From Super-Resolution to Superior Land Cover Detection: Cross-Channel Attention Network for Aerial Image

Yuwei Cai[1], Zhimeng He[1], Meiliu Wu[1], Brian Barrett[1]

[1] University of Glasgow, University Avenue, Glasgow, UK G12 8QQ - (Yuwei.Cai, Zhimeng.He, Meiliu.Wu, Brian.Barrett)@glasgow.ac.uk

Keywords: Super-Resolution, Cross-channel attention, Convnextv2, Land cover detection, Involution.

Abstract

Low-resolution imagery is a major constraint for remote sensing tasks (e.g., urban land cover detection) where accurate classification of buildings, roads, vegetation, and small objects is required. Deep learning-based segmentation models are highly sensitive to image quality, resulting in degraded performance on low-resolution inputs. Super-resolution (SR) techniques offer a promising solution by enhancing image fidelity to support downstream tasks. This work applied MAPSRNet, a Multi-Attention Pyramid SR Network to aerial images used for multi-class land cover detection. Evaluated on the ISPRS Potsdam dataset, MAPSRNet achieves state-of-the-art SR performance with PSNR of 32.92 dB and SSIM of 0.87, outperforming existing methods such as SRCNN (31.54 dB, 0.83) and DRRN (31.03 dB, 0.82) while maintaining competitive inference speed. Beyond image quality, MAPSRNet significantly improves multi-class land cover segmentation when integrated with a ConvNeXtV2-based U-Net, achieving an overall accuracy of 80.60%, mean IoU of 62.54%, and FwIoU of 68.34%, surpassing not only low-resolution inputs (Overall Accuracy: 65.28%, mIoU: 40.20%, FwIoU: 50.12%) but also high-resolution(HR) ones (Overall Accuracy: 80.50%, mIoU: 62.40%, FwIoU: 68.01%), especially in certain classes such as impervious surface and clutter. These results demonstrate that perceptual and structural fidelity, rather than pixel-level similarity, can drive superior performance in urban land cover segmentation. MAPSRNet offers a practical solution for scenarios where HR imagery is limited or unavailable, highlighting its potential for large-scale remote sensing applications.

1. Introduction

Climate change and rapid urbanization are driving the need for accurate and timely urban land cover mapping to support the sustainable planning, infrastructure development, and disaster management (Chen et al., 2023a). High-resolution (HR) aerial imagery enables precise delineation of buildings, roads, and vegetation, which is critical for applications such as digital twins and smart city systems (Chen et al., 2017). In particular, accurate land cover identification is fundamental for large-scale urban analytics. However, acquiring those centimeter-level HR aerial imagery remains costly and often infeasible for large-scale or time-sensitive scenarios (Mao et al., 2025).

Super-resolution (SR) offers a practical and efficient solution by reconstructing high-quality images from low-resolution (LR) inputs, improving visual fidelity and enhancing the performances on downstream tasks such as land cover detection (Razzak et al., 2023). While the existing SR networks improve image quality, they often prioritize perceptual metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM), rather than testing impacts on the downstream task-specific performance, leaving uncertainty about whether SR outputs truly benefit semantic segmentation.

To address this gap, we applied MAPSRNet, a Multi-Attention Pyramid Super-Resolution Network designed as a task-oriented SR solution for remote sensing tasks. The MAPSRNet integrates the spatial, channel, and self-attention mechanisms through a Pyramid Vision Transformer (PVTv2) and an Involution+ module, enabling robust feature representation and reconstruction of fine details critical for multi-class land cover segmentation. Using the ISPRS Potsdam dataset, we evaluated MAPSRNet not only on image quality but also on multi-class land cover detection performance with a ConvNeXtV2 backbone and U-Net decoder. Our findings show that MAPSRNet consistently outperforms other existing SR methods, achieving sharper and more accurate boundary detection for small objects such as vehicles and clutters and delivering segmentation accuracy that rivals or sometimes even surpasses HR imagery in certain classes. These results highlight MAPSRNet’s task-oriented potential to alleviate HR data scarcity and improve urban mapping in large-scale remote sensing applications.

2. Related Work

Land cover mapping traditionally relies on pixel-level classification of medium-resolution satellite imagery, which offers global coverage but suffers from the mixed-pixel problem in heterogeneous urban areas (He et al., 2022a). This limitation motivated the development of super-resolution mapping (SRM) techniques to reconstruct fine-scale spatial patterns from coarse inputs. Early SRM approaches were model-driven, using spatial attraction models, geostatistical priors, or regularization-based methods (He et al., 2022a). While these methods improved spatial details, they depended heavily on handcrafted assumptions and often produced unrealistic blocky patterns, failing to capture irregular urban land cover structures. Research on solving these problems then arised to overcome these limitations and pave the way for more flexible, data-driven approaches.

