remote sensing

MDPI

Article

Spectral-Consistency-Aware Evaluation of Deep Super-Resolution Methods for UAV Five-Band Multispectral Crop Imagery

Whanjo Jung [1], Seung Hwan Wi [2], Jae-Hyun Ryu [3] and Hoonsoo Lee [1,*]

[1] Department of Biosystems Engineering, College of Agriculture, Life, and Environment Sciences, Chungbuk National University, 1 Chungdae-ro, Seowon-gu, Cheongju 28644, Republic of Korea; hwanzo758@chungbuk.ac.kr
[2] Vegetable Research Division, National Institute of Horticultural & Herbal Science, Wanju-gun 55365, Jeollabuk-do, Republic of Korea
[3] National Agricultural Satellite Center, National Institute of Agricultural Sciences, Rural Development Administration, Jeonju 54874, Republic of Korea
* Correspondence: hslee202@chungbuk.ac.kr; Tel.: +82-43-261-2585

Highlights

What are the main findings?

- At the ×4 scale, joint five-channel reconstruction with its spectral losses best preserved spectral relationships (spectral angle and vegetation-index errors), whereas band-wise reconstruction was strongest for per-band spatial fidelity.
- Spectral-angle constraints improved spectral consistency, whereas vegetation-index loss effects were model-dependent in the ×4 super-resolution setting. All model comparisons were made under a deliberately modest, equalized training budget.

What are the implications of the main findings?

- UAV multispectral super-resolution can enhance spatial detail while maintaining key spectral information needed for vegetation monitoring.
- The proposed benchmark provides practical guidance for selecting SR architectures and loss functions for crop-focused remote sensing applications.

Abstract

Unmanned aerial vehicle (UAV)-based multispectral imaging enables flexible, non-destructive crop monitoring. Although UAV imagery offers much higher spatial resolution than satellite platforms, its effective spatial detail at typical operational flight altitudes can still be insufficient for plant-level interpretation and fine canopy structure, which can reduce vegetation-index reliability. Most super-resolution (SR) research targets RGB or satellite imagery and emphasizes perceptual or pixel-wise quality, leaving the spectral fidelity of reconstructed UAV multispectral imagery under-examined. This study benchmarked an SR evaluation framework for UAV-based five-band crop imagery (Blue, Green, Red, Red-edge, and near-infrared) using the open-source AI Hub cabbage dataset, with low-resolution inputs generated by controlled downsampling at ×2, ×3, and ×4. Nine methods (bicubic, SRCNN, EDSR, RCAN, SwinIR-based, ESRGAN-based, HAT-based, DAT-based, and DRCT-based SR) were compared under joint five-channel and band-wise reconstruction on 2170 test scenes using image-quality, spectral-angle, vegetation-index (NDVI, GNDVI, NDRE), band-wise, and efficiency metrics. EDSR and RCAN gave the most balanced performance. At ×4, band-wise reconstruction was strongest for per-band

Academic Editors: Jaromir Krzyszczak and Clement Atzberger

Received: 29 June 2026
Revised: 12 August 2026
Accepted: 18 August 2026
Published: 19 August 2026

spatial fidelity, where EDSR reduced RMSE by 12.6%, and RCAN lowered near-infrared RMSE by about 21% relative to bicubic, whereas joint reconstruction with its spectral-angle and vegetation-index losses best preserved spectral relationships (spectral angle and vegetation-index errors). Learning-based gains were clearest at ×4. The recently proposed HAT-based, DAT-based, and DRCT-based attention models achieved the strongest pixel-wise RMSE and PSNR but did not surpass EDSR or RCAN on spectral angle or vegetation-index preservation under the equalized training budget. These results indicate that UAV multispectral SR should be assessed by spatial fidelity together with spectral consistency and agricultural index reliability.

Keywords: UAV remote sensing; multispectral imagery; super-resolution; precision agriculture; red-edge; near-infrared; vegetation index; spectral consistency

1. Introduction

Precision agriculture increasingly depends on spatially explicit crop information that can be acquired repeatedly and non-destructively. Remote sensing has become a central tool for field crop monitoring because it supports repeated observation of canopy development, crop stress, and within-field variability at the agricultural field scale [1]. Among remote-sensing platforms, unmanned aerial vehicles (UAVs) are particularly useful for agricultural management because acquisition timing, spatial coverage, and ground sampling distance can be controlled more flexibly than in many satellite-based observations [2]. Recent reviews of UAV multispectral remote sensing have reported the continued expansion of crop-monitoring applications, including growth assessment, stress diagnosis, and decision-support mapping [3].

For crop monitoring, RGB information alone is often insufficient because important physiological and structural changes are expressed outside the visible spectrum. Multispectral imagery provides access to red-edge and near-infrared (NIR) responses that are closely related to canopy structure, chlorophyll-related status, and vegetation-index calculation. However, vegetation-index values derived from aerial multispectral imagery are sensitive to illumination, soil background, acquisition geometry, and crop condition, making spectral reliability an essential requirement for UAV-based analysis [4]. UAV canopy-monitoring studies have also shown that red-edge and NIR measurements provide useful information for estimating canopy traits and nitrogen-related crop status [5]. In addition, flight-altitude studies indicate that acquisition geometry can affect multispectral-index accuracy, reinforcing the need to evaluate both spatial and spectral quality in UAV image products [6].

Despite these advantages, UAV-based multispectral imagery remains constrained by sensor specifications, optics, flight altitude, platform motion, and data-processing conditions [7]. Although UAV imagery is spatially detailed relative to satellite data, its effective resolution at higher flight altitudes or wider coverage can still be insufficient to resolve plant rows, canopy texture, leaf boundaries, mixed pixels, and field-edge transitions. Such blurring reduces the reliability of plant-level interpretation, vegetation-index mapping, and downstream field-management decisions, even when the original multispectral bands are radiometrically meaningful.

Image super-resolution (SR) provides a practical approach for improving spatial detail without changing the acquisition platform. In remote sensing, deep learning has substantially improved high-resolution reconstruction [8]. At the same time, it has raised questions about spectral fidelity, degradation assumptions, and evaluation protocols [9]. SR architectures have advanced through several model families: efficient convolutional networks that target pixel-wise fidelity [10], sub-pixel upsampling methods that improve reconstruction efficiency [11], Transformer-based models that capture long-range feature dependencies [12], residual designs that enable deeper networks with stronger fine-detail recovery [13], and attention-based models that improve the selectivity of feature representation [14]. Other work has emphasized perceptual quality and visual realism under more complex image formation conditions [15], and blind and real-world degradation settings have been surveyed in detail [16]. However, these advances do not automatically guarantee reliable spectral reconstruction for agricultural multispectral analysis.

More recently, diffusion-based SR [17] and remote-sensing foundation models [18] have likewise emerged as promising directions. These approaches can generate perceptually detailed reconstructions and transferable image representations, but their usefulness for quantitative multispectral crop analysis depends on whether reconstructed spectral values remain reliable. For agricultural UAV imagery, visual sharpness alone is not sufficient. Small errors in Red, Red-edge, or NIR reflectance can propagate into larger errors in normalized difference vegetation index (NDVI), green NDVI (GNDVI), or normalized difference red-edge (NDRE) maps. Therefore, SR outputs intended for crop monitoring should be evaluated not only by spatial image-quality metrics but also by spectral consistency and vegetation-index reliability [19].

Although multispectral SR studies have begun to include explicit spectral evaluation, much of the existing work has focused on satellite sensors such as Sentinel-2 [20] or on vegetation mapping using multispectral and hyperspectral imagery [21]. Comparatively less attention has been given to UAV-based five-band crop imagery, where Blue, Green, Red, Red-edge, and NIR bands must be reconstructed in a way that preserves inter-band relationships. This distinction is important because agriculturally meaningful information is not contained in a single band, but in the spectral relationships among visible, red-edge, and NIR responses. However, it remains unclear whether joint five-channel reconstruction better preserves spectral relationships and vegetation-index reliability than independent band-wise reconstruction for UAV five-band crop imagery under a unified evaluation protocol.

Another important issue is the degradation assumption used to construct low-resolution and high-resolution image pairs. In this study, low-resolution inputs were generated by controlled downsampling from dataset-derived high-resolution five-band image cubes rather than by pairing images acquired at different flight altitudes or with different sensors. This design provides a reproducible benchmark under known spatial degradation, while limiting the interpretation of the results to controlled downsampling rather than fully realistic cross-altitude or cross-sensor degradation. Three scale factors, ×2, ×3, and ×4, were evaluated to span moderate to severe spatial-resolution loss.

To address these issues, this study evaluates a spectral-consistency-aware SR framework for UAV-based five-band multispectral crop imagery. Bicubic interpolation is used as a deterministic non-learning baseline, and representative learning-based models are compared, including SRCNN, EDSR, RCAN, SwinIR-based SR, ESRGAN-based SR, and the recently proposed HAT-based, DAT-based, and DRCT-based SR. Two reconstruction strategies are examined: joint five-channel reconstruction, which reconstructs all spectral bands simultaneously, and band-wise reconstruction, which reconstructs each band independently. This comparison tests whether preserving inter-band relationships during reconstruction improves spectral consistency and vegetation-index reliability.

This study aims to establish a reproducible evaluation framework that connects SR reconstruction quality with agricultural spectral usability. First, UAV-based five-band crop imagery is reconstructed and evaluated using both spatial image-quality metrics and spectral-similarity metrics, with spectral angle mapper and vegetation-index errors for NDVI, GNDVI, and NDRE used as agricultural task-oriented proxies for spectral consistency. Second, joint five-channel and band-wise reconstruction strategies are compared under loss formulations appropriate for each reconstruction strategy to examine the role of inter-band learning in multispectral SR. Third, reconstruction behavior is quantified across three scale factors and interpreted in relation to the controlled-downsampling setting used in this study. The main contributions are therefore threefold: first, a spectral-fidelity-oriented evaluation protocol for UAV five-band super-resolution that jointly considers spatial quality, spectral angle, and vegetation-index errors; second, a controlled comparison of joint five-channel and band-wise reconstruction under the loss formulations feasible for each mode; and third, a reproducible, scale-dependent benchmark of representative interpolation, residual, attention, Transformer-inspired, and adversarial SR models on an open UAV multispectral crop dataset. Rather than introducing a new network architecture, the study is positioned as an evaluation framework and benchmark for spectral-consistency-aware UAV multispectral super-resolution. Unlike prior SR benchmarks developed on RGB natural images, evaluated by PSNR and SSIM, or on satellite multispectral imagery, none of which quantified inter-band spectral fidelity for UAV five-band crop imagery, this study transfers the spectral-consistency evaluation lens to the UAV five-band agricultural setting.

2. Materials and Methods

2.1. Open-Source Dataset

The multispectral crop imagery used in this study was obtained from the AI Hub Field Crop Growth Condition dataset, provided by the National Information Society Agency of Korea [22]. The dataset contains UAV-based multispectral crop imagery and record-level metadata that were used to construct a reproducible super-resolution experiment on agricultural field scenes.

The dataset was accessed through AI Hub, and no additional field campaign was conducted in this study. The downloaded files were reorganized into a scene-level structure, where one scene denotes a spatially aligned five-band multispectral crop image cube and its corresponding LR versions. The same scene identifiers were used for training, validation, testing, reconstruction-mode comparison, and scale-factor comparison.

According to the accompanying AI Hub data description file (v1.1), the dataset was constructed in 2022 as an AI training dataset for open-field crop monitoring and includes TIFF source images with JSON annotation files. The metadata were used to describe data provenance and acquisition conditions, whereas the SR models used only the five-band image cubes as input.

The data description file specifies five multispectral bands acquired as TIFF imagery: Blue (475 nm center wavelength, 20 nm bandwidth), Green (560 nm, 20 nm), Red (668 nm, 10 nm), Red-edge (717 nm, 10 nm), and NIR (840 nm, 40 nm). These five bands were retained throughout the experiment because the visible, Red-edge, and NIR bands are used for canopy representation and vegetation-index calculation.

The documentation also defines record-level acquisition metadata fields, including camera information, camera model, acquisition institution, acquisition date and time, place name, wind speed, cloud condition, flight altitude, image overlap, calibration-panel use, longitude, latitude, flight speed, flight direction, processing software, radiometric correction level, geometric correction level, coordinate reference system, and ground sampling distance. For example, the data description file provides an image record acquired on 21 September 2022 in Segyo-ri, Baebang-eup, Asan-si, Chungcheongnam-do, using a MicaSense RedEdge-MX camera (MicaSense Inc., Seattle, WA, USA), with a reported flight altitude of 114 m and ground sampling distance of 8.0 cm/pixel. Because such values are record-level metadata rather than fixed dataset-wide conditions, they were used only to document data provenance and acquisition context when explicitly available and were not used as model input variables.

Acquisition Metadata and Documented Imaging Conditions

The AI Hub data description file provides acquisition metadata at the image-record level. Table 1 summarizes the key acquisition metadata fields available for the cabbage UAV multispectral image records. These metadata were used to document data provenance, acquisition context, and sensor characteristics, but they were not used as input variables for the SR models. Because the public dataset contains images from multiple fields and dates, metadata values such as acquisition date, location, flight altitude, and ground sampling distance should be interpreted as record-specific values rather than fixed conditions for the entire dataset.

Table 1. Key acquisition metadata available for the AI Hub cabbage UAV multispectral image records.

Item	AIHub Field (s)	Example/Value
Acquisition date/time	POTO_DATE; BEGIN_TIME; END_TIME	21 September 2022, 11:20–12:20; 24 August 2022, 11:00–11:25
Location	PLACE_NM	Asan-si, Chungcheongnam-do; Gangneung-si, Gangwon-do
Coordinates	ENVRN_LO; ENVRN_LA	127.094971, 36.763391; 128.795011, 37.566350
Flight altitude	POTO_ATTD	114 m; 170 m
GSD	RSOLTN	8.0 cm/pixel; 7.7 cm/pixel
Camera	CMRA_INFO; CMRA_NM	MicaSense RedEdge-MX; MicaSense Altum (both MicaSense Inc., Seattle, WA, USA)
File format	FILE_FRMAT	TIFF
Weather	WS; AMTCLD	5.4 m/s; partly cloudy or cloudy
Flight operation	FLYING_VE; FLYING_DRC	7–14; west-east or northeast-southwest
Image overlap	H_RDCY; V_RDCY	75%/75%
Calibration panel	CRPSHT	Used
Processing	NM; RADLVL; GEOLVL; CNTM	Pix4Dmapper (Pix4D SA, Lausanne, Switzerland; version not specified in the public dataset documentation); reflectance conversion; GNSS/GCP; EPSG:5179
Crop annotation	CTVT_CROPS_SPCHCKN; GRWH_STEP_CODE	Cabbage; leaf expansion or heading stage

Values are examples from the AI Hub data description file and inspected JSON records; they represent record-level metadata, not fixed conditions for the full dataset.

Figure 1 summarizes the overall methodological workflow used in this study. The workflow begins with access-restricted AI Hub UAV multispectral imagery, constructs dataset-derived HR five-band cubes, generates controlled LR cubes for ×2, ×3, and ×4 scale factors, reconstructs SR outputs using joint five-channel and band-wise strategies, and evaluates the outputs using spatial, spectral, vegetation-index, band-wise, and computational metrics.

