A Dynamically Weighted Framework for Adaptive Reference-Based Super-Resolution

Chae-Eun Kim[1], Junhwa Chi [2]

[1]Department of Data Engineering, Pukyong National University, Busan, Republic of Korea - chaeeung@pukyong.ac.kr
[2]Major of Big Data Convergence, Division of Data Information Sciences, Pukyong National University, Busan, Republic of Korea - jchi@pknu.ac.kr

Commission III, (ICWG III/IVb)

Keywords: Reference-Based Super-Resolution, Sentinel-2, Temporal Mismatch, Reference Misuse, Adaptive Framework

Abstract

Satellite remote sensing is inherently constrained by a trade-off between spatial and temporal resolution. As a result, high-temporal-frequency sensors such as Geostationary Ocean Color Imager-II provide operationally valuable observations but at coarse spatial resolution. Reference-Based Super-Resolution (Ref-SR) can address this limitation by transferring high-resolution textures from an external reference image, but temporal mismatch between the target and reference images often leads to unreliable texture transfer and severe artifacts. This problem becomes more critical in extreme low-resolution (LR) settings, where structural information is already severely degraded. To address this issue, we propose the Dynamic Ref-SR Framework, which computes a pixel-wise weight map from intensity differences between the LR and reference images to selectively control reference transfer. The resulting weights promote reference use in stable regions while suppressing it in temporally inconsistent regions. The framework was validated on three backbone architectures—CNN (EDSR), Swin Transformer, and GAN—using a Sentinel-2 dataset for four-band reconstruction (RGB and NIR). Across all metrics and architectures, the proposed Ref-SR framework consistently outperformed the SISR baseline in both structural and spectral evaluations. Among the tested backbones, the GAN-based model achieved the best overall performance, with a PSNR of 35.60 dB, an SSIM of 0.92, a SAM of 2.20°, and an ERGAS of 74.71. These results demonstrate that the proposed framework can improve LR satellite imagery while reducing the risk of reference misuse under temporal mismatch.

1. Introduction

Satellite remote sensing is inherently constrained by a trade-off between spatial and temporal resolution. This trade-off is particularly critical for operational monitoring systems, where frequent observations are required but often come at the cost of coarse spatial resolution. As a result, the direct use of such imagery for detailed surface analysis remain limited despite its high temporal utility.

For instance, geostationary sensors such as the Geostationary Ocean Color Imager-II (GOCI-II) provide high-frequency observations suitable for dynamic monitoring, but their 250 m spatial resolution severely limits detailed spatial analysis, representing an Extreme Low Resolution (Extreme LR) setting. In contrast, polar-orbiting sensors such as Sentinel-2 provide 10 m imagery with much richer spatial detail, but their revisit intervals are insufficient for capturing rapidly evolving phenomena. This discrepancy motivates the need to bridge heterogeneous sensors that complement each other in temporal and spatial resolution.

Super-resolution (SR) has been widely investigated as a software-based approach to enhance the utility of coarse-resolution imagery without changing the physical sensor system. Among SR approaches, Single Image Super-Resolution (SISR) relies solely on the LR input and must therefore infer missing high-frequency details statistically, making the problem inherently ill-posed and limiting reconstruction reliability (Kapilaratne et al., 2022).

Reference-Based Super-Resolution (Ref-SR) addresses this limitation by introducing an external high-resolution (HR) image of the same region as a reference. By transferring valid HR textures from the reference image, Ref-SR can restore details more faithfully than SISR when an appropriate reference is available (Dong et al., 2021).

Sentinel-2 provides a dense and globally available image archive, making it a practical and often continuously accessible reference source for enhancing other LR satellite observations. However, this operational advantage comes with a major challenge: the reference image is rarely acquired at the same time as the target LR image. In optical remote sensing, simultaneous cloud-free HR observations are difficult because of weather constraints and revisit limitations. As a result, archived images acquired at different times must be used as references, inevitable introducing temporal discrepancies such as phenological variation, land cover change, and cloud inconsistency (Agarwal et al., 2025).

