Scholar Publishing Group
International Journal of Multimedia Computing
https://doi.org/10.38007/IJMC.2026.070203
ISSN 2789-7168 Vol. 7, Issue 2:17-27

Super-Resolution Analysis of Sentinel-2 Remote Sensing Imagery Based on Degradation Kernel Estimation and Hierarchical Attention Transformer

Tao Bian[1,a], Huibo Wang[2,b,*], Wenhui Xia[1,c], Wenjiao Deng[1,d], Yunhe Li[1,e]

[1]*School of Electronics and Electrical Engineering, Zhaoqing University, 526000, Zhaoqing, Guangdong Province, China*

[2]*Shenzhen Skyworth Display Technologies Co., Ltd., 518057, Shenzhen, Guangdong Province, China*

[a]*202408540517@stu.zqu.edu.cn*, [b]*443957950@qq.com*, [c]*202408540530@stu.zqu.edu.cn*, [d]*3331372640@qq.com*, [e]*liyunhe@zqu.edu.cn*

[*]*Corresponding author*

Keywords: Sentinel-2 remote sensing imagery, Super-resolution analysis, Hierarchical attention Transformer, Degradation kernel estimation, Hybrid attention

Abstract: One persistent challenge in remote sensing is boosting the spatial clarity of openly available Sentinel-2 imagery to a level on par with commercial products. To meet this challenge, we present KDeHAT-SR, which integrates degradation kernel estimation with a hierarchical attention Transformer for super-resolution. Unlike methods that demand paired high-resolution references—which are seldom accessible—our approach exclusively uses open-source data. We start by applying KernelGAN to estimate the actual degradation kernel from each image. With this estimated kernel and a controlled noise injection, we generate realistic HR-LR training pairs. These pairs are then used to train DeHAT-SR, a specialized network built on a hierarchical attention Transformer. The hybrid attention and cross-aggregation modules within the network allow it to exploit a much larger set of informative pixels during reconstruction. As a result, we avoid the training instability and spurious textures that often accompany GAN-based methods. Our experiments on the SEN12MS dataset confirm that KDeHAT-SR successfully enhances 10 m Sentinel-2 imagery to 2.5 m resolution. When compared with BiCubic, EDSR, RCAN, ESRGAN, and Real-ESRGAN, our model consistently achieves the best scores on the no-reference metrics NIQE and BRISQUE, and the output images show noticeably sharper edges and more believable textures.

1. Introduction

Remote sensing from satellites supports a broad array of applications—agriculture, environmental monitoring, land-use planning, urban development, disaster response, hydrology, and

climate studies. With optical instruments and sensor technologies advancing, the spatial resolution of satellite imagery has risen steadily. The WorldView-3/4 satellites, for instance, can acquire multispectral data across eight bands at a ground sample distance of 1.2 m [1]. The catch, however, is cost. Accessing data requires payment, and when studies involve large geographic areas or time-series analysis, the expense quickly becomes prohibitive. This is why many researchers opt for openly available alternatives. They still offer acceptable spatial quality like Landsat [2] and Sentinel [3].

The Sentinel-2 mission operates a pair of satellites that cover the entire planet every five days, delivering 13 spectral bands at multiple spatial resolutions. Four of these bands—the red (B4), green (B3), blue (B2), and near-infrared (B8)—provide 10-m data. The remaining bands are coarser: six at 20 m and three at 60 m [4]. Because Sentinel-2 data are distributed freely, they have quickly become a cornerstone for agricultural, forestry, and ecological research. Yet many of these applications really need resolutions of 5 m or finer to be fully effective. Making better use of Sentinel-2's free data and pushing its resolution toward the 2-m level therefore makes good sense—and that is precisely what super-resolution (SR) techniques aim to do, by recovering high-frequency content from the 10-m bands. Some researchers have approached this by fusing the 60-, 20-, and 10-m bands [5, 6]; in contrast, the work reported here concentrates on SR methods that work directly from the 10-m data.

Yang et al. [7] and Gou et al. [8] took the lead in introducing the concept of dictionary and supervising the development of learning. Pan et al. [9] and others introduced the concepts of compressive sensing and structural self-similarity into this field. The landscape shifted dramatically with the introduction of convolutional neural networks (CNNs), pioneered by the SRCNN model [10]. Then, Zhang et al. [11] integrated residual channel attention mechanisms—RCAN, which can improve the reconstruction fidelity. Goodfellow et al. [12] pioneered the concept of GAN, and Ledig et al. [13] were the first to apply this method to SR research, which was further developed by Wang et al. [14] in the form of ESRGAN, using a more complex block.