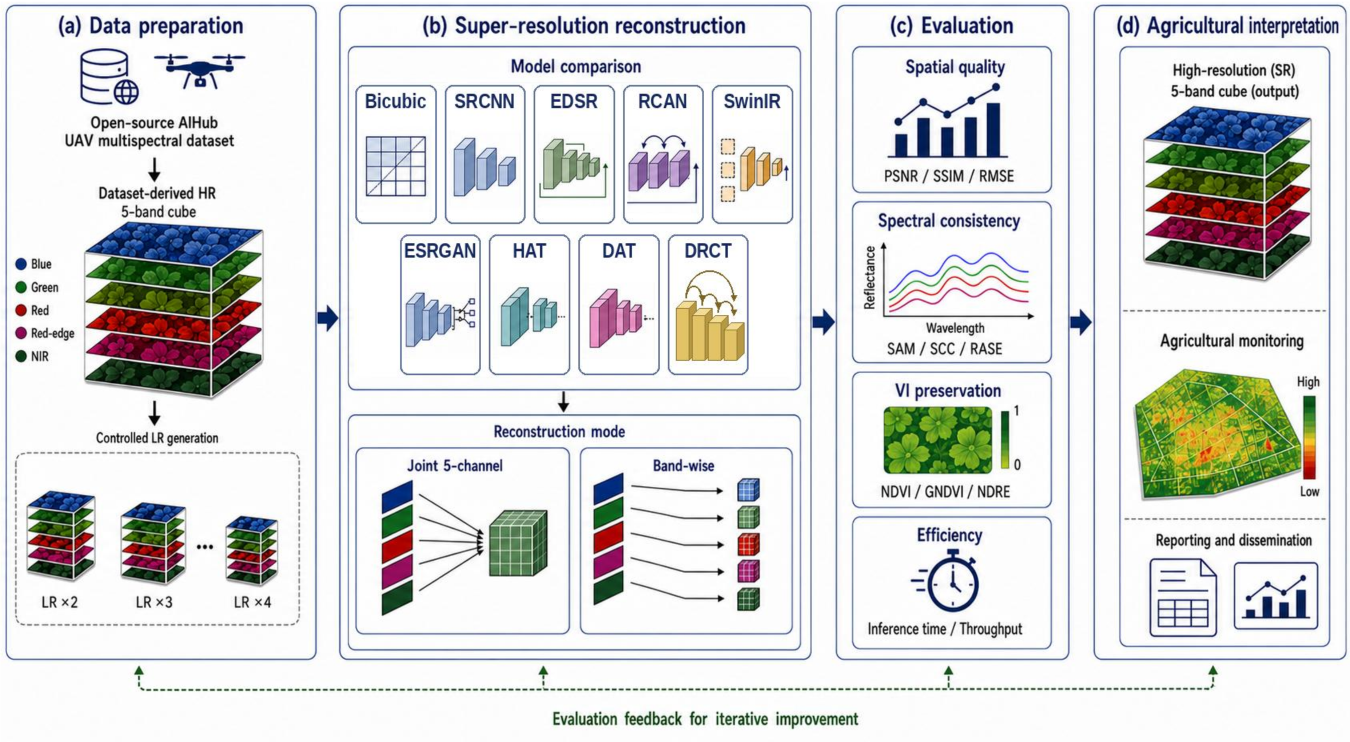

Figure 1. Overall workflow of the AI Hub-based five-band UAV multispectral super-resolution experiment. Starting from dataset-derived HR cubes and controlled LR inputs at ×2, ×3, and ×4, nine SR methods are compared under joint five-channel and band-wise reconstruction and evaluated for spatial quality, spectral consistency, vegetation-index preservation, and computational efficiency. The color legend in panel (**a**) identifies the five spectral bands (Blue, Green, Red, Red-edge, and NIR); solid arrows indicate the main processing sequence, the dashed box marks the controlled low-resolution generation step, and the green dashed arrows denote the evaluation feedback used for iterative refinement.

2.2. Multispectral Image Representation

Each multispectral image was represented as a five-channel image cube consisting of Blue, Green, Red, Red-edge, and near-infrared (NIR) bands. These bands were used because they jointly represent visible canopy information and crop-related spectral responses. The visible bands provide information on canopy color and surface structure, whereas Red-edge and NIR bands are important for vegetation-index calculation and crop vigor assessment.

The HR and LR multispectral image cubes were organized as five-channel arrays, where H and W denote HR spatial dimensions, C is the number of spectral channels, and s is the SR scale factor. In this study, C was fixed to five and s was evaluated at ×2, ×3, and ×4. LR cubes were generated using area-based downsampling with integer LR dimensions floor (H/s) × floor (W/s). Therefore, the 800 × 800 HR scenes produced exact LR dimensions of 400 × 400 for ×2 and 200 × 200 for ×4. For ×3, however, integer downsampling produced a 266 × 266 LR cube, and the natural ×3 SR grid was therefore 798 × 798 rather than 800 × 800. To avoid forced 798-to-800 interpolation during metric calculation, the HR cube and its validity mask were center-cropped to the natural SR grid before ×3 quality checking, training-patch alignment, and final evaluation. The same center crop was applied to all five bands and all reconstruction methods, so ×3 metrics were computed from spatially matched HR-SR pairs. Because two border pixels were excluded in this process, ×3 was reported as a diagnostic condition rather than as a fully equivalent exact-scale condition.

The band order used throughout the experiment was Blue, Green, Red, Red-edge, and NIR. The same order was used for model input, model output, vegetation-index calculation, band-wise evaluation, and visualization.

The super-resolution task was formulated as learning a mapping function from LR multispectral imagery to SR multispectral imagery, as shown in Equation (1). The reconstructed SR cube was compared with the HR cube in both image space and spectral-index space.

$$\hat{X}\_SR\hat{}(s) = f\_\theta(X\_LR\hat{}(s)) \tag{1}$$

where X_hat_SR^(s) is the reconstructed SR cube, f_theta is the super-resolution mapping function, and theta denotes the learnable model parameters.

2.3. Data Organization and File Structure

The downloaded TIFF bands were reorganized into NumPy array files for efficient deep-learning training and evaluation. Each image cube was stored in height-by-width-by-channel format using the fixed band order Blue, Green, Red, Red-edge, and NIR, and HR cubes were stored separately from controlled LR cubes. The five-band cubes were stored as pre-normalized reflectance-scale arrays in the [0, 1] range; consequently, no additional per-patch normalization was applied during training or evaluation, and a 99.5th-percentile stretch was used only for preview-image visualization.

A scene index file was used to manage the relationship between each HR image and its corresponding LR images. For each scene, the index file stored the scene identifier, HR file path, and LR file paths for ×2, ×3, and ×4. This structure ensured that all reconstruction methods were evaluated on identical HR-LR scene pairs within each scale factor.

Controlled LR generation was performed independently for each spectral band using area-based downsampling. For a scale factor s, the LR size was computed using integer division of the HR height and width by s. This procedure provides a reproducible degradation setting.

To test the robustness of the conclusions to the degradation assumption, an additional degradation setting was evaluated at ×4 in which low-resolution inputs were generated by Gaussian blurring followed by bicubic downsampling, as an alternative to area-based downsampling. The two most balanced CNN models, EDSR and RCAN, together with the Transformer representative HAT, were retrained and evaluated under this blur-plus-bicubic degradation in both joint and band-wise modes, and their performance and spectral-consistency trends were compared with the area-downsampling results.

2.4. Data Validation and Valid-Pixel Masking

Before training, all scene records were screened using deterministic data-integrity checks. A scene was retained only when the HR and LR files existed, the image cube had the expected three-dimensional five-channel structure, and the array did not contain NaN or Inf values. Records failing any of these file-existence, dimensionality, or finite-value checks were excluded before model training and evaluation.

Valid-pixel masking was used so that loss calculation and metric evaluation focused on agricultural image regions rather than background or invalid image areas. Pixels identified as nodata or background were removed from loss calculation and metric evaluation using valid-pixel masks. The same masking procedure was applied consistently to all spectral bands, reconstruction methods, scale factors, and evaluation metrics. This design was used to prevent background or invalid regions from dominating the objective function and final quality scores.

For any pixel-wise error term e_i, masked averaging was performed using Equation (2). The small constant epsilon was added to avoid division by zero when a patch contained very few valid pixels.

$$L\_mask(e) = (\sum\_(i=1)^N\ M\_i\ e\_i) \ / \ (\sum\_(i=1)^N\ M\_i + \varepsilon) \quad (2)$$

where L_mask(e) is the masked mean of the pixel-wise error e_i, M_i is the validity mask at pixel i, N is the number of evaluated pixels, and epsilon = 1×10^{-8} is a small constant for numerical stability.

2.5. Data Split and Patch Sampling

After invalid scenes were excluded, the remaining 14,463 valid scenes were randomly divided into training, validation, and test subsets using an approximately 70:15:15 ratio, yielding 10,124 training, 2169 validation, and 2170 test scenes. The training subset was used for parameter optimization, the validation subset was used for checkpoint selection, and the test subset was used only for final evaluation. A fixed random seed was used to improve reproducibility. Because the split was performed at the scene-record level rather than by field-, date-, or flight-level grouping, the evaluation should be interpreted as an internal scene-level benchmark under controlled degradation. The public dataset provides acquisition metadata at the individual image-record level without consistent field- or flight-level identifiers, so a strict acquisition-grouped split could not be constructed reliably; consequently, some train and test scenes may share acquisition-level similarity, and grouped external validation is left for future work.

Patch-based training was adopted to increase the number of training samples and reduce GPU memory usage. During each training iteration, a scene was randomly selected from the training set, and a spatially aligned LR-HR patch pair was extracted. An HR patch size of 192 × 192 pixels was used (selected dynamically from candidate sizes between 24 and 192 pixels to remain compatible with each scale factor), with the corresponding LR patch size of 192/s.

Candidate patches were checked using the HR validity mask, finite-value criteria, and patch-quality filters before being accepted for training. Patches were rejected and resampled when they contained insufficient valid pixels, very low intensity variation, very low dynamic range, or excessive LR-upsampled-to-HR mismatch. Specifically, the patch-level filters required a patch valid-pixel ratio of at least 50% for both the HR and LR patches, an HR validity-mask coverage of at least 10%, a minimum patch standard deviation of 0.003, a minimum 1st-to-99th percentile dynamic range of 0.010, and an LR-upsampled-to-HR RMSE not exceeding 0.20. These deterministic filters were applied before model optimization so that all methods were trained on valid and spatially comparable LR-HR patch pairs; consequently, the reported performance represents the quality-controlled valid-image subset rather than all possible low-information or poorly aligned image regions. These threshold values were established empirically from preliminary inspection of the AI Hub cabbage scenes: the valid-pixel and standard-deviation limits remove near-empty or texture-less patches such as uniform soil or background, the dynamic-range limit removes low-information patches, and the LR-upsampled-to-HR RMSE limit removes spatially misaligned pairs; the same fixed thresholds were applied to all methods.

2.6. Reconstruction Strategies

Two reconstruction strategies were compared to evaluate the effect of inter-band learning. In joint five-channel reconstruction, the model received all five bands simultaneously and produced a five-channel SR cube in a single forward pass. This approach allows the network to learn spatial features together with spectral relationships among

Blue, Green, Red, Red-edge, and NIR bands. The joint reconstruction strategy is defined in Equation (3).

$$\hat{X} = f\_\theta(X\_Blue, X\_Green, X\_Red, X\_Red\text{-}edge, X\_NIR) \quad (3)$$

where the joint five-channel model receives all five spectral bands simultaneously and reconstructs the multispectral SR cube in a single network pass.

In band-wise reconstruction, each spectral band was reconstructed independently using a single-channel model. The reconstructed bands were then concatenated in the original spectral order, as shown in Equation (4). This strategy was used as an ablation experiment to determine whether independent band reconstruction is sufficient or whether joint five-channel learning better preserves agricultural spectral relationships.

$$\hat{X} = concat\_(c \in B)\{\ f\_(\theta\_c)(X\hat{}c)\ \} \quad (4)$$

where concat denotes channel concatenation after independent band-wise reconstruction, X^c is the c-th spectral band, and f_theta_c is the band-specific SR model.

2.7. Super-Resolution Methods

The manuscript-level comparison included bicubic interpolation, SRCNN [23], EDSR [24], RCAN [25], SwinIR-based SR [12], and ESRGAN-based SR [26]. Bicubic interpolation was used as a deterministic non-learning baseline. The five learning-based methods were selected to represent shallow convolutional reconstruction, residual convolutional reconstruction, channel-attention reconstruction, Transformer-inspired reconstruction, and adversarial reconstruction, respectively.

Bicubic interpolation estimates each high-resolution pixel from a 4 × 4 neighborhood of low-resolution pixels using a piecewise cubic convolution kernel. It contains no trainable parameters and was included as a reference for smooth geometric upsampling without learned crop-specific reconstruction.

SRCNN [23] was used as the shallow convolutional baseline. It operates on a bicubic-upsampled low-resolution image and applies three convolutional stages corresponding to patch feature extraction, non-linear feature mapping, and final reconstruction. This compact three-layer design provides a reference model for evaluating the effect of deeper network architectures.

EDSR [24] represented residual CNN super-resolution. It stacks residual blocks without batch-normalization layers and performs upscaling through sub-pixel convolution near the output [11]. EDSR was included as a residual reconstruction model optimized primarily with reconstruction-oriented objectives.

RCAN [25] extended residual reconstruction with a residual-in-residual structure and residual channel-attention blocks. The long and short skip connections allow low-frequency information to bypass deeper transformations, while the channel-attention mechanism re-weights feature channels during reconstruction. This mechanism is relevant for multispectral inputs, where visible, Red-edge, and NIR bands contribute differently to vegetation-index calculation.

SwinIR-based SR [12] represented a Transformer-inspired image-restoration baseline. The implemented model combined a shallow convolutional feature extractor, multi-head self-attention blocks, and a final reconstruction module; therefore, it is referred to as SwinIR-based rather than a full reproduction of the original shifted-window SwinIR architecture. A smaller inference tile size was used for this model during full-image reconstruction because self-attention has higher computational cost for large spatial tokens.

ESRGAN-based SR [26] represented adversarially guided super-resolution. Its generator was built from residual-in-residual dense blocks without batch normalization and adapted to five-band reflectance reconstruction. Generator pre-training used reflectance-space reconstruction with spectral and vegetation-index terms, and an adversarial objective was added after the pixel-wise pre-training stage. ESRGAN-based SR was included to examine how adversarial reconstruction behaves under spectral and vegetation-index evaluation.

HAT-based SR [27] and DAT-based SR [28] were additionally included as recent (2023) attention-based state-of-the-art baselines. HAT combines channel attention with window-based self-attention within a hybrid attention block, whereas DAT alternates spatial self-attention and transposed channel self-attention in a dual-aggregation design. Both were implemented as compact reimplementations configured for five-band input and trained under the same budget as the other learning-based models; they are therefore referred to as HAT-based and DAT-based rather than full reproductions of the original architectures. To keep attention cost bounded, both used a smaller inference tile, as with the SwinIR-based model.

DRCT-based SR [29] was additionally included as a recent (2024) state-of-the-art baseline. It starts from the same Swin-Transformer lineage as the SwinIR-based and HAT-based models but targets a different limitation: the intensity of the feature maps is abruptly suppressed toward the end of a deep sequential stack, which acts as an information bottleneck and discards spatial detail. DRCT removes that bottleneck by connecting the layers of each group densely, so that every layer receives the group input together with all preceding layer outputs, reduced to a narrow working width by a 1 × 1 convolution; this is the residual-dense idea of RRDB applied to attention layers rather than to convolutions. As with the other recent baselines, it was implemented as a compact reimplementation configured for five-band input and trained under the same budget as every other learning-based model, and is therefore referred to as DRCT-based rather than a full reproduction of the original architecture. The shifted-window attention of the original was replaced by the same global attention used by the SwinIR-based and HAT-based models so that the three share one attention implementation, and the model was evaluated on the same reduced inference tile as the other attention-based baselines.

In addition, to separate the effect of model capacity from that of the reconstruction strategy, a widened joint five-channel EDSR variant, denoted EDSR-wide, was trained at ×4 with its total trainable-parameter count matched to that of the five-model band-wise EDSR ensemble (Table 2); comparing EDSR-wide with the standard joint EDSR and the band-wise EDSR indicates how much of the band-wise spatial-accuracy advantage stems from parameter scale rather than independent reconstruction.

Model adaptation. For all learning-based methods, the input channel dimension was set to five in joint five-channel reconstruction and to one in band-wise reconstruction, and the networks were trained from scratch using the loss formulation defined for each reconstruction setting (Section 2.9). The residual-learning methods reconstructed high-frequency detail on top of a bicubic-upsampled base image so that the networks estimated residual spatial detail rather than the entire high-resolution signal.