If reference textures are fused without accounting for such inconsistencies, the model may copy structurally incorrect information from the reference image, resulting in severe visual artifacts and spectral distortions. Recent studies, such as Ref-Diff (Dong et al., 2024), attempted to mitigate this problem using change-aware priors. However, such approaches often depend on pixel-level labels or architecture-specific designs, which limits their practicality in operational settings where labeled change maps are unavailable and flexible backbone selection is desirable .

This issue becomes even more critical in Extreme LR settings. Unlike prior Ref-SR studies that mainly target medium-to-high-resolution enhancement (e.g., 10 m to 2.5 m), restoring imagery from 240 m to 60 m involves far more severe structural degradation. At this scale, a single LR pixel covers a broad spatial extent, and geometric outlines of surface features are already substantially blurred. Consequently, incorrect reference injection is more likely and more damaging, because the LR input itself provides only weak spatial cues for validating where reference textures should be transferred.

Moreover, most SR studies emphasize visually plausible restoration of RGB imagery, whereas operational Earth observation often requires spectrally consistent reconstruction across multiple bands. In particular, the Near-Infrared (NIR) band is essential for applications such as vegetation monitoring and moisture-related analysis. Nevertheless, Ref-SR studies that explicitly address multi-band reconstruction beyond remain limited.

https://doi.org/10.5194/isprs-archives-XLIX-B3-2026-1367-2026

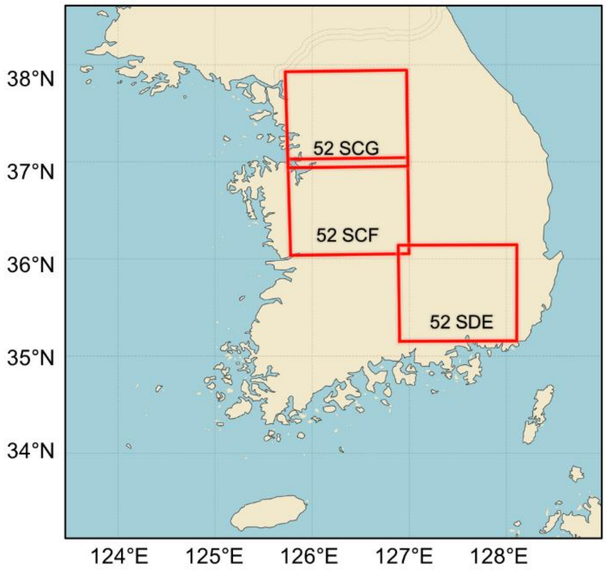

Figure 1. Study areas used for dataset construction.

To address these limitations, this study proposes a Dynamic Ref-SR Framework for operational reference-based SR under Extreme LR conditions. The proposed framework computes a pixel-wise weighting map from physical intensity differences between the LR and reference images, enabling the model to selectively use reference information in stable regions while suppressing it in temporally inconsistent regions. Because the proposed strategy is architecture-agnostic, it can be integrated into diverse backbone models without requiring additional change labels. We validate this framework for four-band (RGB+NIR) reconstruction to improve both spatial detail and spectral reliability in heterogeneous satellite SR scenarios.

2. Methodology

2.1 Dataset Construction and Preprocessing

To evaluate the proposed Dynamic Ref-SR Framework, a dataset was constructed from the Sentinel-2 imagery archive. Three Sentinel-2 tiles covering the central and southern Korean Peninsula were selected as the study areas (Figure 1). The geographic extents of these tiles are 52SCG (36°56′N–37°56′N, 126°44′E–127°58′E), 52SCF (36°03′N–37°02′N, 126°45′E–127° 59′E), and 52SDE (35°08′N–36°08′N, 127°54′E–129°06′E). These regions include diverse land cover types such as urban and forested areas, making them suitable for evaluating texture restoration and reference-control performance. Four 10 m Sentinel-2 bands (B2, B3, B4, and B8) were used.
To simulate an Extreme LR setting, an intentional resolution degradation process was applied to construct the training data. Original 10 m Sentinel-2 images acquired in 2023–2024 were first resampled to 60 m to generate HR images and then degraded to 240 m to produce LR inputs. To reduce cloud contamination and obtain stable surface textures, the reference images were generated using a median composite of Sentinel-2 time-series data from 2021–2022. Consequently, a temporal gap exists between the LR images from 2023–2024 and the reference images derived from the 2021–2022 median composite. This configuration reflects an operational scenario in which pre-existing time-series archives are used as reference data when contemporaneous HR imagery is unavailable. This setup naturally introduces texture inconsistencies caused by seasonal variations and land-cover changes, thereby reflecting conditions prone to reference misuse.
For model training, the HR and reference images were cropped into 256×256 patches, whereas the LR images were cropped into 64×64 patches. In total, 3,528 patch pairs were constructed and split into training, validation, and test sets at a ratio of 70:15:15, resulting in 2,469, 529, and 530 patches, respectively.