The introduction of deep learning significantly advanced SRM. Convolutional neural networks (CNNs) enabled data-driven learning of LR-to-HR mappings, improving visual fidelity and reducing reliance on prior knowledge (Khan et al., 2025). However, most SR networks were optimized for perceptual metrics such as PSNR and SSIM rather than task-specific

objectives like segmentation. This design often resulted in visually sharp outputs that lacked accurate class boundaries, especially for small or complex objects. Furthermore, fixed-kernel upsampling and homogeneous reconstruction suppressed local heterogeneity, and generalization across regions remained a challenge.

To overcome these limitations, hybrid strategies have emerged. Yin et al. proposed a Collaborative Spatial-Spectral Fusion (CSSF) framework that jointly performs image super-resolution and spectral unmixing through multi-task learning (Yin et al., 2025) . By incorporating spatial and channel attention, CSSF enhances feature interaction between spectral and spatial domains, reducing distortion and improving edge preservation. Similarly, Khan et al. (Khan et al., 2025) introduced a fuzzy deep learning architecture combined with SR and chaotic particle swarm optimization (C-PSO) to handle uncertainty and optimize feature selection, achieving competitive accuracy on complex aerial datasets. Despite these advances, most existing designs employ single-level or local attention, limiting their ability to capture global context and multi-scale dependencies essential for reconstructing fine details.

Parallel research on semantic segmentation of very-high-resolution (VHR) imagery demonstrates the effectiveness of attention-based encoder-decoder architectures (Wambugu et al., 2021). However, these methods assume VHR inputs and do not focus on addressing the challenge of reconstructing task-relevant details from LR imagery. Interactive segmentation frameworks (Lenczner et al., 2020) offer refinement through human input but remain impractical for large-scale automated mapping.

In summary, the evolution of SR for land cover mapping highlights a clear trend: attention mechanisms significantly improve reconstruction quality by guiding networks to focus on informative regions. However, most current SR models rely on limited attention types, restricting their capacity to model long-range dependencies and multi-scale features. This motivates the development of SR architectures that integrate multi-attention strategies. By combining spatial, channel, and self-attention, feature representation can be enhanced. Building on these insights, our work applies MAPSRNet to reconstruct fine details and then improve land cover detection performance under constrained HR data availability.

3. Dataset

The ISPRS Potsdam dataset is a high-resolution aerial image benchmark released in 2016 as part of the ISPRS 2D Semantic Labeling Contest (Sherrah, 2016). It covers approximately 3 km $\times$ 3 km of urban area in Potsdam, Germany, captured by airborne sensors with a spatial resolution of 5 cm. The dataset consists of 38 orthorectified image tiles, each measuring 6000 $\times$ 6000 pixels, and provides multi-spectral data (RGB and near-infrared) together with precise ground truth annotations. In this research, only the RGB bands were used. Each pixel is labeled into one of six semantic categories: impervious surfaces, buildings, low vegetation, trees, cars, and clutter/background.

The original RGB images were cliped into 512 $\times$ 512 patches, yielding 16,082 high-resolution (HR) and low-resolution (LR) image pairs for training, and 4,020 image pairs for testing. For the super-resolution (SR) experiments, the LR images were generated by downsampling the HR images to a 20 cm spatial resolution, corresponding to a scaling factor of 4. High-resolution images were first blurred with Gaussian kernels to simulate realistic optical degradations and then contaminated with Gaussian noise to emulate real-world LR conditions.

4. Methods

In this study, we applied MAPSRNet as the SR framework and benchmark its performance against four representative SR models: SRCNN (Super-Resolution Convolutional Neural Network) (Dong et al., 2016), DRRN (Deep Recursive Residual Network) (Tai et al., 2017), SRResNet (Super-Resolution Residual Network) (Ledig et al., 2017), DRRN (Deep Recursive Residual Network), and MSCA-RFANet (Multi-Scale Channel Attention Residual Feature Aggregation Network) (He et al., 2022b). To evaluate the practical benefits of SR for remote sensing, we proceed multi-class land cover detection on the ISPRS Potsdam dataset using a ConvNeXtV2-based U-Net segmentation network. The segmentation results obtained from SR outputs of all SR models are compared with those from original HR images and LR inputs, enabling a comprehensive assessment of how SR quality impacts downstream remote sensing tasks.