All nine methods were evaluated at ×2, ×3, and ×4 scale factors whenever corresponding outputs and checkpoints were available. Bicubic interpolation was evaluated directly without training, whereas the learning-based methods were trained or loaded from saved checkpoints and then evaluated under the same reconstruction-mode and scale-factor settings.

Detailed architecture configurations and trainable parameter counts are summarized in Table 2. The learning-based models were implemented as compact adaptations of their original architectures and configured for five-band input. Identical optimization settings were used for all models (Section 2.10). For band-wise reconstruction, the reported trainable parameter count corresponds to the combined total of five independent single-channel models, one for each spectral band. Because all learning-based models were trained from scratch under an identical and deliberately modest budget (Section 2.10) with compact configurations, the results characterize relative behavior under a common controlled setting rather than the maximum attainable performance of each model family. In particular, the Transformer-inspired and adversarial baselines are compact reimplementations that may benefit more from larger training budgets, pre-training, or architecture-specific tuning; their results should therefore be interpreted as baselines under matched conditions rather than as upper bounds for those approaches.

Table 2. Architecture configuration and trainable parameter counts of the compared models.

Method	Architecture Summary	Joint (5-Channel) Trainable Parameters	Band-Wise Trainable Parameters (5 × 1-Channel)
Bicubic	Deterministic 4 × 4 cubic interpolation (non-learning)	0	0
SRCNN	Three-layer shallow CNN on bicubic-upsampled input	81,221	286,405
EDSR	16 residual blocks, 96 features, sub-pixel upsampling, no batch norm	3,081,221	15,371,525
RCAN	6 residual groups × 6 residual channel-attention blocks, 96 features	6,946,205	34,696,445
SwinIR-based SR	6 Transformer-inspired attention blocks, embed dim 96, 4 attention heads	872,645	4,328,645
ESRGAN-based SR	6 residual-in-residual dense blocks, 64 features, adversarial fine-tuning	4,543,941	22,696,645
HAT-based SR	Compact hybrid attention transformer (channel + window self-attention)	4,962,149	24,776,165
DAT-based SR	Compact dual-aggregation transformer (spatial + channel self-attention)	2,926,917	14,600,005
DRCT-based SR	Compact dense-residual-connected transformer: attention layers joined by dense connections within each group	2,930,693	14,618,885

Parameter counts are trainable parameters as logged during training. Band-wise counts reflect five separate single-channel networks, one per spectral band.

2.8. Vegetation Indices

Because the reconstructed images are intended for agricultural analysis, vegetation-index accuracy was explicitly considered. Three indices were calculated from the HR and SR image cubes: NDVI, GNDVI, and NDRE. NDVI uses Red and NIR bands, GNDVI uses Green and NIR bands, and NDRE uses Red-edge and NIR bands. These indices were treated as agricultural task-oriented proxy metrics because they directly test whether reconstructed spectral relationships remain useful for crop monitoring. All three indices are reported quantitatively for both evaluated reconstruction settings in Section 3.4.

NDVI was included because it is a foundational vegetation index for monitoring vegetation systems [30].

GNDVI was included because green-channel reflectance may support vegetation monitoring [31].

NDRE was included because red-edge information is useful for crop water stress, nitrogen status, and canopy-density-related assessment [32]. Comprehensive reviews summarize the development and ecological applications of optical vegetation indices [33].

https://doi.org/10.3390/rs18162811

$$NDVI = (NIR - Red) / (NIR + Red + \varepsilon) \tag{5a}$$

where NIR and Red denote the near-infrared and red reflectance values, respectively, and epsilon prevents division by zero.

$$GNDVI = (NIR - Green) / (NIR + Green + \varepsilon) \tag{5b}$$

where NIR and Green denote the near-infrared and green reflectance values, respectively, and epsilon prevents division by zero.

$$NDRE = (NIR - Red\text{-}edge) / (NIR + Red\text{-}edge + \varepsilon) \tag{5c}$$

where NIR and Red-edge denote the near-infrared and red-edge reflectance values, respectively, and epsilon prevents division by zero.

The inclusion of vegetation-index terms was intended to prevent the model from optimizing only pixel-wise reflectance similarity. In multispectral crop imagery, small spectral errors in Red-edge or NIR bands may cause large index errors, even when the reconstructed image appears visually sharp. Reconstructed reflectance was bounded to a valid range by the model-output activation prior to index computation, and vegetation-index errors are additionally reported with reflectance clipped to [0, 1] to confirm that any out-of-range values do not distort the index statistics.

2.9. Loss Function

The standard joint five-channel loss function was designed to jointly optimize radiometric accuracy, spatial detail, and spectral consistency. For joint five-channel reconstruction, the total loss combined pixel-wise reconstruction, robust reconstruction, structural similarity, gradient detail, Laplacian edge, frequency-domain, SAM, and vegetation-index terms, as defined in Equation (6).

$$L\_total = \lambda\_1\ L\_1 + \lambda\_char\ L\_char + \lambda\_SSIM\ L\_SSIM + \lambda\_g\ L\_grad + \lambda\_e\ L\_lap + \lambda\_f\ L\_freq + \lambda\_SAM\ L\_SAM + \lambda\_VI\ L\_VI \tag{6}$$

where L_total is the total training objective, L_1, L_char, L_SSIM, L_grad, L_lap, L_freq, L_SAM, and L_VI are the reconstruction, spatial-detail, spectral-consistency, and vegetation-index loss terms, and each lambda denotes the corresponding loss weight.

The loss weights were set to lambda_1 = 1.00, lambda_char = 0.50, lambda_SSIM = 0.20, lambda_g = 0.15, lambda_e = 0.08, lambda_f = 0.03, lambda_SAM = 0.10, and lambda_VI = 0.10 for the standard five-channel SR loss. The masked L1 term was used as the primary reconstruction constraint, and the Charbonnier term complemented it as a robust penalty with reduced sensitivity to residual invalid pixels or extreme reflectance values. The remaining terms were used to constrain edge detail, structural similarity, frequency-domain content, spectral angular consistency, and vegetation-index preservation. These weights were selected from preliminary validation-set tuning so that the auxiliary terms regularize the solution without overriding radiometric accuracy; the empirical influence of the spectral-angle and vegetation-index weights is examined in the loss-term ablation (Section 3.6).

The gradient loss in Equation (7) was used to penalize differences in local spatial transitions such as leaf boundaries, canopy edges, and row patterns. The Laplacian edge and frequency-domain losses in Equation (8) were used to constrain second-order edge responses and high-frequency reconstruction.

Laplacian pyramid super-resolution has previously shown that multi-scale edge/detail constraints may improve reconstruction of high-resolution structure [34].

$$L\_grad = ||\nabla\_x\ \hat{X} - \nabla\_x\ X||\_1 + ||\nabla\_y\ \hat{X} - \nabla\_y\ X||\_1 \tag{7}$$

where nabla_x and nabla_y are the horizontal and vertical first-order gradient operators, respectively, and ||.||_1 denotes the L1 norm.

https://doi.org/10.3390/rs18162811

$$L\_lap = ||\Delta\hat{X} - \Delta X||\_1; \quad L\_freq = ||log(1 + |F(\hat{X})|) - log(1 + |F(X)|)||\_1 \tag{8}$$

where Delta denotes the Laplacian operator, F(.) denotes the Fourier transform, and the two terms penalize edge-response and frequency-domain differences between SR and HR images.

The SAM loss in Equation (9) was used to constrain the angular relationship between HR and SR spectral vectors. Unlike pixel-wise losses, SAM measures spectral direction and was used to evaluate whether the relative relationship among bands was maintained; this follows the use of spectral angle mapping as a spectral-similarity measure in remote sensing [35]. Unlike the radiometric-consistency losses adopted in some satellite super-resolution studies [19], which primarily constrain the absolute reflectance magnitude of the reconstruction, the spectral-angle term used here constrains the angular relationship among the five bands, so that the inter-band ratios underlying NDVI, GNDVI, and NDRE are preserved even when absolute magnitudes shift slightly.

$$L\_SAM = (1/N) \sum\_(i=1)\text{^}N\ arccos((\hat{x}\_i \cdot x\_i) / (||\hat{x}\_i||\_2\ ||x\_i||\_2 + \varepsilon)) \tag{9}$$

where x_hat_i and x_i are the predicted and reference spectral vectors at valid pixel i, the dot operator denotes the inner product, ||.||_2 denotes the L2 norm, and epsilon prevents numerical instability.

The vegetation-index loss in Equation (10) was calculated as the average index reconstruction error of NDVI, GNDVI, and NDRE. This term was intended to constrain agriculturally meaningful spectral relationships; its empirical effect is examined in the loss-term ablation (Section 3.6) using the vegetation-index formulations described in Section 2.8.

$$L\_VI = (1/|V|) \sum\_(v \in V)\ ||v(\hat{X}) - v(X)||\_1, V = \{NDVI, GNDVI, NDRE\} \tag{10}$$

where V is the set of vegetation-index functions used in the loss, v(.) denotes one selected vegetation-index operator, and ||.||_1 denotes the L1 error over valid pixels.

To isolate the effects of spectral and vegetation-index constraints, a loss-term ablation was conducted for the joint five-channel learning-based models at the ×4 scale. Three reduced loss profiles were evaluated: L1, which used only the masked pixel-wise reconstruction term; L1 + SAM, which added the spectral-angle constraint to the L1 term; and L1 + SAM + VI, which further added the vegetation-index constraint based on NDVI, GNDVI, and NDRE. The ablation was restricted to joint five-channel reconstruction because SAM and vegetation-index losses require multi-band spectral vectors. The results of this analysis are reported in Section 3.6.

For single-channel band-wise models, the SAM and vegetation-index terms were omitted because they require multi-band spectral relationships that are not available within an independent single-band forward pass. Band-wise models were therefore trained under their feasible single-channel loss setting, using the reconstruction and spatial-detail terms only, namely the L1, Charbonnier, SSIM, gradient, Laplacian, and frequency-domain losses. Consequently, the joint and band-wise comparison should be interpreted as a practical comparison between reconstruction strategies under feasible loss formulations, rather than as a complete isolation of reconstruction mode from loss design.

To isolate the effect of the reconstruction mode from that of the loss formulation, a joint five-channel model was additionally trained at ×4 using only the spatial and detail loss terms available to the single-channel band-wise models (masked L1, Charbonnier, SSIM, gradient, Laplacian, and frequency losses), with the SAM and vegetation-index terms removed. Comparing this spatial-only joint model with the full-loss joint model and with the band-wise model separates the contribution of the spectral (SAM/VI) loss terms from that of joint reconstruction itself.

https://doi.org/10.3390/rs18162811

2.10. Model Training Configuration

All learning-based models were implemented in PyTorch 2.7 (Linux Foundation, San Francisco, CA, USA) and trained from scratch on a single NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition (96 GB GDDR6; NVIDIA Corporation, Santa Clara, CA, USA) using automatic mixed precision [36]. The AdamW optimizer [37], a decoupled-weight-decay variant of Adam [38], was used with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-6}, and a cosine annealing learning-rate schedule [39] was applied over the training horizon. Models were trained for up to 50 epochs with 1000 randomly sampled patches per epoch and a batch size of 16, and early stopping (patience of 10 epochs, minimum improvement of 1×10^{-5} in validation loss) was used to terminate training once the validation loss stopped improving. Gradient clipping with a maximum norm of 1.0 and a fixed random seed of 42 were applied for numerical stability and reproducibility. For each model, the checkpoint with the lowest validation loss was retained and reloaded for final test evaluation. For the ESRGAN-based model, adversarial training was enabled after a 20-epoch pixel-wise pre-training stage with an adversarial loss weight of 1×10^{-3}.

2.11. Full-Image Inference

During testing, each LR image was reconstructed as a full-size SR image. Because full-scene images may be larger than the training patches, tile-based inference was used. LR tiles of 192 × 192 pixels were extracted with a 12-pixel overlap, passed through the corresponding reconstruction model, and accumulated into the output canvas using overlap averaging. A smaller tile of 64 × 64 pixels with an 8-pixel overlap was used for the SwinIR-based model because self-attention has a higher computational cost for large spatial tokens.

This averaging strategy reduces discontinuities at tile boundaries and allows the trained patch-based model to be applied to full UAV scenes. If the reconstructed output size differed from the HR target size because of scale-factor dimension mismatch, the SR output was aligned to the HR grid before metric calculation.

2.12. Evaluation Metrics

The reconstructed images were evaluated using image-quality, spectral, vegetation-index, band-wise, and computational metrics. General image-quality performance was measured using RMSE, MAE, PSNR, and SSIM. PSNR was computed per scene using a scene-dependent data range (the 99.5th percentile of the valid high-resolution reflectance in each scene) and then averaged across the test scenes. SSIM was included because it measures luminance, contrast, and structural similarity rather than only pixel-wise error [40]. Because this scene-dependent data range departs from the fixed peak value used in most SR studies, a standard PSNR computed over the full [0, 1] reflectance range with a peak value of 1.0 is also reported alongside the scene-dependent PSNR in Table 3; across the test set, it is systematically about 2.2 dB higher than the scene-dependent PSNR, and this standard value is used when comparing with prior work.

Table 3. Core reconstruction performance of the nine methods. Values are mean values over n = 2170 test scenes per scale-mode combination. Metrics: RMSE (reflectance-scale, unitless), PSNR (dB, scene-dependent), PSNR (std) (dB, full [0, 1] range with peak 1.0), SSIM (0–1), SAM (degrees), Mean VI MAE (unitless). Lower is better for RMSE, SAM, and Mean VI MAE; higher is better for PSNR and SSIM. Within each scale and reconstruction mode, the row of the method holding the largest number of best values is lightly shaded; ties are resolved by RMSE.