2.2 Dynamic Ref-SR Framework

Figure 2 illustrates the overall architecture of the proposed Dynamic Ref-SR Framework. The framework adopts a multi-input design that jointly processes an LR input image, a reference image, and a pixel-wise weight map derived from their radiometric differences. A key property of the framework is its architectural flexibility, enabling integration with various deep learning backbones. To demonstrate this generality, the framework was implemented and evaluated on three representative backbones: Enhanced Deep Residual Network (EDSR, Lim et al., 2017), Swin Transformer (Liu et al., 2021), and a GAN-based model (Wang et al., 2018).
The weight map acts as a spatial gating prior that suppresses the injection of unreliable reference textures. To generate the map, the LR image is first upsampled to the spatial resolution of the reference image. Panchromatic (Pan) representations are then constructed from both inputs to quantify their radiometric differences. Each Pan image is obtained by averaging the four Sentinel-2 optical bands, so that no single band dominates the weighting process.
The pixel-wise absolute difference between the upsampled LR image and the reference Pan image is then calculated. This difference map is transformed into a weight map in the range between 0 and 1 using the following exponential decay function:

$$\mathbf{W} = \exp\left(-\alpha \cdot \left|\mathbf{I}^{\uparrow}_{\text{LR, Pan}} - \mathbf{I}_{\text{Ref, Pan}}\right|\right) \tag{1}$$

In Equation (1), α is a scaling parameter controlling the sensitivity of the weight map to inter-image differences and was set to 0.0001. The terms $\mathbf{I}^{\uparrow}_{\text{LR, Pan}}$ and $\mathbf{I}_{\text{Ref, Pan}}$ denote the Pan images derived from the upsampled LR and reference images, respectively. Accordingly, $\mathbf{W}$ approaches 0 in regions with large discrepancies, such as land-cover change or cloud-contaminated areas, and approaches 1 in regions with minimal differences.
Within the network, the LR and reference images are passed through separate feature extraction branches based on the selected backbone (CNN, Swin Transformer, or GAN) to generate their respective feature maps, $\mathbf{F}_{\text{LR}}$ and $\mathbf{F}_{\text{Ref}}$. During feature fusion stage, the reference features are adaptively modulated by the weight map as follows:

$$\mathbf{F}_{\text{fused}} = \mathbf{F}_{\text{LR}} + \mathbf{W} \odot \mathbf{F}_{\text{Ref}} \tag{2}$$

Here, $\mathbf{F}_{\text{fused}}$ denotes the fused feature representation. When $\mathbf{W}$ is closed to 1, $\mathbf{F}_{\text{Ref}}$ is strongly injected to support fine-detail restoration. In contrast, $\mathbf{W}$ is closed to 0, the contribution of $\mathbf{F}_{\text{Ref}}$ is suppressed and the $\mathbf{F}_{\text{fused}}$ is dominated by $\mathbf{F}_{\text{LR}}$.
This mechanism suppresses reference misuse while preserving structures that remain consistent with the LR observation. For training, backbone-specific optimization strategies were applied. All models used were the Adam optimizer with an L_1 reconstruction loss.

