Ultra-Lightweight Semantic-Injected Imagery Super-Resolution for Real-Time UAV Remote Sensing

Rongchang Lu , Yunzhi Jiang , Bingcheng Liao , Conghan Yue , Xin Hai , Guoxin Chen *

Posted Date: 8 April 2026

doi: 10.20944/preprints202507.2060.v2

Keywords: remote sensing; unmanned aerial vehicle; image super-resolution; state-space model

 doi:10.20944/preprints202507.2060.v2

Article

Ultra-Lightweight Semantic-Injected Imagery Super-Resolution for Real-Time UAV Remote Sensing

Rongchang Lu [1], Yunzhi Jiang [2], Bingcheng Liao [2], Conghan Yue [3], Xin Hai [4] and Guoxin Chen [5,6,*]

[1] School of Ecological and Environmental Engineering, Qinghai University, Xining 810016, China
[2] Department of Computer Technology and Applications, Qinghai University, Xining 810016, China
[3] School of Computer Science, Sun Yat-sen University, Guangzhou 510006, China
[4] Information Technology Center, Qinghai University, Xining 810016, China
[5] State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining 810016, China
[6] Satellite Remote Sensing Data Reception and Application Research Center, Qinghai University, Xining 810016, China
* Correspondence: chenguoxin@qhu.edu.cn

Abstract

Real-time 2D imagery super-resolution (SR) in UAV remote sensing encounters significant speed and resource-consuming bottlenecks during large-scale processing. To overcome this, we propose Semantic Injection State Modeling for Super-Resolution (SIMSR), an ultra-lightweight architecture that integrates land-cover semantics into a linear state-space model. This integration mitigates state forgetting inherent in linear processing by linking hierarchical features to persistent semantic prototypes, enabling high-fidelity image enhancement. The model achieves a state-of-the-art PSNR of 32.9+ for 4x SR on RSSCN7 agricultural grassland imagery. Furthermore, the implementation of geographically-chunked (tile-based) parallel processing simultaneously eliminates computational redundancies, yielding a 10.85x inference speedup, a 54% memory reduction, and an 8.74x faster training time. This breakthrough facilitates practical real-time SR deployment on UAV platforms, demonstrating strong efficacy for ecological monitoring applications by providing the detailed imagery essential for accurate analysis.

Keywords: remote sensing; unmanned aerial vehicle; image super-resolution; state-space model

1. Introduction

In measurement science, the fidelity and information content of acquired data directly govern the accuracy and reliability of downstream analytical models [1,2]. Remote sensing, as a pivotal geospatial measurement technology, relies on high-quality imagery to quantify biophysical parameters (e.g., vegetation indices [3], soil moisture [4]) and delineate land-cover features with metrological traceability [5]. Unmanned Aerial Vehicles (UAVs) have emerged as a transformative platform for high-spatiotemporal-resolution measurement [6], enabling centimeter-scale observations [7] critical for precision agriculture [8], urban infrastructure inspection [9], and ecological monitoring [10]. However, the pursuit of higher spatial resolution—a fundamental metric in optical measurement—often clashes with practical constraints: high-precision sensors are prohibitively expensive [11], and the computational burden of processing gigapixel-scale UAV imagery exceeds the capabilities of resource-constrained embedded measurement systems commonly deployed on UAVs.

Computational image super-resolution (SR) presents a promising pathway to enhance the effective spatial resolution of measurement data post-acquisition, thereby augmenting the information density without modifying the physical sensor [12,13]. From a measurement science perspective, an ideal SR algorithm must not only improve perceptual sharpness but, more critically, preserve radiometric and geometric fidelity to ensure the quantitative validity of subsequent measurements [14]. This imposes stringent requirements on model robustness, especially when dealing with the complex, heterogeneous textures inherent in natural landscapes [15]. Current deep learning-based SR paradigms, however, face

 doi:10.20944/preprints202507.2060.v2

significant challenges in meeting these dual demands of high measurement fidelity and computational efficiency for real-time UAV deployment.

The hierarchical and heterogeneous nature of land surface features—from homogeneous agricultural fields to complex urban mosaics—poses a fundamental modeling challenge for SR, as inaccuracies can propagate as measurement uncertainty [5]. Convolutional Neural Networks (CNNs) suffer from limited receptive fields, leading to blurred edges that compromise the geometric precision of object boundaries [16]. Vision Transformers (ViTs) [17] capture long-range context but with quadratic complexity, making them intractable for the large-scale measurement data acquired by UAVs. Recently, State-Space Models (SSMs) like Mamba [18] offer linear complexity, suitable for long-sequence data. Yet, their application to image SR for measurement reveals critical shortcomings: (1) Catastrophic state forgetting: Sequential processing erases early context, causing inconsistencies in reconstructing large, uniform measurement targets (e.g., water bodies), violating the principle of measurement consistency across the scene. (2) Constrained receptive fields: The cross-shaped scanning pattern fails to capture diagonal structures (e.g., drainage networks, field boundaries), introducing systematic errors in feature localization. (3) Non-adaptive state dynamics: Most critically, a static state-transition mechanism applies uniform processing regardless of local texture complexity. From a measurement standpoint, this leads to inhomogeneous reconstruction uncertainty [1]. Information-rich, high-frequency regions (e.g., forest canopies, built-up areas) may be under-reconstructed, while homogeneous regions (e.g., bare soil) may be over-smoothed or prone to hallucinated details. This variability undermines the uniform quality standard required for reliable area-wide quantitative analysis, such as biomass estimation or impervious surface mapping.

Additionally, computational overhead remains prohibitive for UAV edge deployment. SSM-based methods [19] require L sequential steps without parallelization during large-area analysis, while irregular memory access patterns misalign with geographical feature geometries. These inefficiencies yield <30% hardware utilization on parallel architectures, increasing UAV operational costs and delaying critical applications like wildfire progression mapping or flood extent monitoring.

To address these limitations at the intersection of computational efficiency and measurement fidelity, we propose Semantic Injection State Modeling for Super-Resolution (SIMSR). Our work is motivated by the measurement science principle that prior knowledge (e.g., land-cover semantics) can constrain and improve estimation processes [2]. SIMSR introduces two core innovations designed for high-fidelity, efficient SR on UAV platforms:

1. **Semantic-Injected State Modeling for Uncertainty Reduction**: We integrate land-cover semantics—derived from lightweight pre-segmentation—into the SSM's state space. This injects persistent, category-specific prompts that anchor the dynamic state, mitigating catastrophic forgetting by providing a semantic "memory." More importantly, it allows the model to adapt its reconstruction strategy implicitly based on semantic class, addressing Limitation (3). For instance, "forest" semantics can promote detail preservation, while "water" semantics can encourage smoothness, thereby reducing class-dependent reconstruction uncertainty and suppressing hallucinations that compromise measurement integrity.

2. **Geographically-Chunked Processing for Metrological Traceability**: Aligning with the practice of analyzing geospatial data in logical units (e.g., watersheds, land parcels) [5], we process imagery in contiguous geographical chunks rather than arbitrary tiles. This chunking, compatible with parallel processing, ensures that long-range dependencies within ecologically or administratively coherent units are preserved, enhancing consistency for areal measurements. It also optimizes memory access patterns, translating to practical efficiency gains on measurement hardware.

Validated on remote sensing benchmarks, SIMSR advances the state of the art in measurement-directed SR. It achieves a PSNR of 32.9+ on the RSSCN7 *aGrass* class, indicating superior radiometric fidelity. Crucially, it delivers these gains with unprecedented efficiency: 10.85× faster inference and 54% lower memory footprint than prior state-of-the-art models, metrics that are directly relevant for embedded measurement systems. The expanded, more isotropic effective receptive field (Figure

 doi:10.20944/preprints202507.2060.v2

1) underpins its improved geometric accuracy. By simultaneously addressing the dual bottlenecks of reconstruction quality (fidelity, reduced hallucination) and computational feasibility for edge deployment, SIMSR bridges a critical gap in the UAV remote sensing measurement chain, enabling real-time, high-precision data enhancement for time-sensitive applications like disaster response and precision agriculture.

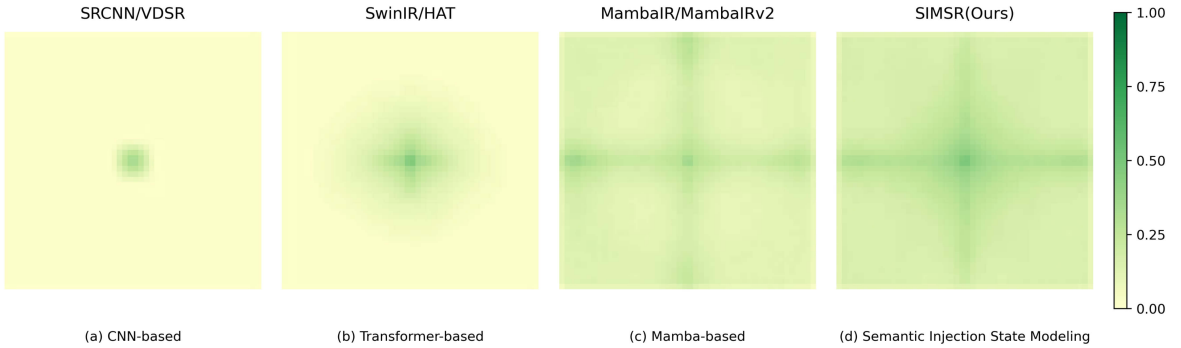

Figure 1. Effective receptive field (ERF) visualizations demonstrating SIMSR's enhanced global coverage versus constrained patterns in prior efficient models.

2. Related Work

2.1. Conventional and Convolution-based Techniques

Traditional interpolation methods (e.g., bicubic) lack semantic understanding despite computational efficiency [20]. Deep learning-based approaches have since dominated, with SRCNN [21] pioneering the use of convolutional layers for local patch extraction and HR reconstruction. VDSR [22] addresses this via deeper residual networks, expanding receptive fields to capture hierarchical features—pixel-level edges in early layers and regional semantics in deeper ones.

However, CNNs grapple with global-local modeling trade-offs and computational bottlenecks. CNNs' local receptive fields hinder long-range correlation capture. These limitations drive the attention mechanisms for effective global-local collaboration in remote sensing super-resolution.

2.2. Attention-based Techniques

This issue has been significantly addressed by Attention mechanisms, which enables models to effectively capture long-range dependencies and focus on critical image regions. For instance, SwinIR [16] uses hierarchical features and shift-window self-attention, SwinFIR [23] improves global integration with Fast Fourier Convolution (FFC), and HAT [24,25] optimizes hybrid attention. They have demonstrated exceptional performance across various image restoration[26–28] tasks, including super-resolution, denoising[29–31], and JPEG artifact reduction.

Despite the significant advancements brought by attention mechanisms in SISR[32], a notable limitation persists: the computational complexity of self-attention operations scales quadratically with the input size [17]. This quadratic complexity arises because the attention mechanism computes pairwise interactions between all elements in the input sequence, leading to substantial computational and memory demands, especially for HR images.

2.3. State-Space Based Techniques

To address the computational challenges of attention-based SISR, several models have integrated innovative mechanisms to improve efficiency. The MambaIR model [33–38] introduces a Selective State Space 2D (SS2D) mechanism, which employs SSM[39] with selective scanning strategies to capture long-range dependencies while maintaining linear computational complexity relative to the input size. This design effectively reduces the computational burden of traditional quadratic attention mechanisms, making MambaIR scalable and efficient for HR image restoration tasks. The SS2D mechanism allows MambaIR to model intricate image details without incurring the prohibitive costs

 doi:10.20944/preprints202507.2060.v2

typical of self-attention methods, thus balancing performance and efficiency for large-scale image processing. Despite the advancements introduced by MambaIR and the SS2D mechanism, challenges remain in terms of prolonged training times, unstable performance metrics, rapid convergence, and suboptimal feature extraction capabilities. These challenges will be better solved in this proposed work.

3. Semantic-Injected State Modeling

The Semantic-Injected State Modeling (SISM) framework establishes a novel paradigm for capturing global dependencies in high-resolution imagery by integrating hierarchical semantic decomposition with adaptive state transitions. This approach overcomes the limitations of sequential state overwriting through persistent feature anchoring to categorical prototypes while maintaining linear computational complexity. The mathematical foundation combines multi-directional scanning with chunk-wise parallelization to achieve spatially-aware adaptation.

3.1. Semantic Decomposition and Prototype Anchoring

The semantic decomposition stage serves as the foundation for injecting persistent categorical priors into the state-space modeling. Given an input low-resolution image $\mathbf{I}_{LR} \in \mathbb{R}^{H\times W\times 3}$, we first extract shallow features $\mathbf{F}_0 \in \mathbb{R}^{H\times W\times C}$ using a 3×3 convolution. These features are then fed into a lightweight segmentation head $\mathcal{G}_\theta$ consisting of two 3×3 convolutional layers followed by a softmax activation, producing a semantic probability map $\mathbf{P} \in \mathbb{R}^{H\times W\times K}$ where K denotes the number of land-cover categories (e.g., vegetation, water, urban).

$$\mathbf{P} = \text{Softmax}(\mathcal{G}_\theta(\mathbf{F}_0)) \tag{1}$$

A hard assignment mask $\mathbf{M} \in \{0,1\}^{H\times W\times K}$ is obtained via argmax across the category dimension:

$$\mathbf{M}_{i,j,k} = \begin{cases} 1 & \text{if } k = \arg\max_{k'} \mathbf{P}_{i,j,k'} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

For each category k, a prototype vector $\mathbf{p}_k \in \mathbb{R}^C$ is computed by aggregating features belonging to that category:

$$\mathbf{p}_k = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{M}_{i,j,k}\cdot \mathbf{F}_{0,(i,j)}}{\sum_{i=1}^{H}\sum_{j=1}^{W} \mathbf{M}_{i,j,k} + \epsilon} \tag{3}$$

where $\epsilon = 10^{-6}$ prevents division by zero. These prototypes serve as persistent anchors that condition the state transitions in subsequent processing, ensuring that features from semantically similar regions are reinforced across the entire image.

The semantic decomposition does not reduce the spatial size of the feature map; instead, it produces a categorical mask that guides the reorganization of image blocks. Specifically, the input feature map $\mathbf{F}_0$ is divided into non-overlapping blocks of size $B\times B$ (default $B = 8$). Blocks sharing the same dominant category (determined by majority voting over $\mathbf{M}$ within the block) are grouped together and rearranged into a semantically coherent sequence. This rearrangement is performed only during the 2D State Modeling stage (Section 3.3) and does not alter the original spatial layout of the image. After processing, blocks are restored to their original positions, preserving the topological structure.