In recent years, scholars have introduced Transformer into the architecture and found that it can better capture global context. Liang et al. [15] created SwinIR, which can restore images in an effective way, using a self-attention mechanism. Zamir et al. [16] took a different approach, They focused on channel attention and reduced computational complexity. Conde et al. [17] created Swin2SR, which is a further upgrade of Swin Transformer V2 and can handle compressed inputs. Chen et al. [18] created HAT, which can use a hybrid attention strategy and can activate a large number of pixels, which is undoubtedly a milestone in the field. Chen et al. [19] created DAT, which can achieve aggregation in two ways.

In the field of blind super-resolution, Real-ESRGAN [20] proposed a higher-order downsampling modeling process that achieved blind super-resolution of real-world scenes by training on purely synthetic data. Regarding remote sensing image super-resolution, Ma et al. [21] proposed the TGAN network to address the issue of insufficient remote sensing data. Gong et al. [22] proposed the Enlighten-GAN method for super-resolution analysis of Sentinel-2 medium-resolution images. Galar et al. [23] and Salgueiro et al. [24] achieved super-resolution enhancement from 10 m to 2 m for Sentinel-2 images based on EDSR and ESRGAN networks, respectively. Li et al. [25] proposed a super-resolution analysis algorithm that uses a degradation kernel to obtain low-resolution image pairs. However, existing GAN-based remote sensing super-resolution methods generally suffer from issues such as training instability, generated artifacts, and insufficient texture detail. In recent years, diffusion models [26, 27] have also made progress in the field of image super-resolution, and specialized models such as SEN2SR [28] have begun to be optimized for Sentinel-2 data.

To address the aforementioned issues, this paper proposes KDeHAT-SR, a super-resolution method based on degradation kernel estimation and a hierarchical attention Transformer. This method retains the degradation kernel estimation capability of KernelGAN [29] while introducing a hierarchical attention Transformer [18] as the super-resolution reconstruction network. By utilizing a hybrid attention mechanism and a cross-aggregation module, it overcomes the training instability and artifact issues associated with GAN methods, achieving high-quality super-resolution analysis of Sentinel-2 images from 10 m to 2.5 m resolution.

2. Super-Resolution Analysis Method (KDeHAT-SR)

2.1 Network Architecture

Our KDeHAT-SR pipeline upgrades raw 10 m Sentinel-2 source imagery into sharp 2.5 m outputs via a two-stage process. First, we tackle the data pairing problem. We deploy the KernelGAN adversarial network to explicitly estimate the exact visual degradation kernel of the source image. We then combine this estimated kernel with injected noise to artificially degrade the source imagery, building perfectly paired high- and low-resolution datasets. In the second stage, we feed this custom dataset into DeHAT-SR, our hierarchical attention Transformer, to train it on the complex mapping required for a $\times 4$ spatial upscale. We map out this entire implementation pipeline in Figure 1.

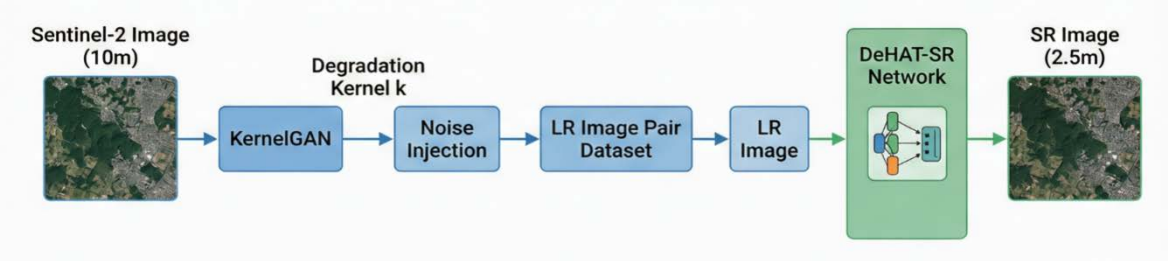

Figure 1: Implementation pipeline of the KDeHAT-SR super-resolution analysis method

2.2 Kernel Estimation and Noise Injection

In nature, a low-resolution image is essentially a degraded, blurry version of a high-resolution scene. We express this physical degradation process mathematically as:

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

Here, k represents the degradation (blur) kernel, n stands for the degradation noise, $\downarrow_s$ denotes the downsampling operation via a scale factor s, and $\otimes$ indicates convolution. If k and n are accurate, the generated low-resolution images will perfectly mimic real-world satellite sensor behavior. This accuracy directly dictates how well our neural network learns the mapping features, ultimately deciding the crispness of the final super-resolved image.