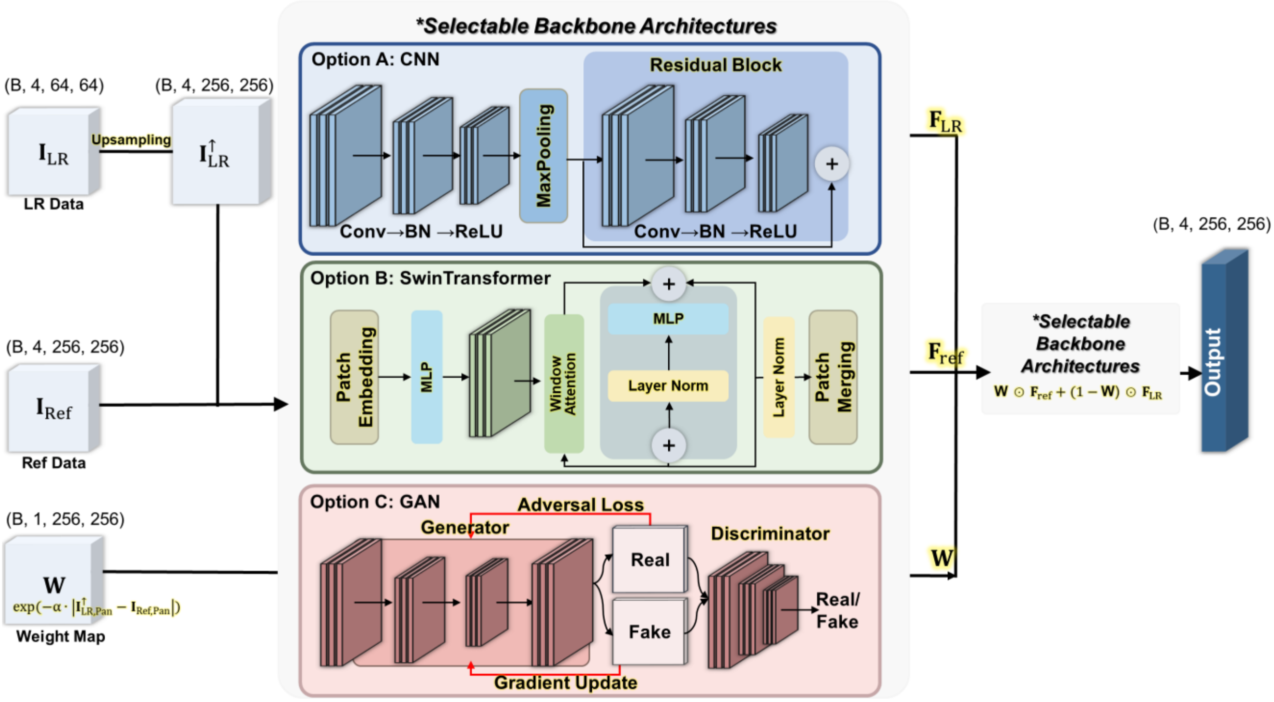

Figure 2. Overall architecture of the proposed Dynamic Ref-SR Framework.

Backbone Model	Method	PSNR(dB) ↑	SSIM ↑	SAM(deg) ↓	ERGAS↓
CNN	SISR	31.23	0.76	4.16	120.12
	Ref-SR	34.35	0.88	2.71	84.74
Swin Transformer	SISR	31.13	0.76	4.22	121.28
	Ref-SR	34.40	0.88	2.71	83.91
GAN	SISR	31.31	0.76	4.09	119.06
	Ref-SR	35.60	0.92	2.20	74.71

Table 1. Quantitative comparison of SISR and Ref-SR across three backbone architectures.

Backbone Model	Method	B2 (Blue)	B3 (Green)	B4 (Red)	B8(NIR)
CNN	SISR	0.0219	0.0226	0.0266	0.04
	Ref-SR	0.0162	0.0166	0.0194	0.0305
Swin Transformer	SISR	0.0222	0.0228	0.0268	0.0405
	Ref-SR	0.0161	0.0164	0.0192	0.0303
GAN	SISR	0.0218	0.0224	0.0263	0.0396
	Ref-SR	0.0142	0.0146	0.0168	0.0258

Table 2. Per-band RMSE comparison of SISR and Ref-SR for RGB and NIR bands.