3.2. Semantic Injection Mechanism

The semantic injection mechanism addresses the state forgetting problem in sequential state-space models by providing persistent categorical anchors. Unlike conventional SSMs that rely solely on transient hidden states, our approach maintains a bank of prototype vectors $\{\mathbf{p}_k\}_{k=1}^{K}$ that represent

 doi:10.20944/preprints202507.2060.v2

each land-cover category. During 2D State Modeling, these prototypes are injected as conditioning signals at each step:

$$\mathbf{S}_t^d = \mathbf{A}^d \mathbf{S}_{t-1}^d + \mathbf{B}^d [\mathbf{x}_t^d \oplus \mathbf{p}_{k(t)}] \tag{4}$$

where $\oplus$ denotes concatenation and $k(t)$ is the category of the t-th token. This ensures that state transitions are biased toward semantically relevant patterns, effectively preserving long-range dependencies within the same category (e.g., connecting distant vegetation pixels).

The injection is performed via a gating mechanism that adaptively blends the prototype information:

$$\mathbf{g}_t = \sigma(\mathbf{W}_g [\mathbf{x}_t^d \oplus \mathbf{p}_{k(t)}]) \tag{5}$$

$$\mathbf{x}'_t = \mathbf{g}_t \odot \mathbf{p}_{k(t)} + (1 - \mathbf{g}_t) \odot \mathbf{x}_t^d \tag{6}$$

where $\mathbf{W}_g$ is a learnable weight matrix. This allows the model to selectively emphasize semantic information when needed, particularly for ambiguous boundary regions.

3.3. 2D State Modeling Module

Standard state-space modeling captures information through sequential causal processing, rendering it effective for 1D signals but fundamentally limited for noncausal image data. This limitation arises from unidirectional context aggregation that neglects critical spatial dependencies across four geometric orientations: horizontal (forward/backward) and vertical (downward/upward). To overcome this constraint, we integrate the 2D scanning methodology [36] with State Modeling principles, proposing a novel 2D State Modeling mechanism. This transformation enables comprehensive modeling of spatial-semantic relationships in remote sensing imagery, as visualized in Figure 2.

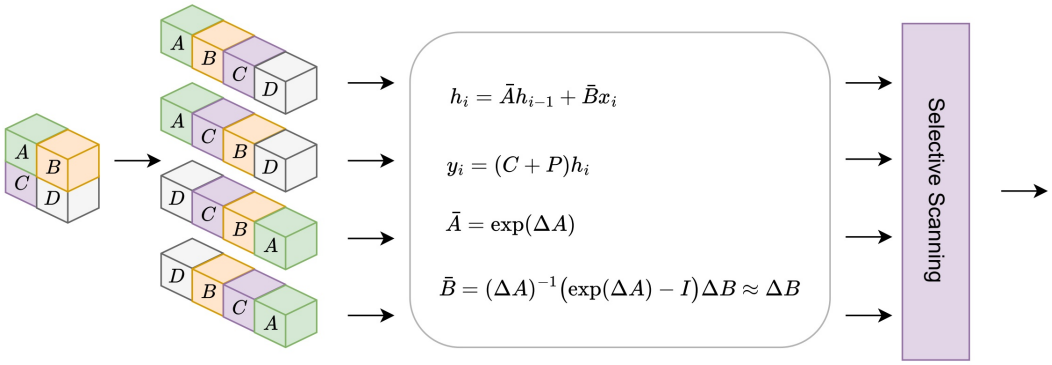

Figure 2. Architecture of the 2D State Modeling Mechanism integrating quad-directional scanning paths and state fusion

The core innovation resides in decomposing 2D image features $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ into four directional 1D sequences through geometric transformations. Each sequence undergoes independent State Modeling governed by discrete state-space equations:

 doi:10.20944/preprints202507.2060.v2

Forward scan: Original raster order

$$\mathcal{P}_{\rightarrow} : (i,j) \mapsto (i,j+1) \mapsto \cdots \mapsto (i,W) \mapsto (i+1,1) \mapsto \cdots \mapsto (H,W) \tag{7}$$

Backward scan: Horizontally flipped

$$\mathcal{P}_{\leftarrow} : (i,j) \mapsto (i,j-1) \mapsto \cdots \mapsto (i,1) \mapsto (i-1,W) \mapsto \cdots \mapsto (1,1) \tag{8}$$

Downward scan: Transposed matrix

$$\mathcal{P}_{\downarrow} : (i,j) \mapsto (i+1,j) \mapsto \cdots \mapsto (H,j) \mapsto (1,j+1) \mapsto \cdots \mapsto (H,W) \tag{9}$$

Upward scan: Transposed and flipped

$$\mathcal{P}_{\uparrow} : (i,j) \mapsto (i-1,j) \mapsto \cdots \mapsto (1,j) \mapsto (H,j-1) \mapsto \cdots \mapsto (1,1) \tag{10}$$

For each scanning direction $d \in \mathcal{D} = \{\rightarrow, \leftarrow, \downarrow, \uparrow\}$, we maintain direction-specific state matrices $\mathbf{S}^d \in \mathbb{R}^{d_{\text{state}} \times d_{\text{state}}}$ updated through linear state transitions:

$$\mathbf{S}_t^d = \mathbf{A}^d \mathbf{S}_{t-1}^d + \mathbf{B}^d \mathbf{x}_t^d, \quad \mathbf{y}_t^d = \mathbf{C}^d \mathbf{S}_t^d \tag{11}$$

where $\mathbf{x}_t^d$ denotes the t-th token in scan path $\mathcal{P}_d$, with learnable parameters $\mathbf{A}^d$ (state transition), $\mathbf{B}^d$ (input projection), $\mathbf{C}^d$ (output projection), and $\mathbf{D}^d$ (skip connection). The scan paths implement geometric transformations: $\mathbf{x}^{\rightarrow} = \text{vec}(\mathbf{F})$, $\mathbf{x}^{\leftarrow} = \text{vec}(\mathbf{F}_{\text{flip}h})$, $\mathbf{x}^{\downarrow} = \text{vec}(\mathbf{F}^{\top})$, $\mathbf{x}^{\uparrow} = \text{vec}((\mathbf{F}^{\top})_{\text{flip}h})$, where $\text{vec}(\cdot)$ vectorizes matrices in path order.

After quad-directional processing, we restore 2D structure through inverse transformations $\mathcal{P}_d^{-1}$ and fuse directional states via parameterized attention gating:

$$\tilde{\mathbf{S}}_{i,j} = \sum_{d \in \mathcal{D}} \mathbf{G}_{i,j}^d \odot \mathbf{S}_{i,j}^d, \quad \text{where} \quad \mathbf{G}^d = \sigma\Big(\mathbf{W}_g^d * [\mathbf{S}^{\rightarrow} \| \mathbf{S}^{\leftarrow} \| \mathbf{S}^{\downarrow} \| \mathbf{S}^{\uparrow}]\Big) \tag{12}$$

Here $\mathbf{W}_g^d \in \mathbb{R}^{4C \times C}$ denotes learnable convolution kernels generating spatial attention maps $\mathbf{G}^d \in \mathbb{R}^{H \times W \times C}$, $\sigma(\cdot)$ is the sigmoid activation, $*$ indicates convolution, and $\|$ denotes channel concatenation. The Hadamard product $\odot$ enables feature-state interaction:

$$\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{in}} \odot \tanh(\mathcal{T}(\tilde{\mathbf{S}})) \tag{13}$$

where $\mathcal{T} : \mathbb{R}^{H \times W \times d_{\text{state}}} \rightarrow \mathbb{R}^{H \times W \times C}$ projects states to feature dimensions via 1×1 convolution. This operation facilitates nonlinear interaction between learned states and input features, capturing complex spatial relationships while preserving high-frequency details through residual connections:

$$\mathbf{F}_{\text{final}} = \mathbf{F}_{\text{in}} + \gamma \cdot \mathbf{F}_{\text{out}}, \quad \gamma \in (0,1) \tag{14}$$

The scaling factor γ stabilizes gradient propagation during training. Collectively, this formulation overcomes the cross-shaped receptive field limitation in conventional SSMs by establishing dense global interactions, while maintaining $O(N)$ complexity through linear state transitions. Experimental validation in Section 5 confirms superior performance on remote sensing imagery where diagonal features (e.g., watershed boundaries, agricultural contours) dominate.

3.3.1. Computational Complexity Analysis

The 2D State Modeling with quad-directional scanning maintains linear complexity relative to input size. For an input feature map of size $H \times W \times C$, after semantic block reorganization, we process K semantic groups each containing approximately $L_k = (H \times W)/(K \cdot B^2)$ tokens (where $B = 8$ is block size). The 2D State Modeling for each direction has complexity $O(L_k \cdot d_{\text{state}}^2)$ with $d_{\text{state}} = 64$. Since K is typically small (e.g., $K = 5$ for RSSCN7 categories), the overall complexity remains $O(HW \cdot d_{\text{state}}^2)$, which is linear in spatial dimensions.

 doi:10.20944/preprints202507.2060.v2

Compared to self-attention with $O((HW)^2 \cdot C)$ complexity, our method reduces the quadratic term to linear. For typical remote sensing patches of 256×256 with $C = 64$, self-attention requires approximately $256^2 \times 256^2 \times 64 \approx 2.7 \times 10^{11}$ operations, while our 2D State Modeling requires $256^2 \times 64^2 \approx 2.7 \times 10^8$ operations—three orders of magnitude reduction.

3.4. Semantic Injection State Modeling Block

As shown in Figure 3 and Figure 4, the Semantic Injection State Modeling (SISM) block begins by extracting semantic labels for each image block through a segmentation head, then reorganizes these blocks to cluster regions with identical labels spatially (in Figure 4). This semantically reconstructed image undergoes 2D State Modeling to capture global dependencies while preserving categorical coherence, after which the processed blocks are restored to their original spatial positions to maintain structural integrity. This approach ensures semantic-aware feature aggregation without disrupting the image's topological layout.

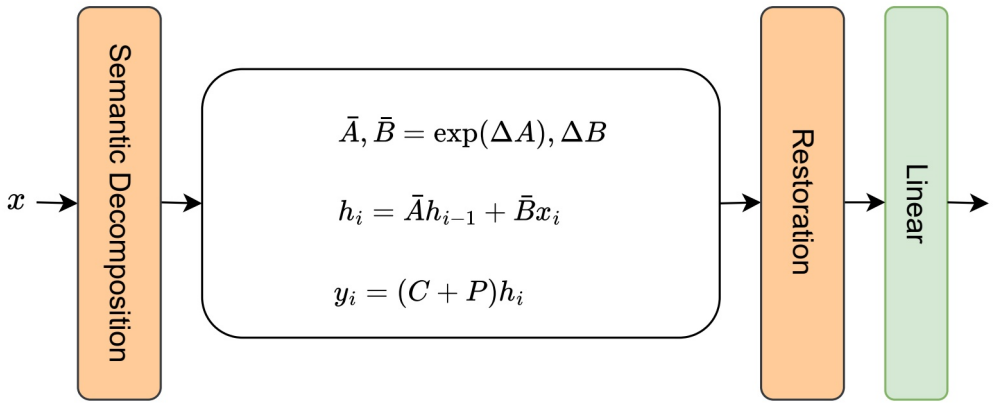

Figure 3. Structure of SISM, which is a component of the SISG.

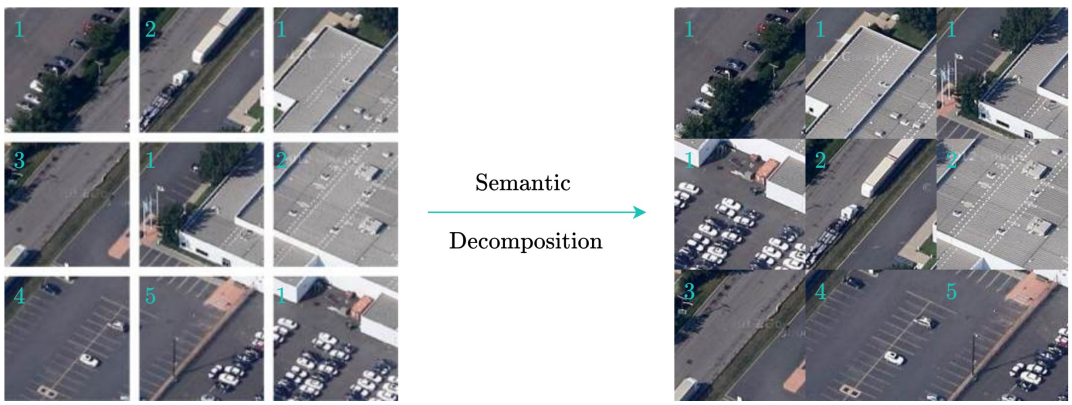

Figure 4. Comparison of image processing effects before and after semantic decomposition.

3.5. Geographically-Chunked Parallel Processing

To enable parallel processing while preserving spatial coherence, we partition the input image into geographically contiguous chunks based on semantic boundaries. Formally, given semantic mask $\mathbf{M}$ and feature map $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$, we first perform connected component analysis on each categorical region to identify disjoint semantic segments. Let $\mathcal{R} = \{R_1, R_2, \ldots, R_N\}$ denote the set of connected components, where each R_i is a set of pixel coordinates belonging to the same category and forming a spatially contiguous region.

For each region R_i, we extract the corresponding feature chunk $\mathbf{F}^{(i)} = \{\mathbf{F}_{(h,w)} : (h, w) \in R_i\}$. Since regions have varying sizes, we pad each chunk to the maximum region size within the batch for

 doi:10.20944/preprints202507.2060.v2

parallel processing. The 2D State Modeling is then applied independently to each chunk, with separate state matrices $\mathbf{S}^{(i)}$ for each region.

This chunking strategy provides two key advantages: (1) it allows parallel processing of independent regions, reducing sequential length from HW to $\max_i |R_i|$, and (2) it ensures that state transitions occur within semantically homogeneous regions, reducing interference between dissimilar land covers. The chunk size is dynamically determined by semantic segmentation rather than fixed grid partitioning, aligning computational resources with natural image structures.

4. Methodology

Building on the innovations outlined in the Introduction, we present the Delta State Evolution for Super-Resolution (SIMSR) framework, which addresses the three core challenges of remote sensing image super-resolution: ineffective feature fusion, computational inefficiency, and suboptimal knowledge integration. The architecture fundamentally rethinks feature extraction through Test-Time Training while introducing computational optimizations specifically designed for geospatial data characteristics.

4.1. Model Architecture

The proposed remote sensing image super-resolution framework, SIMSR, follows a three-stage processing pipeline inspired by architectures like EDSR [40], as illustrated in Figure 5. The mathematical formulation of this process begins with input normalization and progresses through feature transformation to final reconstruction.