4.1 The Super-Resolution Network: MAPSRNet

In this study, we applied Multi-Attention Pyramid Super-Resolution Network (MAPSRNet), a SR network designed to enhance reconstruction quality by integrating cross-channel attention mechanism. Built upon the Residual Feature Aggregation Network (RFANet), MAPSRNet has two key important parts: a multi-stage Pyramid Vision Transformer (PVTv2) (Chen et al., 2023b) for global context modeling and an Involution module for efficient cross-channel attention. These components work together to capture both long-range dependencies and fine-grained spatial details, which are essential for remote sensing imagery.

The architecture comprises three main parts: the head, trunk, and reconstruction parts (Fig. 1). The head part performs shallow feature extraction, while the trunk part refines features through residual aggregation and attention mechanisms, incorporating channel, spatial, and self-attention. A key structure in MAPSRNet is the integration of a multi-stage PVTv2 structure concatenated into the trunk part. This design allows the network to model global semantics more effectively and enhances contextual understanding across multiple scales. The reconstruction part then combines an Involution+ module with a 3 $\times$ 3 convolution layer in a parallel structure, enabling adaptive filtering and cross-channel interaction before upsampling. This design ensures that global context from PVTv2 complements local detail preservation from Involution+, resulting in high-quality SR outputs that benefit downstream land cover segmentation.

Involution is a dynamic operator that addresses the limitations of traditional convolution by generating position-specific kernels for each spatial location (Li et al., 2021) (Fig. 2). Unlike convolution, which applies shared kernels globally, involution adapts its filtering to local neighborhoods, improving spatial detail capture while reducing computational cost. The core operation is defined as:

$$Y_{i,j,k} = \sum_{(u,v)\in\Delta k} \mathcal{H}_{i,j,u+[K/2],v+[K/2],[kG/C]} X_{i+u,j+v,k}, \quad (1)$$

where $\mathcal{H}$ denotes the position-specific kernel, K is the kernel size, G the number of groups, and Δk the local receptive field.

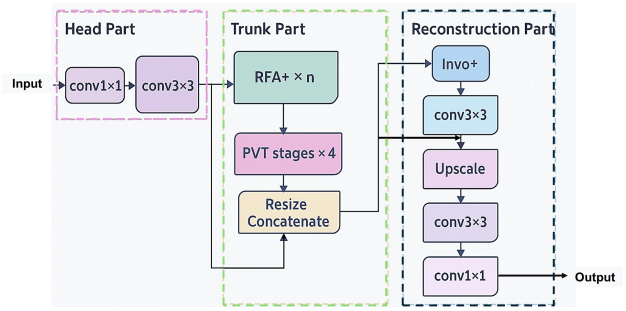

Figure 1. Architecture of the proposed MAPSRNet. “PVT_stage” represents different stages from PVTv2; “RFA” refers to the Residual Feature Aggregation Network; “$conv$” denotes convolution layers; “$Invo$” indicates the Involution module. The parameter n is set to 30.

4.2 The Multi-class land cover Detection Network: ConvNeXtV2 Backbone with U-Net Decoder

For multi-class land cover detection on SR images, we employed a hybrid architecture consisting of a ConvNeXtV2 backbone for feature extraction and a U-Net decoder for pixel-wise prediction. The ConvNeXtV2 backbone is a convolutional network inspired by transformer design principles, which processes input images through multiple stages of depthwise and pointwise convolutions, layer normalization, and GELU activations. Across these stages, the spatial resolution of feature maps is progressively reduced while the channel dimensionality increases, producing multi-scale hierarchical features that capture both fine-grained and high-level information (Pang et al., 2024).

The U-Net decoder reconstructs the output by progressively upsampling the feature maps. At each upsampling stage, skip connections from the corresponding ConvNeXtV2 stage are concatenated to the decoder features, preserving spatial details critical for accurate segmentation. Convolutional layers within each decoder block refine these features, and a final 1×1 convolution produces the segmentation mask. This design leverages the representational power of ConvNeXtV2 while maintaining the strong localization capability of U-Net, allowing effective land cover detection from SR images (Fig. 3).

To optimize segmentation performance, we employed a composite loss strategy combining Dice Loss and Focal Loss. Dice Loss is particularly effective for handling class imbalance by maximizing the overlap between predicted and ground truth masks. It is computed as:

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2 \times |P \cap G| + \epsilon}{|P| + |G| + \epsilon}, \tag{2}$$

where P and G denote predicted and ground truth masks, and ϵ is a smoothing term.

Focal Loss was incorporated to address hard-to-classify pixels by down-weighting well-classified examples and focusing on challenging regions. It is defined as:

$$\mathcal{L}_{\text{Focal}} = -(1 - p_t)^{\gamma} \log(p_t), \tag{3}$$

where p_t is the predicted probability for the true class and γ controls the focusing strength.