Scale	Mode	Method	RMSE	PSNR	PSNR (std)	SSIM	SAM	Mean VI MAE

×2	Joint 5-channel	Bicubic	0.0105	38.17	40.35	0.948	0.656	0.0083
		SRCNN	0.0111	38.38	40.56	0.959	0.635	0.0079
		EDSR	0.0109	38.68	40.86	0.961	0.614	0.0076
		RCAN	0.0115	38.56	40.74	0.962	0.607	0.0075
		SwinIR-based SR	0.0107	38.58	40.76	0.960	0.626	0.0077
		ESRGAN-based SR	0.0128	37.03	39.21	0.939	1.014	0.0121
		HAT-based SR	0.0104	39.19	41.37	0.961	0.623	0.0078
		DAT-based SR	0.0105	39.18	41.36	0.962	0.622	0.0077
		DRCT-based SR	0.0102	39.26	41.43	0.962	0.617	0.0077
	Band-wise	Bicubic	0.0105	38.17	40.35	0.948	0.656	0.0083
		SRCNN	0.0099	39.13	41.31	0.962	0.599	0.0073
		EDSR	0.0098	39.22	41.40	0.963	0.594	0.0072
		RCAN	0.0099	39.23	41.41	0.963	0.595	0.0072
		SwinIR-based SR	0.0096	39.19	41.37	0.961	0.602	0.0074
		ESRGAN-based SR	0.0105	38.69	40.87	0.960	0.665	0.0080
		HAT-based SR	0.0098	39.53	41.71	0.963	0.605	0.0076
		DAT-based SR	0.0097	39.58	41.76	0.963	0.604	0.0076
		DRCT-based SR	0.0096	39.62	41.79	0.963	0.601	0.0075
×3	Joint 5-channel	Bicubic	0.0250	30.51	32.69	0.789	1.339	0.0176
		SRCNN	0.0256	30.25	32.43	0.817	1.320	0.0172
		EDSR	0.0367	27.07	29.25	0.692	1.794	0.0243
		RCAN	0.0372	26.97	29.15	0.692	1.831	0.0245
		SwinIR-based SR	0.0360	27.23	29.41	0.695	1.774	0.0241
		ESRGAN-based SR	0.0385	26.70	28.88	0.674	1.985	0.0260
		HAT-based SR	0.0254	30.58	32.76	0.823	1.326	0.0175
		DAT-based SR	0.0254	30.59	32.77	0.823	1.325	0.0175
		DRCT-based SR	0.0254	30.58	32.75	0.823	1.319	0.0174
	Band-wise	Bicubic	0.0250	30.51	32.69	0.789	1.339	0.0176
		SRCNN	0.0246	30.52	32.70	0.820	1.315	0.0171
		EDSR	0.0360	27.22	29.40	0.693	1.858	0.0246
		RCAN	0.0364	27.14	29.32	0.695	1.859	0.0245
		SwinIR-based SR	0.0361	27.18	29.36	0.693	1.850	0.0246
		ESRGAN-based SR	0.0365	27.10	29.28	0.690	1.895	0.0250
		HAT-based SR	0.0247	30.80	32.98	0.825	1.317	0.0174
		DAT-based SR	0.0250	30.69	32.87	0.825	1.321	0.0174
		DRCT-based SR	0.0249	30.74	32.92	0.824	1.318	0.0174
×4	Joint 5-channel	Bicubic	0.0278	29.60	31.78	0.719	1.455	0.0192
		SRCNN	0.0266	29.93	32.11	0.768	1.375	0.0178
		EDSR	0.0255	30.43	32.61	0.785	1.334	0.0172
		RCAN	0.0257	30.39	32.57	0.786	1.344	0.0173
		SwinIR-based SR	0.0273	29.76	31.94	0.766	1.399	0.0182
		ESRGAN-based SR	0.0301	29.22	31.40	0.744	1.670	0.0206
		HAT-based SR	0.0269	30.22	32.40	0.774	1.408	0.0185
		DAT-based SR	0.0271	30.17	32.35	0.774	1.397	0.0183
		DRCT-based SR	0.0265	30.28	32.45	0.775	1.394	0.0183
	Band-wise	Bicubic	0.0278	29.60	31.78	0.719	1.455	0.0192
		SRCNN	0.0262	30.06	32.24	0.767	1.391	0.0180
		EDSR	0.0243	30.77	32.95	0.782	1.400	0.0179
		RCAN	0.0248	30.67	32.85	0.784	1.399	0.0177
		SwinIR-based SR	0.0244	30.66	32.84	0.772	1.474	0.0190
		ESRGAN-based SR	0.0264	30.16	32.34	0.767	1.463	0.0186

HAT-based SR	0.0239	31.15	33.33	0.782	1.490	0.0193
DAT-based SR	0.0240	31.11	33.29	0.781	1.482	0.0193
DRCT-based SR	0.0244	30.98	33.16	0.780	1.473	0.0192

Spectral consistency was evaluated using SAM in degrees, following spectral angle mapping used for spectral similarity assessment in remote sensing [35]. Lower SAM values indicate smaller angular differences between HR and SR spectral vectors across the five bands. The SAM metric is reported in degrees, whereas the SAM loss term in Equation (9) is computed on the native angular quantity in radians.

Vegetation-index errors were calculated as the MAE between HR and SR index maps, as defined in Equation (11). This metric was calculated separately for NDVI, GNDVI, and NDRE to evaluate whether the SR outputs preserved the vegetation-index information defined in Section 2.8.

$$E\_VI = (1/N)\ \sum\_(i=1)\text{^}N\ |VI(\hat{X})\_i\ -\ VI(X)\_i|, VI\ \in\ \{NDVI, GNDVI, NDRE\} \quad (11)$$

where E_VI is the vegetation-index error, VI(.) denotes NDVI, GNDVI, or NDRE, and N is the number of evaluated valid pixels in the vegetation-index map.

Band-wise RMSE, MAE, and PSNR were computed separately for Blue, Green, Red, Red-edge, and NIR bands to quantify band-specific reconstruction behavior. Computational efficiency was evaluated using wall-clock SR reconstruction time, total evaluation time, and throughput. All methods were evaluated on the same n = 2170 test scenes per scale-mode combination, so method differences were assessed as scene-level paired comparisons within the same evaluation set.

To assess noise robustness, zero-mean Gaussian noise with standard deviation σ = 0.01 and 0.03 in reflectance units was added to the ×4 low-resolution test inputs at inference time, without retraining. EDSR, RCAN, ESRGAN-based SR, and HAT-based SR were evaluated under these noisy inputs, and the resulting degradation in RMSE, SAM, and mean vegetation-index MAE was used to compare the noise robustness of convolutional, adversarial, and Transformer-based reconstruction.

2.13. Statistical Analysis

Paired Wilcoxon signed-rank tests were conducted on the 2170 test scenes. For the primary comparisons of interest, particularly the ×4 scale and agricultural spectral metrics, exact p-values, 95% confidence intervals, and rank-biserial correlation effect sizes were reported. All tests used a two-tailed alpha level of 0.05. Because the full comparison space included multiple methods, scale factors, reconstruction modes, and metrics, the statistical results were interpreted together with effect magnitude and metric-specific relevance rather than by significance thresholds alone.

3. Results

3.1. Evaluation Set

The final evaluation comprised nine reconstruction methods, three scale factors, and two reconstruction modes, producing 54 scale-mode-method combinations. Each combination was evaluated on the same n = 2170 test-scene records, so the reported mean values represent paired comparisons over an identical evaluation set. The compared methods included deterministic interpolation (Bicubic), shallow CNN reconstruction (SRCNN), residual CNN reconstruction (EDSR), attention-based residual CNN reconstruction (RCAN), a Transformer-inspired attention baseline (SwinIR-based SR), adversarial reconstruction (ESRGAN-based SR), two recent attention-based transformers (HAT-based SR and DAT-based SR), and a dense-residual-connected transformer (DRCT-based SR).

Because all methods used the same test scenes within each scale factor and reconstruction mode, differences among methods were not confounded by changes in evaluation sample composition. The results are therefore interpreted primarily from the direction and magnitude of metric differences, with particular attention to whether improvements were consistent across spatial, spectral, vegetation-index, and band-wise criteria.

The ×2 and ×4 conditions provide the cleanest geometric comparison because the original 800 × 800 HR scene size is exactly divisible by two and four. The ×3 condition was retained as a diagnostic condition because it required explicit matching between the HR reference and reconstructed SR grid. Accordingly, ×2 and ×4 are emphasized for model interpretation, while ×3 is reported to show behavior under a less exact scale relationship.

3.2. Spatial Reconstruction Quality

At ×2, reconstruction errors were low, and the differences among the stronger learning-based methods were narrow (Table 3). In the band-wise mode, SwinIR-based and DRCT-based SR produced the lowest RMSE (0.0096), but EDSR (0.0098), SRCNN (0.0099), and RCAN (0.0099) were very close. DRCT-based SR obtained the highest PSNR (39.62 dB), with RCAN highest on SSIM (0.963), indicating that the leading learned models formed a close performance group rather than a clearly separated winner at moderate upscaling. Standard PSNR values, computed over the full [0, 1] reflectance range and reported alongside the scene-dependent PSNR in Table 3, were consistently about 2.2 dB higher across all methods, scales, and reconstruction modes, providing values directly comparable with prior super-resolution studies.

Compared with bicubic interpolation, the ×2 band-wise learned models gave modest but consistent gains in structural quality. Bicubic produced RMSE, PSNR, and SSIM values of 0.0105, 38.17 dB, and 0.948, respectively, whereas the strongest learned models maintained SSIM values above 0.962. These improvements suggest that learning-based reconstruction may recover additional local structure at ×2, although the absolute metric differences were small.

In the joint five-channel mode at ×2, bicubic interpolation retained one of the lowest RMSE values (0.0105), with DRCT-based SR (0.0102) and HAT-based SR (0.0104) marginally lower; the learning-based CNN models did not outperform interpolation on RMSE at ×2 (their ×2 advantage was limited to structural and spectral criteria, with clear RMSE gains emerging only at ×4), but the learned models improved other criteria. DRCT-based SR achieved the highest PSNR (39.26 dB), whereas RCAN obtained the best SSIM (0.962), lowest SAM (0.607 degrees), and lowest mean VI MAE (0.0075). ESRGAN-based SR showed weaker joint ×2 spectral behavior, with a higher SAM of 1.014 degrees, indicating that adversarially guided sharpening was less favorable when spectral fidelity was prioritized.

The ×3 condition showed a different pattern. SRCNN obtained the lowest band-wise RMSE (0.0246) and highest band-wise SSIM (0.820), while bicubic interpolation remained very close in RMSE (0.0250) and PSNR (30.51 dB). Because ×3 required explicit spatial alignment from the 800 × 800 HR grid, this result is interpreted as a diagnostic-condition outcome rather than as evidence that interpolation should be preferred more broadly.

Several deeper models produced larger ×3 errors, with band-wise EDSR increasing to RMSE 0.0360 and SSIM 0.693. The joint five-channel results showed a similar tendency, where SRCNN achieved RMSE 0.0256 while SwinIR-based SR and other deeper models were closer to RMSE 0.0360 or higher.

At ×4, the reconstruction task became more discriminative. In the band-wise mode, HAT-based SR produced the lowest RMSE (0.0239) and highest PSNR (31.15 dB), with EDSR close (0.0243, 30.77 dB), while RCAN achieved the highest SSIM (0.784) and the best overall quality rank. Relative to bicubic interpolation (RMSE 0.0278), EDSR reduced band-

https://doi.org/10.3390/rs18162811

wise RMSE by 12.6% (0.0278→0.0243), demonstrating the clear benefit of learned residual reconstruction under the strongest magnification.

In the joint five-channel mode at ×4, EDSR and RCAN were the leading models. EDSR achieved RMSE 0.0255, PSNR 30.43 dB, and mean VI MAE 0.0172, whereas RCAN achieved the highest SSIM (0.786) and a nearly identical mean VI MAE (0.0173). Relative to joint bicubic RMSE (0.0278), joint EDSR reduced RMSE by 8.3% (0.0278→0.0255). Note that band-wise EDSR achieved a larger absolute improvement (12.6%) than joint EDSR (8.3%), suggesting that preserving inter-band relationships during reconstruction comes at a small cost to individual band fidelity; however, joint reconstruction better preserves spectral consistency for VI applications.

The three recent transformer baselines, HAT-based, DAT-based, and DRCT-based SR, achieved the strongest pixel-wise reconstruction scores, although at different scale factors. At ×2, DRCT-based SR reached the highest PSNR in both modes (39.26 and 39.62 dB), the lowest joint RMSE (0.0102), and matched the lowest band-wise RMSE (0.0096); at ×4 band-wise reconstruction, HAT-based and DAT-based SR reached RMSE values of 0.0239 and 0.0240 and PSNR values of 31.15 and 31.11 dB, the best pixel-fidelity scores among all methods and slightly ahead of EDSR and RCAN. However, none of them improved spectral-angle or vegetation-index preservation: the ×4 band-wise SAM values of HAT-based, DAT-based and DRCT-based SR were 1.490, 1.482 and 1.473 degrees, all higher than the 1.400 and 1.399 degrees of EDSR and RCAN, and at ×2 joint reconstruction the 0.617 degrees of DRCT-based SR likewise exceeded the 0.607 degrees of RCAN. Under the deliberately modest and equalized training budget, these recent attention-based models therefore led on pixel-fidelity metrics but did not surpass the residual and channel-attention CNNs on spectral consistency, consistent with the sensitivity of transformer models to training budget and pre-training.

In summary, the recent attention-based transformers achieved the best pixel-fidelity scores at ×2 and in the ×4 band-wise mode; EDSR was the strongest model in the ×4 joint five-channel mode, and at ×2 the learning-based CNN models did not outperform bicubic interpolation on RMSE. Learning-based gains in spatial reconstruction were therefore clearest at ×4, whereas ×3 remained a diagnostic condition.

3.3. Spectral Consistency and Vegetation-Index Preservation

Vegetation-index preservation showed that the best pixel-wise model was not always the best agricultural model. At ×2 band-wise reconstruction, RCAN produced the lowest mean VI MAE (0.0072), while EDSR produced the lowest SAM (0.594 degrees). The differences among the leading CNN-based models were small, but they were consistently more favorable than bicubic interpolation and ESRGAN-based SR, indicating that moderate SR may improve agricultural index preservation when spectral fidelity is maintained.

At ×4, spectral and vegetation-index preservation became more discriminative (Table 4). In the band-wise mode, RCAN produced the lowest mean VI MAE (0.0177), closely followed by EDSR (0.0179); the difference was small in absolute magnitude (0.0002 in mean VI MAE), and the per-scene paired comparisons reported below indicate that such top-model gaps are statistically detectable but minor in practical terms. In the joint five-channel mode, EDSR produced the lowest mean VI MAE (0.0172), with RCAN nearly identical (0.0173). The small separation between EDSR and RCAN across metrics indicates a shared advantage of residual and attention-enhanced CNN reconstruction, with differences driven more by reconstruction strategy (joint vs. band-wise) than by architectural novelty.

To test whether these differences were consistent across scenes, paired Wilcoxon signed-rank tests with bootstrap 95% confidence intervals (CIs) and rank-biserial effect sizes (r) were computed over the 2170 test scenes (records with RMSE ≥ 1 excluded pairwise). At ×4 band-wise reconstruction, the learning-based models reduced per-scene RMSE relative to bicubic interpolation with a large effect (EDSR vs. bicubic: median ΔRMSE = −0.0038, 95% CI [−0.0040, −0.0037], r = −0.96, $p < 0.001$). In contrast, differences among the leading CNN models were statistically detectable but small in magnitude (band-wise EDSR vs. RCAN: median ΔRMSE = +0.00016, 95% CI [+0.00014, +0.00016], r = +0.77, $p < 0.001$; joint EDSR vs. RCAN: median ΔRMSE = +0.00003, 95% CI [+0.00002, +0.00003], r = +0.16, $p < 0.001$). The per-scene RMSE distributions (Figure 2) overlap substantially among the leading models while remaining clearly below bicubic. Given the large sample size, these p-values should be interpreted together with the effect sizes, which indicate a substantive improvement over interpolation but only a minor practical separation among the top residual and attention-based models.

Across all scale factors and reconstruction modes, records with RMSE ≥ 1 were excluded only for bicubic interpolation, amounting to about 104 of the 2170 test scenes, or 4.8%; all learning-based methods produced no such records, indicating that the exclusions reflect a few high-error boundary scenes that only the non-learning interpolator fails to reconstruct.

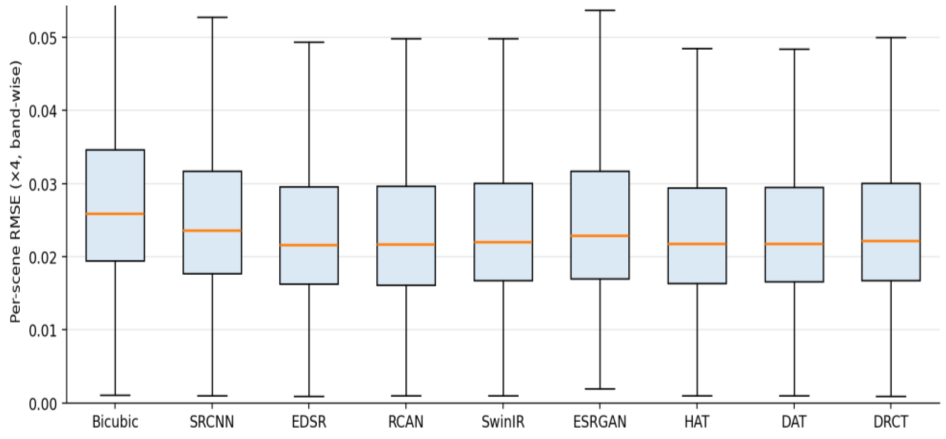

Figure 2. Distribution of per-scene RMSE at ×4 (band-wise reconstruction) for the nine compared methods, excluding records with RMSE ≥ 1. Boxes denote the interquartile range, the orange line marks the median, and the whiskers extend to 1.5 times the interquartile range; scenes beyond the whiskers are not plotted. The learning-based models lower both the median and the spread relative to bicubic interpolation, while the leading CNN models remain closely grouped.