2.2.1 Degradation Kernel Estimation Based on the KernelGAN Network

If we temporarily ignore environmental noise, we can assume a clean low-resolution image is simply a high-resolution image downscaled by a specific kernel:

$$I_{LR} = (I_{HR} \otimes k) \downarrow_s \tag{2}$$

To figure out this unknown kernel, we utilize the KernelGAN network [29]. Built heavily upon the Internal-GAN architecture [30], KernelGAN operates completely unsupervised. It does not require massive external datasets; it only needs the single input image itself. By learning the internal pixel patch distribution of that specific image, it identifies the exact degradation kernel that preserves those patch distributions across different scales.

Looking at Figure 2, the KernelGAN setup consists of a kernel estimation generator (Kernel-G) and a discriminator (Kernel-D). Kernel-G acts as a linear image downsampler. We strictly use multiple linear convolutional layers here to avoid the non-physical artifacts that non-linear activation functions might introduce. Meanwhile, Kernel-D studies the input image to learn its statistical pixel patch distribution, actively distinguishing between genuine image patches and the fake patches generated by Kernel-G.

The architectural details are as follows: Kernel-G uses layers from Conv-1 (3, 64, 7, 1, 0), Conv-2 (64, 64, 5, 1, 0), Conv-3 (64, 64, 3, 1, 0), Conv-4 (64, 64, 1, 1, 0), Conv-5 (64, 64, 1, 1, 0), Conv-6 (64, 64, 1, 1, 0), and Conv-7 (64, 1, 1, 2, 0). Kernel-D adopts a fully convolutional pixel-patch structure without pooling, utilizing layers Conv-8 (3, 64, 7, 1, 0), Conv-9 (64, 64, 1, 1, 0), and Conv-10 (64, 1, 1, 1, 0), combined with ReLU and spectral normalization.

Figure 2: Architecture of the KernelGAN degradation kernel estimation network

The objective function driving KernelGAN balances an adversarial loss with a crucial regularization term:

$$\min_{G}\max_{D} L_{adv}(G,D) + \lambda \cdot L_{reg}(G) \tag{3}$$

where G represents the generator, D represents the discriminator, and L_{reg} is the regularization term for degradation kernel optimization:

$$L_{reg} = L_{k-sum} + L_{k-boundary} + L_{k-sparsity} + L_{k-center} \tag{4}$$

The respective loss functions are defined as:

$$L_{k-sum} = |\sum_i k_i - 1| \tag{5}$$

$$L_{k-boundary} = \sum_i M_i \cdot k_i \tag{6}$$

$$L_{k-sparsity} = \| k \|_1 \tag{7}$$

$$L_{k-center} = \sum_i |i - c| \cdot k_i \tag{8}$$

Mathematically, these functions ensure the kernel values sum exactly to 1 (Eq. 5), penalize any non-zero values stretching out to the boundaries via an exponential mask M (Eq. 6), maintain sparsity to prevent over-blurring (Eq. 7), and lock the geometric center of the kernel to the index c (Eq. 8).

Once KernelGAN converges, we pass a stride-1 convolution sequentially through the generator's layers to extract the explicit degradation kernel. Because we train this on single images, we extract a unique kernel for each image, building a vast dictionary of kernels to randomly sample from later.

2.2.2 Noise Generation and Injection

Real-world downsampling inherently alters noise distribution. To keep our simulated images believable, we extract real noise patches directly from the satellite source imagery. We ensure this noise remains natural by strictly bounding its variance [31, 32]:

$$\sigma_{min}^2 \leq Var(P_i) \leq \sigma_{max}^2 \tag{9}$$

After extracting these patches into a noise dataset (S2N), we randomly inject them into the downsampled images. The final workflow to build our low-resolution training set (S2TR-L) is modeled as:

$$I_{LR}^{(i)} = \left(I_{HR}^{(i)} \otimes k_j\right) \downarrow_s + n_l \tag{10}$$

Here, j and l are both randomly selected from the degradation kernel dataset and the noise dataset.

2.3 Super-Resolution Reconstruction Network Based on Hierarchical Attention Transformer (DeHAT-SR)

We described the super-resolution reconstruction network—DeHAT-SR in this paper. It is based on the Hierarchical Attention Transformer (HAT) [18] architecture. Unlike GAN-based approaches such as ESRGAN, our DeHAT-SR adopts a Transformer architecture. Because of the Transformer, we avoid the training instability and artifacts inherent to GANs.