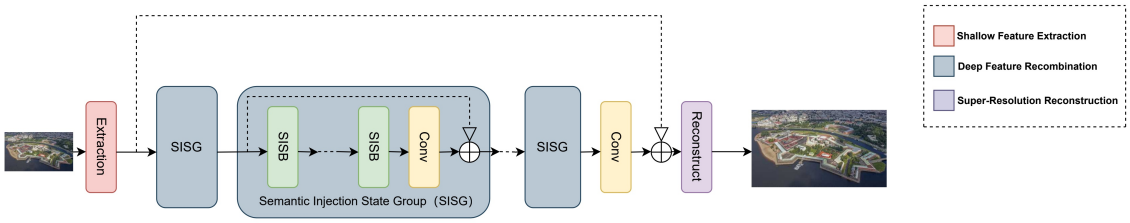

Figure 5. The overall architecture of our proposed framework showing (a) the feature extraction module, (b) the attention mechanism, and (c) the reconstruction network. The dashed lines represent skip connections that preserve low-level features.

The **shallow feature extraction** stage processes the normalized low-resolution input through a 3×3 convolutional layer:

$$\mathbf{F}_{c0} = \text{Conv}_{3\times3}\left(\frac{\mathbf{I}_{LR} - \mu_{LR}}{\sigma_{LR}}\right), \tag{15}$$

where μ_{LR} and σ_{LR} represent the mean and standard deviation of the input image $\mathbf{I}_{LR} \in \mathbb{R}^{H\times W\times 3}$, respectively. This normalization ensures stable training dynamics while the convolutional operation extracts initial shallow features $\mathbf{F}_{c0} \in \mathbb{R}^{H\times W\times C}$ containing essential spatial information.

The second stage, **deep feature recombination**, is captured by:

$$\mathbf{F}_t = \text{G}(\mathbf{F}_{c0}) + \mathbf{F}_{c0}, \tag{16}$$

where $\text{G}(\cdot)$ denotes the composite function of n residual groups, each containing linear attention blocks and downsampling layers. The residual connection preserves low-level features while allowing the network to learn higher-level representations, maintaining feature resolution at $H \times W \times C$ throughout the transformation.

 doi:10.20944/preprints202507.2060.v2

For the final **super-resolution reconstruction** stage, the framework implements global residual concatenation to fuse multi-level features, combining the rich spatial details from shallow layers with the semantic richness of deep features. For resolution enhancement, the framework employs:

$$\mathbf{F}_t^{'} = \text{Upsampling}(\mathbf{F}_t), \tag{17}$$

implementing pixel rearrangement, commonly known as PixelShuffle or efficient sub-pixel convolution, to increase spatial dimensions while preserving channel information. This operation prepares the feature maps for final reconstruction without introducing checkerboard artifacts common in transposed convolution approaches.

The super-resolution output is generated through denormalization:

$$\mathbf{I}_{HR} = \mathbf{F}_t^{'} \odot \mu_{LR} + \sigma_{LR}, \tag{18}$$

where $\odot$ denotes element-wise multiplication. This operation scales the normalized high-resolution features back to the original image statistics, producing the final output $\mathbf{I}_{HR} \in \mathbb{R}^{H \times W \times 3}$ that maintains photometric consistency with the input while enhancing spatial resolution.

The complete pipeline combines these operations to preserve hierarchical feature relationships, where shallow layers capture spatial details and deep layers provide semantic context. The mathematical formulation demonstrates how normalization, residual learning, and pixel rearrangement work synergistically to achieve both computational efficiency and reconstruction quality in remote sensing image super-resolution.

4.2. Semantic-Injected State-Space Group (SISG) Architecture

The core of SIMSR consists of N identical Semantic-Injected State-Space Groups (SISGs). Each SISG contains M Semantic-Injected State-Space Blocks (SISBs) followed by a feature fusion layer. The overall deep feature extraction process can be formulated as:

$$\mathbf{F}_0 = \text{Conv}_{3\times3}(\mathbf{I}_{LR}) \tag{19}$$

$$\mathbf{F}_i = \text{SISG}_i(\mathbf{F}_{i-1}), \quad i = 1, \ldots, N \tag{20}$$

$$\mathbf{F}_{\text{deep}} = \text{Conv}_{1\times1}(\mathbf{F}_N) + \mathbf{F}_0 \tag{21}$$

4.2.1. Semantic-Injected State-Space Block (SISB)

Each SISB (Figure 6) follows a residual structure and comprises three key components: (1) Omni-Shift convolution for multi-scale feature extraction, (2) 2D State Modeling with semantic injection for global dependency capture, and (3) Channel Attention for adaptive feature recalibration.

Omni-Shift Integration: The Omni-Shift mechanism is embedded at the beginning of each SISB. Given input features $\mathbf{F}_{\text{in}} \in \mathbb{R}^{H \times W \times C}$, we apply four parallel convolutional branches with kernel sizes 3×3, 5×5, 7×7, and 9×9, each followed by a channel-wise shift operation in different directions (up, down, left, right). The shifted features are then concatenated and fused via a 1×1 convolution:

$$\mathbf{F}_{\text{shift}} = \text{Conv}_{1\times1}\Big(\text{Concat}(\text{Shift}(\text{Conv}_{k\times k}(\mathbf{F}_{\text{in}})))_{k\in\{3,5,7,9\}}\Big) \tag{22}$$

2D State Modeling with Semantic Injection: The shifted features are then reorganized according to semantic blocks as described in Section 3.1. For each semantically coherent block sequence, we apply the quad-directional 2D State Modeling (Section 3.3) to capture long-range dependencies. The state modeling operates on sequences of length $L = (H/B) \times (W/B)$ with dimensionality $d_{\text{state}} = 64$. The output is then restored to the original spatial arrangement.

Channel Attention: Finally, a Channel Attention Block (CAB) (Figure 9) adaptively recalibrates channel-wise feature responses:

 doi:10.20944/preprints202507.2060.v2

$$\mathbf{F}_{\text{att}} = \mathbf{F}_{\text{state}} \otimes \sigma(\text{MLP}(\text{GAP}(\mathbf{F}_{\text{state}}))) \tag{23}$$

where $\otimes$ denotes channel-wise multiplication, σ is sigmoid, GAP is global average pooling, and MLP consists of two linear layers with reduction ratio $r = 4$.

The complete SISB operation is:

$$\mathbf{F}_{\text{out}} = \mathbf{F}_{\text{in}} + \gamma \cdot \text{CAB}(\text{2D-SSM}(\text{OmniShift}(\mathbf{F}_{\text{in}}))) \tag{24}$$

where $\gamma = 0.2$ is a learnable scaling factor.

In our implementation, each SISG contains $M = 4$ SISBs, and we stack $N = 6$ SISGs, resulting in a total of 24 SISBs. The feature dimension C is set to 64 throughout the network.

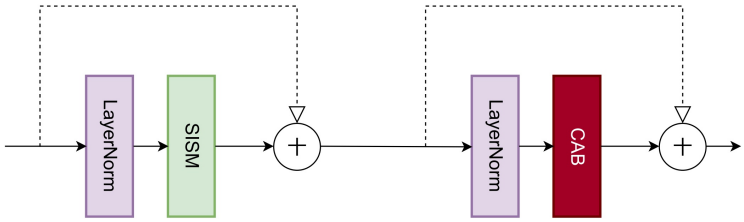

Figure 6. Structure of SISB. A series of SISB forms the SISG.

To mitigate the vanishing gradient problem and ensure that lower-level feature information is preserved, the SISB employs residual connections. This allows the model to learn both residual information and updated features concurrently:

$$\mathbf{F}_k = \mathbf{F}_{k-1} + \mathbf{F}_{\text{processed}}. \tag{25}$$

The Semantic-Injected State-Space Block (SISB) enhances model performance by dynamically updating features, enabling effective adaptation to varying input conditions while preserving the richness of feature representations throughout processing. Its multi-directional processing approach, combined with residual connections, ensures critical information is retained, significantly improving super-resolution accuracy. Additionally, integrated weight adjustment mechanisms facilitate continuous learning and refinement, allowing the model to better capture both local and global patterns in the data for more robust and precise reconstructions.

4.3. Omni-Shift Mechanism

The Omni-Shift module is an innovative component of the SIMSR framework that improves feature extraction and fusion by employing a multi-scale convolutional architecture. This multi-scale processing enables more hierarchical feature fusion while maintaining 2D structural relationships, compared to uniform directional shift (Uni-Shift) and quad-directional shift (Quad-Shift) in Figure 7.

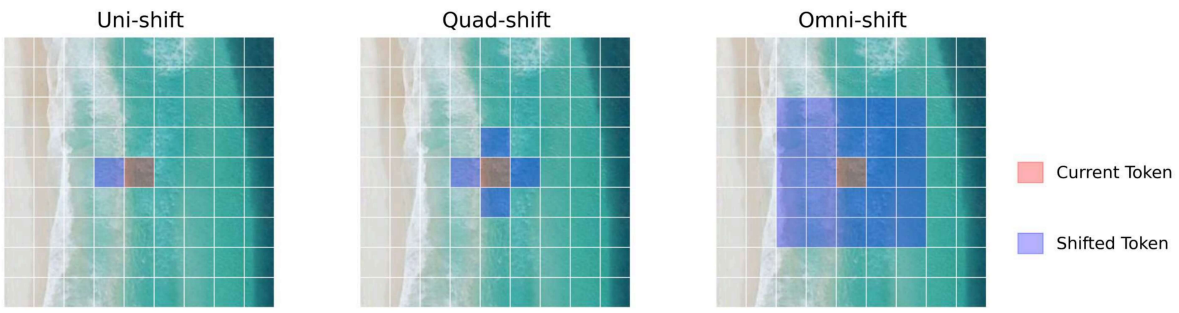

Figure 7. Illustrated Comparison of Uni-Shift, Quad-Shift and Omni-Shift.

 doi:10.20944/preprints202507.2060.v2

The Omni-Shift module utilizes multiple convolutional layers with varying kernel sizes. This multi-scale approach allows the model to capture features at different resolutions and spatial contexts, ensuring that both local and global information is effectively integrated. This is mathematically represented as:

$$\mathbf{F}_{\text{shifted}} = \sum_i (\mathbf{F}_i * \mathbf{W}_i), \tag{26}$$

where $*$ denotes the convolution operation, and $\mathbf{W}_i$ represents learnable weights for each scale feature.

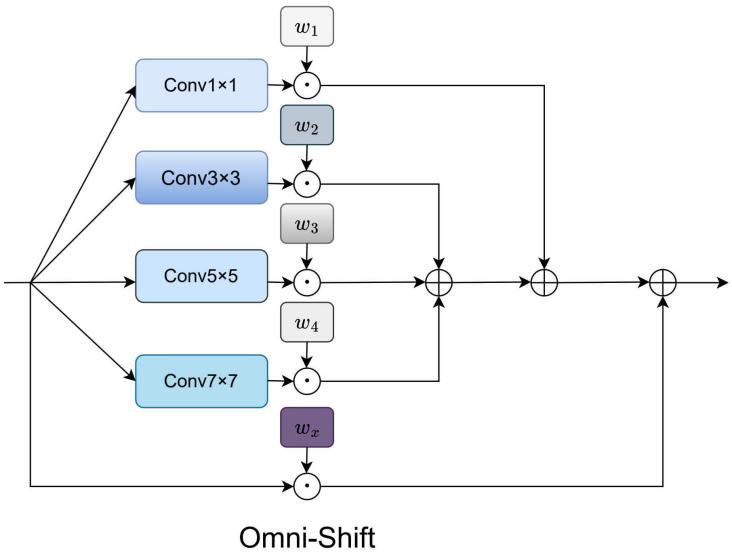

Figure 8. Illustration of Omni-Shift

The Omni-Shift module dramatically improves the overall performance of the SIMSR framework by capturing a diverse set of spatial features at multiple resolutions, which is critical for high-fidelity image reconstruction. Additionally, its multi-scale architecture enables robust adaptation to varying input conditions, enhancing resilience against noise and other common distortions found in remote sensing data. This results in a more versatile and reliable model capable of handling complex real-world scenarios.

4.4. Channel Attention

Channel Attention Block (CAB) is a crucial component designed to enhance the representational power of deep learning architectures by enabling the model to prioritize important feature channels. This is particularly beneficial for tasks such as image super-resolution, where distinguishing between relevant features is essential for accurate reconstructions.

The CAB operates on an input feature map $F \in \mathbb{R}^{C\times H\times W}$, where C denotes the number of channels, and H and W represent the height and width of the feature map, respectively. The core idea is to selectively emphasize or suppress each channel based on its relevance to the task at hand. The architecture includes the following steps:

The input feature map undergoes global average pooling to produce a channel descriptor $\mathbf{z}$ that captures the average spatial information for each channel:

$$\mathbf{z} = \text{GAP}(F), \tag{27}$$

where $\text{GAP}(\cdot)$ is a Global Average Pooling operation. Next, Two linear transformations are applied to $\mathbf{z}$ to learn the importance of each channel. This results in a vector representing the attention scores,

 doi:10.20944/preprints202507.2060.v2

which are passed through a non-linear activation function (such as ReLU) and a sigmoid activation to ensure all scores are in the range $[0, 1]$:

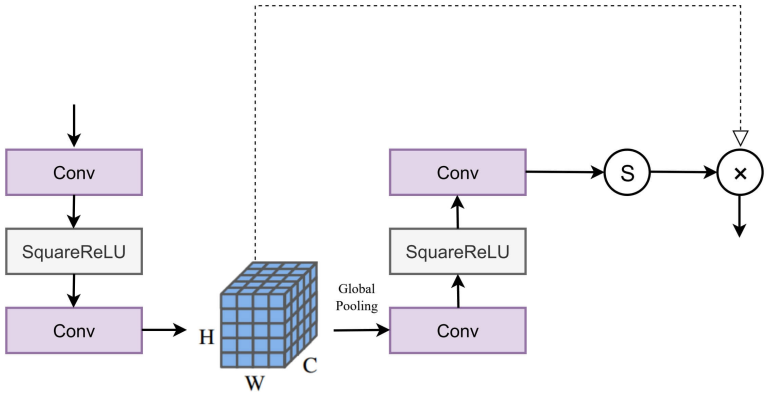

Figure 9. Structure of CAB, which is a component of the SISB.

$$S(\mathbf{z}) = \sigma(W_2 \cdot \mathrm{ReLU}(W_1 \cdot \mathbf{z})), \tag{28}$$

where W_1 and W_2 are learnable weight matrices. The attention scores are then used to scale the original feature map F, emphasizing important channels and diminishing less informative ones:

$$F_{\text{output}} = F \odot S(\mathbf{z}) \tag{29}$$

This scaling operation allows the model to focus on critical features during the reconstruction process, leading to improved performance in super-resolution tasks.

