Optics & Laser Technology 204 (2026) 116189

Contents lists available at ScienceDirect

Optics and Laser Technology

journal homepage: www.elsevier.com/locate/jolt

Full length article

From 30 m to 2.5 m: Cross-sensor reconstruction of landsat-8 imagery using a lightweight multi-source fusion network with downstream task assessment

Hongjie Xie[a], Qiang Zhang[a,*], Yi Xiao[b,*], Zishuo Wang[a], Jianwei Wen[c]

[a] Information Science and Technology College, Dalian Maritime University, Dalian, China
[b] School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China
[c] Inner Mongolia Meteorological Data Center, Hohhot, China

ARTICLE INFO

Keywords:
Landsat-8
Cross-sensor reconstruction
Light weight
Building extraction
Road extraction

ABSTRACT

Landsat-8 multispectral imagery at 30 m spatial resolution underpins a wide range of land-surface monitoring tasks. Nevertheless, it remains insufficient for fine-scale applications such as building and road extraction. To address the spatial-resolution limitation of using single sensor, we construct a unified triple-sensor remote-sensing benchmark at 2.5 m, leveraging auxiliary high-resolution data for reconstruction. PSMNet, a lightweight multi-scale fusion network, balances model complexity (630 K parameters) and inference efficiency (3.66 ms) while delivering well fidelity (PSNR 24.58 dB, SAM 2.69°). To quantify the utility of multi-source reconstruction beyond pixel-wise metrics, building and road extraction are adopted as downstream benchmarks. Experimental results show that PSMNet-reconstructed imagery achieves the highest downstream performance for both building extraction (Recall = 0.7096, IoU = 0.5157) and road extraction (Recall = 0.6830, IoU = 0.5231). It surpasses competing methods by clear margins. These advantages underscore the practical benefits of the proposed framework for large-scale, real-world remote-sensing reconstruction tasks. The dataset is publicly available at https://huggingface.co/datasets/namelesscoder/L8-S2-NAIP/tree/main.

1. Introduction

Launched in 2013, Landsat-8 provides 12 years of uninterrupted 30 m multispectral imagery and functions as the data infrastructure for land-surface dynamic monitoring. Its free and open data policy underpins a broad spectrum of remote-sensing applications, including earthquake damage assessment [31], quantitative estimation of soil content [3,27,41] and diverse mapping [9]. However, the 16 day revisit cycle of Landsat limits its ability to observe ephemeral or rapidly evolving ecological events, such as those critical for disaster response. To address this temporal sparsity, previous studies have often incorporated high temporal resolution MODIS imagery. Nevertheless, the coarse spatial resolution of MODIS, which ranges from 250 m to 1000 m, inherently prevents the retrieval of fine spatial scale land surface details [24,34].

Advanced remote-sensing applications, such as fire detection [33,39,40], building extraction, classification [2,10], and land cover [5], are critically limited by the scarcity of high spatial resolution data. This gap underscores an urgent need for technologies capable of generating high quality, detailed imagery from more readily available sources. The 30 m spatial resolution of Landsat-8 multispectral bands evidently fails to meet such requirements. Consequently, substantial research efforts focus on enhancing the spatial resolution of Landsat-8 imagery. The 15 m panchromatic band of Landsat-8 enables pansharpening as a direct means of increasing the spatial detail of multispectral imagery. Rahaman et al. [15] applied Smoothing Filter-based Intensity Modulation to generate 15 m multispectral composites and subsequently derived vegetation greenness and canopy water content. Mushore et al. [14] evaluated Brovey, ESRI, Gram-Schmidt and Intensity-Hue-Saturation pansharpening algorithms, systematically quantifying their respective contributions to urban land-use/land-cover classification and land surface temperature retrieval. Nevertheless, the spatial-resolution gain delivered by pansharpening is intrinsically bounded by the 15 m panchromatic band. The 15 m panchromatic band inherently limits pansharpened Landsat-8 multispectral imagery to a maximum spatial resolution of 15 m.

The European Space Agency's Sentinel-2 constellation reopens the prospect of transcending the 15 m barrier. The Harmonized Landsat and Sentinel-2 (HLS) initiative [8] supplies spatiotemporally aligned, analysis ready imagery, increasing the effective revisit frequency of medium-resolution data to 2–3 days. HLS, however, concentrates on radiometric and geometric harmonization; its pixels remain 30 m. Consequently, the product functions as a gap filling layer rather than a genuine spatial-resolution enhancement.

* Corresponding authors.
Email addresses: qzhang95@dlmu.edu.cn (Q. Zhang), yixiao@zzu.edu.cn (Y. Xiao).

https://doi.org/10.1016/j.optlastec.2026.116189
Received 20 January 2026; Received in revised form 11 May 2026; Accepted 11 August 2026
Available online 19 August 2026

Sigurdsson et al. [18] introduce Sentinel-2 Landsat 8 Sharpening, a model-based fusion framework that extends Sentinel-2 Sharpening Using a Reduced-Rank Method to jointly sharpen Sentinel-2 and Landsat-8 multispectral imagery, elevating every band to 10 m spatial resolution. Wang et al. [23] propose a multi-scale smoothing-sharpening filter that preprocesses the inputs for quality enhancement and transfers high frequency detail from Sentinel-2 to Landsat-8, yielding 10 m fusion products. Song et al. [19] address temporal continuity through the Temporal Sequence Image Fusion algorithm, which assimilates Landsat-8/9 (30 m) and Sentinel-2 (10 m) observations to generate a spectrally consistent, 10 m land surface reflectance time series. Such approaches often struggle to balance detail injection with spectral fidelity, frequently incurring spectral distortion or over smoothing.

Benefiting from the rapid advances in deep learning [7,35–38], especially in super-resolution [28–30], cross-sensor reconstruction has demonstrated significant potential in remote sensing. These approaches leverage complementary information from multiple sensors to enhance spatial detail, yielding outputs that preserve the spectral characteristics of the target sensor while incorporating spatial information from higher-resolution sources. Unlike traditional single-image super-resolution techniques that aim to recover hypothetical high-frequency details from a single sensor's observations, these methods focus on information integration rather than information recovery. Shao et al. [17] introduce the Extended Super-Resolution Convolutional Neural Network to synergize Landsat-8 Operational Land Imager (OLI) and Sentinel-2 Multi-Spectral Instrument (MSI) observations, generating a 10 m Landsat-like surface-reflectance composite through cross-sensor information transfer. By ingesting multi-temporal Sentinel-2 acquisitions as auxiliary predictors, the approach enhances the spatio-temporal coherence of the Landsat-8 time series while maintaining spectral consistency. Sambandham et al. [16] present a UNet-driven pipeline that couples Landsat-8 OLI and Sentinel-2 MSI imagery, reconstructing Landsat-8 observations at 10 m resolution and additionally incorporating the 15 m panchromatic band to refine spatial details. Their work represents a cross-sensor reconstruction approach rather than conventional super-resolution, as it fuses information from multiple sensors to overcome the physical resolution limitations of individual systems. Wang et al. [25] devise the Multi-Scale Cross-Attention Embedding CNN, a deep learning framework that addresses inter-sensor data gaps by reconstructing Landsat-8 imagery at enhanced spatial resolution. By introducing spatio-temporal embedding and multi-scale feature extraction, the method effectively integrates spectral differences between sensors, markedly improves spatial detail while preserving spectral fidelity, and fills data voids caused by cloud cover. Michel and Inglada [13] present Temporal Attention Multi-Resolution Fusion of Satellite Image Time Series, a generic deep-learning framework for cross-sensor reconstruction that generates spectral bands at the desired spatial resolution for any time step. Integrating a residual convolutional backbone with a temporal attention module, the model demonstrates strong performance in spatial enhancement through cross-sensor fusion rather than traditional super-resolution.

Nevertheless, these approaches still exhibit three main limitations. First, their output is fixed at the 10 m Sentinel-2 native pixel spacing, which is insufficient for applications demanding finer spatial detail. Second, sophisticated components such as attention mechanisms markedly raise computational complexity, restricting operational deployment. Third, evaluations rely predominantly on pixel level metrics, while downstream task validations (e.g., building or road extraction) remain scarce, leaving the practical utility of the methods largely unverified.

In summary, although existing methods for remote sensing imagery have achieved certain advances in spatial resolution enhancement, they still suffer from conspicuous shortcomings: the absence of real high resolution labels, inefficient model architectures, and the lack of downstream task validation. To address these gaps, we construct a unified three-source dataset targeting a 2.5 m Landsat-8 imagery and train with Sentinel-2 MSI and Landsat-8 OLI samples, while adopting the National Agriculture Imagery Program (NAIP) 2.5 m imagery as ground truth. A lightweight multi-scale fusion network, Poly Sensor Multi-scale Net (PSMNet), is proposed to reconstruct the Red, Blue, Green, Near Infrared (RGBN) bands of Landsat-8 to 2.5 m, and its superiority and practicality are demonstrated on realistic tasks such as building and road extraction. Our contributions are summarized in the following three aspects:

1. **Dataset**: To the best of our knowledge, this study presents the first 2.5 m spatial-resolution-aligned dataset that jointly incorporates Landsat-8 OLI, Sentinel-2 MSI, and NAIP imagery. It bridges the data gap for multi-scale and cross-sensor learning; the dataset is publicly available at https://huggingface.co/datasets/namelesscoder/L8-S2-NAIP/tree/main.
2. **Model**: We develop PSMNet and its core PSMBlock, which efficiently fuses multi-source and multi-scale information from Landsat-8 OLI and Sentinel-2 MSI to enhance spatial-detail recovery. The architecture achieves a favorable trade-off among parameters, computational cost, and inference speed: only 630 K parameters, 12.716 GFLOPs, and 3.66 ms runtime, outperforming most existing methods.
3. **Validation**: We incorporate downstream remote-sensing tasks: building and road extraction, to validate reconstruction utility. Experiments show that imagery enhanced by PSMNet yields the highest Recall, F1 and IoU scores on both extraction tasks, demonstrating that the proposed method delivers not only superior visual quality, but also tangible practical advantages.

2. Data description and preprocessing

2.1. Landsat-8/ sentinel-2/ NAIP data

To improve the spatial resolution of the Landsat-8 RGBN bands to 2.5 m, we collect cross-sensor imagery for joint reconstruction. Landsat-8 OLI data consist of the 30 m RGBN multispectral bands and the 15 m panchromatic band; Sentinel-2 MSI provides the native 10 m RGBN bands, subsequently upsampled to 2.5 m in this study. NAIP aerial imagery, originally at 0.6 m ground-sample distance, is acquired in its RGBN configuration and resampled to 2.5 m to serve as ground truth for training and evaluation. All datasets are extracted on the Google Earth Engine (GEE) platform.