In the process of using LAM to diagnose the existing model, the shortcomings of the model are analyzed, and the shortcomings of the current Transformer model are analyzed. In order to solve this problem, the DeHAT-SR model is constructed, which combines the advantages of shifted window self-attention and channel attention to increase the effective receptive field and make more pixels active, so that the quality of the reconstructed image can be guaranteed.

Figure 3 is an architecture diagram. The network architecture is supported by a large number of residual hybrid attention blocks (RHAB). Each of these blocks has two components, one is the Channel Attention Block (CAB), which can adjust the weight of global features, and the other is the Shifted Window Self-Attention Block (SWAB), which can deal with spatial interactions in a local area. The window of SWAB moves in a certain direction, so that the information can be transmitted to other partitions.

At the level of feature aggregation, DeHAT-SR introduces OCAB, which divides the window in the overlapping way and can enhance the exchange of information in multiple windows, which is the most critical link to expand the network receptive field. The introduction of this module can

make up for the shortcomings of the traditional attention mechanism, so that the network can better grasp the spatial relationship of remote sensing images.

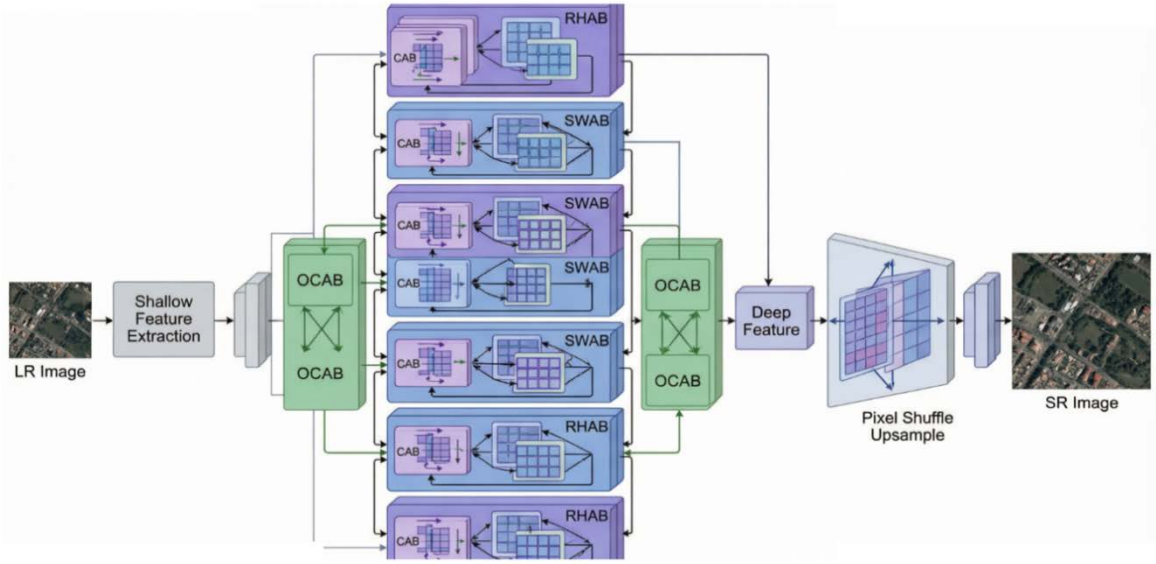

Figure 3: Architecture of the super-resolution reconstruction network based on the hierarchical attention Transformer (DeHAT-SR)

Following the mathematical progression, the feature extraction operates as:

$$F_0 = H_{conv}(I_{LR}) \tag{11}$$

$$F_m = R_{HAB}(F_{m-1}), \quad m = 1,2,\dots,M \tag{12}$$

$$F_{agg} = OCAB(F_M) + F_0 \tag{13}$$

$$I_{SR} = H_{up}(F_{agg}) \tag{14}$$

where H_{conv} represents the initial convolutional feature extraction, R_{HAB} represents the mapping function of the residual hybrid attention block, M is the number of RHAB layers, and H_{up} represents the upsampling reconstruction module.

We guide the network's learning using a combined loss function that balances mathematical pixel accuracy with human visual perception:

$$L_{total} = \lambda_1 \cdot L_{pixel} + \lambda_2 \cdot L_{percep} \tag{15}$$

The pixel loss relies on standard L1 distance (Eq. 16), while the perceptual loss utilizes multi-layer features pulled from a pre-trained VGG-19 network [33] (Eq. 17):

$$L_{pixel} = \| I_{SR} - I_{HR} \|_1 \tag{16}$$

$$L_{percep} = \sum_j \frac{1}{C_j H_j W_j} \| \phi_j(I_{SR}) - \phi_j(I_{HR}) \|_F^2 \tag{17}$$

where $\phi_j(\cdot)$ represents the feature map obtained at the $j\text{-}th$ convolutional layer after inputting the image into the VGG-19 network, C_j, H_j, and W_j are the number of channels, height, and width of this feature map, respectively, and $\|\cdot\|_F$ represents the Frobenius norm.