To further mitigate class imbalance, we computed class weights based on pixel frequency across the training dataset. Following a median-frequency balancing approach, weights were derived as:

$$w_c = \frac{\text{median}(f)}{f_c}, \tag{4}$$

where f_c is the frequency of class c and median(f) is the median of non-zero frequencies. These weights were applied in the cross-entropy component of Focal Loss to ensure rare classes contribute proportionally to the optimization process.

Furthermore, urban land cover datasets often exhibit severe class imbalance, where large homogeneous regions (e.g., impervious surfaces) dominate while small objects (e.g., vehicles) occupy only a tiny fraction of pixels. Dice Loss (Eq. 2) directly optimizes spatial overlap, improving boundary delineation for fragmented classes such as vegetation and clutter. Focal Loss (Eq. 3) emphasizes hard examples, reducing misclassification of small or visually ambiguous objects. Finally, class weighting (Eq. 4) ensures that rare classes like vehicles and clutter have sufficient influence during training. Together, these strategies enhance segmentation robustness, leading to higher IoU and F1 scores across all land cover categories.

4.3 SR Evaluation Metrics

We evaluate SR performance using three metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Frames Per Second (FPS) (Hore and Ziou, 2010). PSNR and SSIM measure image reconstruction quality, while FPS reflects testing speed. Higher PSNR and SSIM indicate better fidelity, and higher FPS means faster processing. The metrics are computed as:

$$\text{PSNR} = 20 \log_{10} \frac{L}{\sqrt{\frac{\sum_{i=1}^{N} (\widehat{g}_i - g_i)^2}{N}}} \tag{5}$$

$$\text{SSIM} = \frac{(2\mu_{\widehat{g}}\mu_g + C_1)(2\sigma_{\widehat{g}g} + C_2)}{(\mu_{\widehat{g}}^2 + \mu_g^2 + C_1)(\sigma_{\widehat{g}}^2 + \sigma_g^2 + C_2)} \tag{6}$$

$$\text{FPS} = \frac{\text{Number of Frames}}{\text{Total Time (seconds)}} \tag{7}$$

Here, $\widehat{g}_i$ and g_i are pixel values in the SR and HR images, N is the total number of pixels, and L is the maximum pixel value (e.g., 255 for 8-bit images). μ and σ denote mean and standard deviation, and C_1, C_2 are constants for stability.

4.4 Implementation Details

All experiments in this study, including super-resolution (SR) reconstruction and multi-class land-cover detection, were conducted under an identical hardware environment to ensure a fair and consistent evaluation of model performance. Specifically, both training and validation processes were executed on a High-Performance Computing (HPC) platform equipped with an NVIDIA RTX 6000 Ada Generation GPU and CUDA version 13.0. This unified setup eliminates variability introduced by differing computational resources and ensures that performance differences are attributable solely to model design.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-515-2026 | © Author(s) 2026. CC BY 4.0 License.