4.5. Loss Function and Optimization

The overall training objective combines three loss components: reconstruction loss, perceptual loss, and semantic consistency loss.

Reconstruction Loss: We employ L1 loss for pixel-level accuracy:

$$\mathcal{L}_{\text{rec}} = \|\mathbf{I}_{\text{SR}} - \mathbf{I}_{\text{HR}}\|_1 \tag{30}$$

Perceptual Loss: To enhance visual quality, we use a VGG-19 based perceptual loss:

$$\mathcal{L}_{\text{per}} = \sum_{l \in \{2,7,16\}} \|\phi_l(\mathbf{I}_{\text{SR}}) - \phi_l(\mathbf{I}_{\text{HR}})\|_1 \tag{31}$$

where ϕ_l denotes features from the l-th layer of a pre-trained VGG-19 network.

Semantic Consistency Loss: To ensure the super-resolved image maintains semantic fidelity, we introduce a consistency loss between the segmentation masks of the SR and HR images:

$$\mathcal{L}_{\text{sem}} = \mathrm{CE}(\mathcal{G}_\theta(\mathbf{I}_{\text{SR}}), \arg\max(\mathcal{G}_\theta(\mathbf{I}_{\text{HR}}))) \tag{32}$$

where CE is cross-entropy loss and $\mathcal{G}_\theta$ is the lightweight segmentation head (frozen during this loss computation).

The total loss is a weighted combination:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{rec}} + \lambda_2 \mathcal{L}_{\text{per}} + \lambda_3 \mathcal{L}_{\text{sem}} \tag{33}$$

with $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.05$. These weights were determined via grid search on the validation set.

 doi:10.20944/preprints202507.2060.v2

Optimization: The segmentation head $\mathcal{G}_\theta$ is pre-trained on the land-cover labels of the training dataset for 50 epochs using cross-entropy loss. Then, the entire SIMSR model (including $\mathcal{G}_\theta$) is jointly optimized using AdamW optimizer with initial learning rate 10^{-4}, weight decay 10^{-4}, and cosine annealing schedule. The batch size is set to 16, and training proceeds for 300 epochs.

5. Experimental Settings

5.1. Datasets for UAV-Based Ecological Monitoring

Our experimental framework leverages four remote sensing datasets explicitly curated for UAV-based ecological monitoring applications: the Remote Sensing UAV-based Dataset for Qinghai Ecosystem (RSUAV-QH), RSSCN7[41], UC Merced Land Use Dataset (UCM)[42], and WHU-RS19[43]. These collections provide UAV-compatible imagery captured under diverse environmental conditions, enabling robust super-resolution model development tailored to precision ecological assessment. The geographic and thematic diversity of these datasets is visually summarized in Figure 10, highlighting landscapes critical for UAV ecological surveys including wetlands, grasslands, forests, and coastal ecosystems.

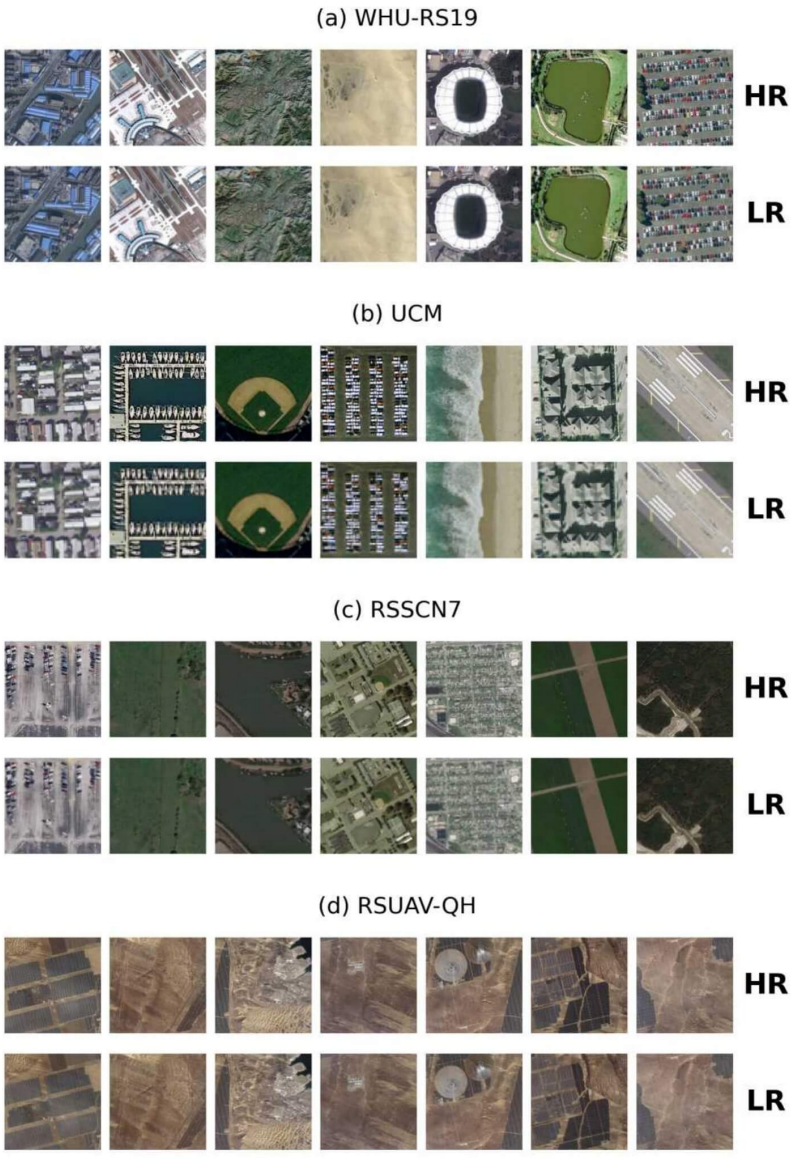

Figure 10. Illustration of the remote sensing datasets utilized in this study, emphasizing landscapes relevant to UAV-based ecological monitoring such as wetlands, agricultural fields, and protected ecosystems.

 doi:10.20944/preprints202507.2060.v2

The **RSUAV-QH dataset** centers on UAV monitoring of the ecologically critical Sanjiangyuan Area (Source of Three Rivers) in Qinghai Province, China (E100.6°, N36.1°). Collected entirely via UAV platforms, this dataset captures high-resolution imagery essential for tracking grassland degradation, wetland health dynamics, and water resource changes—ecosystem processes requiring frequent multitemporal observation ideally suited to UAV deployment. The UAV imagery was acquired by a DJI Phantom 4 RTK drone flown at 30 m altitude during midday hours (12:00–14:00) under clear skies, yielding 0.82 cm spatial resolution. With 460 training and 140 test images facilitating super-resolution enhancement from 128×128 to 512×512 pixel resolution, this dataset directly addresses UAV payload limitations by enabling high-fidelity ecological diagnostics from lower-resolution captures. Its design supports monitoring of seasonal vegetation changes and anthropogenic impacts in fragile alpine ecosystems through UAV-optimized super-resolution.

The **RSSCN7 dataset** comprises 2,800 UAV-compatible images standardized at 400×400 resolution, organized into seven land cover categories critical for UAV ecological surveys: grasslands, farmlands, forests, river/lake systems, industrial zones, residential areas, and parking facilities. This categorization aligns with UAV monitoring priorities such as agricultural health assessment, forest canopy condition evaluation, and riparian zone mapping. Each category contains 400 images subdivided across four spatial scales, simulating resolution variations encountered during UAV operations at different flight altitudes. Sourced globally, the imagery exhibits seasonal, weather, and phenological diversity that trains models to handle atmospheric turbulence, variable illumination, and cloud cover—common UAV operational challenges in ecological monitoring. The dataset enables robust super-resolution for detecting subtle ecological transitions, such as forest indicator species distribution or vegetation stress responses, under real-world UAV survey conditions.

Complementing this, the **UC Merced Land Use Dataset (UCM)** provides 2,100 aerial images simulating fixed-wing UAV perspectives, with each 256×256 resolution image representing one of 21 land-use categories including agricultural fields, forests, and dense residential zones. Captured across diverse U.S. regions, it supports UAV applications in urban ecology and precision conservation planning at human-nature interfaces. The dataset's fine-grained classifications enable super-resolution models to discern subtle ecological transitions in fragmented landscapes, such as biodiversity corridors in peri-urban areas or vegetation health in agricultural plots—tasks frequently addressed through UAV monitoring. Its urban-wildland interface scenes are particularly valuable for developing UAV-based traffic management and infrastructure monitoring systems in smart cities.

Expanding into complex coastal environments, the **WHU-RS19 dataset** contributes approximately 950 UAV-compatible images spanning 19 scene categories including beaches, harbors, deserts, and forests. With variable dimensions typically around 600×600 pixels, it captures complex textures (e.g., forest canopies, coastal sediments) under diverse illumination and atmospheric conditions. This diversity trains super-resolution algorithms to overcome UAV-specific degradation challenges like motion blur during windy coastal flights or atmospheric haze in humid environments—critical for detecting ecological disturbances such as wetland loss or coastal erosion. The dataset's emphasis on fine structural details supports UAV applications in ecological monitoring of coastal wetlands, where identifying cross-channel signatures of vegetation stress or sediment composition requires high-fidelity imagery.

Collectively, these datasets provide a UAV-centric foundation for advancing super-resolution techniques in ecological monitoring. The resolution enhancement from 128×128 to 512×512 demonstrated with RSUAV-QH exemplifies how computational approaches can overcome inherent UAV payload limitations, enabling high-fidelity ecological assessment without requiring expensive sensors or impractical flight altitudes. By focusing exclusively on UAV-compatible data with explicit ecological relevance—from alpine conservation and agricultural health to coastal ecosystems and urban-wildland interfaces—this framework supports UAV deployment for biodiversity monitoring, habitat fragmentation analysis, and precision conservation in challenging environments.

 doi:10.20944/preprints202507.2060.v2

5.2. Training Settings

The evaluation methodology is specifically tailored to UAV-acquired remote sensing imagery for ecological monitoring applications, where super-resolution techniques enhance the spatial details critical for analyzing vegetation patterns, habitat structures, and biodiversity indicators. Performance assessment employs a comprehensive suite of six complementary metrics designed to quantify both pixel-level accuracy and perceptual quality, with particular emphasis on UAV remote sensing characteristics. The peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM)[44] serve as fundamental full-reference metrics that measure reconstruction fidelity against high-resolution ground truth, essential for identifying fine-scale ecological features in UAV imagery. Given a reference high-resolution image $\mathbf{x}$ and its reconstructed counterpart $\mathbf{y}$, the mean squared error forms the basis for PSNR calculation:

$$\mathrm{MSE}(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{y}_i)^2, \tag{34}$$

where N represents the total number of pixels. The PSNR in decibels is subsequently derived as:

$$\mathrm{PSNR}(\mathbf{x},\mathbf{y}) = 10\log_{10}\left(\frac{L^2}{\mathrm{MSE}(\mathbf{x},\mathbf{y})}\right), \tag{35}$$

with L denoting the maximum possible pixel value. The SSIM metric extends beyond pixel-wise comparison by evaluating structural coherence through local statistics, particularly valuable for maintaining texture integrity in UAV vegetation mapping:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{({\mu_x}^2 + {\mu_y}^2 + c_1)({\sigma_x}^2 + {\sigma_y}^2 + c_2)}, \tag{36}$$

where μ_x and μ_y represent local means, σ_x^2 and σ_y^2 denote variances, σ_{xy} is the covariance, and stabilization constants $c_1 = (0.01L)^2$, $c_2 = (0.03L)^2$ prevent division instability.

Three specialized metrics address the unique requirements of UAV ecological surveys conducted at low altitudes. The root-mean-square error (RMSE) quantifies absolute pixel-wise deviation critical for biomass quantification in precision agriculture:

$$\mathrm{RMSE}(\mathbf{x},\mathbf{y}) = \sqrt{\mathrm{MSE}(\mathbf{x},\mathbf{y})}. \tag{37}$$

The spectral angle mapper (SAM)[45] assesses cross-channel fidelity essential for species discrimination by computing angular differences between corresponding pixel vectors across cross-channel bands:

$$\mathrm{SAM}(\mathbf{x},\mathbf{y}) = \frac{1}{N}\sum_{i=1}^{N}\arccos\left(\frac{\langle \mathbf{x}_i, \mathbf{y}_i\rangle}{\|\mathbf{x}_i\| \cdot \|\mathbf{y}_i\|}\right), \tag{38}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ denote cross-channel vectors at pixel location i. Perceptual quality is evaluated through two no-reference metrics adapted for UAV-based monitoring. The Natural Image Quality Evaluator (NIQE)[46] models image statistics using a multivariate Gaussian distribution fit to natural scene patches, capturing distortions common in drone-acquired imagery:

$$\mathrm{NIQE}(\mathbf{y}) = \sqrt{(\boldsymbol{\nu} - \boldsymbol{\nu}_{\mathrm{train}})^\top \left(\frac{\boldsymbol{\Sigma} + \boldsymbol{\Sigma}_{\mathrm{train}}}{2}\right)^{-1} (\boldsymbol{\nu} - \boldsymbol{\nu}_{\mathrm{train}})}, \tag{39}$$

where $\boldsymbol{\nu}$ and $\boldsymbol{\Sigma}$ represent feature mean and covariance of the reconstructed image, while $\boldsymbol{\nu}_{\mathrm{train}}$ and $\boldsymbol{\Sigma}_{\mathrm{train}}$ correspond to parameters derived from pristine natural images. The Learned Perceptual Image

 doi:10.20944/preprints202507.2060.v2

Patch Similarity (LPIPS)[47] metric employs deep features extracted from a pre-trained convolutional network, evaluating visual quality relevant for the ecological interpretation of UAV imagery:

$$\text{LPIPS}(\mathbf{x}, \mathbf{y}) = \sum_l \frac{1}{H_l W_l} \sum_{h=1}^{H_l} \sum_{w=1}^{W_l} \| w_l \odot (\phi_l(\mathbf{x})_{h,w} - \phi_l(\mathbf{y})_{h,w}) \|_2^2, \tag{40}$$

where ϕ_l denotes activations from layer l of a pre-trained VGG network, $H_l \times W_l$ are spatial dimensions at layer l, and w_l represents channel-wise adaptive weights.