2.2. Cross-sensor data acquisition and preliminary alignment

To establish a 2.5 m reference benchmark over the contiguous United States, this study first batches cloud-free (≤10%) quadruple-image pairs acquired within a 10-day temporal window for the period 2020–2023 on the GEE platform. Fig. 1 shows the distribution of the 22 selected sites.

Fig. 2 illustrates the end-to-end workflow for data acquisition and preliminary processing. After co-registering all layers to guarantee identical sampling locations, we impose a temporal filter that retains only Sentinel-2 and Landsat-8 images acquired within ±10 days of the NAIP acquisition date. A cloud filter simultaneously excludes scenes whose cloud fraction exceeds 10%. Finally, to reconcile the differing digital number ranges between NAIP and Landsat-8, a histogram matching is applied to calibrate the NAIP data range.

Table 1 summarizes the key specifications of the four co-registered remote sensing datasets utilized in this section. The foundational data source is the NAIP imagery, which is resampled from its native 0.6 m resolution to a consistent 2.5 m benchmark grid for the entire dataset. Satellite-derived surface reflectance data are incorporated through Sentinel-2 Level-2A (10 m spatial resolution) and Landsat-8 Level-2 Tier 1 (30 m spatial resolution) products. To augment the spatial detail, the panchromatic band from the Landsat-8 Tier 1 Top-of-Atmosphere product (15 m resolution) is also integrated. A critical preprocessing step involved spectral harmonization; all four sources are carefully processed

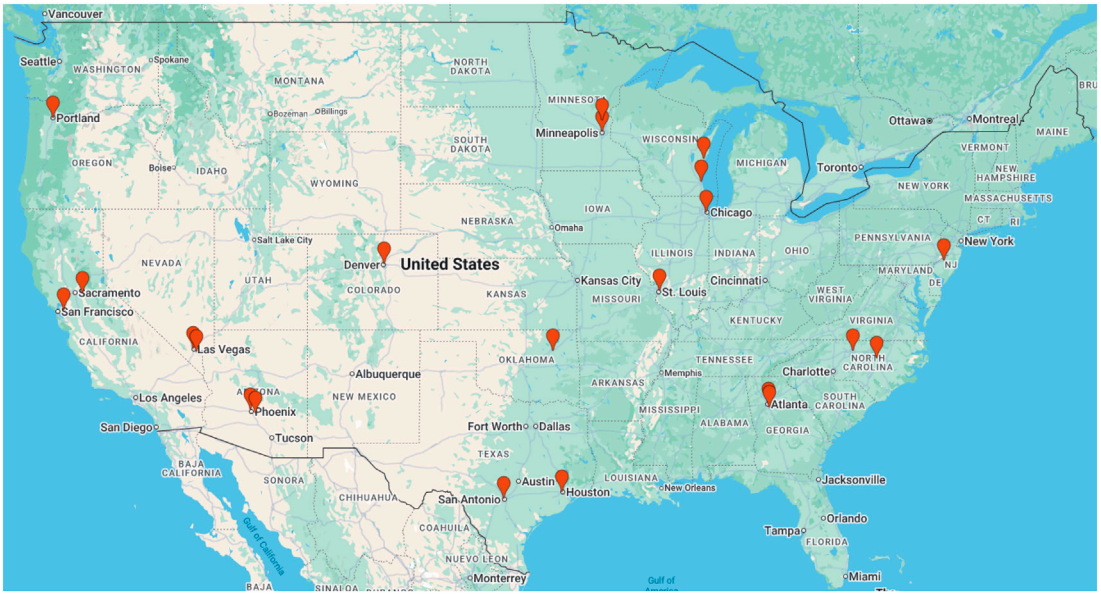


Fig. 1. Geographic distribution of the selected data points across the contiguous United States and adjacent regions. The locations of the data points (indicated by red pins) show a wide geographic spread, encompassing major urban areas and regions from the West Coast to the East Coast.

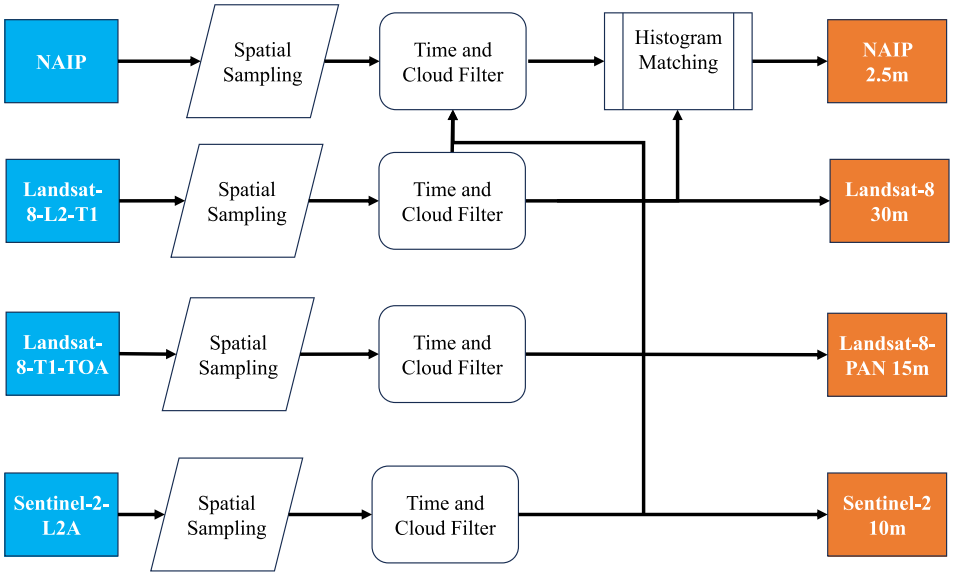


Fig. 2. Parallel preprocessing flowchart for NAIP, Landsat 8, and Sentinel-2 imagery, detailing common and sensor-specific steps to produce coregistered, quality-filtered data at native resolutions.

Table 1
Specifications of the cross-sensor data after co-registration and preprocessing.

Data	Resolution	Bands
NAIP	0.6 m → 2.5 m	RGBN
Sentinel-2-L2A	10 m	RGBN
Landsat-8-L2-T1	30 m	RGBN
Landsat-8-T1-TOA	15 m	PAN

to provide consistent RGBN spectral bands, ensuring spectral comparability essential for robust multi-sensor analysis. This multi-scale data framework, combining high resolution NAIP data with medium resolution multispectral and panchromatic satellite data, provides a solid foundation for generating a consistent 2.5 m reference benchmark.

2.3. Tile generation

Section 2.2 yields 22 large-scene quadruplets that surpass GPU memory. The open-source Sentinel-2 model [4] also enforces strict tile size limits. Aligned cropping is therefore applied to generate medium patches before the Sentinel-2 spatial resolution is enhanced. Non-overlapping sliding windows slice each large scene into tiles. After visual screening for cloud shadows, striping artifacts, and land–water boundary anomalies, 905 medium tiles remain. These are split 1:1:8 into 91 test, 91 validation, and 723 training patches; details are listed in Table 2.

2.4. Training and testing patch generation

To mitigate the redundancy of flat-region patches (farmland, bare soil) and the shortage of high frequency urban samples, a Texture-Score Probability Retention (TSPR) strategy is introduced to modulate patch

Table 2
Spatially aligned tile sizes for each data source after cropping.

Data Source	Resolution	Tile Size	Ratio to NAIP
NAIP	2.5 m	1536 × 1536	1×
Sentinel-2-L2A	2.5 m	1536 × 1536	4×
Landsat-8-L2-T1	15 m	256 × 256	6×
Landsat-8-T1-TOA	30 m	128 × 128	12×

selection probability during random cropping. The score is computed on the 2.5 m NAIP tile and combines three terms: gray level entropy, gradient energy, and Canny edge density. Gray level entropy is derived in three steps: the NAIP tile is first converted to a gray-level image. Its histogram is computed as $\{h_k\}_{k=0}^{255}$, and the probability of each gray value is then calculated as $p_k = \frac{h_k}{\sum_{i=0}^{255} h_i}$. Finally, gray-level entropy is evaluated as:

$$Entropy = -\sum_{k=0}^{255} \mathrm{p}_k \log_2(\mathrm{p}_k + \epsilon), \quad \epsilon = 10^{-8} \tag{1}$$

Gradient energy is computed by applying Sobel operators to the gray-level image to extract horizontal gradients and vertical gradients, and then taking the mean of the resulting gradient magnitudes. The detailed formulation is given below:

$$g_x = \mathrm{Sobel_x} \cdot \mathrm{gray}, \quad g_y = \mathrm{Sobel_y} \cdot \mathrm{gray}, \quad G = \frac{\sum(g_x^2 + g_y^2)}{\mathrm{N}} \tag{2}$$

Canny edge density is computed as follows: Canny detection is first applied to the grayscale image with a low threshold of 50 and a high threshold of 150 to obtain a binary edge map $E(u, v) \in \{0255\}$, where $u \in [0, H-1]$ and $v \in [0, W-1]$ denote the pixel coordinates and $H \times W$ is the image size. The Canny density is then calculated by:

$$C = \frac{\sum_{u=0}^{H-1} \sum_{v=0}^{W-1} E(u, v)}{255 \cdot H \cdot W} \tag{3}$$

The weighting coefficients and selection threshold are empirically tuned on 100 validation patches and are finally set to: $\mathrm{w_{ent}} = 1.0, \mathrm{w_{grad}} = 10.0, \mathrm{w_{canny}} = 15.0, \mathrm{threshold} = 5.0$.

Hence, the score is calculated as:

$$score = \mathrm{w_{ent}} \cdot Entropy + \mathrm{w_{grad}} \cdot G + \mathrm{w_{canny}} \cdot C \tag{4}$$

Patches whose $score \leq$ threshold are excluded from the dataset. To preserve diversity, the TSPR strategy is invoked with a fixed probability p at every random crop. If fewer than 30% of the images added so far bypassed screening, the next tile is subjected to TSPR with probability p; once the unscreened fraction reaches or exceeds 30%, TSPR is activated automatically for every subsequent crop. The probability p is set to 0.7.

For training and validation construction, each medium patch is randomly divided into 32 small patches. After cropping, the spatial dimensions are 144 × 144 pixels for NAIP and Sentinel-2, 24 × 24 for Landsat-8 PAN, and 12 × 12 for Landsat-8 multispectral bands. Under the TSPR policy, this yields 2 912 validation patches and 23 136 training patches.