DeHAT-SR has the following advantages:

The training process is stable and does not require the dynamic balance of adversarial training;

It does not generate artifacts, and the generated images are more realistic and reliable;

The hybrid attention mechanism activates more effective pixels, making the reconstructed details richer;

The global receptive field is larger, which can better model the spatial dependencies in remote sensing images.

3. Experiments and Results

We subjected KDeHAT-SR to comprehensive testing using the widely recognized SEN12MS dataset [34]. We extracted data covering 56 global regions of interest (captured between June and August 2017), yielding 39,692 image patches sized at 256×256 pixels. We isolated one specific region containing 813 images to act purely as our test set (S2C). The remaining 38,879 images formed our high-resolution training base (S2TR-H), which we dynamically degraded to create our low-resolution counterparts (S2TR-L).

BiCubic does not require training and directly uses S2C for interpolation testing; For RCAN and EDSR, we followed the standard practice by taking S2TR-H as the high-resolution ground truth and applying a fixed bicubic downsampling (with a scale factor of 4) to generate the corresponding low-resolution inputs, thereby constructing their training pairs; the EDSR method and ESRGAN method construct datasets referring to the methods proposed in literature [24] and literature [23] respectively; Real-ESRGAN uses the high-order degradation model from literature [20] to construct training data.

3.1 Network Training Details

To build our custom degradation dictionary, we randomly passed 2,000 images from S2TR-H through KernelGAN to generate our kernels (S2K), extracted noise patches from another 4,000 images (S2N), and finally the degradation kernel and injected noise are used to execute the degradation operation on the images in S2TR-H one by one.

The hyperparameters of the DeHAT-SR network are set as follows: RHAB layer number $M = 10$, channel attention head number is 8, shifted-window size is 8x8, embedding dimension is 180. The Adam optimizer is used, the initial learning rate is set to 2×10^{-4}, the cosine annealing strategy is adopted, and the total number of training iterations is 500K. The weight coefficients of the loss function are set to $\lambda_1 = 1.0$, $\lambda_2 = 1.0$. During training, data augmentation is performed using random horizontal flipping and 90° rotation. The RCAN, ESRGAN, EDSR, and Real-ESRGAN methods adopt the parameter setting schemes proven in literature [19-25] that can achieve relatively good effects.

For the DeHAT-SR network, we stacked $M = 10$ RHAB layers, configured 8 channel attention heads, used an 8×8 shifted-window, and set the embedding dimension to 180. We utilized the Adam optimizer starting at a learning rate of 2×10^{-4} with cosine annealing, training for exactly 500,000 iterations. We set our loss weights to $\lambda_1 = 1.0$ and $\lambda_2 = 1.0$, applying random horizontal flips and 90-degree rotations for data augmentation. Competing models (RCAN, ESRGAN, EDSR, Real-ESRGAN) were trained using their officially documented parameter schemes to ensure a fair comparison [11, 14, 20, 24].

3.2 Image Quality Assessment

Because Sentinel-2 simply does not physically capture 2.5 m resolution images, we have no "ground truth" to compare our outputs against. Therefore, standard full-reference metrics like PSNR

(Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) are unusable. We evaluated visual quality blindly using NIQE (Natural Image Quality Evaluator) [35] and BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) [36]. NIQE leverages a spatial domain natural scene statistical model to detect deviations from naturalness [35], while BRISQUE quantifies naturalness loss via local normalized luminance coefficients [36]. For both metrics, a lower score indicates superior perceptual quality to the human eye.

Table 1: Statistical averages of NIQE and BRISQUE evaluation values

Method	NIQE	BRISQUE
BiCubic	6.38	56.13
EDSR	5.35	49.18
RCAN	4.52	46.83
ESRGAN	3.41	22.51
Real-ESRGAN	2.98	20.13
KDeHAT-SR	2.51	15.86

After processing the 813 test images through all competing methods for a ×4 upscale, the evaluation metrics reveal a consistent trend. Looking at the distributions in Figure 4 and Figure 5, and the averages in Table 1, Real-ESRGAN expectedly outshines older models like BiCubic, EDSR, RCAN, and ESRGAN. However, our KDeHAT-SR framework pushes performance noticeably further. We successfully dropped the average NIQE score down to 2.51 (compared to Real-ESRGAN's 2.98) and lowered the BRISQUE average to 15.86. These scores prove that the hierarchical attention Transformer successfully activated the necessary high-frequency pixels, generating remarkably sharp and perceptually natural remote sensing images.