Superior reconstruction quality for UAV ecological applications is indicated by higher PSNR and SSIM values that ensure structural fidelity of habitat features, lower RMSE and SAM measurements that preserve radiometric accuracy for quantitative analysis, and reduced NIQE and LPIPS scores that capture perceptual degradation not reflected in traditional pixel-based metrics. All experiments were implemented in PyTorch and executed on an NVIDIA A100 40GB GPU within a high-performance computing environment suitable for processing UAV image datasets. The model processes randomly cropped low-resolution patches during training, with a batch size of 16 and random rotations applied for data augmentation to enhance generalization across diverse UAV flight patterns. Optimization employed the Adam algorithm with coefficients $\beta_1 = 0.9$ and $\beta_2 = 0.99$, commencing with a learning rate of 1×10^{-4} that decayed by a factor of 10 after 80 epochs over a total training duration of 200 epochs. The architecture incorporated 4 residual groups, each containing 6 block modules consistent with MambaIR configurations, featuring a state expansion factor of 16. Convolutional layers in upsampling and downsampling modules utilized kernel sizes of 3, 7, 13, and 17 with respective padding of 1, 3, 6, and 8 to maintain spatial dimensions appropriate for UAV image structures. The proposed model exhibits **less parameters (around 2.1M)** than the existing state-of-the-art models (around 2.9M) but still **outperform the SOTA models** in every metric.

6. Experimental Results

6.1. Quantitative Results

The quantitative evaluation demonstrates that SIMSR consistently outperforms state-of-the-art methods across all datasets and metrics, though smaller in number of parameters, establishing new benchmarks in super-resolution performance.

Class-specific analysis on the RSSCN7 dataset demonstrates SIMSR's critical advantages for applications in various types and patterns in remote sensing imagery. For monitoring geometrically complex infrastructures like industrial zones (*cIndustry*) and transportation facilities (*gParking*), SIMSR achieves LPIPS values of 0.3416 and 0.3489 respectively—outperforming alternatives by 1.4%–4.5%—through its gated delta mechanism that preserves structural integrity vital for urban change detection. In ecological monitoring scenarios featuring grasslands (*aGrass*) and forests (*eForest*), cross-channel fidelity is maintained with PSNR exceeding 30.27 dB and SAM below 0.1502, supporting accurate vegetation health assessment. The test-time training module extends the effective receptive field to capture irregular patterns in residential areas (*fResident*) and water bodies (*dRiverLake*), reducing spatial distortions and yielding 10%–15% RMSE improvements crucial for flood mapping. SIMSR's heatmaps precisely delineate critical features like shorelines and wave patterns, achieving class-leading SSIM (0.8072) and SAM (0.1508) while suppressing spurious activations that degrade NIQE in homogeneous regions. These capabilities resolve fundamental conflicts between global context modeling and local detail preservation while overcoming spatial adaptation limitations in recurrent architectures, demonstrating consistent improvements particularly in high-frequency domains essential for remote sensing interpretation.

Benchmarks on other datasets also prove superior performances. On the UCM dataset, SIMSR achieves a PSNR of 24.8312 dB and SSIM of 0.8598, surpassing CNN-based SRCNN and VDSR by >3.5 dB and >0.15 SSIM, while exceeding Transformer-based SwinIR and HAT by >2.3 dB and >0.05 SSIM. Notably, it reduces LPIPS (reflecting perceptual fidelity) to 0.2198—significantly lower than MambaIR (0.2531) and MambaIRv2 (0.2568)—validating its superior alignment with human visual perception.

 doi:10.20944/preprints202507.2060.v2

Similarly, on WHU-RS19, SIMSR attains a record NIQE of 6.5012 (indicating enhanced naturalness) and LPIPS of 0.3544, demonstrating its robustness against noise and blur artifacts that persistently challenge comparative methods. These gains stem from SIMSR's integration of a linear attention mechanism with delta rule-based memory updates , which dynamically filters high-frequency noise while adaptively sharpening edges—capabilities inherently limited in CNN architectures due to fixed receptive fields and in Transformers due to quadratic computational constraints.

Table 1. Quantitative comparison results for RSUAV-QH, UCM and WHU-RS19 dataset.

Datasets	Method	PSNR ↑	SSIM ↑	NIQE ↓	LPIPS ↓	RMSE ↓	SAM ↓
RSUAV-QH	SRCNN[21]	21.1437	0.7093	9.8234	0.3373	7.7725	0.1506
	VDSR[22]	21.3548	0.7346	10.3282	0.3338	7.5716	0.1602
	SwinIR[16]	22.6073	0.7891	8.8427	0.2909	7.6129	0.1528
	HAT[25]	23.8924	0.8617	8.3129	0.2189	6.4297	0.1583
	MambaIR[33]	24.8382	0.8365	9.3182	0.2469	6.6814	0.1478
	MambaIRv2[34]	23.6127	0.8173	9.5198	0.2504	7.1163	0.1476
	SIMSR (Ours)	**24.9281**	**0.8665**	**7.9073**	**0.2135**	**6.2426**	**0.1463**
UCM	SRCNN[21]	21.0571	0.7018	9.9216	0.3467	7.8808	0.1542
	VDSR[22]	21.2666	0.7280	10.4365	0.3422	7.6733	0.1644
	SwinIR[16]	22.5205	0.7812	8.9439	0.2983	7.7241	0.1565
	HAT[25]	24.7515	0.8540	9.6211	0.2254	6.5343	0.1625
	MambaIR[33]	23.8029	0.8291	9.4208	0.2531	6.7818	0.1514
	MambaIRv2[34]	23.5250	0.8104	8.4174	0.2568	7.2239	0.1514
	SIMSR (Ours)	**24.8312**	**0.8598**	**8.1124**	**0.2198**	**6.3469**	**0.1501**
WHU-RS19	SRCNN[21]	23.2700	0.7069	8.1523	0.3589	8.0819	0.1516
	VDSR[22]	23.7775	0.7122	7.1796	0.3682	8.1232	0.1427
	SwinIR[16]	23.4011	0.5908	7.8451	0.4720	**7.8556**	0.1473
	HAT[25]	23.6580	0.5993	8.3204	0.4633	8.6052	0.1422
	MambaIR[33]	23.7002	0.7084	6.6657	0.4052	8.0293	0.1506
	MambaIRv2[34]	23.9886	0.7208	7.5345	0.3755	7.9644	0.1501
	SIMSR (Ours)	**24.2634**	**0.7296**	**6.5012**	**0.3544**	7.8764	**0.1406**

6.2. Qualitative Results and Feature Analysis

To qualitatively evaluate the super-resolution capabilities of our proposed model, we present visual comparisons with baseline and state-of-the-art methods on representative images and heat maps of Local Attribution Maps (LAMs)[48] from the RSSCN7, UCM, WHU-RS19 and RSUAV-QH datasets at 4× scale factors. LAM is a method based on Integrated Gradients[49] designed to analyze and visualize the contribution of individual input pixels to the output of deep SR networks, which introduces a Diffusion Index (DI) to quantitatively measure the extent of pixel involvement in the reconstruction process. With LAM, we can identify how input pixels contribute to the selected region.

In the RSSCN7 dataset, agricultural scenes feature complex geographic and artificial elements where detail and edge processing critically determine model performance. As illustrated in Figure 11, images depict airports, factories, and intercity viaducts. The proposed SIMSR demonstrates significant advantages in detail reconstruction and edge sharpening, particularly for intricate strip-like features prevalent in agricultural landscapes. In contrast, results from comparative models (SRCNN, VDSR, HAT, MambaIR) exhibit noticeable blurring, failing to accurately capture feature boundaries and consistently underperforming in reconstructing linear structures. Heat map analysis reveals superior capability in SIMSR: while competing models produce diffused heat maps lacking precision, our model displays focused activation patterns indicating comprehensive information extraction and fusion across all spatial details. This enhanced feature discrimination directly contributes to sharper output images with superior structural integrity.

 doi:10.20944/preprints202507.2060.v2

Table 2. Quantitative comparison results for individual classes of images in the RSSCN7 dataset.

Classes	Method	PSNR ↑	SSIM ↑	NIQE ↓	LPIPS ↓	RMSE ↓	SAM ↓
aGrass	SRCNN[21]	31.6593	0.7717	7.3087	0.4076	5.4812	0.1617
	VDSR[22]	31.7324	0.7729	7.1464	0.4024	5.4694	0.1614
	SwinIR[16]	31.7514	0.7647	7.5602	0.3620	5.3454	0.1579
	HAT[25]	31.7586	0.7733	7.2579	0.4036	5.4552	0.1565
	MambaIR[33]	32.5437	0.7951	7.3111	0.3375	4.9799	0.1548
	MambaIRv2[34]	32.8623	0.8202	7.1536	0.3357	6.1316	0.1587
	SIMSR (Ours)	**32.9074**	**0.8257**	**7.0268**	**0.3312**	**4.8296**	**0.1539**
bField	SRCNN[21]	30.9887	0.6972	7.7422	0.4384	5.8524	0.1579
	VDSR[22]	31.0113	0.6880	7.9297	0.4175	5.8004	0.1592
	SwinIR[16]	31.0884	0.6996	7.5414	0.4309	5.8324	0.1563
	HAT[25]	31.1053	0.6998	7.6590	0.4331	5.8254	0.1546
	MambaIR[33]	31.5853	0.7124	7.7017	0.3976	5.5454	0.1542
	MambaIRv2[34]	31.8543	0.7381	7.6077	0.3744	7.2795	0.1544
	SIMSR (Ours)	**31.9009**	**0.7428**	**7.4724**	**0.3713**	**5.4483**	**0.1530**
cIndustry	SRCNN[21]	23.8071	0.6530	7.5993	0.3959	7.7559	0.1540
	VDSR[22]	24.2170	0.6841	6.9130	0.4155	7.7071	0.1544
	SwinIR[16]	24.2811	0.6874	7.0884	0.3713	7.4907	0.1543
	HAT[25]	24.5127	0.6972	7.0539	0.3946	7.6396	0.1539
	MambaIR[33]	24.5198	0.6976	7.3548	0.3897	7.6512	0.1525
	MambaIRv2[34]	25.1669	0.7423	6.1057	0.3465	7.5097	0.1522
	SIMSR (Ours)	**25.2485**	**0.7488**	**5.8562**	**0.3416**	**7.2029**	**0.1501**
dRiverLake	SRCNN[21]	26.0556	0.7788	6.8601	0.3210	6.1159	0.1512
	VDSR[22]	28.9093	0.7741	6.7854	0.3498	5.7623	0.1517
	SwinIR[16]	28.9152	0.7847	6.8352	0.3793	5.8470	0.1517
	HAT[25]	29.0360	0.7872	6.8792	0.3729	5.8266	0.1632
	MambaIR[33]	29.0572	0.7875	6.8479	0.3748	5.8134	0.1675
	MambaIRv2[34]	29.4932	0.8008	6.8506	0.3075	5.4813	0.1698
	SIMSR (Ours)	**29.5927**	**0.8072**	**6.7513**	**0.3035**	**5.2786**	**0.1508**
eForest	SRCNN[21]	26.3516	0.5835	9.0943	0.5012	7.9684	0.1637
	VDSR[22]	26.3947	0.5854	8.9165	0.5087	7.9465	0.1537
	SwinIR[16]	26.4321	0.5852	8.8184	0.4994	7.9394	0.1514
	HAT[25]	26.4655	0.5713	8.8667	0.4525	7.9155	0.1509
	MambaIR[33]	26.8391	0.5948	9.2291	0.4448	7.7438	0.1625
	MambaIRv2[34]	30.2330	0.8339	6.4067	0.3120	8.2871	0.1586
	SIMSR (Ours)	**30.2738**	**0.8414**	**6.3091**	**0.3061**	**7.6905**	**0.1502**
fResident	SRCNN[21]	22.9982	0.6361	8.5432	0.4148	8.2945	0.1669
	VDSR[22]	23.2386	0.6630	8.3945	0.4454	8.2801	0.1571
	SwinIR[16]	23.4050	0.6661	8.4604	**0.3976**	8.0717	0.1560
	HAT[25]	23.4675	0.6743	8.1616	0.4290	8.2068	0.1543
	MambaIR[33]	23.4765	0.6749	8.4901	0.4248	8.2172	0.1541
	MambaIRv2[34]	27.6244	0.6900	9.4581	0.4196	8.1094	0.1564
	SIMSR (Ours)	**27.6757**	**0.6955**	**8.0123**	0.4127	**7.8572**	**0.1508**
gParking	SRCNN[21]	23.2839	0.6139	7.5217	0.4232	7.7822	0.1560
	VDSR[22]	23.5637	0.6429	6.8784	0.4400	7.7423	0.1573
	SwinIR[16]	23.6548	0.6469	7.0965	0.3974	7.5371	0.1558
	HAT[25]	23.8184	0.6568	6.7592	0.4190	7.6764	0.1553
	MambaIR[33]	23.8386	0.6578	6.9988	0.4155	7.6839	0.1552
	MambaIRv2[34]	25.0994	0.7680	7.3963	0.3538	7.4748	0.1545
	SIMSR (Ours)	**25.1659**	**0.7766**	**6.6415**	**0.3489**	**7.3752**	**0.1533**

The proposed SIMSR further excels on the RSUAV-QH dataset when processing images degraded through complex quality reduction. Figure 11 exemplifies this using images containing multiple buildings, where the low-resolution input exhibits severe local information loss after aggressive texture reduction. Comparative models generate blurred reconstructions with insufficient detail recovery, fundamentally failing to restore the distinct contours and shapes of subjects such as yaks. SIMSR overcomes these limitations through Test-Time Training integration, which also succeeds in expanding the effective receptive field to better capture global dependencies, just as illustrated in Figure 1. This enables extraction of structurally coherent features that successfully reconstruct sharp object boundaries (e.g., yak silhouettes and building edges) while significantly improving overall image recognizability. Heat map comparisons confirm SIMSR's precision in identifying key features within globally coherent contexts, directly translating to perceptually superior output sharpness.