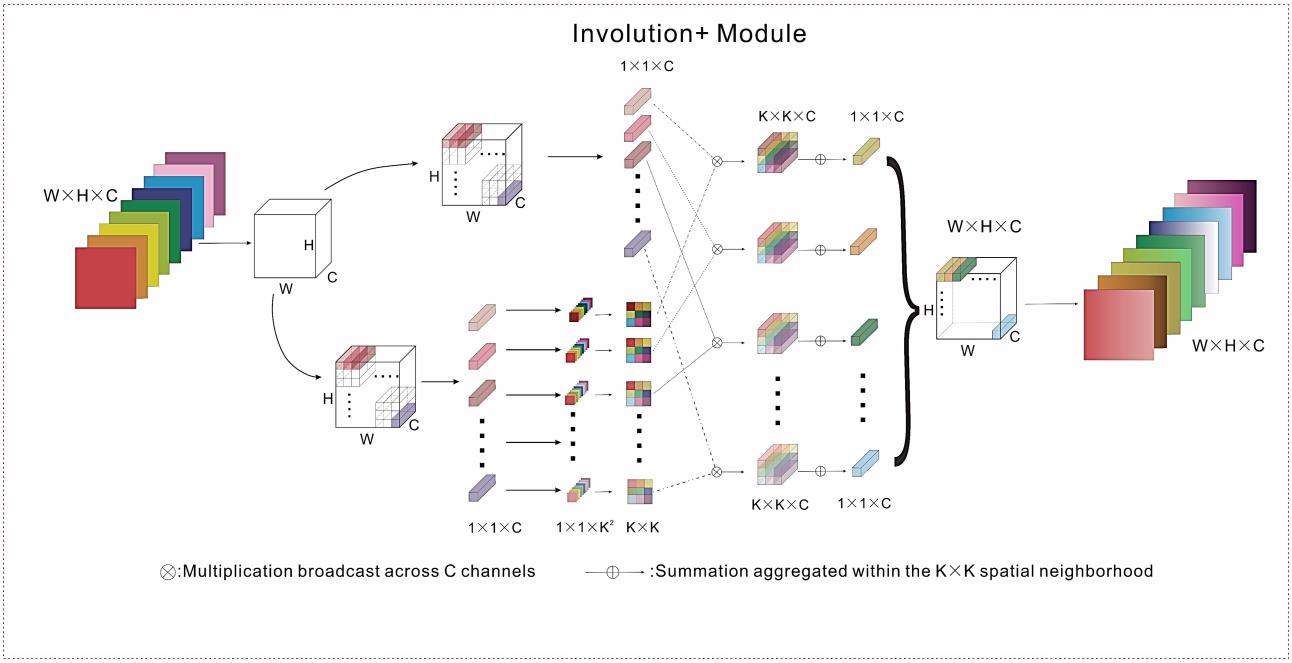

Figure 2. Overview of the Involution module architecture. The figure illustrates the kernel generation and application process, where position-specific kernels are dynamically created and applied to enhance spatial adaptivity and channel interaction.

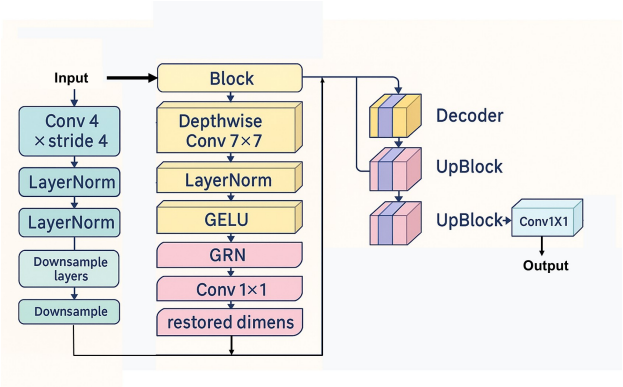

Figure 3. Architecture of Multi-class land cover Detection Network: ConvNeXtV2 backbone with a U-Net decoder.

For the SR experiments, the initial learning rate was set to 1×10^{-5} and progressively reduced by a factor of 0.5 every 10 epochs to facilitate stable convergence and prevent oscillations during later training stages. The Mean Squared Error loss function was adopted for all SR models due to its strong capability in minimizing pixel-wise reconstruction errors and preserving fine-grained spatial details, which is critical for downstream tasks such as segmentation (Muhammad and Laaksonen, 2025). Each SR network was trained for 100 epochs with a batch size of 2, balancing computational efficiency with memory constraints imposed by HR imagery.

For the multi-class land-cover detection task, the ConvNeXtV2-UNet model was trained for 300 epochs with a batch size of 2. Class weighting was incorporated to mitigate the impact of severe class imbalance across land-cover categories. This training strategy enhances the model’s ability to accurately delineate diverse classes, including small-scale objects such as vehicles and clutter, and improves overall segmentation robustness and generalization performance.

5. Results

5.1 SR Qualitative Evaluation

Figure. 4 presents qualitative comparisons between HR, LR, and SR outputs generated by different super-resolution networks. All SR methods provide clear visual improvements over LR inputs, particularly in recovering structural details such as building edges and road surfaces. These enhancements are especially noticeable for small and complex objects, including vehicles and cluttered regions, where SR outputs exhibit reduced blurring and more coherent boundaries.

Among the compared approaches, MAPSRNet produces results that are visually closest to the HR references, with sharper roof delineation and more distinguishable object shapes. This suggests that the proposed multi-attention design is more effective in preserving fine-grained spatial details and structural consistency. Nevertheless, the visual differences between methods remain relatively subtle, particularly in homogeneous regions. Therefore, quantitative evaluation is necessary to rigorously assess whether these perceptual improvements translate into measurable gains in reconstruction quality and downstream task performance.

5.2 SR Quantitative Evaluation

For quantitative evaluation,Bicubic Interpolation (BI) achieved the fastest inference (133.33 FPS) but produced the lowest quality (PSNR: 23.22 dB, SSIM: 0.66). SRCNN and DRRN significantly improved PSNR (31.54 dB and 31.27 dB) and SSIM (0.83 and 0.82), though at slower speeds (29.35 FPS and 22.76 FPS). SRResNet delivered moderate results (PSNR: 26.19 dB, SSIM: 0.78), while MSCA-RFANet, leveraging multi-scale channel attention, achieved higher SSIM (0.85), indicating the benefit of attention mechanisms. The proposed MAPSRNet outperformed all baselines with the highest PSNR (32.92 dB) and SSIM (0.87) while maintaining competitive speed (28.60 FPS), demonstrating its effectiveness in balancing

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-515-2026 | © Author(s) 2026. CC BY 4.0 License.