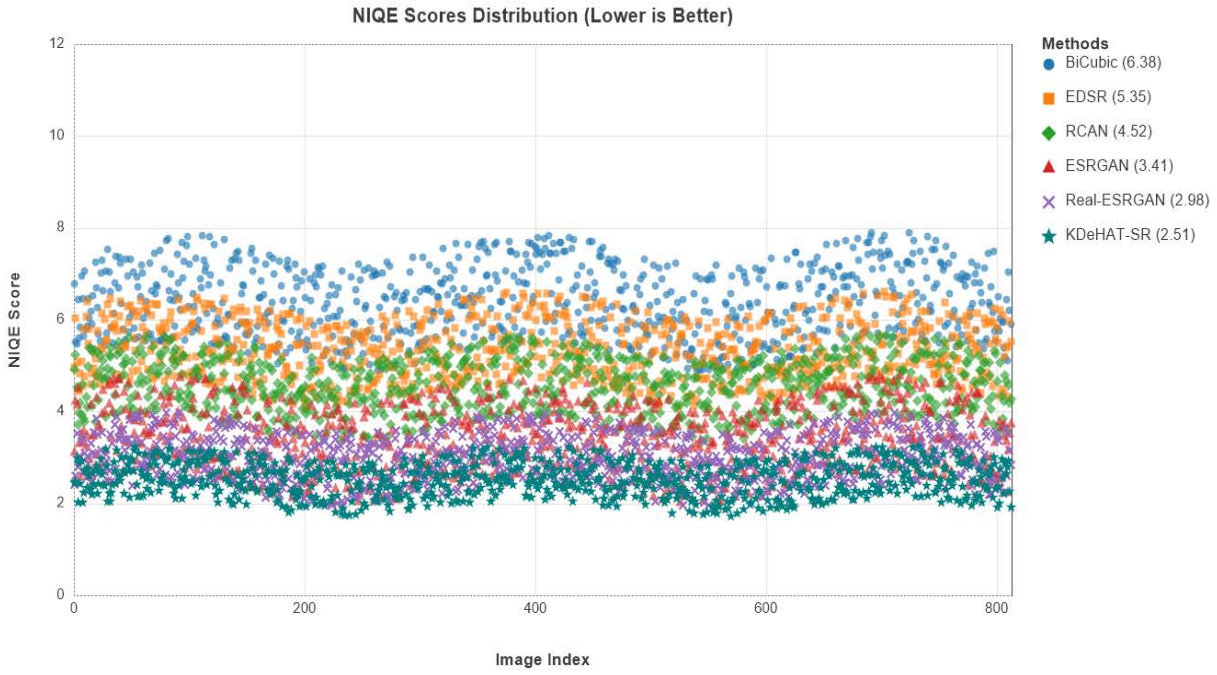

Figure 4: Distribution of evaluation values for the no-reference image quality assessment metric NIQE