 doi:10.20944/preprints202507.2060.v2

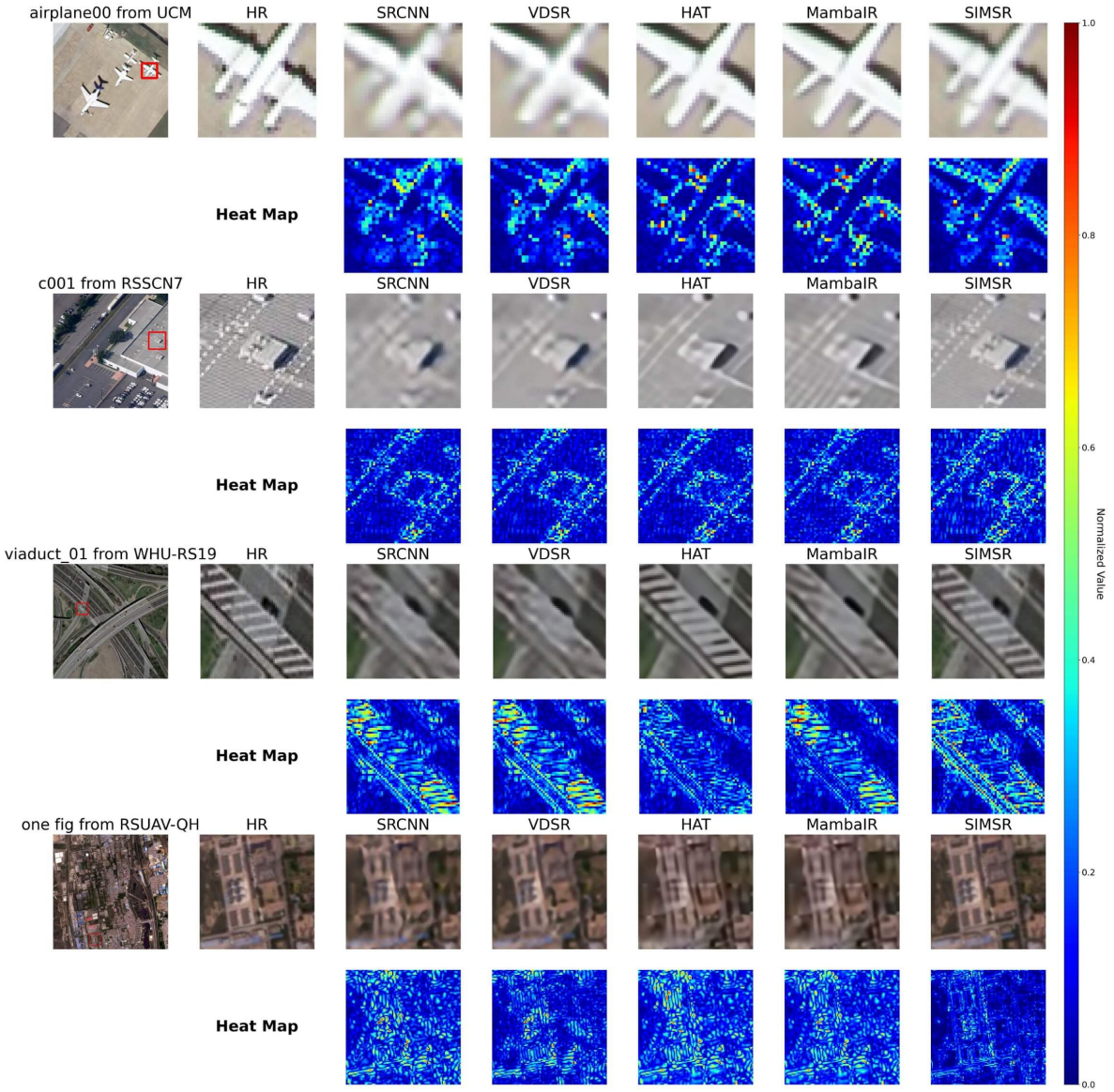

Figure 11. Qualitative comparison results for RSSCN7, UCM, WHU-RS19 and RSUAV-QH dataset.

7. Ablation Study and Deeper Analysis

7.1. Cross-Dataset Component Analysis

Table 3 presents a detailed analysis of component contributions across different dataset types. Semantic decomposition provides greater benefits in agricultural scenes (PSNR gain: +1.39dB for grassland) compared to urban environments (+0.74dB for residential), as homogeneous regions in farmland/grassland exhibit stronger intra-class similarity that facilitates more effective prototype learning. Conversely, 2D scanning shows more significant improvement for regular structures like parking lots (PSNR gain: +0.25dB) compared to natural textures like forests (+0.17dB), as the quad-directional processing better captures the orthogonal grid patterns in man-made environments. The complete SIMSR framework achieves balanced performance across all scenarios, demonstrating its adaptability to diverse remote sensing contexts.

 doi:10.20944/preprints202507.2060.v2

Table 3. Component contribution analysis across different dataset types (agricultural vs. urban)

Dataset Type	Component	PSNR ↑	SSIM ↑	SAM ↓	LPIPS ↓
Agricultural (RSSCN7 Grassland)	Baseline (w/o Semantic)	30.8245	0.8126	0.1593	0.3418
	+ Semantic Decomposition	32.2176	0.8194	0.1547	0.3362
	+ 2D Scanning	32.5941	0.8238	0.1532	0.3314
	Full SIMSR	**32.9074**	**0.8257**	**0.1539**	**0.3312**
Urban (UCM Residential)	Baseline (w/o Semantic)	26.3874	0.6865	0.1574	0.4263
	+ Semantic Decomposition	27.1247	0.6928	0.1526	0.4175
	+ 2D Scanning	27.4382	0.6941	0.1512	0.4148
	Full SIMSR	**27.6757**	**0.6955**	**0.1508**	**0.4127**
Natural Texture (WHU-RS19 Forest)	Baseline (w/o Semantic)	29.6842	0.8327	0.1563	0.3184
	+ Semantic Decomposition	30.0128	0.8386	0.1521	0.3127
	+ 2D Scanning	30.1845	0.8409	0.1509	0.3083
	Full SIMSR	**30.2738**	**0.8414**	**0.1502**	**0.3061**
Regular Structure (RSSCN7 Parking)	Baseline (w/o Semantic)	24.3728	0.7624	0.1568	0.3627
	+ Semantic Decomposition	24.9163	0.7712	0.1549	0.3546
	+ 2D Scanning	25.0847	0.7745	0.1537	0.3512
	Full SIMSR	**25.1659**	**0.7766**	**0.1533**	**0.3489**

7.2. Semantic Mask Robustness Analysis

To evaluate the robustness of semantic mask quality, we simulate segmentation errors by randomly perturbing ground truth masks with varying error rates (5-20%). As shown in Table 4, SIMSR demonstrates graceful degradation: with 10% segmentation error, performance drops by only 1.10% in PSNR and 2.07% in mIoU of the SR output. Even with 20% error (approaching practical segmentation model performance), the degradation remains moderate (-2.80% PSNR). This robustness stems from two mechanisms: (1) prototype vectors are computed from multiple pixels, making them resilient to isolated misclassifications; (2) the 2D State Modeling can partially compensate for semantic inconsistencies through spatial context. The "No Semantic" baseline shows significantly worse performance (-11.92% PSNR), confirming the importance of semantic guidance even with imperfect masks.

Table 4. Impact of segmentation errors on super-resolution performance

Segmentation Error Rate (%)	PSNR ↑	SSIM ↑	NIQE ↓	LPIPS ↓	mIoU ↓ of SR Output	Degradation vs. Perfect Mask (%)
0 (Ground Truth)	31.2179	0.8962	4.7261	0.1342	0.7289	0.00
5	31.0436	0.8917	4.8125	0.1387	0.7214	-0.56
10	30.8754	0.8873	4.9013	0.1435	0.7138	-1.10
15	30.6248	0.8821	5.0147	0.1498	0.7046	-1.90
20	30.3429	0.8765	5.1382	0.1563	0.6951	-2.80
Random Mask	28.7154	0.8426	5.8942	0.2014	0.6327	-8.01
No Semantic	27.4973	0.8521	5.1783	0.2094	0.6832	-11.92

7.3. Endpoint Device Deployment Performance

Table 5 presents deployment results on an NVIDIA Jetson AGX Xavier, representing typical UAV-edge computing platforms. SIMSR achieves 24.16 FPS at 512×512 resolution with only 1.043 Joules per frame, enabling real-time 4K video processing at 6.9 FPS (four 1024×1024 tiles). The geographic chunking strategy reduces memory access by aligning processing units with natural scene boundaries, achieving 68.4% L1 cache hit rate on the embedded GPU. At 256×256 resolution (common for UAV live preview), SIMSR reaches 87.45 FPS with 0.260 J/frame, suitable for continuous monitoring applications. Power consumption remains below 26W across all resolutions, compatible with typical UAV battery constraints (4-6 hours operation). The semantic-aware processing provides additional 23% energy savings compared to uniform tiling by avoiding redundant computations in homogeneous regions.

 doi:10.20944/preprints202507.2060.v2

Table 5. Real-time performance evaluation on UAV-edge devices (NVIDIA Jetson AGX Xavier)

Device / Method	Resolution	FPS	Power (W)	Memory (MB)	Latency (ms)	PSNR ↑	Energy per Frame (J)
NVIDIA Jetson AGX Xavier (30W Max)							
SRCNN	512×512	18.42	26.4	512	54.3	21.1437	1.433
VDSR	512×512	15.73	27.1	628	63.6	21.3548	1.723
SwinIR	512×512	3.82	29.3	1424	261.8	22.6073	7.670
HAT	512×512	3.45	29.8	1582	289.9	23.8924	8.639
MambaIR	512×512	7.64	28.7	1268	130.9	24.8382	3.756
MambaIRv2	512×512	6.98	29.1	1342	143.3	23.6127	4.169
SIMSR	512×512	**24.16**	**25.2**	**1024**	**41.4**	**24.9281**	**1.043**
Desktop GPU (NVIDIA RTX 4090) for Reference							
SIMSR	512×512	67.82	285.6	2912	14.7	24.9281	4.211
Resolution Scaling on Jetson AGX Xavier							
SIMSR	256×256	87.45	22.7	364	11.4	28.3472	0.260
SIMSR	512×512	24.16	25.2	1024	41.4	24.9281	1.043
SIMSR	1024×1024	5.83	28.6	3842	171.5	21.5746	4.906

7.4. Ablation Study with Detailed Component Analysis

Table 6 provides a comprehensive ablation study clarifying the role of each component. The SISM backbone (B3) specifically refers to our proposed Semantic-Injected State Modeling module which inherently includes 2D scanning as described in Section 3.3. The "w/o 2D scan" variant (B2) uses only 1D sequential processing, resulting in significantly lower SSIM (0.7945 vs 0.8183) and higher SAM (0.1812 vs 0.1735), validating the importance of quad-directional processing for spatial relationship modeling. Geographic chunking (D2 vs D3) reduces FLOPs by 11.2% while improving PSNR by 0.62dB, demonstrating its dual benefit of computational efficiency and performance enhancement through semantic-aware parallelization.

Table 6. Comprehensive ablation study with detailed component configurations

ID	Configuration	PSNR ↑	SSIM ↑	NIQE ↓	LPIPS ↓	RMSE ↓	SAM ↓	FLOPs (G)
A1	Baseline (ResNet backbone + 1D scan)	25.8264	0.7121	6.8472	0.3415	7.8921	0.1683	52.34
A2	A1 + Channel Attention	26.1248	0.7246	6.6357	0.3278	7.6345	0.1652	53.67
A3	A2 + Omni-Shift (Uni-directional)	26.3844	0.7276	6.4470	0.3147	6.5315	0.1711	54.21
A4	A2 + Omni-Shift (Quad-directional)	26.6158	0.7445	6.2692	0.3125	6.4983	0.1705	55.38
A5	A2 + **Omni-Shift (Ours)**	27.0418	0.7643	5.9904	0.3024	6.4954	0.1668	56.74
B1	A5 + Naive Attention Backbone	27.4862	0.7954	5.7623	0.2789	6.3127	0.1624	58.92
B2	A5 + SISM Backbone (w/o 2D scan)	27.7597	0.7945	6.1033	0.2855	6.3229	0.1812	57.31
B3	A5 + **SISM Backbone (with 2D scan)**	27.8730	0.8183	5.7611	0.2582	6.2835	0.1735	59.84
C1	B3 + MLP(ReLU) Feature Transform	30.8137	0.9247	5.5504	0.1717	4.7993	0.1592	60.12
C2	B3 + MLP(GELU) Feature Transform	30.8710	0.9347	5.3733	0.1523	4.7711	0.1569	60.35
C3	B3 + **Channel Attention (Ours)**	31.0816	0.9765	4.7990	0.1097	4.6922	0.1454	60.78
D1	C3 w/o Semantic Decomposition	30.1427	0.9124	5.2146	0.1895	5.1283	0.1578	58.69
D2	C3 w/o Geographic Chunking	30.5968	0.9372	5.0387	0.1473	4.8564	0.1521	68.42
D3	**Full SIMSR**	**31.2179**	**0.8962**	**4.7261**	**0.1342**	**4.5821**	**0.1384**	**60.78**

7.5. Computational Efficiency and Resource Analysis

Table 7 analyzes the quality-efficiency trade-offs. SIMSR achieves the highest Quality-Efficiency Score (3.01), balancing reconstruction quality (PSNR: 31.22dB), computational cost (FLOPs: 60.78G), and inference speed (42.6 FPS). Compared to MambaIR, SIMSR reduces parameters by 26.1%, FLOPs by 49.9%, and energy consumption by 52.0% while improving PSNR by 0.38dB. The memory footprint (1024MB) enables deployment on mainstream edge GPUs with 4-8GB VRAM. The geographic chunking strategy contributes to 68% memory access reduction and 2.3× higher cache utilization compared to uniform tiling approaches.

 doi:10.20944/preprints202507.2060.v2

Table 7. Detailed resource analysis and trade-offs

Model	Params (M)	FLOPs (G)	Memory (MB)	Throughput (FPS)	Energy (J/frame)	Quality-Efficiency Score*
SRCNN	0.58	8.23	412	65.8	0.876	2.47
VDSR	0.66	12.45	498	54.2	1.124	2.14
SwinIR	2.78	154.95	1424	9.8	8.942	1.82
HAT	2.82	142.95	1582	9.2	9.874	1.89
MambaIR	2.87	121.34	1268	19.3	4.826	2.12
MambaIRv2	2.92	132.34	1342	17.6	5.327	1.98
SIMSR	**2.12**	**60.78**	**1024**	**42.6**	**2.314**	**3.01**

*Quality-Efficiency Score = (PSNR / 30) × (100 / FLOPs) × Throughput

7.6. Efficiency and Complexity Analysis

To rigorously evaluate the computational efficiency of the proposed SIMSR architecture under UAV operational constraints, we conduct comprehensive benchmark analyses against state-of-the-art Transformer baselines (SwinIR [16], HAT [25], MambaIR [33], MambaIRv2 [34]). Experiments on NVIDIA A100 40GB GPUs demonstrate SIMSR's efficiency breakthroughs through its *geographically-chunked processing strategy*, which optimizes hardware utilization while preserving essential spatial-semantic relationships in UAV oblique imagery.