For the test set, TSPR is disabled to ensure comprehensive and unbiased evaluation; only random cropping is applied. Each medium patch yields 32 tiles, resulting in patch dimensions of 768 × 768 pixels for NAIP and Sentinel-2, 128 × 128 for Landsat-8 PAN, and 64 × 64 for Landsat-8 multispectral bands. In total, 2 912 test patches are generated.

3. Method

3.1. Overall architecture

As shown in Fig. 3, this work proposes an efficient cross-sensor reconstruction network PSMNet. PSMNet composes of three stages: (1) shallow-feature extraction via a single 3 × 3 convolution on the low-resolution input; (2) deep-feature extraction by stacking multiple attention-based PSMBlocks for fusion and enhancement; (3) a reconstruction module that employs a 3 × 3 convolution to project the refined features into the target high-resolution space.

Before feeding the network, the spatial resolution of Y_1, Y_2 is raised to 2.5 m via bilinear interpolation, yielding $Y_1 \uparrow, Y_2 \uparrow$. A 3 × 3 convolution then extracts the initial feature map:

$$x_{\mathrm{SF}} = \mathrm{Conv}_{3\times3}\left(\mathrm{Concat}\left(X, Y_1 \uparrow, Y_2 \uparrow\right)\right) \tag{5}$$

Let x_{SF} denote the shallow feature and Concat denote channel-wise concatenation. x_{SF} is then fed to a stack of N PSMBlocks for deep extraction, denoted by F_{DF}, yielding the deep feature x_{DF}:

$$x_{\mathrm{DF}} = F_{\mathrm{DF}}(x_{\mathrm{SF}}) \tag{6}$$

The final image pred is reconstructed by aggregating the deep feature x_{DF} with the shallow feature x_{SF} via a residual connection, as expressed in Eq. 7. This residual pathway mitigates gradient vanishing and preserves low frequency information throughout training.

$$\mathrm{pred} = \mathrm{Conv}_{3\times3}(x_{\mathrm{SF}} + x_{\mathrm{DF}}) \tag{7}$$

3.2. PSMBlock

As illustrated in Fig. 4, a PSMBlock comprises three components: GroupNorm (GN), Multi-Scale Chunk Aggregation (MSCA) and Convolutional Channel Mixer (CCM). To accommodate the distinct feature distributions across multi-source channels, GN replaces LayerNorm with the group number set to 4. The PSMBlock is formulated as:

$$\begin{aligned} x_{\mathrm{t}} &= \mathrm{MSCA}(\mathrm{GN}(x_{\mathrm{in}})) + x_{\mathrm{in}} \\ x_{\mathrm{out}} &= \mathrm{CCM}(\mathrm{GN}(x_{\mathrm{t}})) + x_{\mathrm{t}} \end{aligned} \tag{8}$$

where x_{in} is the input feature map; x_{t} is an intermediate result obtained by aggregating multi-scale information from the normalized input and adding it back via a skip connection; and x_{out} is the final output after further channel mixing and another residual connection.

3.2.1. Multi-scale chunk aggregation

The MSCA module extracts and aggregates multi-scale information at three discrete scales; its structure is shown in Fig. 4. The input channels are first split into three branches. Branch 1 applies a 3 × 3 depth-wise convolution for small scale features. Branch 2 downsamples the feature map to half its size via max pooling, processes it with a 3 × 3 depth-wise convolution, and upsamples it back to the original resolution using nearest neighbour interpolation, yielding medium scale cues. Branch 3 captures large scale context directly with a 7 × 7 depth-wise convolution. Finally, a 1 × 1 convolution fuses the three scaled outputs. This procedure is expressed as:

$$\begin{aligned} y_1, y_2, y_3 &= \mathrm{Chunk}(y_{\mathrm{in}}) \\ y_{\mathrm{out1}} &= \mathrm{DwConv}_{3\times3}(y_1) \\ y_{\mathrm{out2}} &= \mathrm{Upsample}\left(\mathrm{DwConv}_{3\times3}\left(\mathrm{Maxpooling}(y_2)\right)\right) \\ y_{\mathrm{out3}} &= \mathrm{DwConv}_{7\times7}(y_3) \\ y_{\mathrm{out}} &= \mathrm{Conv}_{1\times1}\left(\mathrm{Concat}(y_{\mathrm{out1}}, y_{\mathrm{out2}}, y_{\mathrm{out3}})\right) \end{aligned} \tag{9}$$

where y_{in} is the input feature map; y_1, y_2 and y_3 are its three equally split channel groups, corresponding to the three branches; y_{out1}, y_{out2} and y_{out3} are the outputs of the small scale, medium scale, and large scale processing branches, respectively; and is the final aggregated output of the MSCA module.

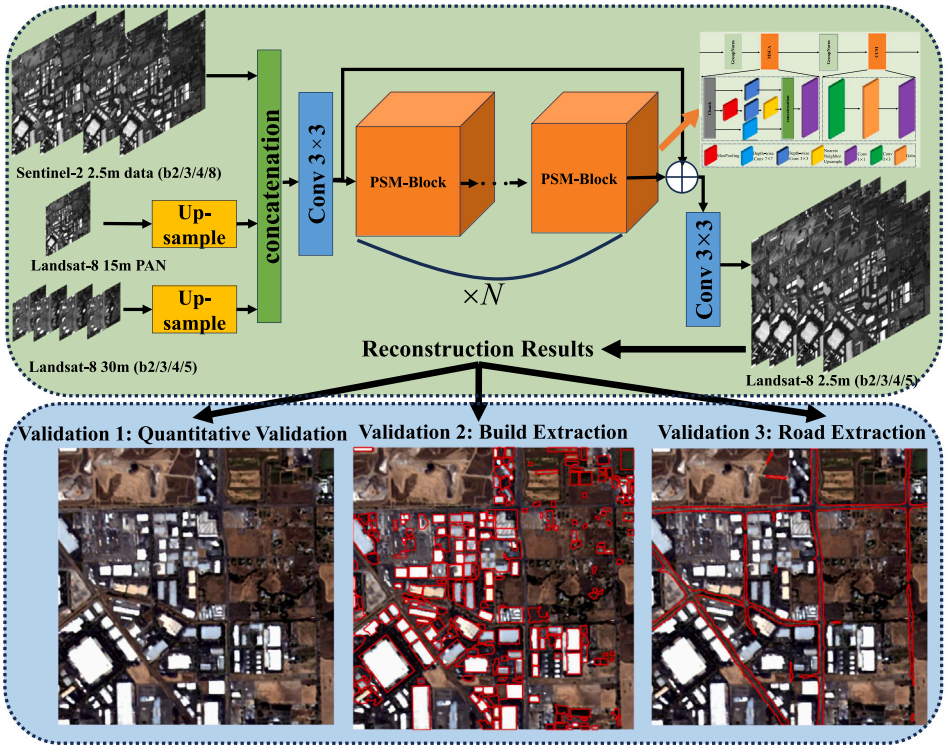


Fig. 3. Schematic of the proposed deep learning model for enhancing spatial-spectral information from cross-sensor data. The core network fuses Sentinel-2 (2.5 m), Landsat-8 panchromatic (15 m), and Landsat-8 multispectral (30 m) inputs. The fused features are processed through a series of customized PSMBlocks to generate high-resolution (2.5 m) output. The bottom panels demonstrate the model's application, in building extraction and road extraction tasks, validating its performance for both enhancement and semantic feature extraction.

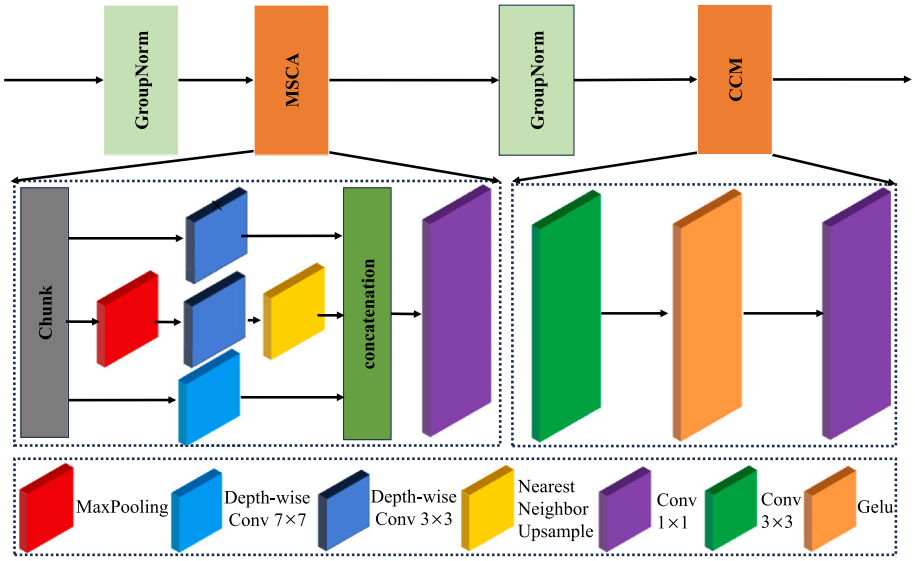


Fig. 4. Schematic of the PSMBlock. The module consists of a GroupNorm (GN) layer, a Multi-Scale Chunk Aggregation (MSCA) module, and a Convolutional Channel Mixer (CCM). The MSCA module employs a parallel structure (shown expanded) to extract and fuse multi-scale features via pooling and depth-wise convolutions, which are then integrated by the CCM. The data flow is indicated by connecting arrows.

3.2.2. Convolutional channel mixer

This work adopts the Convolutional Channel Mixer (CCM) [20] for channel-wise feature shuffling. CCM consists of one 3 × 3 convolution followed by one 1 × 1 convolution: the former encodes local spatial context and expands the channel dimension to twice the input, enabling richer inter channel information exchange, while the latter projects the channels back to the original dimension. A Gelu activation [6] is applied in the intermediate hidden layer to introduce non-linearity. This procedure is described by the following equation:

$$z_{\text{out}} = \text{Conv}_{1\times1}\left(\text{Gelu}\left(\text{Conv}_{3\times3}\left(z_{\text{in}}\right)\right)\right) \tag{10}$$

where z_{in} is the input feature map and z_{out} is the output feature map of the CCM module.