The CNN and Swin Transformer models were trained in a single-stage. In contrast, the GAN-based model employed a two-stage strategy to improve training stability. In the first stage, the generator was pre-trained using only the L_1 loss to reconstruct coarse structures and overall spatial layouts. This stage reduces the instability and artifacts that can arise when adversarial training begins from scratch. In the second stage, the model was fine-tuned using a combination of the L_1 reconstruction loss and an adversarial loss derived from a VGG-style discriminator.
All models were trained on an NVIDIA GeForce RTX 5090 GPU with a learning rate of 1×10^{-4}, and gradient accumulation was used to improve computational efficiency.

3. Results and Discussion

3.1 Quantitative Analysis

The quantitative evaluation compared the proposed Ref-SR method with an SISR baseline that uses only the LR input. Both methods were evaluated on the Sentinel-2 test set of 530 patches using three backbone architectures: CNN (EDSR), Swin Transformer, and GAN.
Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) were used to assess spatial quality, whereas Spectral Angle Mapper (SAM) and Error Relative Global

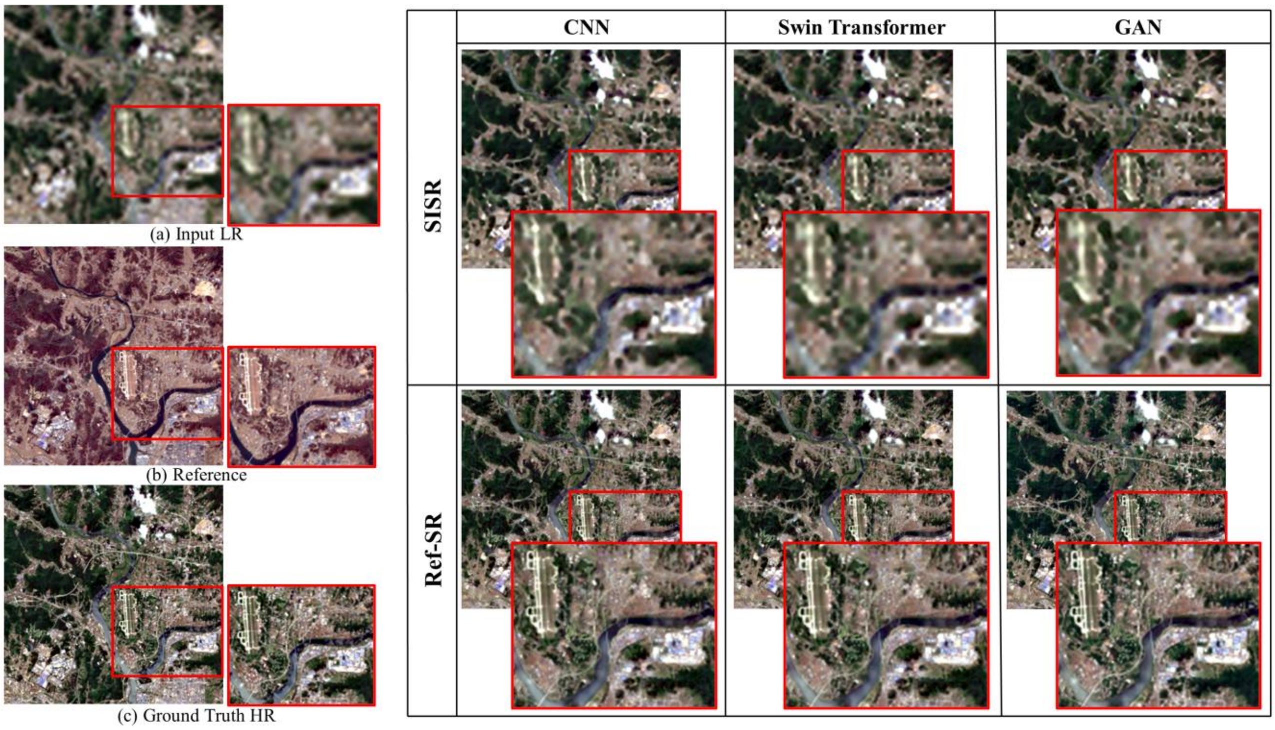

Figure 3. Visual comparison of super-resolution results between the SISR baseline and the proposed Ref-SR method across three backbone architectures (CNN, Swin Transformer, and GAN). The left panel shows (a) the input LR image, (b) the reference image, and (c) the ground truth HR image. The right grid presents SISR results in the top row and Ref-SR results in the bottom row.