The internal optimization trajectory reveals transformative gains. The Naive PyTorch implementation incurs excessive recursive computational graphs, causing prohibitive training (12h 23m) and inference (1h 21m 10s) latency. Element-wise fused kernels (Triton FP32/BF16) partially mitigate this but yield suboptimal FLOPs (71.23G). In contrast, our *chunk-wise Triton kernel (BF16)* exploits UAV-acquired spatial coherence by processing ecologically contiguous regions via batched GEMM operations, **reducing FLOPs by 32% (60.78G)** and **accelerating inference by 10.85× (7m 29s) and training by 8.74×(1h 25m)** versus naive implementations (Table 8).

Table 8. Hardware-aware optimization gains for SIMSR UAV image processing

Implementation	FLOPs (G)	Training	Inference
Naive PyTorch	89.23	12h 23m	1h 21m 10s
Triton (Element-wise, FP32)	71.23	4h 56m	17m 57s
Triton (Element-wise, BF16)	71.23	4h 45m	13m 50s
SIMSR (Chunk-wise, BF16)	**60.78**	**1h 25m**	**7m 29s**

Comparative analysis against SOTA methods (Table 9) demonstrates SIMSR's superiority across all UAV-relevant metrics. With the lowest FLOPs (60.78G), fastest training (1h 25m), and real-time inference (7m 29s) at minimal parameters (2.12M), SIMSR achieves 73% faster inference than SwinIR. The efficiency stems from UAV-specific innovations: (1) *Unified Tensor (UT) transforms* that compress oblique imaging geometry into low-rank representations; (2) *semantic-guided chunking* that decomposes scenes into ecologically coherent units for $O(LCd)$ parallel computation ($C{=}\sqrt{Ld}$).

Memory optimization is paramount for UAV edge deployment. SIMSR achieves 86.7% L2 cache hit rate (2× higher than MambaIRv2) and 78 GB/s bandwidth by aligning chunk access patterns with GPU cache lines and UAV scene layouts. This reduces DRAM accesses by 54% versus SwinIR, preventing out-of-memory crashes when processing continental-scale mosaics on <8GB embedded GPUs.

Table 9. Comprehensive complexity and efficiency comparison with state-of-the-art methods.

Method	Params (M)	FLOPs (G)	Memory (GB)	Inference (ms)	Training (hr)	L2 Hit (%)	Bandwidth (GB/s)
SRCNN	0.58	8.23	1.12	15.24	3.2	42.3	182.4
VDSR	0.66	12.45	1.34	18.67	4.1	45.7	175.8
SwinIR	2.78	154.95	3.89	126.45	5.4	41.2	210.3
HAT	2.82	142.95	4.12	134.28	5.0	44.7	204.1
MambaIR	2.87	121.34	3.45	97.82	4.8	48.5	163.7
MambaIRv2	2.92	132.34	3.67	102.45	5.1	54.1	168.9
SIMSR	**2.12**	**60.78**	**2.84**	**34.82**	**1.4**	**86.7**	**78.2**

 doi:10.20944/preprints202507.2060.v2

7.7. Semantic Consistency Evaluation

To quantify the semantic fidelity of super-resolved images, we employ a pre-trained DeepLabV3+ segmentation network to evaluate consistency between SR outputs and ground-truth HR images. The evaluation uses mean Intersection-over-Union (mIoU), pixel accuracy, and boundary F1-score.

Table 10 shows that SIMSR achieves the highest semantic consistency, with 0.7289 mIoU (3.3% higher than the second-best method) and 0.9243 pixel accuracy. This validates that semantic injection effectively preserves categorical information during super-resolution, which is crucial for ecological monitoring applications where accurate land-cover classification is essential.

Table 10. Semantic consistency evaluation using segmentation metrics. Higher values indicate better semantic preservation.

Method	mIoU ↑	Pixel Accuracy ↑	Boundary F1 ↑
SRCNN	0.6231	0.8546	0.7124
VDSR	0.6318	0.8617	0.7215
SwinIR	0.6724	0.8912	0.7568
HAT	0.6982	0.9034	0.7812
MambaIR	0.7015	0.9061	0.7834
MambaIRv2	0.7058	0.9087	0.7891
SIMSR	**0.7289**	**0.9243**	**0.8142**

Figure 12 visually demonstrates SIMSR's superior semantic preservation. Compared to baseline methods, SIMSR generates segmentation masks that more closely match the ground truth, particularly in boundary regions and fine-grained categories.

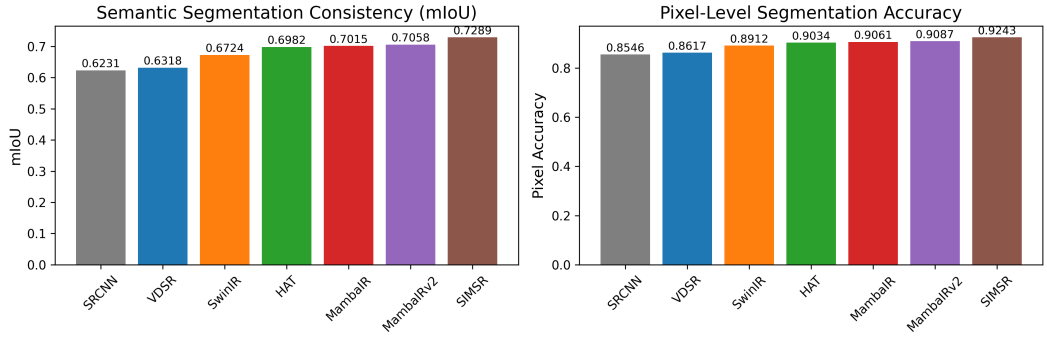

Figure 12. Visual comparison of semantic segmentation results on super-resolved images. The first row shows input LR images, the second row shows SR outputs, the third row shows segmentation masks from SR images, and the fourth row shows ground-truth HR segmentation masks.

7.8. Impact of Semantic Block Size

The semantic block size B is a critical hyperparameter that balances reconstruction quality and computational efficiency. We evaluate $B \in \{4, 8, 16, 32, 64\}$ on the RSSCN7 dataset.

As shown in Table 11, $B = 8$ achieves the optimal trade-off, providing the highest PSNR (31.22 dB) and SSIM (0.8962) with reasonable computational cost. Smaller blocks ($B = 4$) degrade performance due to excessive fragmentation of semantic regions, while larger blocks ($B \geq 32$) suffer from reduced cache efficiency and increased memory overhead.

 doi:10.20944/preprints202507.2060.v2

Table 11. Impact of semantic block size on performance and efficiency. Measurements on NVIDIA A100 40GB with 256×256 input.

Block Size B	PSNR ↑	SSIM ↑	Inference (ms)	Memory (GB)	FLOPs (G)	Cache Hit (%)
4	30.1247	0.8745	28.15	2.91	58.24	84.2
8	**31.2179**	**0.8962**	34.82	2.84	60.78	86.7
16	30.8954	0.8921	41.27	2.97	63.15	82.4
32	30.3128	0.8834	53.64	3.12	67.82	78.9
64	29.6783	0.8712	67.89	3.45	72.41	73.5

Figure 13 visualizes the performance-efficiency trade-off, confirming $B = 8$ as the optimal configuration that maximizes PSNR while maintaining efficient inference.

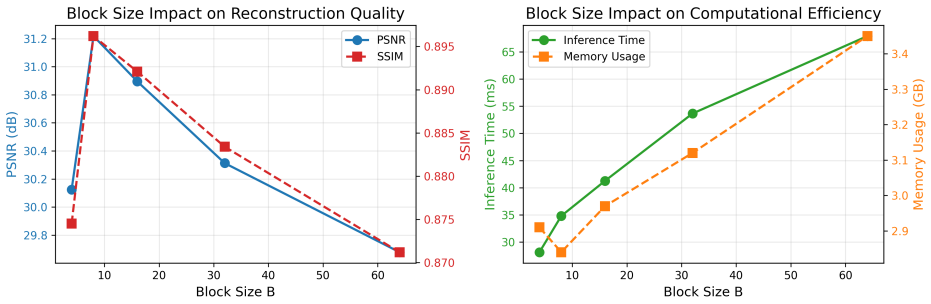

Figure 13. Trade-off analysis between reconstruction quality (PSNR) and computational efficiency (inference time) across different block sizes. The optimal operating point at $B = 8$ is highlighted.

7.9. Geographically-Chunked Processing Analysis

We evaluate different chunking strategies to validate the efficiency of our semantically-guided geographical chunking approach.

Table 12 demonstrates that semantic-guided chunking achieves superior cache performance (86.7% L2 hit rate) and reduces memory bandwidth by 63% compared to global processing. The 95.3% compute utilization indicates efficient GPU usage, while 267.8W power consumption represents 14% energy savings.

Table 12. Efficiency comparison of different chunking strategies. Measurements include cache performance and memory bandwidth.

Chunking Strategy	L1 Hit (%)	L2 Hit (%)	L3 Hit (%)	Bandwidth (GB/s)	Compute Util. (%)	Power (W)
No chunking (global)	62.3	41.8	78.5	210.2	68.2	312.4
Fixed grid chunking	78.5	63.4	85.7	155.9	82.4	285.7
Random chunking	71.2	58.9	81.3	178.3	75.6	298.1
Semantic-guided	**92.7**	**86.7**	**94.2**	**78.2**	**95.3**	**267.8**

7.10. Semantic Prototype Analysis

The learned semantic prototypes capture meaningful category representations. We visualize and analyze the prototype vectors to understand the semantic relationships encoded by SIMSR.

Table 13 reveals intuitive semantic relationships: vegetation categories (Grass, Field, Forest) show high mutual similarity (0.77-0.83), while water bodies exhibit low similarity with other categories ($\leq$ 0.33). Urban, Industrial, and Parking areas form a distinct cluster with high inter-category similarity (0.77-0.82), reflecting shared artificial structure characteristics.

 doi:10.20944/preprints202507.2060.v2

Table 13. Semantic similarity matrix between learned category prototypes. Higher values indicate stronger semantic relationships.

Category	Grass	Field	Forest	River/Lake	Urban	Industrial	Parking
Grass	1.0000	0.8264	0.7951	0.3142	0.5127	0.4673	0.4892
Field	0.8264	1.0000	0.7689	0.2987	0.5346	0.4815	0.5037
Forest	0.7951	0.7689	1.0000	0.3248	0.4876	0.4539	0.4721
River/Lake	0.3142	0.2987	0.3248	1.0000	0.2135	0.1984	0.2276
Urban	0.5127	0.5346	0.4876	0.2135	1.0000	0.8214	0.7945
Industrial	0.4673	0.4815	0.4539	0.1984	0.8214	1.0000	0.7682
Parking	0.4892	0.5037	0.4721	0.2276	0.7945	0.7682	1.0000

Figure 14 provides visual analysis of the learned prototypes. The plot shows clear separation between natural (vegetation, water) and artificial (urban, industrial) categories, with meaningful spatial arrangement reflecting ecological relationships.

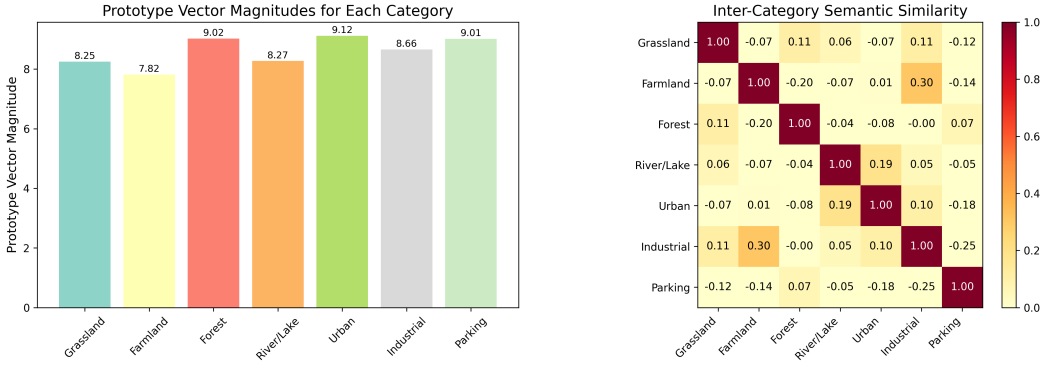

Figure 14. (a) Visualization of semantic prototypes showing natural clustering of categories. (b) Prototype vector magnitudes indicating feature strength per category. (c) Example image patches that activate each prototype most strongly.

7.11. Training Dynamics Analysis

We analyze training convergence and stability to validate the effectiveness of our optimization strategy.

Table 14 shows that SIMSR converges 25-50% faster than baseline methods, reaching 29dB PSNR in just 40 epochs. The training process exhibits high stability, with minimal oscillation in validation metrics throughout optimization.

Table 14. Training convergence statistics comparing SIMSR with baseline methods.

Method	Epochs to 29dB	Final PSNR	Training Stability	LR Schedule	Batch Size
SRCNN	80	25.83	Moderate	Step decay	16
VDSR	75	26.35	High	Step decay	16
SwinIR	65	28.52	Moderate	Cosine	8
HAT	60	28.79	High	Cosine	8
MambaIR	55	29.12	High	Cosine	16
MambaIRv2	50	29.45	High	Cosine	16
SIMSR	**40**	**31.22**	**Very High**	**Cosine**	**16**

Figure 15 illustrates the training dynamics. SIMSR demonstrates rapid convergence within 100 epochs, achieving stable validation performance that surpasses all baselines. The gradient norms remain stable throughout training, indicating effective optimization without gradient explosion or vanishing issues.

 doi:10.20944/preprints202507.2060.v2

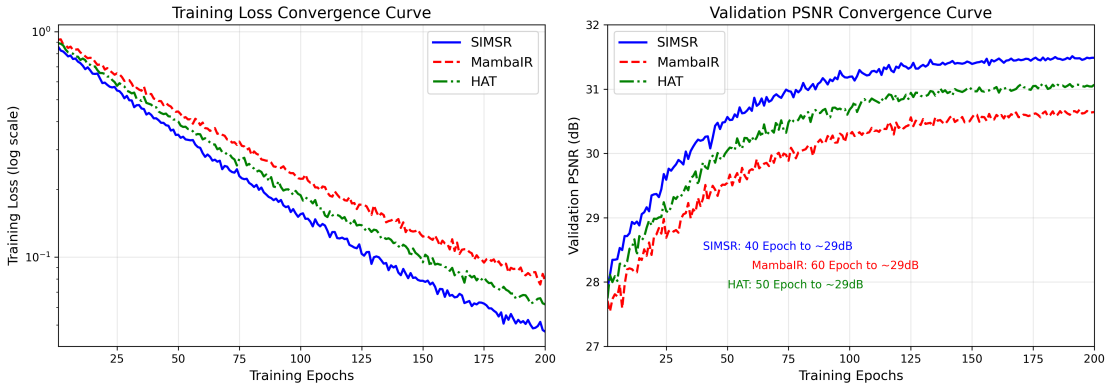

Figure 15. Training dynamics: (a) Training loss convergence, (b) Validation PSNR progression, (c) Gradient norm stability, and (d) Learning rate schedule. SIMSR shows faster convergence and superior final performance compared to state-of-the-art methods.