3.3. Loss function and parameter settings

This section details the loss function, optimization algorithm, learning-rate schedule, and other key training hyper-parameters. L1Loss is adopted as the training objective:

$$\mathrm{L1Loss}(\mathrm{pred}, Z) = \frac{\sum_{i=1}^{n} |\mathrm{pred}_i - Z_i|}{n} \tag{11}$$

where pred is the reconstructed Landsat-8 image, Z is the NAIP Ground Truth (GT). Optimization is performed with the Adam optimizer, whose momentum and adaptive learning rate estimates yield fast, stable convergence; the initial learning rate is set to 1×10^{-4}. Training is further refined by the Reduce Learning Rate On Plateau scheduler: after every epoch the validation's Loss is monitored, and if no improvement greater than 1×10^{-4} is observed for 3 consecutive epochs the learning rate is halved (factor = 0.5) until a lower bound of 1×10^{-6} is reached. Gradient norm clipping (max = 1.0) is applied to prevent explosion, and the total budget is fixed at 100 epochs. These settings are used for all deep learning methods compared in Section 4, except interpolation baselines. PSMNet stacks 8 PSMBlocks to balance accuracy and efficiency, and follows the channel configuration of SAFMN [20], ShuffleMixer [21] and SPAN [22] by fixing the number of channels to 60.

4. Experiments

4.1. Quantitative evaluation

To quantitatively evaluate the reconstruction results, we adopted three widely used metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Spectral Angle Mapper (SAM). The definitions of the three quantitative metrics are given as follows:

$$\begin{aligned}
\mathrm{PSNR}(x, y) &= 10 \log_{10} \left(\frac{\mathrm{MAX}_\mathrm{I}^2}{\mathrm{MSE}} \right) \\
\mathrm{SSIM}(x, y) &= \frac{(2\mu_\mathrm{x}\mu_\mathrm{y} + C_1)(2\sigma_\mathrm{xy} + C_2)}{(\mu_\mathrm{x}^2 + \mu_\mathrm{y}^2 + C_1)(\sigma_\mathrm{x}^2 + \sigma_\mathrm{y}^2 + C_2)} \\
\mathrm{SAM}(x, y) &= \cos^{-1} \left(\frac{x \cdot y}{|x||y|} \right)
\end{aligned} \tag{12}$$

where x and y denote the predicted image and the GT image, respectively. In the PSNR formula, MAX_I is the maximum possible pixel value of the image and MSE is the mean squared error between x and y. In the SSIM formula, μ_x and μ_y are the local pixel means, σ_x^2 and σ_y^2 are the variances, σ_xy is the cross-covariance, C_1 and C_2 are small constants added to stabilize the division. In the SAM formula, $x \cdot y$ represents the dot product of the spectral vectors, and $|x|$ denotes the Euclidean norm (magnitude) of a vector.

For the two downstream tasks of building extraction and road extraction, Precision, Recall, F1 and IoU are employed for quantitative evaluation. These four metrics can be formulated as below:

$$\begin{aligned}
\mathrm{Precision} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \\
\mathrm{Recall} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \\
\mathrm{F1} &= \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \\
\mathrm{IoU} &= \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}}
\end{aligned} \tag{13}$$

where TP (True Positive) denotes the number of pixels correctly predicted as the target class (e.g., building or road); FP (False Positive) represents the number of pixels incorrectly predicted as the target class when they actually belong to the background; and FN (False Negative) indicates the number of target-class pixels that are missed by the prediction. Based on these terms, Precision measures the correctness of the predicted positives, Recall quantifies the detection completeness, F1 provides their harmonic mean, and IoU evaluates the spatial overlap between the prediction and the GT.

Table 3
Model Parameters, FLOPs, and Average inference Time.

Model	Parameters (K)	FLOPs (G)	Peak VRAM (MB)	Average Inference Time (ms)
CARN	722	16.962	126.59	1.9422
EDSR	1008	22.316	24.53	1.8012
SPAN	638	13.209	118.60	11.7693
SAFMN	620	12.716	48.33	6.8209
ShuffleMixer	540	11.074	61.41	12.2730
PSMNet	630	12.716	66.23	3.6640

Table 4
Inference performance comparison on 1024×1024 patches.

Model	Parameters(K)	FLOPs (G)	Peak VRAM (MB)	Average Inference Time (s)
CARN	722	754.79	6058.91	0.1218
EDSR	1080	1130.20	1256.26	0.1166
SPAN	638	667.96	4874.65	0.0975
SAFMN	620	643.04	1981.33	0.1768
ShuffleMixer	540	560.00	1876.78	0.2910
PSMNet	630	651.65	2009.28	0.2566

Table 5
Inference performance comparison on 2048×2048 patches. "–" denotes out-of-memory or unsupported configuration.

Model	Parameters	FLOPs (T)	Peak VRAM (MB)	Average Inference Time (s)
CARN	722	–	–	–
EDSR	1080	4.51	5012.39	0.4586
SPAN	638	2.67	19,430.65	0.3733
SAFMN	620	2.57	7897.33	0.7277
ShuffleMixer	540	2.24	7432.78	1.1918
PSMNet	630	2.61	7925.28	0.9996

4.2. A comparative analysis with lightweight deep learning models

To validate the efficiency and quality trade-off of PSMNet in practical deployment scenarios, this study selected five efficient deep learning models with parameters $\leq$ 1 M—EDSR [12], CARN [1], SAFMN [20], ShuffleMixer [21], and SPAN [22] as comparisons, with Bicubic interpolation serving as the baseline. All methods were evaluated under identical conditions on an RTX 4090 GPU, with an input size of $9 \times 144 \times 144$ and a batch size of 1. FLOPs are measured using the thop library, with 10 warm up runs and 100 timed repetitions. Additionally, we measured the Peak VRAM usage during the inference. The results are summarized in Table 3. It can be observed that the proposed model has a parameter count comparable to that of SAFMN, while reducing inference time by 46%, demonstrating its deployment friendly nature.

In practical applications, a larger patch size is typically chosen to improve algorithm efficiency. We measured the peak VRAM consumption during inference on large-scale remote-sensing patches, specifically at resolutions of 1024×1024 and 2048×2048 pixels (Tables 4 and 5).

Due to hardware limitations (RTX 4090, 24 GB VRAM), these models encountered out-of memory errors during inference and could not complete the test. Key Observations: Competitive Memory Efficiency: Our method demonstrates comparable peak VRAM usage to SAFMN and ShuffleMixer (all around 2 GB for 1024×1024 and 7.5–8 GB for 2048×2048).

4.3. Quantitative evaluation of cross-sensor reconstruction and visualization

Table 6 reports the averaged results over the 2912 test patches. Bicubic interpolation, which performs pure resampling without any information fusion, ranks last in PSNR, SSIM and SAM. Classical residual-stacking models like CARN and EDSR deliver 24.06 dB and 24.39 dB

Table 6
Quantitative evaluation of cross-sensor reconstruction performance. Red indicates the best performance, and blue indicates the second-best performance.

Model	PSNR	SSIM	SAM
Bicubic	15.8865	0.5809	14.3197
CARN	24.0616	0.7011	3.0129
EDSR	24.3959	0.7168	2.8167
SPAN	24.1124	0.7037	2.9783
SAFMN	24.5012	0.7212	2.7533
ShuffleMixer	24.5157	0.7239	2.7339
PSMNet (Ours)	24.5861	0.7225	2.6965

PSNR, yet their fixed 3 × 3 kernels provide limited receptive fields and fail to capture multi-scale geographic objects (e.g., large building blocks vs. small detached houses).

SPAN introduces free parameters, attention and re-parameterisation, but the overly simplistic attention maps cannot model complex spatial dependencies, and the re-param step incurs a slight accuracy drop (24.11 dB PSNR), even underperforming the lighter EDSR.

SAFMN employs multi-scale max-pooling plus depth-wise convolutions, attaining 24.50 dB PSNR and 2.75° SAM, yet its upsampling relies on nearest neighbour interpolation, which amplifies high frequency noise and produces edge jaggies, thereby explaining its sub-optimal SSIM (0.7212).

ShuffleMixer adopts 7 × 7 large kernel depth-wise convolutions and channel shuffling, achieving the second best PSNR (24.51 dB) and best SSIM (0.7239). The large kernel and shuffle, however, raise the inference time to 12.27 ms which is the longest among all methods.

The proposed PSMNet strikes the best overall trade-off: it records the highest PSNR (24.58 dB) and lowest SAM (2.69°), while its SSIM (0.7225) is within 0.0014 of the best. With 630 k parameters and 12.72 G FLOPs (comparable to SAFMN), PSMNet infers in only 3.66 ms which is 1.86 × and 3.35 × faster than SAFMN and ShuffleMixer, respectively demonstrating superior balance between performance and efficiency.

Fig. 5 visualises a typical dense urban area (768 × 768, RGB composite) together with two 100 × 100 Regions of Interest (ROI) enlarged 2× and placed at the top-left and top-right corners. The visual inspection corroborates the quantitative rankings: bicubic is the blurriest, with indistinct building edges and severe detail loss in both ROIs.

CARN and EDSR improve sharpness slightly, yet the magnified patches still exhibit smoothed, fuzzy and uncertain edge responses. SPAN performs marginally better in the top-left ROI, but its free parameter's attention and re-parameterisation loss fail to restore coherent building outlines in the top-right ROI.

SAFMN captures multi-scale cues and strengthens the top-right building edges, yet maxpooling followed by nearest neighbour upsampling introduces conspicuous jagged artifacts and noticeable brightness blotches which are likely caused by over-enhanced local extrema.

ShuffleMixer partly recovers edges in the top-left ROI, but the top-right patch still contains ambiguous contours. These defects stem from its coarse, single stage "multi-scale" scheme; although the 7 × 7 depth-wise kernel enlarges the receptive field, it lacks adaptive, fine grained aggregation.

PSMNet produces the most visually plausible result. The embedded Multi-Scale Chunk Aggregation exploits hierarchical context and infers the most consistent and structurally coherent building edges, yielding the closest match to the real-world reference among all competitors.

4.4. Building extraction results and analysis

The ultimate value of reconstruction transcends mere pixel level fidelity; it is truly gauged by its tangible benefits to downstream remote-sensing applications. Among these, building extraction stands out as a critical task with profound implications for urban planning [32], population estimation, disaster management, and sustainable development. The accuracy of building extraction relies on high-quality imagery, as it demands precise spatial detail to delineate building contours and robust spectral fidelity to distinguish architectural materials from similar looking impervious surfaces. This heavy reliance makes it an ideal benchmark for evaluating the practical utility of reconstruction techniques. Consequently, this section quantitatively and visually assesses how different reconstruction methods impact real world building extraction performance on the AI-Earth platform (https://engine-aiearth.aliyun.com/), validating whether enhanced spatial resolution translates into superior application outcomes. To implement the building extraction evaluation, we first exported the reconstructed images in GeoTIFF format with preserved geospatial projection information and uploaded them to the AI-Earth platform. Subsequently, we navigated to the *Data Analysis* page, selected the *Single-Object Extraction* tool under the AIE-SEG module, and specifically initiated the **Building Extraction** pipeline for automated segmentation and quantitative assessment.