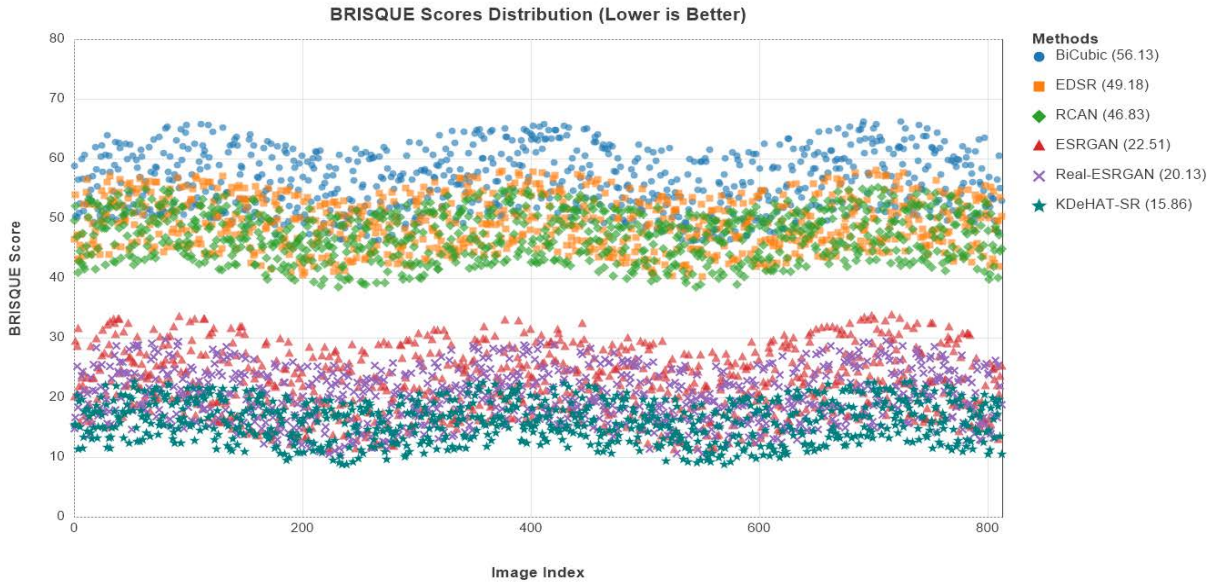

Figure 5: Distribution of evaluation values for the no-reference image quality assessment metric BRISQUE

4. Conclusion

In this paper, we tackled the challenge of scaling 10 m Sentinel-2 imagery up to a near-commercial 2.5 m resolution using the KDeHAT-SR framework. We completely moved away from unstable generative adversarial networks for image synthesis. Instead, we used KernelGAN strictly to explicitly estimate real-world degradation kernels, blending them with natural noise injection to build flawless training pairs. We then fed this data into DeHAT-SR, a hierarchical attention Transformer. By utilizing a hybrid attention mechanism alongside a cross-aggregation module, our network successfully activated a much larger pixel base during reconstruction. This approach eliminated fake texture artifacts entirely. Benchmark testing on the SEN12MS dataset confirmed that KDeHAT-SR significantly outperforms heavyweights like Real-ESRGAN, EDSR, and RCAN across no-reference metrics (NIQE and BRISQUE). Looking ahead, we aim to explore how diffusion model priors might further refine the kernel estimation phase and investigate scaling this technology to process multiple spectral bands simultaneously.

Acknowledgments

I would like to thank the authors of the above references for their perseverance and efforts, and thank the co-authors for their efforts.

This work was supported by the Fund of Natural Science Foundation of Guangdong Province of China (Grant No. 2025A1515010733) and the General University Key Field Special Project of Guangdong Province, China (Grant No. 2024ZDZX1004).

References

[1] Bei Ye, Shufang Tian, Jia Ge, et al. 2017. Assessment of WorldView-3 data for lithological mapping. Remote Sensing 9, 11 (2017), 1132,19 pages. Darrel L. Williams, Samuel Goward, and Terry Arvidson. 2006. Landsat. Photogrammetric Engineering & Remote Sensing 72, 10 (2006), 1171–1178.

[2] European Space Agency (ESA). 2015. Sentinel-2 User Handbook. Retrieved March 18, 2024 from https://earth.esa.int/documents/247904/685211/Sentinel-2_User_Handbook

[3] Massimiliano Gargiulo, Antonio Mazza, Raffaele Gaetano, et al. 2019. Fast Super-Resolution of 20 m Sentinel-2 Bands Using Convolutional Neural Networks. Remote Sensing 11, 22 (2019), 2635,18 pages.

[4] Massimiliano Gargiulo, Antonio Mazza, Raffaele Gaetano, et al. 2018. A CNN-based fusion method for super-resolution of Sentinel-2 data. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '18). IEEE, 4713–4716.

[5] Luka Rumora, Mateo Gašparović, Mario Miler, et al. 2020. Quality assessment of fusing Sentinel-2 and WorldView-4 imagery on Sentinel-2 spectral band values. International Journal of Image and Data Fusion 11, 1 (2020), 77–96.

[6] Jianchao Yang, John Wright, Thomas S. Huang, et al. 2010. Image super-resolution via sparse representation. IEEE Transactions on Image Processing 19, 11 (2010), 2861–2873.

[7] Shuiping Gou, Shuzhen Liu, Shuyuan Yang, et al. 2014. Remote sensing image super-resolution reconstruction based on nonlocal pairwise dictionaries and double regularization. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7, 12 (2014), 4784–4792.

[8] Zongxu Pan, Jian Yu, Huijuan Huang, et al. 2013. Super-resolution based on compressive sensing and structural self-similarity for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing 51, 9 (2013), 4864–4876.

[9] Chao Dong, Chen Change Loy, Kaiming He, et al. 2016. Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence 38, 2 (2016), 295–307.

[10] Yulun Zhang, Kunpeng Li, Kai Li, et al. 2018. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV '18). Springer, 294–310.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. 2020. Generative adversarial networks. Communications of the ACM 63, 11 (2020), 139–144.