The Red-edge and NIR bands were examined separately because they directly influence NDRE, NDVI, and crop-vigor interpretation. At ×4 band-wise reconstruction, RCAN reduced Red-edge RMSE to 0.0193 and NIR RMSE to 0.0428, compared with bicubic values of 0.0215 and 0.0542, respectively. This corresponds to a 10.2% reduction in Red-edge RMSE and a 21.0% reduction in NIR RMSE, showing that model choice can materially affect agriculturally important spectral bands. Even so, the NIR band retained the highest band-wise reconstruction error, indicating that NIR reconstruction remained the most challenging band.

ESRGAN-based SR was generally less stable for spectral and vegetation-index preservation, especially in joint five-channel reconstruction where its ×2 SAM reached 1.014 degrees. This pattern is consistent with adversarial reconstruction emphasizing visually plausible detail while allowing larger deviations in physically meaningful spectral relationships. For multispectral crop monitoring, such deviations are important because small band-ratio changes may propagate into NDVI, NDRE, and field-level interpretation.

In summary, the ranking by spectral criteria differed from the ranking by pixel fidelity. At ×4, RCAN and EDSR produced the lowest spectral angle and mean vegetation-index error in both reconstruction modes, whereas the models that led on RMSE and PSNR, including the 2024 DRCT-based baseline, showed higher spectral-angle and vegetation-index errors. RCAN also gave the largest Red-edge and NIR improvements over bicubic (10.2% and 21.0% at ×4 band-wise). Model selection for vegetation-index work therefore cannot rely on pixel-fidelity metrics alone.

Table 4. Red-edge and NIR reconstruction accuracy (reflectance-scale, unitless). Lower values are better. These bands are reported separately because they directly affect NDRE, NDVI, and crop-vigor interpretation. Within each scale and reconstruction mode, the row of the method holding the largest number of best values is lightly shaded; ties are resolved by Red-edge RMSE.

Scale	Mode	Method	Red-Edge RMSE	NIR RMSE	NDVI MAE	GNDVI MAE	NDRE MAE
×2	Joint 5-channel	Bicubic	0.0087	0.0198	0.0082	0.0084	0.0082
		SRCNN	0.0097	0.0175	0.0076	0.0081	0.0079
		EDSR	0.0085	0.0175	0.0072	0.0078	0.0076
		RCAN	0.0104	0.0177	0.0071	0.0078	0.0076
		SwinIR-based SR	0.0089	0.0174	0.0074	0.0080	0.0078
		ESRGAN-based SR	0.0110	0.0201	0.0138	0.0117	0.0109
		HAT-based SR	0.0082	0.0181	0.0076	0.0079	0.0078
		DAT-based SR	0.0083	0.0183	0.0076	0.0078	0.0078
		DRCT-based SR	0.0078	0.0181	0.0076	0.0078	0.0077
	Band-wise	Bicubic	0.0087	0.0198	0.0082	0.0084	0.0082
		SRCNN	0.0075	0.0158	0.0067	0.0078	0.0075
		EDSR	0.0074	0.0155	0.0065	0.0077	0.0075
		RCAN	0.0074	0.0154	0.0065	0.0077	0.0075
		SwinIR-based SR	0.0075	0.0160	0.0070	0.0078	0.0075
		ESRGAN-based SR	0.0079	0.0162	0.0074	0.0081	0.0086
		HAT-based SR	0.0073	0.0172	0.0073	0.0077	0.0077
		DAT-based SR	0.0072	0.0172	0.0073	0.0077	0.0077
		DRCT-based SR	0.0072	0.0172	0.0073	0.0077	0.0076
×3	Joint 5-channel	Bicubic	0.0194	0.0482	0.0199	0.0166	0.0163
		SRCNN	0.0203	0.0464	0.0194	0.0164	0.0159
		EDSR	0.0283	0.0693	0.0287	0.0224	0.0217
		RCAN	0.0285	0.0696	0.0288	0.0226	0.0221
		SwinIR-based SR	0.0273	0.0684	0.0286	0.0223	0.0214
		ESRGAN-based SR	0.0298	0.0715	0.0305	0.0238	0.0236
		HAT-based SR	0.0196	0.0480	0.0197	0.0165	0.0161
		DAT-based SR	0.0196	0.0480	0.0197	0.0165	0.0162
		DRCT-based SR	0.0197	0.0480	0.0196	0.0164	0.0161
	Band-wise	Bicubic	0.0194	0.0482	0.0199	0.0166	0.0163
		SRCNN	0.0189	0.0452	0.0190	0.0164	0.0159
		EDSR	0.0271	0.0684	0.0286	0.0228	0.0224
		RCAN	0.0273	0.0680	0.0285	0.0228	0.0224
		SwinIR-based SR	0.0272	0.0681	0.0286	0.0227	0.0223
		ESRGAN-based SR	0.0273	0.0677	0.0291	0.0230	0.0228
		HAT-based SR	0.0187	0.0471	0.0195	0.0165	0.0160
		DAT-based SR	0.0187	0.0471	0.0196	0.0165	0.0161

		DRCT-based SR	0.0187	0.0472	0.0196	0.0165	0.0161
×4	Joint 5-channel	Bicubic	0.0215	0.0542	0.0219	0.0180	0.0177
		SRCNN	0.0210	0.0494	0.0198	0.0170	0.0166
		EDSR	0.0207	0.0462	0.0189	0.0165	0.0161
		RCAN	0.0203	0.0464	0.0190	0.0166	0.0162
		SwinIR-based SR	0.0212	0.0501	0.0205	0.0173	0.0169
		ESRGAN-based SR	0.0252	0.0527	0.0225	0.0196	0.0197
		HAT-based SR	0.0211	0.0505	0.0210	0.0174	0.0171
		DAT-based SR	0.0213	0.0507	0.0208	0.0173	0.0169
		DRCT-based SR	0.0202	0.0500	0.0207	0.0173	0.0169
	Band-wise	Bicubic	0.0215	0.0542	0.0219	0.0180	0.0177
		SRCNN	0.0201	0.0482	0.0195	0.0173	0.0171
		EDSR	0.0194	0.0433	0.0184	0.0177	0.0175
		RCAN	0.0193	0.0428	0.0183	0.0177	0.0172
		SwinIR-based SR	0.0202	0.0438	0.0197	0.0182	0.0191
		ESRGAN-based SR	0.0204	0.0453	0.0194	0.0181	0.0183
		HAT-based SR	0.0197	0.0434	0.0201	0.0185	0.0194
		DAT-based SR	0.0197	0.0438	0.0203	0.0184	0.0193
		DRCT-based SR	0.0197	0.0446	0.0203	0.0183	0.0191

3.4. Visual Comparison

Representative visual comparisons were used to check whether the quantitative trends also appeared as spatial patterns relevant to crop monitoring. The selected scene contains repeated crop rows, soil–background transitions, and canopy texture, making it suitable for inspecting row continuity, field-boundary detail, and vegetation-index consistency across scale factors (Figures 3–7).

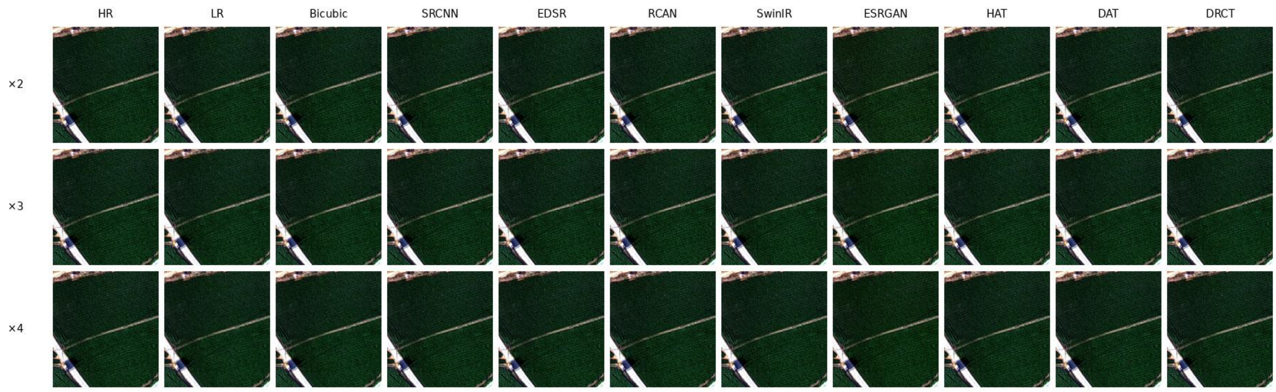

Figure 3. Joint five-channel RGB reconstruction for the selected scene at ×2, ×3, and ×4. Rows correspond to the scale factor and columns to the HR reference, the LR input, and the nine methods (Bicubic, SRCNN, EDSR, RCAN, SwinIR-based SR, ESRGAN-based SR, HAT-based SR, DAT-based SR, and DRCT-based SR). At the full-scene scale, the methods look broadly similar; their differences are clearest in the zoomed comparison (Figure 5).

The full-scene RGB and false-color views (Figures 3 and 4) mainly provide scene-level qualitative confirmation of the quantitative trends, as the leading methods show no large visible differences at this scale. Such differences become clearest in the labeled zoomed ×4 comparison (Figure 5), where bicubic interpolation produces overly smooth crop-row transitions, whereas EDSR and RCAN preserve finer canopy texture and row

https://doi.org/10.3390/rs18162811

definition; at ×2, the learning-based methods are visually close to the HR reference. Visual appearance alone, however, does not fully reflect spectral or vegetation-index reliability, as shown by the joint and band-wise VI error maps in Figures 6 and 7. This disconnect between RGB appearance and spectral fidelity underscores the need for evaluation beyond visual inspection.

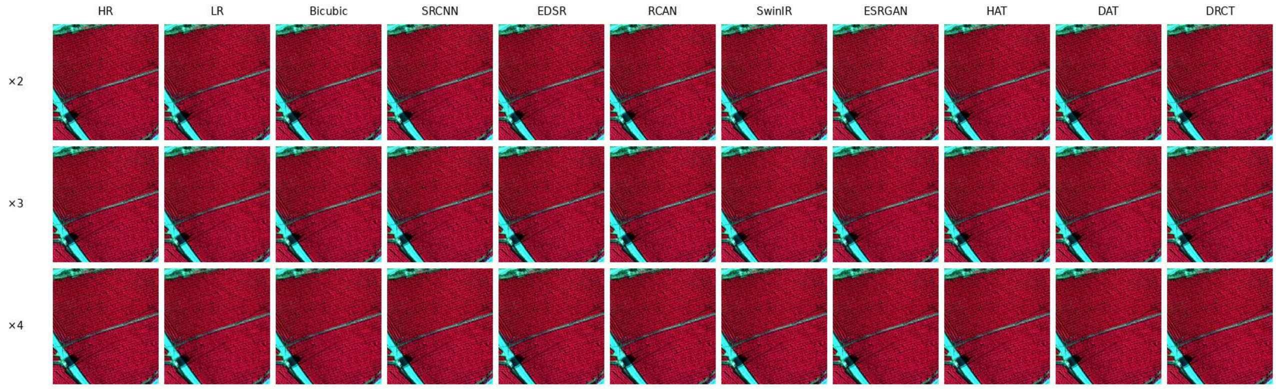

Figure 4. Joint five-channel false-color reconstruction for the selected scene at ×2, ×3, and ×4, with the same row and column arrangement as Figure 3. The false-color view emphasizes Red-edge and NIR canopy responses, and differences among the leading methods remain subtle at this full-scene scale.

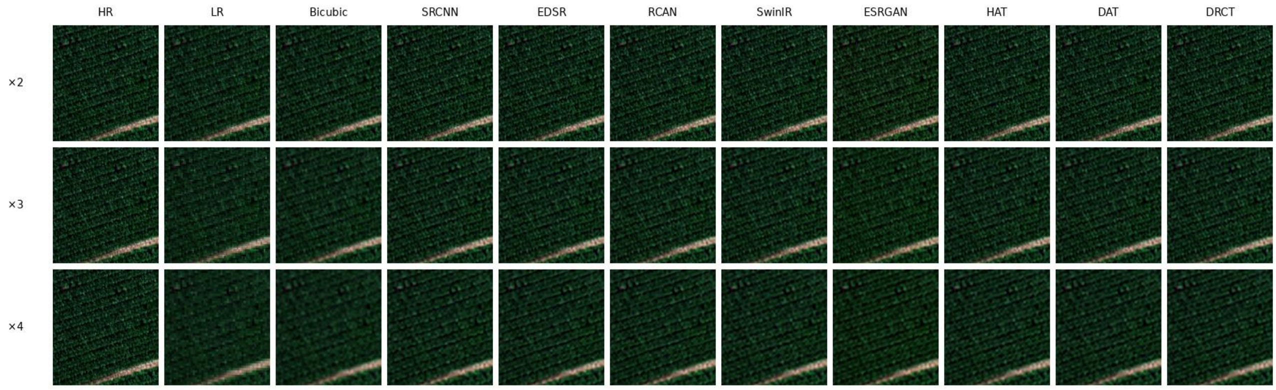

Figure 5. Labeled zoomed RGB comparison for the selected scene at ×2, ×3, and ×4, using the same layout as Figure 3. This closer view most clearly reveals the differences in edge sharpness, canopy texture, and crop-row recovery, particularly at ×4.

The joint five-channel per-band error maps (Figure 6) show where the reconstructed reflectance deviates from the HR reference across the Blue, Green, Red, Red-edge, and NIR bands. The errors are generally small and concentrate along crop rows, soil gaps, and field boundaries, with the largest residuals in the NIR band.

The band-wise per-band error maps (Figure 7), plotted on the same color scale, allow a direct comparison with the joint mode (Figure 6); both modes yield small and spatially similar errors, with the NIR band showing the largest deviations. Together, Figures 6 and 7 provide a spatial counterpart to the quantitative band-wise reconstruction errors reported in Tables 3 and 4.

In summary, the visual comparisons confirm the quantitative trends but cannot replace them. Differences among the leading methods are clearly visible only in the zoomed ×4 view, and the per-band error maps show that the largest residuals occur in the NIR band, which the RGB appearance of the reconstructions does not reveal.