Dimensionless Synthesis (ERGAS) were used to assess spectral fidelity. SAM calculates the mean angular difference between pixel spectral vectors. ERGAS represents a global relative error weighted by individual bands.

Table 1 summarizes the quantitative comparison between SISR and Ref-SR. Across the three backbones, the SISR baseline yielded PSNR values of 31.13–31.31 dB and a consistent SSIM of 0.76. The proposed Ref-SR method outperformed the SISR baseline across all metrics and backbone architectures. For the CNN and Swin Transformer backbones, Ref-SR increased PSNR by 3.12–3.27 dB to 34.35–34.40 dB, and improved SSIM from 0.76 to 0.88. Spectral fidelity also improved, with SAM decreasing from 4.16–4.22° to 2.71° and ERGAS decreasing from 120.12–121.28 to 83.91–84.74. Among the three backbones, the GAN-based model achieved the best overall performance and the largest gains from reference guidance. Its PSNR increased by 4.29 dB to 35.60 dB, and its SSIM increased from 0.76 to 0.92. It also produced the lowest spectral errors, reducing SAM from 4.09° to 2.20° and ERGAS from 119.06 to 74.71.

Table 2 compares the per-band radiometric errors of the two methods using Root Mean Square Error (RMSE). Across all three architectures, Ref-SR consistently reduced RMSE in both the RGB and NIR bands relative to the SISR baseline. For the RGB bands (B2, B3, and B4), SISR baseline produced RMSE values of 0.0218–0.0268, whereas Ref-SR reduced them to 0.0142–0.0194. The lowest RGB errors were achieved by the GAN-based Ref-SR model, with values of 0.0142–0.0168. In the NIR band (B8), RMSE decreased from 0.0400–0.0405 to 0.0303–0.0305 for the CNN and Swin Transformer models. The GAN-based model showed the largest NIR improvement, reducing RMSE from 0.0396 to 0.0258.

Taken together, the results in Tables 1 and 2 show the spatial improvements achieved by Ref-SR were accompanied by consistent per-band error reductions. This indicates that the proposed framework improved both structural quality and spectral accuracy across all evaluated bands.

3.2 Qualitative Analysis

Figure 3 visually compares the restoration results of SISR and Ref-SR across the three backbone architectures. In the comparison grid, the columns correspond to the CNN, Swin Transformer, and GAN backbones, and the top and bottom rows show the outputs of SISR and Ref-SR, respectively. In the input LR image (a), fine spatial details such as building outlines and river boundaries are largely lost under the 240 m Extreme LR condition. The reference image (b) provides HR textures, but its land-cover pattern along the riverbank differs from that of the ground truth HR (c) because of temporal mismatch.

Across all three architectures, SISR restored the coarse surface structure but remained limited in recovering high-frequency detail. The CNN- and Swin Transformer-based SISR models exhibited blurred edges and smoothered local details. The GAN-based SISR generated relatively sharper outputs, but building boundaries and river edges were still not accurately recovered.

In contrast, Ref-SR produced results that were visually closer to the ground truth HR by selectively incorporating reliable textures from the reference image through the dynamic weight map. This approach restored building outlines and river boundaries while reducing reference misuse in inconsistent regions affected by land-cover change or clouds. Among the three backbones, the GAN-based Ref-SR achieved the sharpest visual result. The CNN- and Swin Transformer-based Ref-SR models produced slightly smoother outputs than the GAN-based model, but they still showed clear texture improvements over their SISR counterparts.

