the effective spatial resolution of freely available Sentinel-2 imagery from 10m to 2.5m, our method enables detection of smaller photovoltaic installations and more precise delineation of existing systems.

Cost efficiency: Super-resolving freely available Sentinel-2 data reduces reliance on expensive high-resolution commercial imagery. For continental-scale monitoring, this could represent substantial cost savings.

Temporal frequency: Sentinel-2's 5-day revisit time (with both satellites) combined with super-resolution enables frequent monitoring for change detection.

Scalability: Once trained, the model can process large image amounts efficiently using GPU acceleration. Inference on a typical Sentinel-2 tile could be completed in minutes on modern hardware, supporting operational workflows.

7 Conclusions

We have developed and demonstrated a Transformer-based approach to 4× super-resolution of multispectral Sentinel-2 satellite imagery, focusing on the four highest-resolution bands (R, G, B, NIR). Our adaptation of the Progressive Focused Transformer architecture successfully handles multispectral data while maintaining spectral integrity and computational efficiency.

Despite significant computational constraints as a student team, we have achieved promising preliminary results that validate the feasibility of Transformer-based methods for operational satellite image enhancement. The focused attention mechanism proves particularly well-suited to capturing the long-range spatial dependencies characteristic of remote sensing data, while the progressive refinement strategy enables efficient processing.

Our work demonstrates clear potential value for Enel's renewable energy monitoring applications, with promising implications for detection capability, cost efficiency, and scalability. However, we acknowledge that as a research prototype, there is substantial room for improvement given additional resources. With access to professional-grade computational infrastructure, larger training datasets, and extended development time, we are confident that the performance and robustness of this approach could be significantly enhanced.

The methodology presented here represents a solid foundation for advancing the state of the art in satellite image super-resolution, with direct applicability to operational challenges in renewable energy infrastructure monitoring and broader Earth observation domains.

You can find code and additional information here: https://github.com/marcociama/SEN2PFT-MSSR

References

1. W. Long, X. Zhou, L. Zhang, S. Gu: Progressive Focused Transformer for Single Image Super-Resolution. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2279-2288, June 2025.
2. X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, C. Change Loy: BasicSR: Open Source Image and Video Restoration Toolbox. `https://github.com/XPixelGroup/BasicSR`, 2020.
3. European Space Agency: Sentinel-2 User Handbook. ESA Standard Document, Issue 1, Rev 2, 2015.
4. C. Dong, C. C. Loy, K. He, X. Tang: Learning a Deep Convolutional Network for Image Super-Resolution. In: *European Conference on Computer Vision (ECCV)*, pp. 184-199, 2014.
5. B. Lim, S. Son, H. Kim, S. Nah, K. Mu Lee: Enhanced Deep Residual Networks for Single Image Super-Resolution. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 136-144, 2017.