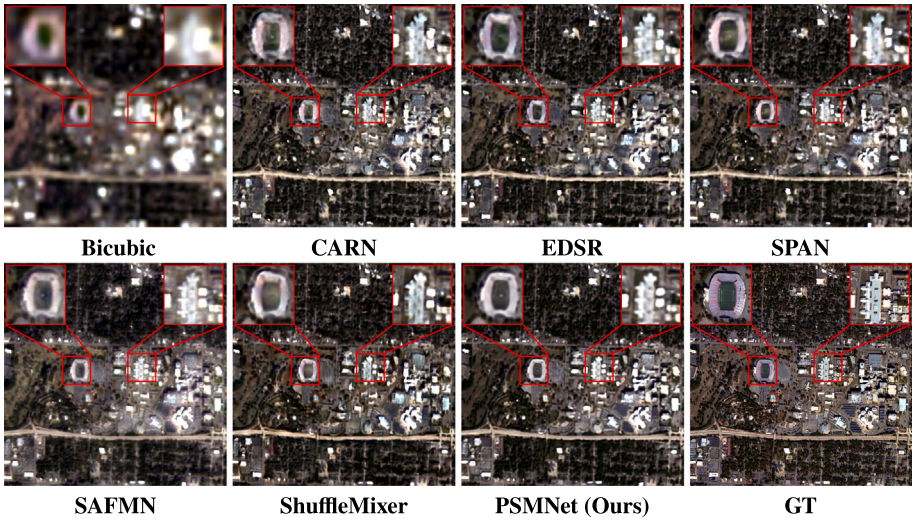


Fig. 5. The proposed PSMNet yields the most visually accurate reconstruction. This figure compares the reconstruction results of various methods (including Bicubic, CARN, EDSR, etc.) on a dense urban image, with two 2× magnified ROIs.

Table 7
Quantitative evaluation of building extraction. Red indicates the best performance, and blue indicates the second-best performance.

Model	Precision	Recall	F1	IoU
Bicubic	0.5512	0.5169	0.5140	0.3459
CARN	0.6634	0.3819	0.4847	0.3199
EDSR	0.6527	0.4733	0.5487	0.3781
SPAN	0.6472	0.4136	0.5046	0.3375
SAFMN	0.6660	0.6650	0.6655	0.4987
ShuffleMixer	0.6610	0.5868	0.6217	0.4511
PSMNet (Ours)	0.6537	0.7096	0.6805	0.5157

Table 8
Areas corresponding to different reconstruction methods in building extraction. Red indicates the best performance, and blue indicates the second-best performance.

Model	Intersection_Area	Prediction_Area	Union_Area	True_Area
Bicubic	6,400,943	12,520,433	18,501,877	12,382,387
CARN	4,729,203	7,127,761	14,780,945	12,382,387
EDSR	5,861,176	8,978,754	15,499,964	12,382,387
SPAN	5,121,481	7,913,027	15,173,933	12,382,387
SAFMN	8,235,433	12,364,352	16,511,306	12,382,387
ShuffleMixer	7,266,765	10,992,508	16,108,130	12,382,387
PSMNet (Ours)	8,786,857	13,441,265	17,036,795	12,382,387

Quantitative results are reported in Table 7. Overall, segmentation models fed with higher quality reconstructed imagery achieve superior scores, confirming the importance of enhanced spatial resolution for land-cover mapping. Two noteworthy phenomena emerge from the table: 1. High Precision yet low Recall of CARN: CARN achieves high Precision (0.6634) but the lowest Recall (0.3819), yielding poor F1 and IoU. This apparent contradiction stems from its limited reconstruction quality. As noted in Section 4.3, CARN images suffer from blurred edges and missing details. When fed to the segmentation model, this ambiguity forces an extremely conservative strategy—only the most evident, highly confident regions (usually building interiors) are labeled as buildings, while complex edges are discarded. Consequently, the predicted building area is far smaller than the reference (Table 8: 7.12 M vs. 12.38 M), giving high Precision but severe omission (low Recall). 2. Misleading metrics of Bicubic interpolation: although Bicubic outperforms CARN and SPAN in several indicators, visual inspection (Fig. 6) and area statistics (Table 8) reveal this superiority to be illusory. The uniformly blurred reconstruction lacks high frequency cues, preventing the segmentation model from discriminating fine textures. Consequently, large spectrally homogeneous areas are erroneously labeled as buildings, producing an inflated prediction area (12.52 M, close to the reference) and high Recall. Yet this “bulk guess” behaviour yields coarse, inaccurate masks with low practical value; the poor Precision (0.5512) confirms extensive false positives.

EDSR, benefiting from its deeper stack of residual blocks, delivers a respectable PSNR of 24.40 dB; however, the rigid 3×3 kernels restrict the receptive field, yielding slightly blurred edges. Transferred to the segmentation stage, this limitation surfaces as coarse boundary localization: the model roughly detects building areas yet misplaces their outlines. Hence, EDSR’s predicted footprint exceeds CARN’s, but its intersection area (5.86 M) and IoU (0.3781) remain low, evidencing widespread boundary errors and omissions. SPAN, whose reconstruction quality is marginally inferior (PSNR 24.11 dB), trails EDSR across all extraction metrics (F1 0.5046, IoU 0.3375), further confirming the direct link between reconstruction fidelity and downstream performance.

SAFMN and ShuffleMixer push reconstruction performance above 24.50 dB PSNR, and this gain is immediately reflected in the extraction results. SAFMN’s multi-scale features produce sharper images and highly accurate segmentation, closely matching reference data with top precision (0.6660) and balanced recall (0.6650). This confirms that SAFMN reconstructions supply rich, trustworthy cues. ShuffleMixer, with its large kernel depth-wise convolutions, also performs strongly, yet its Recall (0.5868) and F1 (0.6217) lag slightly behind SAFMN. As visualized in Section 4.3, structural errors around complex edges mislead the segmentation model, producing a smaller predicted area and indicating minor deficiencies in structural coherence.

The superior PSNR and SAM achieved by the proposed PSMNet in the reconstruction are fully echoed in the building extraction experiment, confirming that its reconstructions deliver the highest practical value.

1. Best overall extraction metrics: Table 7 shows that PSMNet attains the highest Recall (0.7096), IoU (0.5157) and F1 (0.6805). The leading Recall indicates that the segmentation model locates the largest fraction of real buildings, markedly reducing omission errors, while the top IoU demonstrates that the predicted footprints best match the true shapes and positions.

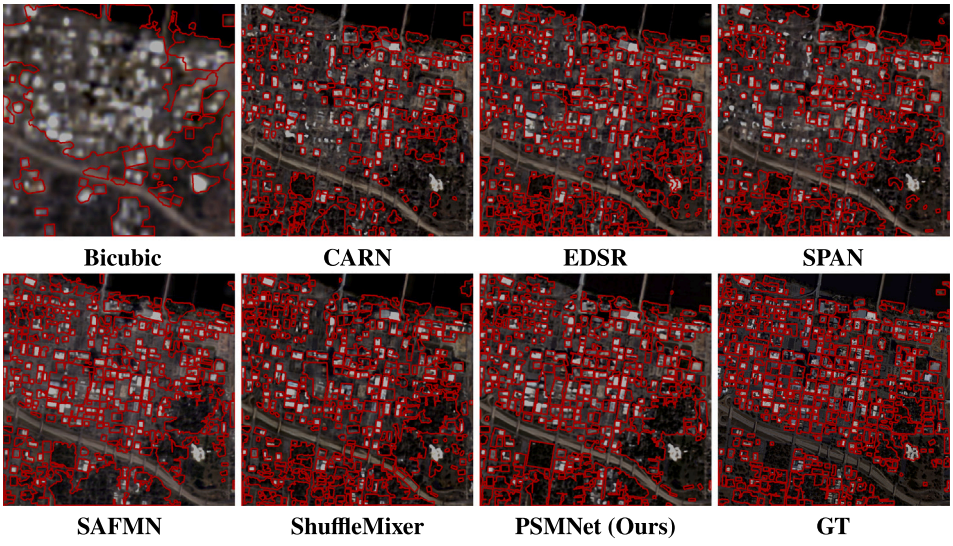


Fig. 6. Comparative visualization of building extraction as a downstream task to evaluate different reconstruction methods. Results from Bicubic, CARN, EDSR, SPAN, SAFMN, ShuffleMixer, and the proposed method PSMNet are shown alongside the GT. The red outlines highlight the extracted buildings, facilitating a direct assessment of how structural continuity is recovered by each reconstruction approach.

2. Largest intersection area: According to Table 8, PSMNet yields the greatest intersection area, directly evidencing the largest overlap with the reference mask. Rich spatial detail in the result enables the detector to capture numerous small structures and intricate boundaries that other methods miss.
3. Strong alignment between reconstruction quality and downstream utility: The success stems from the synergy of the MSCA block and the CCM. MSCA harvests context from local textures to global outlines, producing sharp and accurately located building edges, while CCM efficiently integrates these cues. Consequently, the segmentation network can simultaneously identify more true objects (high Recall) and delineate their boundaries with high Precision and high IoU.

To visually assess how different methods influence downstream building extraction, Fig. 6 overlays the predicted footprints (red outlines) on the segmentation maps; the HR reference result serves as GT. The visual ranking mirrors the quantitative scores, underscoring that reconstruction quality governs application performance.

Low fidelity reconstruction yields conspicuous flaws. Bicubic produces bulky, indistinct red swaths: absent high frequency cues, the segmenter performs "bulk" guessing, merging spectrally similar roofs into one blob—high Recall yet abysmal Precision. Conversely, CARN generates sparse, fragmented outlines that cover only the most evident building cores, confirming its overly conservative strategy (high Precision, critically low Recall).

EDSR and SPAN improve footprint localization, but their edges remain smooth and irregularly offset from the HR mask, reflecting insufficient edge sharpness. SAFMN and ShuffleMixer deliver markedly fuller, more continuous contours; however, ShuffleMixer exhibits slight warping or breakage at dense corners, whereas SAFMN shows minor fluctuations along large roofs.

PSMNet's results are closest to the HR benchmark: red lines closely follow almost every building, including small ancillary structures, with sharp boundaries. In crowded blocks, adjacent roofs are cleanly separated, eliminating the adhesion phenomenon seen in other methods. This visual evidence corroborates that the rich, structurally faithful detail produced by PSMNet enables the segmenter to issue high confidence, pixel accurate decisions, highlighting its strong potential for operational urban mapping and change detection.