Figure 4 compares the per-band reflectance histograms of the SISR and Ref-SR outputs against those of the Input LR, reference,

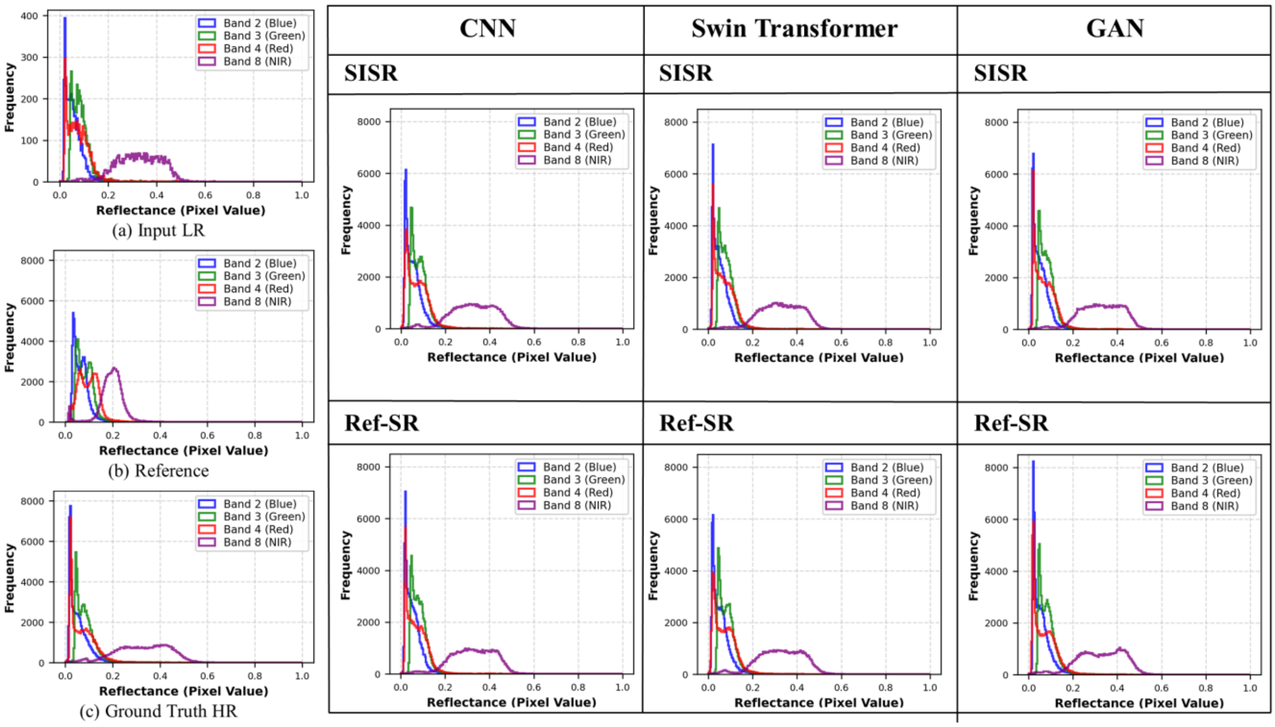

Figure 4. Per-band reflectance histogram comparison of the SISR baseline and the proposed Ref-SR method across three backbone architectures. The left panel shows the histograms of (a) the input LR image, (b) the reference image, and (c) the ground truth HR image, while the right grid compares SISR (top row) and Ref-SR (bottom row).

and ground truth HR images. As in Figure 3, the columns correspond to the CNN, Swin Transformer, and GAN backbones, and the top and bottom rows show the SISR and Ref-SR results, respectively. The ground truth HR exhibits sharp peaks in the Blue, Green, and Red bands within the low-reflectance region (0.0–0.2), whereas the NIR band shows a broader distribution over 0.2 and 0.5. The input LR image follows a broadly similar distribution pattern, although its absolute frequencies are lower because of the reduced pixel count. In contrast, the reference image shows clear distributional differences caused by temporal mismatch; notably, the Red-band peak is shifted and the NIR band exhibits a different shape from that of the ground truth HR. In the SISR results, the CNN- and Swin Transformer-based models produced Blue-band peaks closer to the ground truth HR than the GAN-based SISR model. Across all three architectures, the SISR outputs kept the RGB bands within the correct low-reflectance range, but their distributions were less concentrated than those of the ground truth HR. Likewise, although the NIR band remained within the correct value range, its dispersed shape was not accurately reproduced.