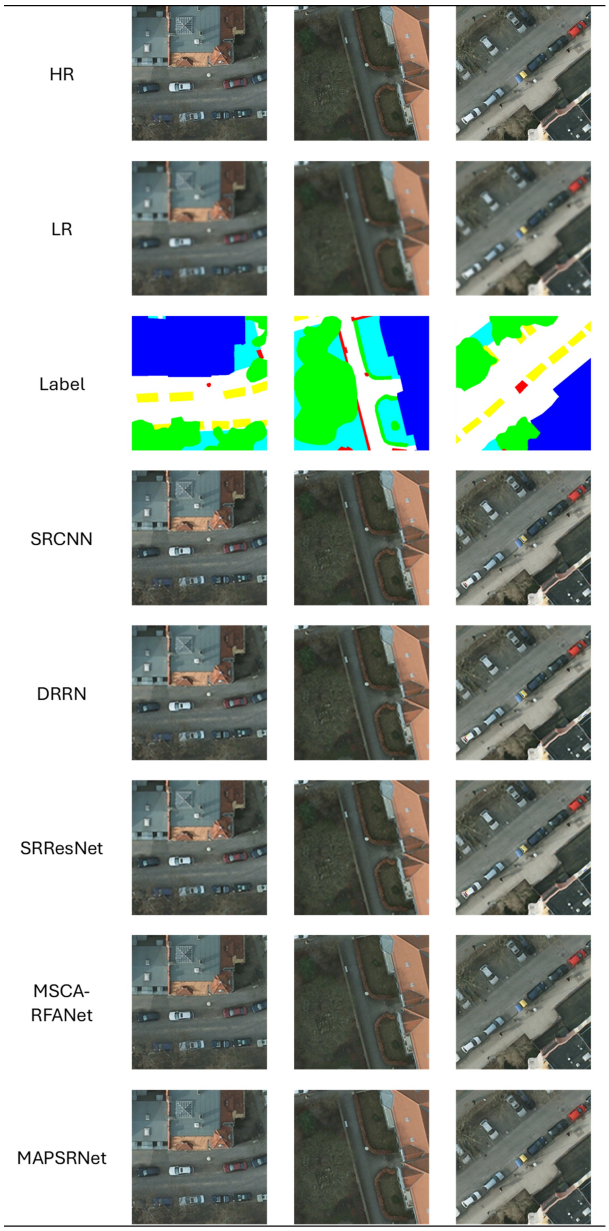

Figure 4. Output Examples of SR Results on the Potsdam Dataset (Spatial Resolution of HR inputs: 5 cm/pixel; Spatial Resolution of LR inputs: 20 cm/pixel).

reconstruction quality and computational efficiency for aerial image super-resolution. Although the SR networks all show good performances on PSNR and SSIM, their impacts on remote sensing down-stream tasks still remain unknown, Therefore, HR, LR and all the SR outputs were then applied to multi-class land cover detection (Table. 1).

Table 1. Performance comparison of super-resolution methods in terms of PSNR and SSIM.
Best values are highlighted in bold.

Method	PSNR(dB)	SSIM	Testing Speed (FPS)
BI	23.22	0.66	133.33
SRCNN	31.54	0.83	29.35
DRRN	31.27	0.82	22.76
SRResNet	26.19	0.78	22.09
MSCA-RFA	30.76	0.85	22.35
MAPSRNet	**32.92**	**0.87**	28.60

5.3 Multi-class Land Cover Detection Results

According to Figure. 5, LR images exhibit significant misclassification, particularly along building edges and small objects such as cars, where boundaries are blurred and object coverage is incomplete. While among all SR images, MAPSRNet still delivers the most visually accurate segmentation, closely matching ground truth with sharper delineation of buildings and improved detection of small objects like vehicles and clutters. However, qualitative differences between SR images remain subtle, emphasizing the need for quantitative evaluation.

Quantitatively, Table 2 demonstrates that MAPSRNet consistently outperforms all other SR methods across every class. For impervious surfaces, MAPSRNet achieves an IoU of 75.08%, not only 13.74% higher than LR but also slightly surpassing HR images (74.93%). Building segmentation benefits from MAPSRNet’s ability to preserve structural details, reaching an IoU of 80.37%, 18.8% higher than LR. Similarly, low vegetation and tree classes exhibit substantial gains, with IoUs of 54.29% and 54.98%, respectively. This proves the enhancing green plants detection. Small-object categories such as vehicles and clutter show the largest relative improvements, achieving IoUs of 76.22% and 34.31%, compared to 36.98% and 24.50% for LR, demonstrating MAPSRNet’s ability to recover fine details lost during downsampling.