4.5. Road extraction results and analysis

Road extraction is another key urban remote sensing task whose accuracy hinges on spatial detail and structural fidelity. As with building extraction, the quality of reconstruction governs the completeness and geometric correctness of the derived road network. All road extraction experiments are conducted on the https://engine-aieearth.aliyun.com/. For the road extraction task, we followed an identical workflow: the reconstructed imagery was saved as GeoTIFF files retaining their geospatial metadata and uploaded to the AI-Earth platform. Within the AIE-SEG module's *Single-Object Extraction* interface, we selected the **Road Extraction** task to generate vectorized road networks and evaluate the reconstruction fidelity for linear infrastructure.

Quantitative metrics and area statistics are used to systematically evaluate the downstream benefits of each method. Bicubic ranks last in Precision (0.5130), Recall (0.2833), F1 (0.3650) and IoU (0.2232). Simple interpolation fails to reconstruct fine details, so the segmenter cannot separate narrow streets or intersections, resulting in both heavy omission and commission errors. Its predicted area is close to the reference, yet the intersection is the smallest, confirming coarse "bulk" predictions with poor localization.

CARN attains the highest Precision (0.7117); however its low Recall (0.5732) leads to only moderate F1 (0.6374) and IoU (0.4677) scores. This pattern suggests that blurred edges compel the network to be highly conservative, prioritizing confident predictions for clear arterial roads at the expense of missing complex or faint ones. Consequently, the predicted area and intersection are considerably less than the reference, verifying significant underdetection.

EDSR improves Recall (0.6116) and IoU (0.4840) over CARN, but Precision drops slightly (0.6987) because the fixed small kernel architecture cannot recover sharp boundaries. SPAN underperforms EDSR across all metrics (F1 0.6269 vs. 0.6523), consistent with its lower PSNR (24.11 dB), again underscoring the link between reconstruction fidelity and downstream accuracy.

SAFMN yields good Recall (0.6438) and IoU (0.4914), yet lower Precision (0.6748) owing to serrated artifacts introduced by multi-scale max-pooling. ShuffleMixer, benefiting from pixel-shuffle upsampling, achieves better Precision (0.7022) and F1 (0.6649), but Recall (0.6273) still trails PSMNet, indicating incomplete recovery of intricate road networks.

PSMNet records the highest Recall (0.6830), IoU (0.5231) and F1 (0.6869); Precision (0.6908) is slightly behind CARN and ShuffleMixer, demonstrating the best trade-off. High Recall means more true roads, especially narrow or occluded segments are detected. Meanwhile, the leading IoU confirms that predicted central lines and edges coincide closely with the reference. Area statistics corroborate this: PSMNet attains the largest intersection and a predicted area almost identical to the GT, proving that its multi-scale, multi-sensor reconstruction supplies the richest spatial cues for accurate road mapping (Table 10).

Consistent with Section 4.4, Fig. 7 visually compares the road masks (red outlines) produced by each method after semantic segmentation, using the HR reference as the baseline. The observations mirror those of the building experiment and reaffirm that reconstruction quality is decisive for downstream utility.

CARN, EDSR, SPAN: The three models behave conservatively and suffer from severe omission. Red lines are scarce, appearing mainly on a couple of wide roads at the bottom of the image. Blurred edges and lacking detail force the network to assign high confidence only to the most salient segments, neglecting the rest of the network. It explains the phenomenon of "high Precision, low Recall" pattern quantified in Table 9.

SAFMN: The extracted network is relatively complete; even some complex upper area roads are recovered. Nevertheless, artifacts introduced by multi-scale max pooling and nearest neighbour upsampling

Table 9
Quantitative evaluation of road extraction. Red indicates the best performance, and blue indicates the second-best performance.

Model	Precision	Recall	F1	IoU
Bicubic	0.5130	0.2833	0.3650	0.2232
CARN	0.7117	0.5732	0.6374	0.4677
EDSR	0.6987	0.6116	0.6523	0.4840
SPAN	0.6765	0.5841	0.6269	0.4566
SAFMN	0.6748	0.6438	0.6590	0.4914
ShuffleMixer	0.7022	0.6273	0.6649	0.4980
PSMNet (Ours)	0.6908	0.6830	0.6869	0.5231

Table 10
Areas corresponding to different reconstruction methods in road extraction. Red indicates the best performance, and blue indicates the second-best performance.

Model	Intersection_Area	Prediction_Area	Union_Area	True_Area
Bicubic	661,572	1,289,549	3,099,496	2,471,518
CARN	1,416,763	1,973,918	3,028,673	2,471,518
EDSR	1,511,766	2,163,601	3,123,352	2,471,518
SPAN	1,443,774	2,133,968	3,161,712	2,471,518
SAFMN	1,591,329	2,357,997	3,238,185	2,471,518
ShuffleMixer	1,550,475	2,192,248	3,113,292	2,471,518
PSMNet (Ours)	1,688,199	2,443,555	3,226,874	2,471,518

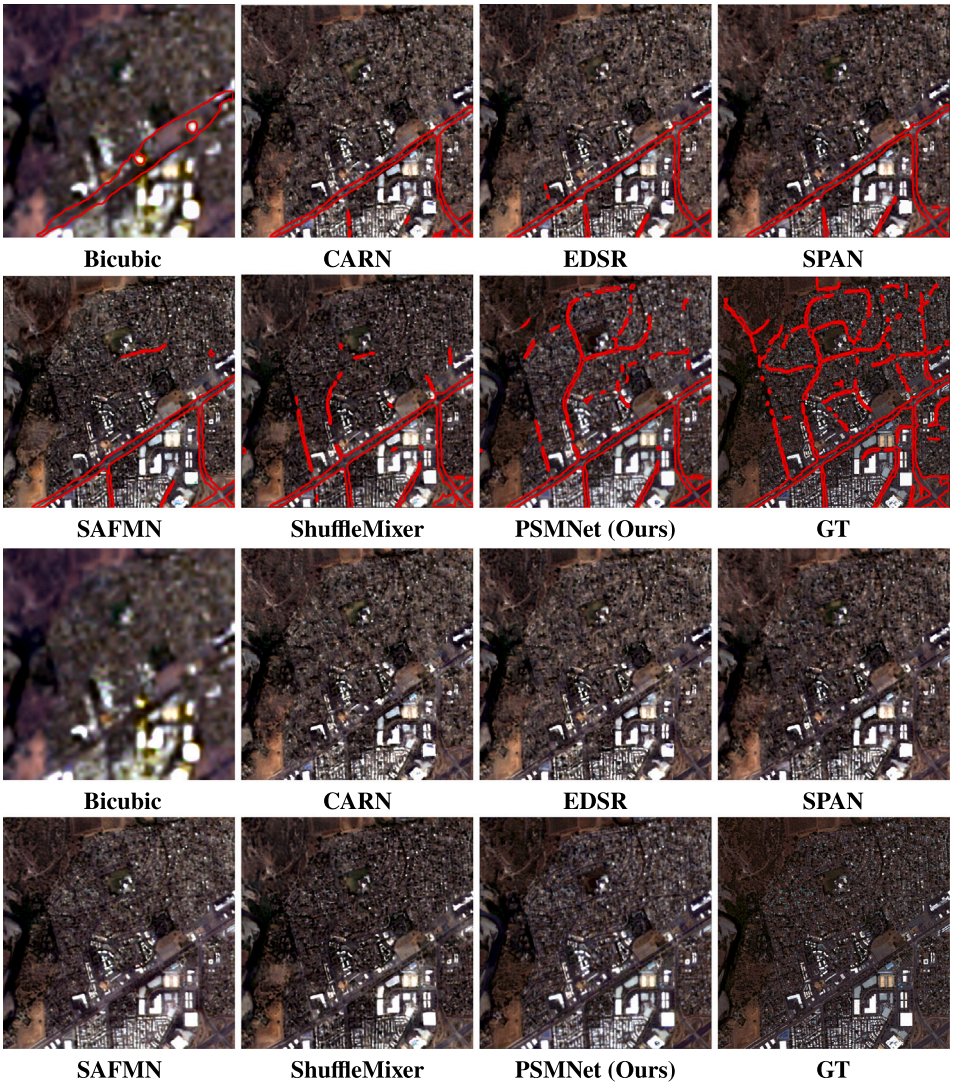


Fig. 7. Visual comparison of road extraction results from different reconstruction methods (Bicubic, CARN, EDSR, SPAN, SAFMN, ShuffleMixer and the proposed PSMNet) against the high resolution GT. The 2 × 4 grid facilitates a direct assessment of how each method recovers road structures, with red lines denoting the extracted road networks. The lower panel displays the corresponding pure RGB reconstructions without any overlays, allowing for a direct evaluation of intrinsic texture fidelity and structural continuity.

reduce boundary accuracy, lowering overall Precision in agreement with the numeric scores.

ShuffleMixer: Long roads are generally well captured, but curved or extended segments occasionally break (e.g., the road at the top). Although large kernel convolutions enlarge the receptive field, global consistency along complex topologies is still limited, so the segmenter fails to maintain long range continuity.

PSMNet (Ours): The result is closest to the HR reference. The network is highly complete and connected: arterials, minor alleys and intricate interchange structures are all retrieved. Road edges are sharp and accurately aligned with the true outlines, visually confirming that PSMNet furnishes the richest spatial detail and the most faithful structural cues for the downstream segmentation model.

In Fig. 8, the ROI reveals obvious differences in the capability of each method to recover fine linear features and maintain the connectivity of road segments. Bicubic interpolation yields heavily blurred outlines, causing the road network to appear fragmented and leading to significant omissions during segmentation. While CARN, EDSR, SPAN and SAFMN show slightly sharper edges, their reconstructions remain conservative, often failing to restore thin roads or complex intersections, which aligns with their moderate Recall scores. ShuffleMixer, employing large kernel depth-wise convolutions, achieves better recovery of long roads but struggles with maintaining continuity at curved segments, resulting in occasional breaks in the network due to its coarse multi-scale scheme. The proposed PSMNet, by contrast, demonstrates superior performance in this region, producing continuous and well-defined road boundaries. This visual superiority is consistent with its leading Recall (0.6830) and IoU (0.5231), as reported in Table 9, underscoring the practical impact of its multi-scale feature fusion strategy on preserving long range spatial dependencies.

5. Discussion

5.1. Operational limitations and input dependencies

PSMNet is designed as a cross-sensor reconstruction framework that fuses Landsat-8 and Sentinel-2 to generate 2.5 m Landsat-like imagery. The concatenation-based fusion reflects our core hypothesis that spatial detail enhancement benefits from complementary multi-source observations. Consequently, the model assumes triple-source availability during inference and does not natively support arbitrary modality dropout.