Under Ref-SR, the GAN-based model most closely reproduced the sharp RGB distribution of the ground truth HR, with the Green and Red bands also showing well-aligned peaks. The CNN- and Swin Transformer-based Ref-SR models showed lower RGB peak heights than their SISR counterparts, but they preserved stable distributional proportions. Notably, none of the three Ref-SR backbones inherited the distorted Red- and NIR-band distributions observed in the reference image. Instead, they avoided reproducing the mismatched distributional tendencies of the reference image and preserved a broadly dispersed NIR shape similar to that of the ground truth HR.

4. Conclusion

This study proposed the Dynamic Ref-SR Framework to mitigate reference misuse caused by temporal mismatch in Extreme LR environments. The framework computes a pixel-wise weight map from intensity differences between the LR and reference images. This weight map enables adaptive fusion by selectively using reference textures in stable regions while suppressing unreliable reference transfer in inconsistent regions. The framework is architecture-agnostic and was validated on three backbone architectures: CNN (EDSR), Swin Transformer, and GAN.

Quantitative evaluation on the Sentinel-2 test set showed that the Ref-SR consistently outperformed the SISR baseline across all metrics and backbone architectures. Among the tested backbones, the GAN-based model achieved the best overall performance, with a PSNR of 35.60 dB, an SSIM of 0.92, a SAM of 2.20°, and an ERGAS of 74.71. Per-band RMSE comparisons also showed consistent error reductions across both the RGB and NIR bands, indicating improved spectral accuracy in multi-band restoration. Qualitative analysis further corroborated that Ref-SR restored structural boundaries and textures more clearly than SISR, while its per-band reflectance distributions remained closer to those of the ground truth HR.

These results were achieved without relying on large-scale external models or additional labeled data. By super-resolving Extreme LR imagery with severely degraded structural information, the proposed framework improves the practical utility of satellite observations that offer high temporal frequency but limited spatial detail. Future work will extend this framework to real heterogeneous sensor data, including direct application to GOCI-II, to validate its generalizability and operational usefulness across diverse Extreme LR environments.

Acknowledgements

This work was supported by the Korea Institute of Marine Science & Technology Promotion (KIMST) grant funded by the Ministry of Oceans and Fisheries (RS-2024-00415504), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00242528).

References

Agarwal, A., Jain, A., Deshpande, I., Annkah, A., 2025. Segmentation-aware latent diffusion for satellite image super-resolution: Enabling smallholder farm boundary delineation. *arXiv preprint arXiv:2511.14481*. https://doi.org/10.48550/arXiv.2511.14481.

Dong, R., Yuan, S., Luo, B., Chen, M., Zhang, J., Zhang, L., Li, W., Zheng, J., Fu, H., 2024. Building Bridges across Spatial and Temporal Resolutions: Reference-Based Super-Resolution via Change Priors and Conditional Diffusion Model. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, 7153-7162. doi.org/10.1109/CVPR52733.2024.02614.

Dong, R., Zhang, L., Fu, H., 2021. RRSGAN: Reference-based super-resolution for remote sensing image. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-17. doi.org/10.1109/TGRS.2020.3046045.

Kapilaratne, R.G.C.J., Kakuta, S., Kaneta, S., 2022. Enhanced super resolution for remote sensing imageries. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, V-3-2022, 53-60. doi.org/10.5194/isprs-annals-V-3-2022-53-2022.

Lim, B., Son, S., Kim, H., Nah, S., Lee, K.M., 2017. Enhanced Deep Residual Networks for Single Image Super-Resolution. *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, 1132-1140. doi.org/10.1109/CVPRW.2017.151.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 10012-10022. doi.org/10.1109/ICCV48922.2021.00986.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C.C., Qiao, Y., Tang, X., 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. *Proc. European Conf. Comput. Vis. (ECCV) Workshops*, 0-0. doi.org/10.1007/978-3-030-11021-5_5.