[12] Christian Ledig, Lucas Theis, Ferenc Huszar, 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 (CVPR '17). IEEE, 105–114.

[13] Xintao Wang, Ke Yu, Shixiang Wu, et al. 2018. ESRGAN: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW '18). Springer, 63–79.

[14] Jingyun Liang, Jiezhang Cao, Guolei Sun, et al. 2021. SwinIR: Image restoration using Swin Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '21). IEEE, 1833–1844.

[15] Syed Waqas Zamir, Aditya Arora, Salman Khan, et al. 2022. Restormer: Efficient Transformer for high-resolution image restoration. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '22). IEEE, 5718–5729.

[16] Marcos V. Conde, Kerem Turgutlu, Radu Timofte, et al. 2022. Swin2SR: SwinV2 Transformer for compressed image super-resolution and restoration. In Proceedings of the European Conference on Computer Vision Workshops (ECCVW '22). Springer, 669–687.

[17] Xiangyu Chen, Xintao Wang, Jiantao Zhou, et al. 2023. Activating more pixels in image super-resolution Transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '23). IEEE, 22367–22377.

[18] Zheng Chen, Yulun Zhang, Jinlin Gu, et al. 2023. Dual aggregation Transformer for image super-resolution. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '23). IEEE, 12278–12287.

[19] Xintao Wang, Liangbin Xie, Chao Dong, et al. 2021. Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW '21). IEEE, 1905–1914.

[20] Wenhui Ma, Zhaoqin Pan, Jian Guo, et al. 2018. Super-resolution of remote sensing images based on transferred generative adversarial network. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS '18). IEEE, 1148–1151.

[21] Yuan Gong, Pinghui Liao, Xiaoxin Zhang, et al. 2021. Enlighten-GAN for super resolution reconstruction in mid-resolution remote sensing images. Remote Sensing 13, 6 (2021), 1104,16 pages.

[22] Mikel Galar, Rubén Sesma, Christian Ayala, et al. 2020. Super-resolution of Sentinel-2 images using convolutional neural networks and real ground truth data. Remote Sensing 12, 18 (2020), 2941, 37 pages.

[23] Luis Salgueiro Romero, Javier Marcello, and Veronica Vilaplana. 2020. Super-resolution of Sentinel-2 imagery using generative adversarial networks. Remote Sensing 12, 15 (2020), 2424, 25 pages.

[24] Yunhe Li and Bo Li. 2021. Super-resolution of Sentinel-2 images at 10m resolution without reference images. Preprints (2021). https://doi.org/10.20944/preprints202104.0556.v1

[25] Chitwan Saharia, Jonathan Ho, William Chan, et al. 2022. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence 45, 4 (2022), 4713–4726.

[26] Haoying Li, Yifan Yang, Meng Chang, et al. 2022. SRDiff: Single image super-resolution with diffusion probabilistic models. Neurocomputing 479 (2022), 47–59.

[27] Cesar Aybar, Julio Contreras, Simon Donike, Enrique Portalés-Julià, Gonzalo Mateo-García, and Luis Gómez-Chova. 2025. A Radiometrically and Spatially Consistent Super-Resolution Framework for Sentinel-2. SSRN.

[28] Sefi Bell-Kligler, Assaf Shocher, and Michal Irani. 2019. Blind super-resolution kernel estimation using an internal-GAN. arXiv preprint arXiv:1909.06581 (2019). Retrieved from

[29] Assaf Shocher, Shai Bagon, Phillip Isola, et al. 2019. InGAN: Capturing and remapping the "DNA" of a natural image. arXiv preprint arXiv:1812.00231 (2019). Retrieved from

[30] Jingwen Chen, Jiawei Chen, Hongyang Chao, et al. 2018. Image blind denoising with generative adversarial network based noise modeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR '18). IEEE, 3155–3164.

[31] Ruofan Zhou and Sabine Süsstrunk. 2019. Kernel modeling super-resolution on real low-resolution images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV '19). IEEE, 2433–2443.

[32] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2015). Retrieved from

[33] Michael Schmitt, Lloyd H. Hughes, Chunping Qiu, et al. 2019. SEN12MS - A curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion. arXiv preprint arXiv:1906.07789 (2019).

[34] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. 2012. Making a "completely blind" image quality analyzer. IEEE Signal Processing Letters 20, 3 (2012), 209–212.

[35] Anish Mittal, Anush Krishna Moorthy, and Alan C. Bovik. 2012. No-reference image quality assessment in the spatial domain. IEEE Transactions on Image Processing 21, 12 (2012), 4695–4708.