Beyond class-specific improvements, MAPSRNet also achieves the best overall segmentation metrics among all SR methods. It records an overall accuracy of 80.60%, mean IoU of 62.54%, fwIoU of 68.34%, and an Overall Kappa of 71.51. These values not only surpass LR baselines (Overall Accuracy: 65.28%, mIoU: 40.20%, fwIoU: 50.12%, and an Overall Kappa of 51.71) but also exceed HR performance (Overall Accuracy: 80.50%, mIoU: 62.40%, fwIoU: 68.01%, and an Overall Kappa of 71.16), especially in certain classes, such as impervious surface, buildings and clutter. This confirms that MAPSRNet’s multi-attention design enhances task-oriented feature reconstruction, delivering superior performance for urban mapping.

These results reveal two important insights. First, for impervious surfaces, clutters, and also some evaluation metrics on buildings, MAPSRNet achieves segmentation performance that slightly exceeds raw HR images, indicating its capability to reconstruct task-oriented features beyond simple resolution recovery. Second, although PSNR and SSIM values for SRCNN, DRRN, SRResNet, and MSCA-RFANet are only slightly lower than those of MAPSRNet, the downstream task of multi-class land cover detection reveals a clear gap. MAPSRNet delivers substantially better segmentation accuracy across all classes. This confirms that perceptual and structural fidelity, rather than pixel-level similarity, drives superior performance in urban mapping applications.

In summary, with MAPSRNet applied to the aerial images in the ISPRS Potsdam dataset, not only overall segmentation performance was enhanced but also the class-specific detection performance was enhanced, particularly for complex and small-scale land cover. These all further validate its role as a task-oriented SR solution for multi-class land cover detection on urban mapping.