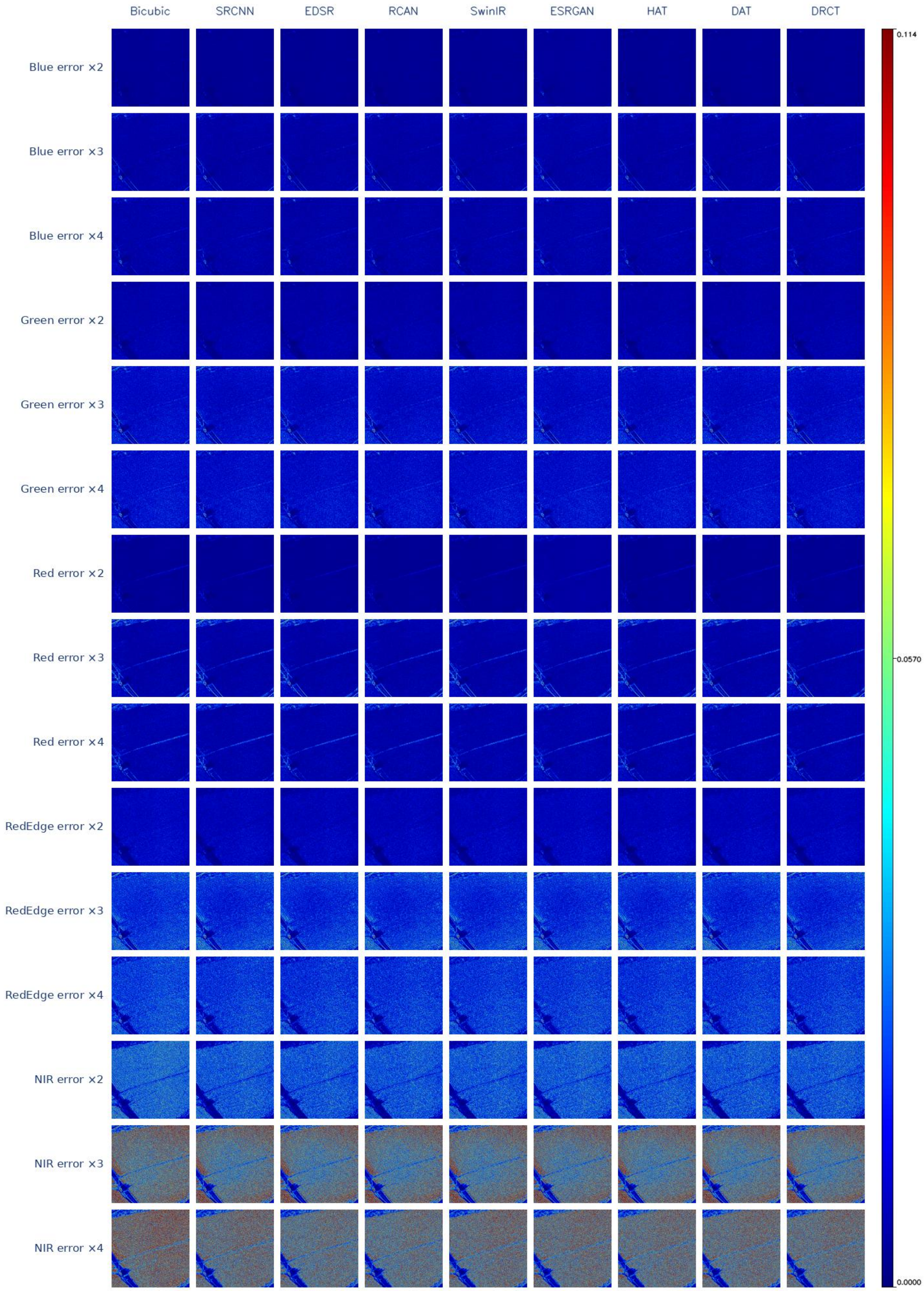

Figure 6. Joint five-channel per-band reflectance error maps for the Blue, Green, Red, Red-edge, and NIR bands (|SR − HR|) for the selected scene at ×2, ×3, and ×4. Rows correspond to the spectral band and scale factor, and columns to the nine reconstruction methods, each shown as the absolute difference from the HR reference. All maps share one error color scale (0 to 0.114), the same as in Figure 7.

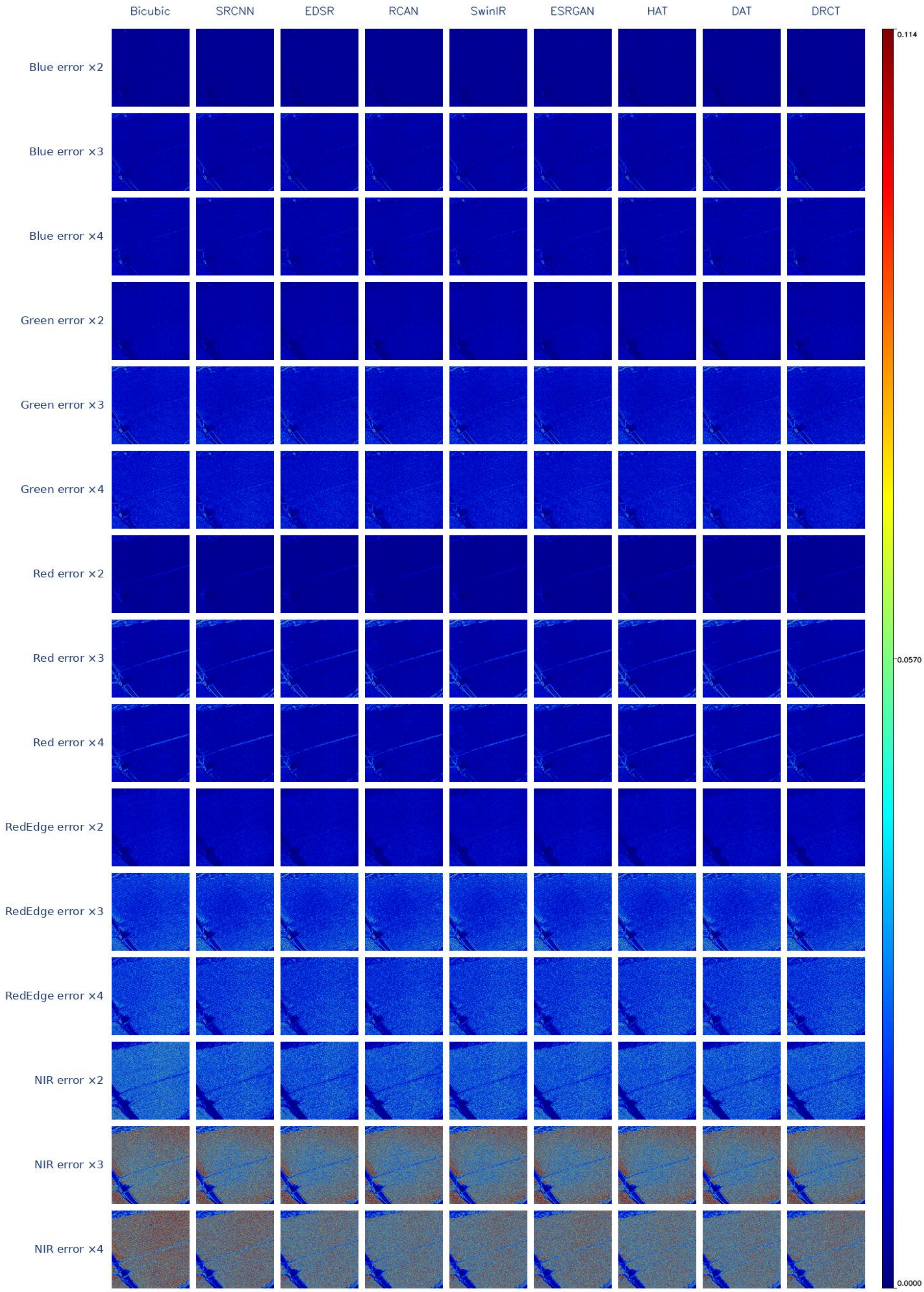

Figure 7. Band-wise per-band reflectance error maps for the Blue, Green, Red, Red-edge, and NIR bands (|SR − HR|) for the selected scene at ×2, ×3, and ×4. Rows correspond to the spectral band and scale factor, and columns to the nine reconstruction methods, each shown as the absolute difference from the HR reference. All maps share one error color scale (0 to 0.114), indicated by the colorbar.

3.5. Model Ranking and Computational Efficiency

The metric-wise ranking indicates that no single method dominated every criterion (Table 5). SwinIR-based SR occasionally achieved the lowest RMSE, particularly at ×2 band-wise reconstruction, but it did not consistently provide the best SAM, Red-edge/NIR accuracy, or vegetation-index preservation. SRCNN performed strongly under the ×3 diagnostic condition, but this advantage was less evident under the cleaner ×2 and ×4 conditions. DRCT-based SR, the most recent architecture evaluated, attained the best pixel-fidelity scores at ×2 but did not lead on SAM, Red-edge/NIR accuracy, or vegetation-index preservation. ESRGAN-based SR was visually plausible in RGB panels but was generally not preferred when spectral consistency and VI errors were considered.

Across the primary ×2 and ×4 conditions, EDSR and RCAN were the most consistently supported choices. Among the CNN models, EDSR was most favorable when RMSE and PSNR were emphasized, especially in the ×4 joint mode, whereas RCAN was consistently competitive for SSIM, Red-edge/NIR reliability, and mean VI MAE. The recent transformer baselines led on pixel fidelity at different scale factors—DRCT-based SR led at ×2, and HAT-based and DAT-based SR in the ×4 band-wise mode—but their larger spectral-angle and vegetation-index errors meant EDSR and RCAN remained preferable when spectral fidelity was prioritized. This complementary pattern suggests that residual CNN reconstruction is suitable for spatial fidelity, while attention-based residual reconstruction is useful when spectral-index stability is prioritized.

Computational efficiency was summarized using wall-clock SR reconstruction time, total evaluation time, and throughput (Table 6). Bicubic interpolation was the fastest approach, requiring 0.016 to 0.029 s for SR reconstruction depending on scale and mode. Among the learned models, SRCNN was the fastest, with joint five-channel reconstruction times of 0.126 s at ×2, 0.110 s at ×3, and 0.160 s at ×4.

Band-wise reconstruction increased reconstruction time because each spectral band was processed separately. For example, RCAN required 0.231 s in ×2 joint five-channel mode but 0.638 s in ×2 band-wise mode. SwinIR-based SR showed the highest band-wise cost at ×2 (1.275 s), consistent with the heavier computation of attention-based reconstruction when applied independently to each band. These timing values are interpreted as practical wall-clock measurements for the implemented evaluation pipeline rather than as hardware-independent measures of computational complexity. The recent transformer baselines were the most computationally expensive overall, with the lowest throughput among all methods at every scale. DRCT-based SR was the slowest of all, and its band-wise ×2 reconstruction time of 2.877 s was the highest recorded for any model, followed by DAT-based SR (1.815 s) and HAT-based SR (1.771 s). Because the dense residual connections of DRCT-based SR progressively widen the feature tensor within each group, its accuracy advantage at ×2 was obtained at the highest computational cost in this benchmark.

In summary, no single method was best on every criterion. EDSR and RCAN offered the most favorable balance of spatial and spectral quality at moderate computational cost, the recent attention-based transformers led on pixel fidelity but were the slowest methods at every scale, and band-wise reconstruction required roughly three to five times the inference time of joint five-channel reconstruction.

Table 5. Best-performing method by metric for each scale and reconstruction mode. Directions: RMSE (lower), PSNR (higher), SSIM (higher), SAM (lower), Mean VI MAE (lower). Bicubic is the deterministic baseline.

Scale	Mode	RMSE	PSNR	SSIM	SAM	Mean VI MAE
×2	Joint 5-channel	DRCT-based SR	DRCT-based SR	RCAN	RCAN	RCAN

Scale	Mode	RMSE	PSNR	SSIM	SAM	Mean VI MAE
	Band-wise	SwinIR-based SR	DRCT-based SR	RCAN	EDSR	RCAN
×3	Joint 5-channel	Bicubic	DAT-based SR	DAT-based SR	DRCT-based SR	SRCNN
	Band-wise	SRCNN	HAT-based SR	HAT-based SR	SRCNN	SRCNN
×4	Joint 5-channel	EDSR	EDSR	RCAN	EDSR	EDSR
	Band-wise	HAT-based SR	HAT-based SR	RCAN	SRCNN	RCAN

Table 6. Computational efficiency (wall-clock time in seconds). SR recon.: super-resolution reconstruction time. Total eval.: includes loading, metric calculation, and output saving. Throughput: megapixels per second. Lower is better for SR reconstruction time and total evaluation time; higher is better for throughput. Note: Band-wise reconstruction requires roughly 3 to 5x more time than joint five-channel; SRCNN remains fastest.

Scale	Mode	Method	SR Recon. (s)	Total Eval. (s)	Throughput (Mpx/s)
×2	Joint 5-channel	Bicubic	0.029	1.226	26.47
		SRCNN	0.126	1.323	5.22
		EDSR	0.155	1.348	4.16
		RCAN	0.231	1.434	2.80
		SwinIR-based SR	0.343	1.533	1.87
		ESRGAN-based SR	0.204	1.415	3.18
		HAT-based SR	0.403	1.601	1.59
		DAT-based SR	0.433	1.631	1.48
		DRCT-based SR	0.618	1.395	1.04
	Band-wise	Bicubic	0.028	1.232	23.46
		SRCNN	0.134	1.314	4.83
		EDSR	0.319	1.488	2.01
		RCAN	0.638	1.807	1.00
		SwinIR-based SR	1.275	2.452	0.51
		ESRGAN-based SR	0.508	1.706	1.26
		HAT-based SR	1.771	2.954	0.36
		DAT-based SR	1.815	2.998	0.35
		DRCT-based SR	2.877	3.551	0.22
×3	Joint 5-channel	Bicubic	0.018	1.186	36.92
		SRCNN	0.110	1.265	5.82
		EDSR	0.134	1.278	4.83
		RCAN	0.159	1.321	4.05
		SwinIR-based SR	0.182	1.345	3.52
		ESRGAN-based SR	0.142	1.287	4.52
		HAT-based SR	0.237	1.393	2.68
		DAT-based SR	0.242	1.398	2.63
		DRCT-based SR	0.333	1.016	1.91
	Band-wise	Bicubic	0.024	1.200	26.18
		SRCNN	0.113	1.288	5.69
		EDSR	0.204	1.353	3.12
		RCAN	0.351	1.507	1.82
		SwinIR-based SR	0.545	1.703	1.17
		ESRGAN-based SR	0.279	1.443	2.29
		HAT-based SR	0.937	2.100	0.68
		DAT-based SR	0.958	2.121	0.66
		DRCT-based SR	1.429	2.121	0.45
×4	Joint 5-channel	Bicubic	0.016	1.179	40.21

	SRCNN	0.160	1.329	4.06
	EDSR	0.177	1.335	3.64
	RCAN	0.208	1.374	3.10
	SwinIR-based SR	0.152	1.316	4.23
	ESRGAN-based SR	0.196	1.362	3.29
	HAT-based SR	0.184	1.348	3.48
	DAT-based SR	0.187	1.351	3.42
	DRCT-based SR	0.247	0.934	2.60
Band-wise	Bicubic	0.023	1.191	27.64
	SRCNN	0.156	1.335	4.13
	EDSR	0.254	1.430	2.53
	RCAN	0.402	1.579	1.60
	SwinIR-based SR	0.386	1.558	1.66
	ESRGAN-based SR	0.344	1.571	1.86
	HAT-based SR	0.616	1.799	1.04
	DAT-based SR	0.628	1.811	1.02
	DRCT-based SR	0.934	1.628	0.69

3.6. Loss-Term Ablation (Joint Five-Channel, ×4)

A loss-term ablation was conducted at ×4 in the joint five-channel mode to isolate the contribution of the spectral-angle (SAM) and vegetation-index (VI) loss terms. The ablation started from a masked-L1 base and then added SAM and VI under the same training and evaluation configuration. All loss-ablation configurations were evaluated under the ×4 joint five-channel setting, allowing the values in Table 7 to be compared directly across the reduced loss configurations. Adding the SAM term generally reduced spectral angle among the non-adversarial models, with the clearest change observed for EDSR (SAM 1.439 to 1.407 degrees) and the same direction observed for the transformer baselines, including DRCT-based SR (1.385 to 1.378 degrees), while RMSE changed only slightly. Adding the VI term produced a model-dependent effect on vegetation-index error: mean VI MAE decreased for SwinIR-based SR, ESRGAN-based SR, and HAT-based SR, increased slightly for EDSR, RCAN, SRCNN, and DAT-based SR, and was essentially unchanged for DRCT-based SR relative to the L1 + SAM setting. These results indicate that SAM and VI constraints affected different metrics differently, and that vegetation-index loss weights should be tuned and validated rather than assumed to improve all models uniformly. Because the ablation profiles omit the additional spatial-detail loss terms used in the deployed models, Table 7 isolates the relative effect of the SAM and VI terms rather than reproducing the full-loss performance, and its absolute errors are therefore higher than those of the corresponding full-loss models in Table 3. For training stability, the ESRGAN-based ablation used lower spectral and vegetation-index loss weights and a more gradual training schedule than the other models.

In summary, adding the SAM term reduced spectral angle for the non-adversarial models, whereas the vegetation-index term changed mean vegetation-index error in a model-dependent direction. Spectral and vegetation-index loss weights should therefore be validated for each architecture rather than assumed to be uniformly beneficial.

Table 7. Loss-term ablation at ×4 (joint five-channel reconstruction) for three reduced loss configurations (L1; L1 + SAM; and L1 + SAM + VI). Values are mean values from the ×4 joint five-channel loss-ablation evaluation. Records with RMSE ≥ 1 were excluded before summary calculation. Lower is better for RMSE, SAM, mean VI MAE, Red-edge RMSE, and NIR RMSE; higher is better for PSNR and SSIM. Within each model, the row of the loss configuration holding the largest number of best values is lightly shaded; ties are resolved by RMSE.