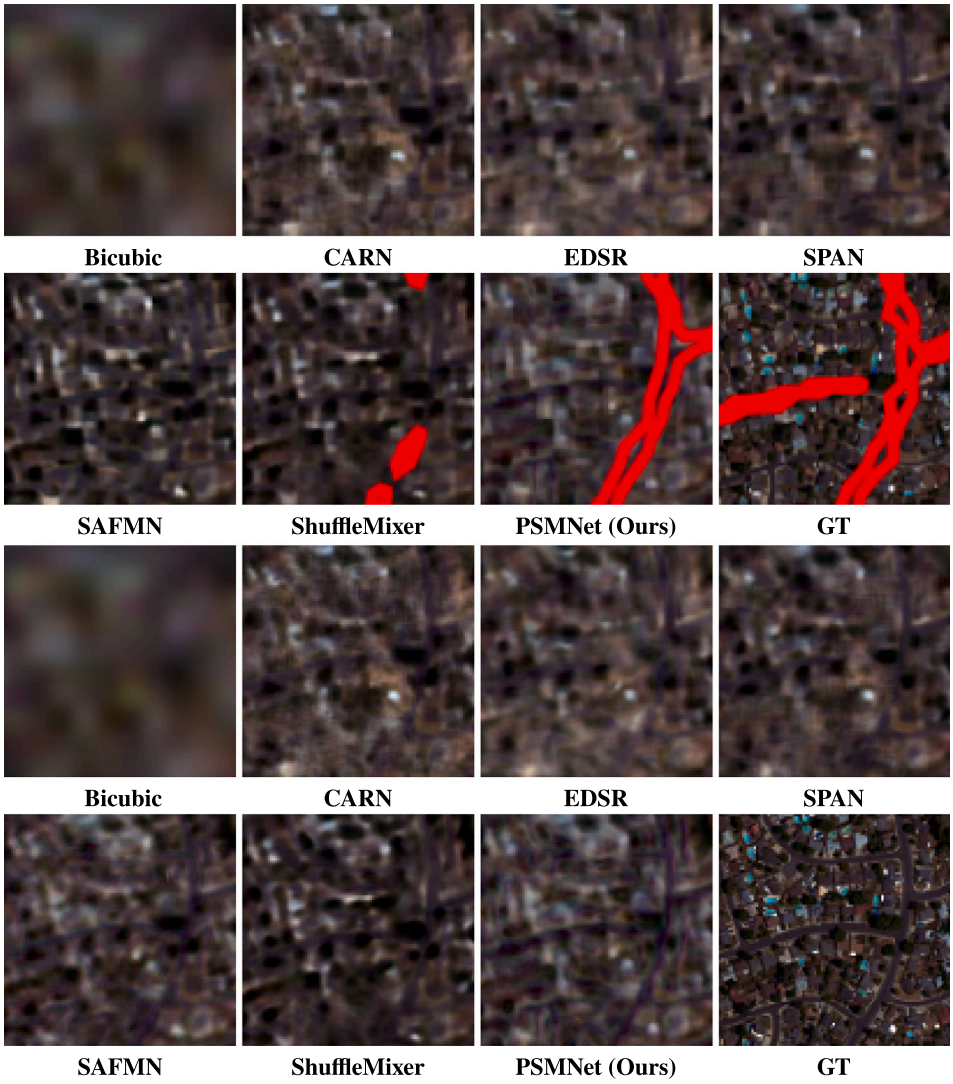


Fig. 8. The proposed PSMNet yields the most visually accurate reconstruction. This figure compares the reconstruction results of various methods (including Bicubic, CARN, EDSR, etc.) on a dense urban image. The lower block presents the full-scene pure RGB outputs, free from segmentation masks or annotations, clearly demonstrating each method's capability in preserving fine linear features and complex road connectivity.

While not designed for missing inputs, we empirically observe graceful degradation rather than catastrophic failure:

- *Landsat-8 obscured, Sentinel-2 clear* (Fig. 9): Sentinel-2's higher resolution provides sufficient spatial cues to maintain plausible reconstruction, though spectral fidelity may be affected.
- *Sentinel-2 obscured, Landsat-8 clear* (Fig. 10): Cloud artifacts tend to persist, as Sentinel-2 serves as the primary spatial detail donor in our framework.

We acknowledge that persistent adverse conditions affecting both sensors simultaneously remain a limitation. Future extensions could include adaptive input gating, training with simulated dropout, or temporal memory modules to enhance robustness in the following.

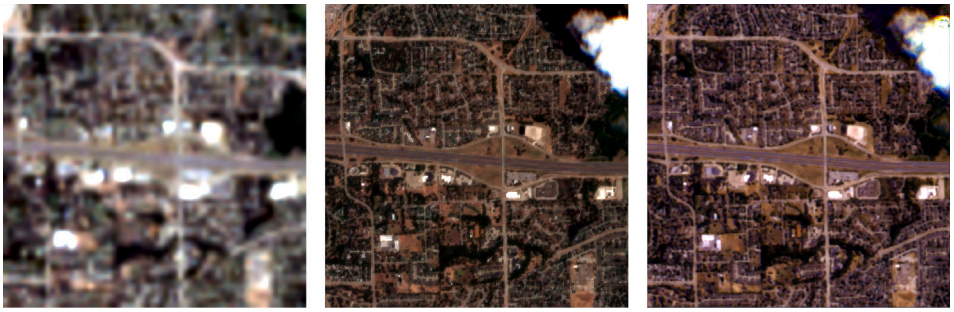
5.2. Performance compared with advanced baselines

To establish a more comprehensive performance context and identify the true performance upper bound, we have incorporated two state-of-the-art Transformer-based super-resolution methods into our comparative evaluation: SwinIR [11] and CAMixer [26]. The quantitative results and visual comparisons are presented below.

Our method achieves competitive performance compared to Transformer-based approaches while maintaining lower computational overhead. From the quantitative results (Table 11) and visual comparisons (Fig. 11), we make several key observations. First, SwinIR underperforms our method on both PSNR and SAM metrics. Visually, the magnified region in the upper left corner fails to recover the original line structures, and the zoomed area in the upper right corner also lacks fine detail preservation. Furthermore, SwinIR exhibits significantly higher parameter counts and computational complexity than our model, indicating that naive sliding window attention is inefficient for this task. In contrast, CAMixer achieves the best results across PSNR, SSIM, and SAM. CAMixer successfully restores details in the upper left magnified region and demonstrates superior performance in reconstructing the square patterns in the upper right corner. While other methods produce somewhat blurry results, CAMixer generates noticeably sharper outputs. This success stems from its efficient sliding window attention mechanism, which divides the image into patches and computes complexity scores. Patches with rich details are routed to sophisticated attention modules, while simpler patches are directed to lightweight convolutional operations. This design elegantly balances CNN and Transformer architectures.

(a) Landsat-8 input **(b)** Sentinel-2 input **(c)** PSMNet reconstruction

Fig. 9. Case study: Landsat-8 input cloud-contaminated while Sentinel-2 remains clear. The reconstructed output (c) retains plausible spatial structures by leveraging Sentinel-2's high-resolution cues.



(a) Landsat-8 input **(b)** Sentinel-2 input **(c)** PSMNet reconstruction

Fig. 10. Case study: Sentinel-2 input is cloud-contaminated while Landsat-8 remains clear. Cloud artifacts persist in the reconstruction (c) because Sentinel-2 serves as the primary spatial detail donor.

Table 11
Quantitative comparison with Transformer-based methods on the L8-S2-NAIP benchmark.

Model	Type	PSNR	SSIM	SAM (°)
PSMNet (Ours)	Lightweight CNN	24.5861	0.7225	2.6965
SwinIR	Transformer	24.4960	0.7233	2.7319
CAMixer	Hybrid (CNN + Transformer)	**24.6498**	**0.7259**	**2.6612**

Table 12
Ablation study on normalization and channel mixing strategies for spectral preservation.

Variant	Normalization	Channel Mixing	SAM (°)
Ours (Full)	GroupNorm	CCM	2.69
w/o CCM	GroupNorm	Standard 3×3 Conv	3.01
w/ LayerNorm	LayerNorm	CCM	2.72

Nevertheless, CAMixer remains comparable to SwinIR in terms of resource consumption, requiring substantial memory and computational resources.

5.3. Spectral preservation via GroupNorm and convolutional channel mixer

The proposed PSMNet achieves strong spectral preservation (SAM = 2.69°), benefiting from the synergy between GroupNorm (GN) and the Convolutional Channel Mixer (CCM). An ablation study isolates their contributions, as summarized in Table 12.

Replacing CCM with a standard 3×3 convolution raises SAM to 3.01°, indicating that CCM's depthwise-separable convolutions constrain channel interactions within physically meaningful groups. This localized mixing avoids propagating cross-sensor statistical mismatches into the spectral domain during 12× upsampling. Replacing GN with

(a) PSMNet (Ours) **(b)** SwinIR **(c)** CAMixer

Fig. 11. Visual comparison of reconstructed RGB images. Magnified regions highlight structural fidelity and texture preservation.

LayerNorm leads to a slight SAM increase to 2.72°, suggesting that GN's group-wise normalization better respects the distinct radiometric characteristics of different sensors by preserving inter-sensor intensity ratios. The full GN + CCM configuration achieves the lowest SAM (2.69°), confirming their complementary roles in cross-sensor spectral fidelity.

6. Conclusion

Targeting the limited resolution gain, high computational cost, and lack of real downstream validation that plague current cross-sensor reconstruction schemes, this study presents a complete solution: a 2.5 m benchmark dataset that synergizes Landsat-8, Sentinel-2 and NAIP imagery, and a lightweight multi-scale reconstruction network: PSMNet. Our main contributions can be summarized below:

1. **Dataset**: The first Landsat-8 2.5 m aligned, cross-sensor open dataset, offering an essential data foundation for future research.
2. **Model**: PSMNet couples 630 K parameters with 3.66 ms inference, achieving a well-balanced performance and efficiency.
3. **Validation**: Building and road extraction tasks are, for the first time, systematically used to quantify real world utility, moving evaluation beyond pixel statistics.

In the future work, we will extend single scene reconstruction to a spatio-temporal fusion framework and explore end to end optimization of the reconstruction module and downstream networks, promoting deeper integration of reconstruction into operational remote sensing systems.