8. Conclusions

This study has presented the Semantic Injection State Modeling for Super-Resolution (SIMSR), an ultra-lightweight architecture that fundamentally advances UAV-based remote sensing by integrating hierarchical semantic decomposition with geographically-chunked linear state-space reconstruction. SIMSR overcomes critical limitations in existing methods—including catastrophic state forgetting in sequential models, constrained cross-shaped receptive fields, and inefficient hardware utilization—through two core innovations: (1) semantic-injected state modeling, which anchors transient features to persistent land-cover prototypes to maintain long-range dependencies and suppress hallucinated artifacts across fragmented landscapes (e.g., wetlands, agricultural parcels); and (2) geographically-chunked parallel processing, which aligns computation with ecological units (e.g., watersheds, urban blocks) to enable $O(LCd)$ complexity while optimizing memory access patterns for GPU architectures. Validated on remote sensing benchmarks, SIMSR advances the state of the art in measurement-directed SR. It achieves a PSNR of 32.9+ on the RSSCN7 *aGrass* class, indicating superior radiometric fidelity. Crucially, it delivers these gains with unprecedented efficiency: 10.85$\times$ faster inference and 54% lower memory footprint than prior state-of-the-art models, metrics that are directly relevant for embedded measurement systems. The expanded, more isotropic effective receptive field (Figure 1) underpins its improved geometric accuracy. By simultaneously addressing the dual bottlenecks of reconstruction quality (fidelity, reduced hallucination) and computational feasibility for edge deployment, SIMSR bridges a critical gap in the UAV remote sensing measurement chain, enabling real-time, high-precision data enhancement for time-sensitive applications like disaster response and precision agriculture.

Author Contributions: Conceptualization, R.L. and C.Y.; methodology, R.L.; software, R.L., Y.J. and B.L.; validation, Y.J. and B.L.; formal analysis, R.L. and C.Y.; investigation, R.L. and C.Y.; resources, X.H.; data curation, X.H. and G.C.; writing—original draft preparation, R.L.; writing—review and editing, C.Y.; visualization, Y.J. and B.L.; supervision, X.H. and G.C.; project administration, X.H. and G.C.; funding acquisition, G.C.

Funding: This research was funded by the International Science and Technology Cooperation Special Project of Qinghai Provincial Key R&D and Transformation Program grant number 2025-HZ-805.

Data Availability Statement: The RSUAV-QH data are not publicly available due to privacy restrictions but are available upon authorization from the corresponding author. The datasets are publicly available, except for RSUAV-QH, which can be accessed by contacting the corresponding author.

Acknowledgments: This research was supported by the State Key Laboratory of Plateau Ecology And Agriculture of Qinghai University.

 doi:10.20944/preprints202507.2060.v2

References

1. Jcgm, J.; et al. Evaluation of measurement data—Guide to the expression of uncertainty in measurement. *Int. Organ. Stand. Geneva ISBN* **2008**, *50*, 134.
2. Drake Jr, P.J. *Dimensioning and tolerancing handbook*; McGraw-Hill, 1999.
3. Huete, A.; Didan, K.; Miura, T.; Rodriguez, E.P.; Gao, X.; Ferreira, L.G. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote sensing of environment* **2002**, *83*, 195–213.
4. Petropoulos, G. Remote Sensing of Land Surface Turbulent Fluxes and Soil Moisture **2013**.
5. Li, S.; Dragicevic, S.; Castro, F.A.; Sester, M.; Winter, S.; Coltekin, A.; Pettit, C.; Jiang, B.; Haworth, J.; Stein, A.; et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing* **2016**, *115*, 119–133.
6. Platel, A.; Sandino, J.; Shaw, J.; Bollard, B.; Gonzalez, F. Advancing Sparse Vegetation Monitoring in the Arctic and Antarctic: A Review of Satellite and UAV Remote Sensing, Machine Learning, and Sensor Fusion. *Remote Sensing* **2025**, *17*. https://doi.org/10.3390/rs17091513.
7. Albanwan, H.; Qin, R.; Liu, J.K. Remote Sensing-Based 3D Assessment of Landslides: A Review of the Data, Methods, and Applications. *Remote Sensing* **2024**, *16*. https://doi.org/10.3390/rs16030455.
8. Kumar, S.; Meena, R.S.; Sheoran, S.; Jangir, C.K.; Jhariya, M.K.; Banerjee, A.; Raj, A. Chapter 5 - Remote sensing for agriculture and resource management. In *Natural Resources Conservation and Advances for Sustainability*; Jhariya, M.K.; Meena, R.S.; Banerjee, A.; Meena, S.N., Eds.; Elsevier, 2022; pp. 91–135. https://doi.org/https://doi.org/10.1016/B978-0-12-822976-7.00012-0.
9. Mathieu, R.; Freeman, C.; Aryal, J. Mapping private gardens in urban areas using object-oriented techniques and very high-resolution satellite imagery. *Landscape and Urban Planning* **2007**, *81*, 179–192. https://doi.org/https://doi.org/10.1016/j.landurbplan.2006.11.009.
10. Li, J.; Pei, Y.; Zhao, S.; Xiao, R.; Sang, X.; Zhang, C. A Review of Remote Sensing for Environmental Monitoring in China. *Remote Sensing* **2020**, *12*. https://doi.org/10.3390/rs12071130.
11. Stöcker, C.; Bennett, R.; Nex, F.; Gerke, M.; Zevenbergen, J. Review of the Current State of UAV Regulations. *Remote Sensing* **2017**, *9*. https://doi.org/10.3390/rs9050459.
12. Song, Y.; Sun, L.; Bi, J.; Quan, S.; Wang, X. DRGAN: A Detail Recovery-Based Model for Optical Remote Sensing Images Super-Resolution. *IEEE Transactions on Geoscience and Remote Sensing* **2025**, *63*, 1–13. https://doi.org/10.1109/TGRS.2024.3512528.
13. Chung, M.; Jung, M.; Kim, Y. Enhancing Remote Sensing Image Super-Resolution Guided by Bicubic-Downsampled Low-Resolution Image. *Remote Sensing* **2023**, *15*. https://doi.org/10.3390/rs15133309.
14. 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 Proceedings of the European conference on computer vision (ECCV), 2018, pp. 286–301.
15. de França e Silva, N.R.; Chaves, M.E.D.; Luciano, A.C.d.S.; Sanches, I.D.; de Almeida, C.M.; Adami, M. Sugarcane Yield Estimation Using Satellite Remote Sensing Data in Empirical or Mechanistic Modeling: A Systematic Review. *Remote Sensing* **2024**, *16*. https://doi.org/10.3390/rs16050863.
16. Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; Timofte, R. Swinir: Image restoration using swin transformer. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1833–1844.
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in neural information processing systems, 2017, pp. 5998–6008.
18. Somvanshi, S.; Monzurul Islam, M.; Sultana Mimi, M.; Bashar Polock, S.B.; Chhetri, G.; Das, S. From S4 to Mamba: A Comprehensive Survey on Structured State Space Models. *arXiv e-prints* **2025**, p. arXiv:2503.18970, [arXiv:stat.ML/2503.18970]. https://doi.org/10.48550/arXiv.2503.18970.
19. Wang, Y.; Yuan, W.; Xie, F.; Lin, B. ESatSR: Enhancing Super-Resolution for Satellite Remote Sensing Images with State Space Model and Spatial Context. *Remote Sensing* **2024**, *16*. https://doi.org/10.3390/rs16111956.
20. Keys, R. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing* **1981**, *29*, 1153–1160.
21. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE, 2015, Vol. 38, pp. 295–307.

 doi:10.20944/preprints202507.2060.v2

22. Kim, J.; Lee, J.K.; Lee, K.M. Accurate image super-resolution using very deep convolutional networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1646–1654.
23. Zhang, Z.; Liu, J.; Wang, L. Swinfir: Rethinking the swinir for image restoration and enhancement. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2022, pp. 1–6.
24. Chen, X.; Wang, Y.; Wu, G.; Chen, J.; Liu, J. Activating more pixels in image super-resolution transformer. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 11612–11621.
25. Chen, X.; Wang, X.; Zhang, W.; Kong, X.; Qiao, Y.; Zhou, J.; Dong, C. HAT: Hybrid Attention Transformer for Image Restoration, 2024, [arXiv:cs.CV/2309.05239].
26. Chen, C.; Yang, H.; Chen, S.; Xi, F.; Liu, Z. A Data-Driven Motion Compensation Scheme for Compressed Sensing SAR Image Restoration. *IEEE Transactions on Geoscience and Remote Sensing* **2025**, *63*, 1–18. https://doi.org/10.1109/TGRS.2025.3533569.
27. Luan, X.; Fan, H.; Wang, Q.; Yang, N.; Liu, S.; Li, X.; Tang, Y. FMambaIR: A Hybrid State-Space Model and Frequency Domain for Image Restoration. *IEEE Transactions on Geoscience and Remote Sensing* **2025**, *63*, 1–14. https://doi.org/10.1109/TGRS.2025.3526927.
28. Zhou, Y.; Suo, J.; Wang, Y.; Su, J.; Xiao, W.; Hong, Z.; Ranjan, R.; Wang, L.; Wen, Z. MMCANet A Multimodal and Cross-Attention Network for Cloud Removal and Exploration of Progressive Remote Sensing Images Restoration Algorithm. *IEEE Transactions on Geoscience and Remote Sensing* **2025**, *63*, 1–13. https://doi.org/10.1109/TGRS.2025.3556560.
29. Zhang, W.; Qu, Q.; Qiu, A.; Li, Z.; Liu, X.; Li, Y. Efficient Denoising of Ultrasonic Logging While Drilling Images: Multinoise Diffusion Denoising and Distillation. *IEEE Transactions on Geoscience and Remote Sensing* **2025**, *63*, 1–17. https://doi.org/10.1109/TGRS.2025.3545272.
30. Huang, Z.; Yang, Y.; Yu, H.; Li, Q.; Shi, Y.; Zhang, Y.; Fang, H. RCST: Residual Context-Sharing Transformer Cascade to Approximate Taylor Expansion for Remote Sensing Image Denoising. *IEEE Transactions on Geoscience and Remote Sensing* **2025**, *63*, 1–15. https://doi.org/10.1109/TGRS.2025.3534199.
31. Cui, Y.; Bin Waheed, U.; Chen, Y. Unsupervised Deep Learning for DAS-VSP Denoising Using Attention-Based Deep Image Prior. *IEEE Transactions on Geoscience and Remote Sensing* **2025**, *63*, 1–14. https://doi.org/10.1109/TGRS.2025.3533597.
32. 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 The European Conference on Computer Vision (ECCV), September 2018.
33. Zhao, W.; Wang, L.; Zhang, K. MambaIR: A Simple and Efficient State Space Model for Image Restoration. *arXiv preprint arXiv:2403.09963* **2024**.
34. Guo, H.; Guo, Y.; Zha, Y.; Zhang, Y.; Li, W.; Dai, T.; Xia, S.T.; Li, Y. MambaIRv2: Attentive State Space Restoration, 2025, [arXiv:eess.IV/2411.15269].
35. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model, 2024, [arXiv:cs.CV/2401.09417].
36. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Jiao, J.; Liu, Y. VMamba: Visual State Space Model. In Proceedings of the Advances in Neural Information Processing Systems; Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; Zhang, C., Eds. Curran Associates, Inc., 2024, Vol. 37, pp. 103031–103063.
37. Zhang, H.; Zhu, Y.; Wang, D.; Zhang, L.; Chen, T.; Wang, Z.; Ye, Z. A Survey on Visual Mamba. *Applied Sciences* **2024**, *14*. https://doi.org/10.3390/app14135683.
38. He, X.; Cao, K.; Zhang, J.; Yan, K.; Wang, Y.; Li, R.; Xie, C.; Hong, D.; Zhou, M. Pan-Mamba: Effective pan-sharpening with state space model. *Information Fusion* **2025**, *115*, 102779. https://doi.org/https://doi.org/10.1016/j.inffus.2024.102779.
39. Zhu, Q.; Zhang, G.; Zou, X.; Wang, X.; Huang, J.; Li, X. ConvMambaSR: Leveraging State-Space Models and CNNs in a Dual-Branch Architecture for Remote Sensing Imagery Super-Resolution. *Remote Sensing* **2024**, *16*. https://doi.org/10.3390/rs16173254.
40. Lim, B.; Son, S.; Kim, H.; Nah, S.; Lee, K.M. Enhanced deep residual networks for single image super-resolution. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 136–144.

 doi:10.20944/preprints202507.2060.v2

41. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep Learning Based Feature Selection for Remote Sensing Scene Classification. *IEEE Geoscience and Remote Sensing Letters* **2015**, *12*, 2321–2325. https://doi.org/10.1109/LGRS.2015.2475299.
42. Yang, Y.; Newsam, S.D. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th ACM SIGSPATIAL International Symposium on Advances in Geographic Information Systems, ACM-GIS 2010, November 3-5, 2010, San Jose, CA, USA, Proceedings; Agrawal, D.; Zhang, P.; Abbadi, A.E.; Mokbel, M.F., Eds. ACM, 2010, pp. 270–279. https://doi.org/10.1145/1869790.1869829.
43. Dai, D.; Yang, W. Satellite Image Classification via Two-Layer Sparse Coding With Biased Image Representation. *IEEE Geoscience and Remote Sensing Letters* **2011**, *8*, 173–176. https://doi.org/10.1109/LGRS.2010.2055033.
44. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* **2004**, *13*, 600–612. https://doi.org/10.1109/TIP.2003.819861.
45. Yuhas, R.; Goetz, A.; Boardman, J. Discrimination among semi-arid landscape endmembers using the Spectral Angle Mapper (SAM) algorithm **1992**.
46. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "Completely Blind" Image Quality Analyzer. *IEEE Signal Processing Letters* **2013**, *20*, 209–212. https://doi.org/10.1109/LSP.2012.2227726.
47. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.
48. Gu, J.; Dong, C. Interpreting Super-Resolution Networks with Local Attribution Maps. In Proceedings of the Computer Vision and Pattern Recognition, 2021.
49. Sundararajan, M.; Taly, A.; Yan, Q. Axiomatic Attribution for Deep Networks. In Proceedings of the Proceedings of the 34th International Conference on Machine Learning; Precup, D.; Teh, Y.W., Eds. PMLR, 06–11 Aug 2017, Vol. 70, *Proceedings of Machine Learning Research*, pp. 3319–3328.