Model	Loss	RMSE	PSNR	SSIM	SAM	Mean VI MAE	Red-Edge RMSE	NIR RMSE
EDSR	L1	0.0294	30.44	0.790	1.439	0.01869	0.0231	0.0528
	L1 + SAM	0.0295	30.38	0.788	1.407	0.01825	0.0224	0.0525
	L1 + SAM + VI	0.0304	30.12	0.781	1.455	0.01869	0.0210	0.0533
RCAN	L1	0.0288	30.51	0.789	1.453	0.01892	0.0223	0.0532
	L1 + SAM	0.0308	30.21	0.787	1.445	0.01862	0.0214	0.0538
	L1 + SAM + VI	0.0311	30.11	0.784	1.447	0.01874	0.0240	0.0540
SwinIR-based SR	L1	0.0299	30.00	0.775	1.485	0.01943	0.0229	0.0544
	L1 + SAM	0.0308	29.81	0.771	1.468	0.01917	0.0220	0.0560
	L1 + SAM + VI	0.0310	29.75	0.770	1.457	0.01898	0.0235	0.0568
SRCNN	L1	0.0304	29.93	0.774	1.468	0.01914	0.0235	0.0554
	L1 + SAM	0.0306	29.89	0.773	1.440	0.01876	0.0235	0.0554
	L1 + SAM + VI	0.0305	29.86	0.769	1.457	0.01900	0.0223	0.0555
ESRGAN-based SR	L1	0.0345	29.44	0.759	1.743	0.02190	0.0277	0.0584
	L1 + SAM	0.0340	29.43	0.757	1.756	0.02182	0.0272	0.0586
	L1 + SAM + VI	0.0341	29.43	0.758	1.741	0.02155	0.0272	0.0584
HAT-based SR	L1	0.0265	30.31	0.772	1.415	0.01861	0.0201	0.0497
	L1 + SAM	0.0269	30.19	0.768	1.393	0.01829	0.0202	0.0515
	L1 + SAM + VI	0.0272	30.10	0.767	1.395	0.01826	0.0209	0.0513
DAT-based SR	L1	0.0262	30.37	0.771	1.416	0.01864	0.0198	0.0498
	L1 + SAM	0.0272	30.14	0.769	1.393	0.01828	0.0206	0.0512
	L1 + SAM + VI	0.0272	30.12	0.767	1.396	0.01829	0.0211	0.0506
DRCT-based SR	L1	0.0258	30.57	0.780	1.385	0.01820	0.0203	0.0476
	L1 + SAM	0.0271	30.16	0.770	1.378	0.01810	0.0205	0.0508
	L1 + SAM + VI	0.0269	30.17	0.768	1.381	0.01810	0.0203	0.0509

3.7. Parameter Scaling, Degradation, and Noise Robustness

At ×4, widening the joint five-channel EDSR to a trainable-parameter count comparable to that of the band-wise EDSR ensemble reduced its RMSE to 0.0239, on par with band-wise reconstruction, while lowering the spectral angle to 1.27°. This indicates that the band-wise per-band spatial advantage reflects model capacity rather than the reconstruction strategy. To separate the reconstruction mode from the loss formulation, a joint model trained without the SAM and vegetation-index terms was compared with the full-loss joint model and with band-wise reconstruction. Removing the spectral losses increased the mean spectral angle from 1.33° to 1.37°, yet this configuration still preserved spectral angle slightly better than band-wise reconstruction at 1.40°. This control narrows the confound between reconstruction mode and loss formulation, but it does not by itself establish a general causal separation across architectures or training regimes; the result is therefore reported as controlled support for the joint-reconstruction interpretation under the present architecture, dataset, and training budget rather than as a universal conclusion. Retraining EDSR, RCAN, and HAT under a Gaussian-blur-plus-bicubic degradation reproduced the area-downsampling accuracy and preserved the joint-versus-band-wise tradeoff (Table 8). Under additive Gaussian noise on the low-resolution inputs, the convolutional and transformer models degraded sharply in spectral angle and vegetation-index error, whereas the adversarial ESRGAN-based model was more noise-robust (Table 9).

In summary, the band-wise advantage in per-band spatial accuracy is largely attributable to model capacity rather than to independent reconstruction; the joint mode retains a small spectral-angle advantage when the spectral losses are removed, which is reported as controlled support rather than as a general causal separation; and the model rankings were stable under a Gaussian-blur-plus-bicubic degradation. Additive input noise was the one factor that reordered the methods, with the adversarial model degrading least.

Table 8. Reconstruction accuracy of EDSR, RCAN, and HAT at ×4 under area-based downsampling and Gaussian-blur-plus-bicubic degradation. This experiment requires full retraining under a second degradation model, so it was deliberately scoped to three representative architectures—a residual CNN (EDSR), a channel-attention CNN (RCAN), and a hybrid-attention transformer (HAT)—and is not intended as a full nine-method comparison. Within each reconstruction mode, the row of the method holding the largest number of best values is lightly shaded.

Mode	Method	Area RMSE	Blur + Bicubic RMSE	Area SAM (°)	Blur SAM (°)
Joint	EDSR	0.0248	0.0249	1.332	1.344
	RCAN	0.0249	0.0252	1.337	1.341
	HAT	0.0269	0.0270	1.408	1.408
Band-wise	EDSR	0.0239	0.0238	1.399	1.373
	RCAN	0.0240	0.0239	1.394	1.377
	HAT	0.0239	0.0238	1.490	1.474
—	Bicubic	0.0279	0.0378	1.439	1.516

Table 9. Effect of additive Gaussian noise on the ×4 low-resolution test inputs for EDSR, RCAN, ESRGAN-based, and HAT reconstruction. The noise evaluation reuses the ×4 checkpoints of four representative architectures spanning the residual-CNN, channel-attention, adversarial, and transformer families, and is likewise not intended as a full nine-method comparison.

Method	σ	RMSE	SAM (°)	Mean VI MAE
EDSR	0.00	0.0248	1.332	0.0173
	0.01	0.0263	2.914	0.0357
	0.03	0.0395	6.857	0.0839
RCAN	0.00	0.0249	1.337	0.0173
	0.01	0.0265	2.958	0.0363
	0.03	0.0403	6.986	0.0858
ESRGAN	0.00	0.0284	1.674	0.0208
	0.01	0.0276	2.283	0.0296
	0.03	0.0356	4.372	0.0584
HAT	0.00	0.0269	1.408	0.0185
	0.01	0.0279	2.970	0.0369
	0.03	0.0402	6.863	0.0847

3.8. Summary of Results

Overall, the results support four main findings. First, ×2 and ×4 provide the most geometrically reliable scale-factor comparisons because they are exactly compatible with the 800 × 800 HR scene size, whereas ×3 is better interpreted as a diagnostic condition. Second, deep SR improved over bicubic interpolation most clearly at ×4, where EDSR and RCAN reduced reconstruction error while maintaining stronger Red-edge, NIR, NDVI, GNDVI, and NDRE consistency.

Third, the comparison between joint five-channel and band-wise reconstruction showed that the best mode depends on the evaluation target and on the feasible loss formulation used for each mode. Band-wise reconstruction may improve band-specific spatial accuracy, whereas joint reconstruction can better preserve the five-band cube as a coupled spectral input. Fourth, the visual comparisons reinforced the conclusion that agricultural usefulness cannot be judged from RGB appearance alone. For UAV-based crop monitoring, SR models should be evaluated using spatial, spectral, vegetation-index, band-wise, and computational criteria together.

4. Discussion

4.1. Interpretation of Model Performance

The comparative results show that UAV multispectral SR should be interpreted through both spatial reconstruction quality and spectral reliability. RMSE, PSNR, and SSIM describe pixel-wise and structural reconstruction, but they do not fully indicate whether the reconstructed Red, Green, Red-edge, and NIR relationships remain usable for vegetation-index calculation [9,19]. This distinction explains why methods with favorable visual or structural behavior were not always the most reliable for agricultural spectral interpretation. Under the controlled degradation setting used here, EDSR and RCAN provided the most balanced behavior across spatial quality, spectral similarity, band-wise accuracy, and vegetation-index preservation. These differences reflect the underlying reconstruction assumptions rather than model novelty alone: residual CNN reconstruction (EDSR) recovers high-frequency spatial detail stably and is therefore strong on RMSE/PSNR; channel-attention reconstruction (RCAN) re-weights spectral channels and is thus more favorable for Red-edge/NIR and vegetation-index consistency; whereas the adversarial model (ESRGAN-based) optimizes perceptual realism and can deviate more in physically meaningful band ratios, and the compact Transformer-inspired model is comparatively data- and budget-hungry under the present matched budget.

The nine compared methods represent different reconstruction assumptions rather than a simple ladder of model novelty. Bicubic interpolation provides a deterministic non-learning reference for smooth resizing. SRCNN represents a shallow CNN baseline [23], EDSR represents residual CNN reconstruction with stable high-frequency recovery [24], and RCAN adds channel-attention mechanisms that are relevant for multi-band feature weighting [25]. The SwinIR-based model functions here as a compact Transformer-inspired attention baseline derived from image-restoration SR concepts [12], whereas ESRGAN-based SR represents adversarially guided reconstruction, where apparent sharpness may be favored over strict pixel-wise or spectral fidelity [26]. The HAT-based, DAT-based, and DRCT-based models are recent attention-based transformers that extend the Transformer-inspired family [27–29]; DRCT-based SR additionally connects the attention layers of each group densely, so that early spatial detail remains reachable at the group output instead of being squeezed through a plain sequential stack. In this benchmark, they achieved the strongest pixel-wise RMSE and PSNR, DRCT-based SR at ×2 and HAT-based and DAT-based SR in the ×4 band-wise mode, yet none of them improved spectral-angle or vegetation-index preservation, so their contribution was concentrated in spatial rather than spectral fidelity. That a 2024 architecture reproduces the same pattern suggests it reflects the pixel-oriented optimization target shared by these methods rather than a limitation of one model generation.

The comparatively weaker spectral-fidelity behavior of the Transformer-inspired and adversarial baselines should therefore be interpreted cautiously, as perceptually and adversarially oriented reconstruction is known to prioritize visual realism over strict pixel-wise and spectral fidelity [15,16]. It does not imply that these model families are unsuitable for agricultural multispectral SR in general. Rather, within the present training

https://doi.org/10.3390/rs18162811

budget, compact implementation, controlled degradation setting, and spectral-fidelity-oriented evaluation, residual CNN and channel-attention CNN reconstruction were more stable [13,14]. This distinction indicates that agricultural SR methods should be selected not only by architecture novelty or visual sharpness, but by whether they preserve quantitative spectral relationships.

4.2. Implications of Joint Five-Channel and Band-Wise Reconstruction

The joint five-channel and band-wise results indicate that the preferred reconstruction strategy depends on the evaluation target. Joint reconstruction is conceptually attractive because vegetation indices are computed from relationships among spectral bands, and a joint model can process the five-band cube as a coupled spectral object [19,20]. Band-wise reconstruction can be advantageous when the objective is band-specific spatial accuracy or when independent bands have different contrast and noise behavior. The results therefore support reporting both modes rather than assuming that one strategy is universally superior.

At the same time, the comparison does not isolate reconstruction mode alone. Joint five-channel models can use SAM and vegetation-index terms during training, whereas single-channel band-wise models cannot directly use losses that depend on multi-band spectral relationships. The joint and band-wise results should therefore be read as a practical comparison under feasible loss formulations. This limitation is methodologically important because differences between modes may reflect both the reconstruction strategy and the loss terms available to each strategy. To disentangle these factors, a joint model trained without the SAM and vegetation-index terms was evaluated in Section 3.7; it still preserved spectral angle slightly better than band-wise reconstruction. This narrows the confound but does not by itself establish a general causal separation, so the result is read as controlled support for the joint-reconstruction interpretation under the present architecture, dataset, and training budget.

The loss-ablation results further support this interpretation. Adding SAM modestly improved spectral-direction consistency [35], whereas the VI term produced model-dependent effects, indicating that spectral constraints should be tuned and validated rather than assumed to improve all agricultural metrics.

4.3. Vegetation-Index Preservation as an Agricultural Proxy

The VI analysis links the reconstruction metrics to agricultural interpretation while remaining a proxy rather than a complete agronomic validation. NDVI reflects Red-NIR relationships, GNDVI reflects Green-NIR relationships, and NDRE reflects Red-edge-NIR relationships that are closely connected to crop-monitoring interpretation [30–32]. Because SR can alter local band ratios even when RGB appearance improves, VI errors provide a useful test of whether reconstructed images preserve agriculturally interpretable spectral structure.

Nevertheless, low VI error should not be interpreted as direct evidence of improved crop diagnosis, yield estimation, or field management decisions. The VI metrics used here bridge image reconstruction and agricultural interpretation, but they do not replace downstream validation using biomass measurements, stress labels, segmentation masks, or yield-related observations [1,33]. The main value of the VI evaluation is therefore to screen whether SR outputs remain spectrally plausible enough for subsequent agricultural analysis.

https://doi.org/10.3390/rs18162811

4.4. Scale-Factor and Degradation Assumptions

The controlled degradation design is both a strength and a limitation. It provides paired LR-HR data with known scale factors and allows all methods to be compared under identical reference conditions. However, LR inputs generated by downsampling dataset-derived HR cubes do not fully represent operational UAV degradation, where motion blur, wind-induced canopy movement, atmospheric variation, exposure differences, radiometric calibration uncertainty, lens effects, and sensor modulation characteristics may affect both spatial detail and spectral response [7,16]. The results should therefore be interpreted as an internal controlled benchmark rather than as direct evidence of performance on real multi-altitude or multi-sensor LR-HR image pairs.

The ×3 condition further illustrates this geometric sensitivity. Because 800 pixels cannot be exactly divided by three, explicit alignment was required before metric calculation, and small residual shifts may affect sharper learned outputs differently from smoother interpolation. A likely mechanism is that the residual-learning models reconstruct high-frequency detail on top of a bicubic-upsampled base at a non-integer scale (266->798 pixels, followed by a two-pixel crop), so a small systematic resampling offset is amplified by the learned high-frequency branch, whereas a smooth interpolator (bicubic) or a shallow model (SRCNN) is far less sensitive; models with larger receptive fields are therefore more strongly affected. For this reason, ×3 was retained as a diagnostic condition, whereas ×2 and ×4 provide the cleaner basis for model interpretation, and the ×3 results should not be used to rank relative model performance.

A second related limitation is that the data split was performed at the scene-record level rather than by field, date, or flight. This improves sample size for model comparison but does not completely exclude acquisition-level similarity between training and test records. Consequently, this benchmark should be read as a comparison of relative method performance under matched acquisition conditions; the stability of the model rankings across geographically independent study sites remains an open question, and grouped cross-validation by field or acquisition date is left for future work to address this limitation.

4.5. Relationship to Recent Remote-Sensing SR Trends

Recent remote-sensing SR research increasingly includes Transformer-based [12], diffusion-based [17,41,42], and foundation-model [18] approaches. These directions are important because they can model long-range context or generate detailed spatial structures. For agricultural multispectral SR, however, visually realistic detail is not sufficient; reconstructed reflectance relationships must remain reliable enough for spectral indices and downstream agronomic interpretation [9,19]. The present study therefore emphasizes a reproducible comparison of representative CNN, attention, Transformer-inspired, and adversarial baselines under spectral-fidelity-oriented evaluation.

This positioning is consistent with recent reviews that emphasize both the rapid expansion of deep learning-based remote-sensing SR and the need for application-specific evaluation [8]. Diffusion models have become a major direction in image super-resolution because they may generate detailed high-resolution structures from degraded inputs [17]. Diffusion-based remote-sensing SR has been proposed for aerial and satellite imagery [41]. Efficient diffusion probabilistic models have also been developed for remote-sensing image reconstruction [42]. Vision foundation models are also increasingly discussed in remote sensing because they offer transferable representations across heterogeneous tasks and sensors [18].