CRediT authorship contribution statement

Hongjie Xie: Writing – original draft, Methodology, Data curation. **Qiang Zhang:** Writing – review & editing, Validation, Supervision. **Yi Xiao:** Writing – review & editing, Validation, Supervision. **Zishuo Wang:** Writing – review & editing, Visualization, Data curation. **Jianwei Wen:** Writing – review & editing, Resources, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study is supported in part by the National Natural Science Foundation of China under Grant 62401095 and 62602006; and in part by the Natural Science Foundation of Heilongjiang Province of China for Doctoral Students under Grant BS2025F001; and in part by the Natural Science Foundation of Inner Mongolia Autonomous Region under Grant 2025ZDLH007; and in part by the Dalian Science and Technology Talent Innovation Supporting Project under Grant 2024RQ028; and in part by the China Postdoctoral Science Foundation under Grant BX2026298, 2023M740460 and 2025T180065; and in part by the Natural Science Foundation of Liaoning Province under Grant 2025-BS-0236; and in part by the Fundamental Research Funds for the Central Universities under Grant 3132026239.

Data availability

Data will be made available on request.

References

[1] N. Ahn, B. Kang, K.-A. Sohn, Fast, accurate, and lightweight super-resolution with cascading residual network, version: 5 arXiv:1803.08664, 2018.
[2] W. Chen, Y. Wen, J. Zheng, J. Huang, H. Fu, Ban: a universal paradigm for cross-scene classification under noisy annotations from RGB and hyperspectral remote sensing images, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–13.
[3] S. Cui, G. Jiang, J. Lu, Quantitative estimation of soil lithium content using fractional derivative spectroscopy and machine learning algorithms, Opt. Laser Technol. 192 (2025) 114057.
[4] S. Donike, C. Aybar, L. Gómez-Chova, F. Kalaitzis, Trustworthy super-resolution of multispectral Sentinel-2 imagery with latent diffusion, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 18 (2025) 6940–6952.
[5] J. Guo, J. Zheng, Y. Liang, Y. Wen, B. Yu, J. Hu, J. Huang, H. Fu, Toward cross-regional land cover mapping using regional consistency and certainty-based active domain adaptation with limited annotations, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 19 (2026) 5900–5910.
[6] D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), arXiv:1606.08415, 2023.
[7] M. Hu, Y. Pan, D. Xue, X. Xu, K. Zhang, Z. Wang, S. Sun, Bidirectional cross-attention mechanism based super-resolution for full polarization characteristic imaging, Opt. Laser Technol. 193 (2026) 114223.
[8] J. Ju, Q. Zhou, B. Freitag, D.P. Roy, H.K. Zhang, M. Sridhar, J. Mandel, S. Arab, G. Schmidt, C.J. Crawford, F. Gascon, P.A. Strobl, J.G. Masek, C.S. Neigh, The harmonized landsat and Sentinel-2 version 2.0 surface reflectance dataset, Remote Sens. Environ. 324 (2025) 114723.
[9] N.L. Kazanskiy, L.L. Doskolovich, N.V. Golovastikov, S.N. Khonina, The power of fusion: LiDAR meets hyperspectral imaging in a new era of exploration, Opt. Laser Technol. 192 (2025) 114080.
[10] Q. Li, Y. Zhang, J. Zheng, Y. Zhang, J. Huang, H. Fu, Boosting universal domain adaptation in remote sensing with dual-classifiers consistency discrimination and cross-domain feature mixup, IEEE Trans. Geosci. Remote Sens. 63 (2025) 1–15.
[11] J. Liang, J. Cao, G. Sun, K. Zhang, L.V. Gool, R. Timofte, Swinir: image restoration using swin transformer, 2021.
[12] B. Lim, S. Son, H. Kim, S. Nah, K.M. Lee, Enhanced deep residual networks for single image super-resolution, arXiv:1707.02921, 2017.
[13] J. Michel, J. Inglada, Temporal attention multi-resolution fusion of satellite image time-series, applied to Landsat-8/9 and Sentinel-2: all bands, any time, at best spatial resolution, Remote Sens. Environ. 334 (2026) 115159.
[14] T.D. Mushore, O. Mutanga, J. Odindi, V. Sadza, T. Dube, Pansharpened landsat 8 thermal-infrared data for improved land surface temperature characterization in a heterogeneous urban landscape, Remote Sens. Appl. Soc. Environ. 26 (2022) 100728.
[15] K. Rahaman, Q. Hassan, M. Ahmed, Pan-sharpening of Landsat-8 images and its application in calculating vegetation greenness and canopy water contents, ISPRS Int. J. Geo-Inf. 6 (6) (2017) 168.
[16] V.T. Sambandham, K. Kirchheim, F. Ortmeier, S. Mukhopadhaya, Deep learning-based harmonization and super-resolution of Landsat-8 and Sentinel-2 images, ISPRS J. Photogramm. Remote Sens. 212 (2024) 274–288.
[17] Z. Shao, J. Cai, P. Fu, L. Hu, T. Liu, Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product, Remote Sens. Environ. 235 (2019) 111425.
[18] J. Sigurdsson, S.E. Armannsson, M.O. Ulfarsson, J.R. Sveinsson, Fusing Sentinel-2 and landsat 8 satellite images using a model-based method, Remote Sens. 14 (13) (2022) 3224. Publisher: Multidisciplinary Digital Publishing Institute.
[19] K. Song, Z. Zhu, S. Qiu, P. Olofsson, C.S. Neigh, J. Ju, Q. Zhou, TIF: a time-series-based image fusion algorithm, Remote Sens. Environ. 331 (2025) 115035.
[20] L. Sun, J. Dong, J. Tang, J. Pan, Spatially-adaptive feature modulation for efficient image super-resolution, arXiv:2302.13800, 2023.
[21] L. Sun, J. Pan, J. Tang, ShuffleMixer: an efficient ConvNet for image super-resolution, arXiv:2205.15175, 2023.
[22] C. Wan, H. Yu, Z. Li, Y. Chen, Y. Zou, Y. Liu, X. Yin, K. Zuo, Swift parameter-free attention network for efficient super-resolution, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, Seattle, WA, USA, 2024, pp. 6246–6256.
[23] P. Wang, M. Huang, S. Shi, B. Huang, B. Zhou, G. Xu, L. Wang, H. Leung, Landsat-8 and Sentinel-2 image fusion based on multiscale smoothing-sharpening filter, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 17 (2024) 17957–17970.
[24] Q. Wang, Q. Zhang, T. Yang, W. Sun, Q. Yuan, SGD-SST 2.0: seamless global daily sea surface temperature products cross-sensors generating from 2003 to 2025, Expert Syst. Appl. 322 (2026) 132259.
[25] X. Wang, S. Chen, C. Zhou, S.-B. Duan, Z. Shi, Cross-sensor data reconstruction for optical remote sensing gap-filling with attention-enhanced multi-scale fusion network, Int. J. Appl. Earth Obs. Geoinf. 140 (2025) 104571.
[26] Y. Wang, Y. Liu, S. Zhao, J. Li, L. Zhang, Camixersr: only details need more "attention", 2024.
[27] D. Xiao, L. Wan, X. Sun, Remote sensing retrieval of saline and alkaline land based on reflectance spectroscopy and rv-melm in Zhenlai county, Opt. Laser Technol. 139 (2021) 106909.
[28] Y. Xiao, Q. Yuan, K. Jiang, Y. Chen, S. Wang, C.-W. Lin, Multi-axis feature diversity enhancement for remote sensing video super-resolution, IEEE Trans. Image Process. 34 (2025) 1766–1778.
[29] Y. Xiao, Q. Yuan, K. Jiang, Y. Chen, Q. Zhang, C.-W. Lin, Frequency-assisted Mamba for remote sensing image super-resolution, IEEE Trans. Multimed. 27 (2025) 1783–1796.
[30] Y. Xiao, Q. Yuan, K. Jiang, J. He, C.-W. Lin, L. Zhang, Ttst: a top-k token selective transformer for remote sensing image super-resolution, IEEE Trans. Image Process. 33 (2024) 738–752.
[31] S. Xie, L. Sun, L. Liu, X. Liu, Global cross-sensor transformation functions for Landsat-8 and Sentinel-2 top of atmosphere and surface reflectance products within Google earth engine, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–9.

[32] M. Xu, X. Zhong, R. Zhong, A method for automatic extraction and individual segmentation of urban street trees from laser point clouds, Opt. Laser Technol. 180 (2025) 111431.
[33] B. Zhang, Q. Zhang, W. Gao, B. Liu, J. Wen, 10-minute level and near real-time wildfire detection: integrating multiple features for Himawari-8/9 satellite, Int. J. Remote Sens. 47 (8) (2026) 3285–3309.
[34] B. Zhang, Q. Zhang, Z. Wang, T. Yang, Immediate remote sensing: dynamic context-adaptive fusion for himawari-8/9 10-minute wildfire detection, Remote Sens. Environ. 344 (2026) 115524.
[35] Q. Zhang, Y. Dong, Y. Zheng, H. Yu, M. Song, L. Zhang, Q. Yuan, Three-dimension spatial–spectral attention transformer for hyperspectral image denoising, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–13.
[36] Q. Zhang, Q. Yuan, M. Song, H. Yu, L. Zhang, Cooperated spectral low-rankness prior and deep spatial prior for HSI unsupervised denoising, IEEE Trans. Image Process. 31 (2022) 6356–6368.
[37] Q. Zhang, X. Zhang, Y. Xiao, H. Xie, Deep low-rank tensor embedded network for hyperspectral image super-resolution, Expert Syst. Appl. 299 (2026) 129864.
[38] Q. Zhang, Y. Zheng, Q. Yuan, M. Song, H. Yu, Y. Xiao, Hyperspectral image denoising: from model-driven, data-driven, to model-data-driven, IEEE Trans. Neural Netw. Learn. Syst. 35 (10) (2024) 13143–13163.
[39] Q. Zhang, J. Zhu, Y. Dong, E. Zhao, M. Song, Q. Yuan, 10-minute forest early wildfire detection: fusing multi-type and multi-source information via recursive transformer, Neurocomputing 616 (2025) 128963.
[40] Q. Zhang, J. Zhu, Y. Huang, Q. Yuan, L. Zhang, Beyond being wise after the event: combining spatial, temporal and spectral information for Himawari-8 early-stage wildfire detection, Int. J. Appl. Earth Obs. Geoinf. 124 (2023) 103506.
[41] R. Zhou, Z. Tang, K. Liu, W. Zhang, K. Liu, X. Li, P. Yang, G. Li, High sensitivity analysis of soil by laser-induced breakdown spectroscopy with AG nanoparticles, Opt. Laser Technol. 155 (2022) 108386.