The absence of diffusion and foundation-model baselines should therefore be viewed as a scope decision rather than as a claim that these methods are unsuitable. Future work can extend the same evaluation framework to generative and foundation-model approaches, but those methods should be judged using spectral angle, band-wise error, vegetation-index preservation, and downstream agricultural utility, not visual sharpness alone.

4.6. Practical Implications and Limitations

From a practical perspective, the findings suggest that SR may be useful for multispectral crop imagery acquired by UAV when spatial enhancement is coupled with explicit spectral evaluation. EDSR and RCAN provided the most balanced performance in this dataset by combining spatial reconstruction quality with Red-edge, NIR, and VI reliability under the controlled benchmark. However, the results are most directly applicable to quality-controlled AI Hub cabbage scenes reconstructed from dataset-derived HR references. They should not be generalized uncritically to other crops, sensors, flight altitudes, radiometric workflows, or real LR-HR acquisition conditions.

Several steps would strengthen future validation. First, grouped field-, date-, or flight-level splits should be used to test whether model rankings remain stable when acquisition-level similarity is reduced. Second, real multi-altitude or multi-sensor LR-HR image pairs should be included to evaluate realistic degradation. Third, SR outputs should be connected to downstream tasks such as plant segmentation, stress detection, biomass estimation, or yield-related prediction. Finally, broader hyperparameter tuning and longer training budgets may be needed to determine whether deeper attention, Transformer-inspired, or adversarial models can become more competitive under multispectral spectral-fidelity constraints.

Taken together, these findings support an evaluation practice in which UAV crop-image SR is reported through spatial quality, spectral consistency, band-wise reconstruction, VI preservation, computational efficiency, and clearly stated degradation assumptions. This framing keeps the emphasis on whether enhanced imagery remains quantitatively meaningful for agricultural analysis, rather than on visual sharpness alone.

5. Conclusions

This study evaluated a super-resolution framework for UAV-based five-band multispectral cabbage imagery under a controlled downsampling benchmark. Nine reconstruction methods, including bicubic interpolation, SRCNN, EDSR, RCAN, SwinIR-based SR, ESRGAN-based SR, HAT-based SR, DAT-based SR, and DRCT-based SR, were compared under joint five-channel and band-wise reconstruction settings at ×2, ×3, and ×4 scale factors. The evaluation combined spatial image-quality metrics, spectral similarity, band-wise reconstruction errors, vegetation-index errors, per-band reflectance error maps, and computational efficiency so that SR performance could be assessed beyond RGB visual sharpness alone.

No single method dominated all criteria. Across the geometrically cleaner ×2 and ×4 conditions, EDSR and RCAN showed the most balanced performance in the present dataset. EDSR was generally strong for pixel-wise reconstruction and RMSE/PSNR-oriented performance, whereas RCAN was competitive for structural similarity, Red-edge/NIR reliability, and vegetation-index preservation. SwinIR-based SR and ESRGAN-based SR served as useful attention-based and adversarial baselines, but they did not consistently improve spectral or vegetation-index reliability under the current implementation and training budget. Among the recently proposed attention-based models, DRCT-based SR attained the strongest pixel-fidelity scores at ×2, and HAT-based and DAT-based SR did so in the ×4 band-wise mode, but none of them improved spectral-angle or vegetation-index preservation, and therefore they did not surpass the residual and channel-attention CNNs on spectral consistency under the equalized budget. This ranking should be read in the context of the deliberately modest and equalized training budget (50 epochs, compact from-scratch configurations); the Transformer-inspired and adversarial baselines may become more competitive under larger training budgets, pre-training, or architecture-specific tuning.

The vegetation-index results support the need to evaluate UAV multispectral SR as a spectral reconstruction problem. NDVI, GNDVI, and NDRE errors showed that visually sharper reconstructions were not necessarily more reliable for crop-monitoring interpretation, because small distortions in Red, Green, Red-edge, or NIR bands may propagate into vegetation-index maps. Therefore, vegetation-index error and per-band reflectance consistency should be treated as agricultural proxy measures alongside conventional image-quality metrics.

The ×3 scale factor provided useful diagnostic information but remained alignment-sensitive because the 800 × 800 HR scene size is not exactly divisible by three. Although explicit alignment was applied, the ×3 results should be interpreted more cautiously than the exact ×2 and ×4 scale factors. Accordingly, ×2 and ×4 provide the clearest basis for model interpretation in this controlled benchmark.

Overall, the results indicate that UAV multispectral SR for precision agriculture should be optimized and evaluated with explicit attention to spectral consistency, not only spatial detail. The loss-term ablation further showed that adding SAM modestly improved spectral-direction consistency, whereas the VI term produced model-dependent effects. Parameter-matched and alternative-degradation analyses further indicated that the band-wise spatial advantage largely reflects model capacity rather than the reconstruction strategy, while the joint reconstruction mode itself contributed to spectral preservation beyond the spectral loss terms, and that the model rankings were robust to the degradation assumption and to input noise. The evaluation code has been released to support reproducibility. Within the controlled degradation and internal scene-level benchmark used here, residual CNN and channel-attention CNN approaches provided the most balanced performance for preserving both spatial structure and agricultural spectral usability in the AI Hub cabbage dataset. Future work should test these findings using grouped field/date splits, real multi-altitude or multi-sensor LR-HR image pairs, additional crops and sensors, uncertainty-aware spectral evaluation, and downstream agricultural tasks such as plant segmentation, stress detection, biomass estimation, and yield-related prediction.

Author Contributions: Conceptualization, W.J. and H.L.; methodology, W.J., J.-H.R. and H.L.; software, W.J.; validation, W.J., S.H.W. and J.-H.R.; formal analysis, W.J.; investigation, W.J. and S.H.W.; resources, S.H.W. and H.L.; data curation, W.J. and S.H.W.; writing—original draft preparation, W.J.; writing—review and editing, S.H.W., J.-H.R. and H.L.; visualization, W.J.; supervision, H.L.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Rural Development Administration, Republic of Korea (Project No. RS-2026-02283021).

Data Availability Statement: The multispectral image dataset analyzed in this study is available through the AI Hub public platform (https://aihub.or.kr/, accessed on 24 July 2026) under dataset code 71484. Access to the dataset requires registration and approval through the AI Hub portal (data description file v1.1). The preprocessed dataset, controlled downsampled images, model outputs, and supplementary statistics are available from the corresponding author upon reasonable request with appropriate institutional data-use agreements. To support reproducibility, the training and evaluation code (data-integrity filters, patch-quality thresholds, valid-pixel masking, metric implementations, and the fixed random seed) together with the scene-index file are publicly available on

GitHub at https://github.com/hwanzo758-creator/UAV-Multispectral-Super-Resolution (accessed on 28 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Weiss, M.; Jacob, F.; Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **2020**, *236*, 111402. https://doi.org/10.1016/j.rse.2019.111402.
2. Phang, S.K.; Chiang, T.H.A.; Happonen, A.; Chang, M.M.L. From satellite to UAV-based remote sensing: A review on precision agriculture. *IEEE Access* **2023**, *11*, 127057–127076. https://doi.org/10.1109/access.2023.3330886.
3. Omia, E.; Bae, H.; Park, E.; Kim, M.S.; Baek, I.; Kabenge, I.; Cho, B.-K. Remote sensing in field crop monitoring: A comprehensive review of sensor systems, data analyses and recent advances. *Remote Sens.* **2023**, *15*, 354. https://doi.org/10.3390/rs15020354.
4. Wang, Y.; Yang, Z.; Kootstra, G.; Khan, H.A. The impact of variable illumination on vegetation indices and evaluation of illumination correction methods on chlorophyll content estimation using UAV imagery. *Plant Methods* **2023**, *19*, 51. https://doi.org/10.1186/s13007-023-01028-8.
5. Hunt, E.R., Jr.; Daughtry, C.S. What good are unmanned aircraft systems for agricultural remote sensing and precision agriculture? *Int. J. Remote Sens.* **2018**, *39*, 5345–5376.
6. Li, Y.; Guo, S.; Jia, S.; Yan, Y.; Jia, H.; Zhang, W. Quantifying the Effects of UAV Flight Altitude on the Multispectral Monitoring Accuracy of Soil Moisture and Maize Phenotypic Parameters. *Agronomy* **2025**, *15*, 2137. https://doi.org/10.3390/agronomy15092137.
7. Aasen, H.; Honkavaara, E.; Lucieer, A.; Zarco-Tejada, P.J. Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: A review of sensor technology, measurement procedures, and data correction workflows. *Remote Sens.* **2018**, *10*, 1091. https://doi.org/10.3390/rs10071091.
8. Wang, P.; Bayram, B.; Sertel, E. A comprehensive review on deep learning based remote sensing image super-resolution methods. *Earth-Sci. Rev.* **2022**, *232*, 104110. https://doi.org/10.1016/j.earscirev.2022.104110.
9. Qi, Y.; Lou, M.; Liu, Y.; Li, L.; Yang, Z.; Nie, W. Advancing image super-resolution techniques in remote sensing: A comprehensive survey. *ISPRS J. Photogramm. Remote Sens.* **2026**, *231*, 68–100. https://doi.org/10.1016/j.isprsjprs.2025.10.024.
10. Lan, R.; Sun, L.; Liu, Z.; Lu, H.; Pang, C.; Luo, X. MADNet: A fast and lightweight network for single-image super resolution. *IEEE Trans. Cybern.* **2020**, *51*, 1443–1453. https://doi.org/10.1109/tcyb.2020.2970104.
11. Sun, L.; Liu, Z.; Sun, X.; Liu, L.; Lan, R.; Luo, X. Lightweight image super-resolution via weighted multi-scale residual network. *IEEE/CAA J. Autom. Sin.* **2021**, *8*, 1271–1280. https://doi.org/10.1109/jas.2021.1004009.
12. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*; IEEE: Piscataway, NJ, USA, 2021.
13. Anwar, S.; Barnes, N. Densely residual laplacian super-resolution. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 1192–1204. https://doi.org/10.1109/tpami.2020.3021088.
14. Li, W.; Li, J.; Li, J.; Huang, Z.; Zhou, D. A lightweight multi-scale channel attention network for image super-resolution. *Neurocomputing* **2021**, *456*, 327–337. https://doi.org/10.1016/j.neucom.2021.05.090.
15. Chen, H.; He, X.; Qing, L.; Wu, Y.; Ren, C.; Sheriff, R.E.; Zhu, C. Real-world single image super-resolution: A brief review. *Inf. Fusion* **2022**, *79*, 124–145. https://doi.org/10.1016/j.inffus.2021.09.005.
16. Liu, A.; Liu, Y.; Gu, J.; Qiao, Y.; Dong, C. Blind image super-resolution: A survey and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5461–5480. https://doi.org/10.1109/tpami.2022.3203009.
17. Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D.J.; Norouzi, M. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4713–4726. https://doi.org/10.1109/tpami.2022.3204461.
18. Huo, C.; Chen, K.; Zhang, S.; Wang, Z.; Yan, H.; Shen, J.; Hong, Y.; Qi, G.; Fang, H.; Wang, Z. When Remote Sensing Meets Foundation Model: A Survey and Beyond. *Remote Sens.* **2025**, *17*, 179. https://doi.org/10.3390/rs17020179.
19. Razzak, M.T.; Mateo-García, G.; Lecuyer, G.; Gómez-Chova, L.; Gal, Y.; Kalaitzis, F. Multi-spectral multi-image super-resolution of Sentinel-2 with radiometric consistency losses and its effect on building delineation. *ISPRS J. Photogramm. Remote Sens.* **2023**, *195*, 1–13. https://doi.org/10.1016/j.isprsjprs.2022.10.019.
20. Lanaras, C.; Bioucas-Dias, J.; Galliani, S.; Baltsavias, E.; Schindler, K. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 305–319. https://doi.org/10.1016/j.isprsjprs.2018.09.018.

21. Adão, T.; Hruška, J.; Pádua, L.; Bessa, J.; Peres, E.; Morais, R.; Sousa, J.J. Hyperspectral Imaging: A Review on UAV-Based Sensors, Data Processing and Applications for Agriculture and Forestry. *Remote Sens.* **2017**, *9*, 1110. https://doi.org/10.3390/rs9111110.
22. National Information Society Agency. *Field Crop (Cabbage,* etc.*) Growth Condition Data (노지작물(배추 등) 작황 데이터)*; Dataset code 71484; AI Hub: Daegu, Republic of Korea, 2022. Available online: https://aihub.or.kr/aihubdata/data/view.do?dataSetSn=71484 (accessed on 24 July 2026).
23. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 295–307. https://doi.org/10.1109/tpami.2015.2439281.
24. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops*; IEEE: Piscataway, NJ, USA, 2017. https://doi.org/10.1109/CVPRW.2017.151.
25. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018.
26. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; Loy, C.C. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*; Springer: Cham, Switzerland, 2018.
27. Chen, X.; Wang, X.; Zhou, J.; Qiao, Y.; Dong, C. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2023.
28. Chen, Z.; Zhang, Y.; Gu, J.; Kong, L.; Yang, X.; Yu, F. Dual aggregation transformer for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: Piscataway, NJ, USA, 2023.
29. Hsu, C.-C.; Lee, C.-M.; Chou, Y.-S. DRCT: Saving image super-resolution away from information bottleneck. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*; IEEE: Piscataway, NJ, USA, 2024; pp. 6133–6142.
30. Rouse, J.W., Jr.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring vegetation systems in the Great Plains with ERTS. In Proceedings of the Third Earth Resources Technology Satellite-1 (ERTS) Symposium, NASA SP-351, Goddard Space Flight Center, Washington, DC, USA, 10–14 December 1973; Volume 1, pp. 309–317.
31. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. https://doi.org/10.1016/s0034-4257(96)00072-7.
32. Barnes, E.M.; Clarke, T.R.; Richards, S.E.; Colaizzi, P.D.; Haberland, J.; Kostrzewski, M.; Waller, P.; Choi, C.; Riley, E.; Thompson, T.; et al. Coincident detection of crop water stress, nitrogen status and canopy density using ground-based multispectral data. In Proceedings of the Fifth International Conference on Precision Agriculture, Bloomington, MN, USA, 16–19 July 2000.
33. Zeng, Y.; Hao, D.; Huete, A.; Dechant, B.; Berry, J.; Chen, J.M.; Joiner, J.; Frankenberg, C.; Bond-Lamberty, B.; Ryu, Y.; et al. Optical vegetation indices for monitoring terrestrial ecosystems globally. *Nat. Rev. Earth Environ.* **2022**, *3*, 477–493. https://doi.org/10.1038/s43017-022-00298-5.
34. Lai, W.-S.; Huang, J.-B.; Ahuja, N.; Yang, M.-H. Fast and accurate image super-resolution with deep laplacian pyramid networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2599–2613. https://doi.org/10.1109/tpami.2018.2865304.
35. Kruse, F.A.; Lefkoff, A.B.; Boardman, J.W.; Heidebrecht, K.B.; Shapiro, A.T.; Barloon, P.J.; Goetz, A.F.H. The spectral image processing system (SIPS)—Interactive visualization and analysis of imaging spectrometer data. *Remote Sens. Environ.* **1993**, *44*, 145–163. https://doi.org/10.1016/0034-4257(93)90013-n.
36. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 8024–8035.
37. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
38. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
39. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
40. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. https://doi.org/10.1109/tip.2003.819861.

41. Han, L.; Zhao, Y.; Lv, H.; Zhang, Y.; Liu, H.; Bi, G.; Han, Q. Enhancing Remote Sensing Image Super-Resolution with Efficient Hybrid Conditional Diffusion Model. *Remote Sens.* **2023**, *15*, 3452. https://doi.org/10.3390/rs15133452.
42. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Jin, X.; Zhang, L. EDiffSR: An efficient diffusion probabilistic model for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2023**, *62*, 5601514. https://doi.org/10.1109/tgrs.2023.3341437.