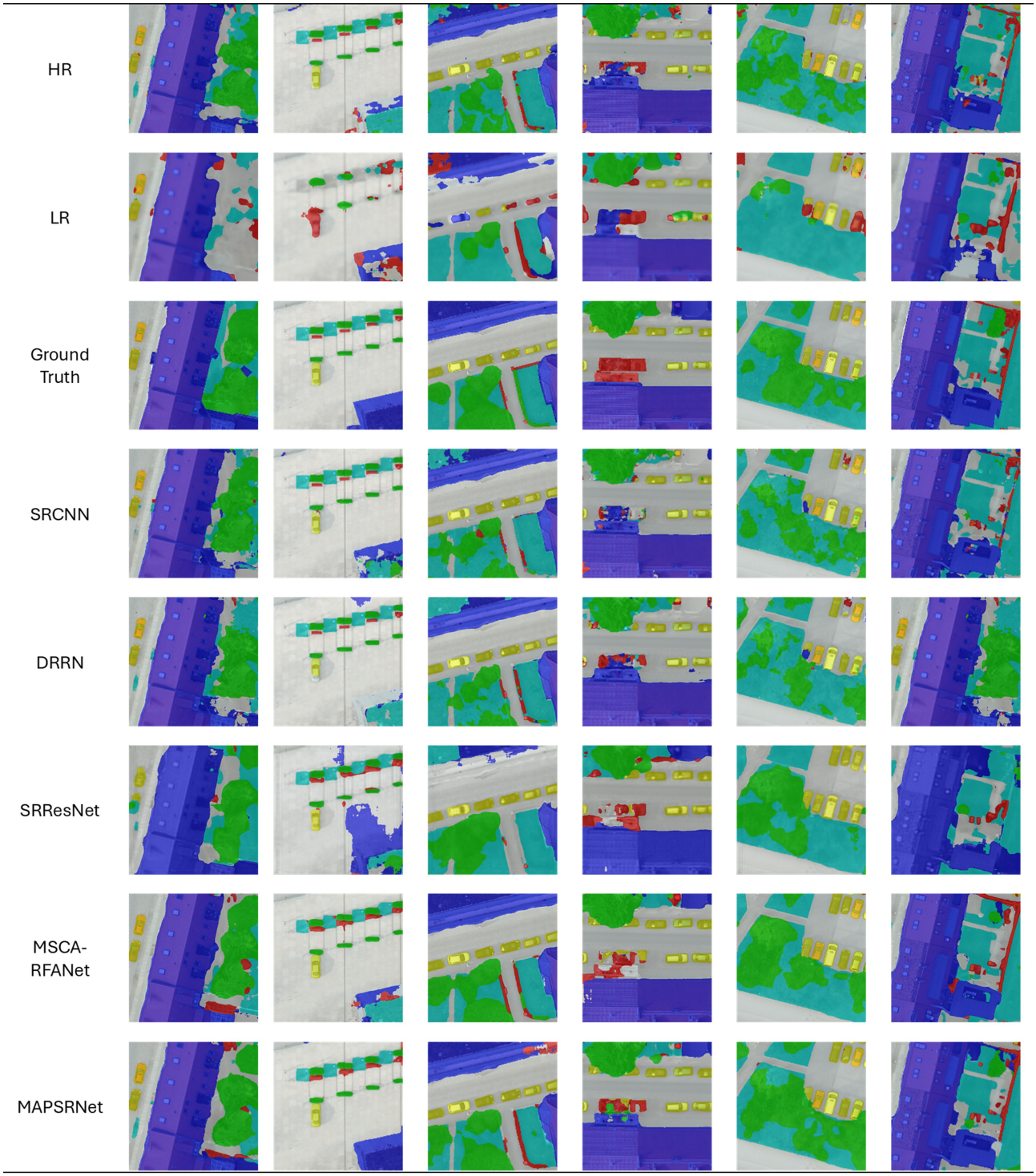

Figure 5. Output examples of land cover detection results using ConvNeXtV2 on the ISPRS Potsdam dataset. The example images show detection results overlaid on the SR outputs with 50% transparency.white indicates impervious surfaces, blue represents buildings, cyan denotes low vegetation, green corresponds to trees, yellow marks cars, and red shows Clutter.

https://doi.org/10.5194/isprs-annals-XI-3-2026-515-2026

Table 2. Quantitative Multi-class Land Cover Detection Results of MAPSRNet Compared with Baselines across the Potsdam Dataset.

Method	land cover Category	TP	TN	FP	FN	Acc.	IoU	Prec.	Recall	F1	Overall Acc.	mIoU	fwIoU	Overall Kappa
HR	Impervious surface	336,008,114	578,877,201	78,529,676	33,912,793	89.05	74.93	81.06	**90.83**	85.67	80.50	62.40	68.01	71.16
	Building	251,446,455	712,769,736	19,902,234	43,209,359	**93.86**	79.94	**92.67**	85.34	88.85				
	Low Vegetation	114,063,477	817,675,671	59,115,808	36,472,828	**90.70**	**54.41**	**65.86**	75.77	**70.47**				
	Tree	84,204,259	878,462,002	27,188,975	37,472,548	**93.71**	**56.56**	**75.56**	**69.20**	**72.26**				
	Vehicles	17,769,391	1,005,289,865	3,162,977	1,105,551	**99.58**	**80.63**	**84.89**	94.14	**89.28**				
	Clutter	23,475,776	943,204,133	12,460,642	48,187,233	94.10	27.91	65.33	32.76	43.64				
LR	Impervious surface	314,607,340	514,471,990	142,934,887	55,313,567	80.70	61.34	68.67	85.05	76.04	65.28	40.20	50.12	51.71
	Building	188,620,461	720,954,397	28,921,508	106,035,353	88.54	61.57	94.15	64.01	76.21				
	Low Vegetation	93,305,415	811,528,868	65,262,611	57,230,890	88.08	43.24	58.84	61.98	60.37				
	Tree	18,276,823	896,241,161	9,409,816	103,399,984	89.02	13.94	66.01	15.02	24.47				
	Vehicles	7,431,020	1,007,231,216	1,221,626	11,443,922	98.77	36.98	85.88	39.37	53.99				
	Clutter	48,441,985	829,566,548	126,098,227	23,221,024	58.47	24.50	27.75	27.60	39.35				
SRCNN	Impervious surface	327,057,684	552,259,526	105,147,351	42,863,223	85.59	68.84	75.67	88.41	81.55	75.95	55.74	61.72	65.07
	Building	234,364,083	713,373,256	19,298,714	60,291,731	92.25	74.65	92.39	79.54	85.48				
	Low Vegetation	120,271,902	792,787,550	84,003,929	30,264,403	88.88	51.28	58.88	**79.90**	67.79				
	Tree	68,415,979	876,028,056	29,622,921	53,260,828	91.93	45.22	69.78	56.23	62.28				
	Vehicles	16,125,165	1,006,287,089	2,165,753	2,749,777	99.52	76.64	**88.16**	85.43	86.77				
	Clutter	14,000,247	948,810,719	6,854,056	6,854,056	93.72	17.83	67.13	19.54	30.27				
DRRN	Impervious surface	324,207,458	561,677,236	95,729,641	45,713,449	86.23	69.62	77.20	87.64	82.09	76.10	56.17	62.09	65.39
	Building	234,447,679	713,515,794	19,156,176	60,208,135	92.27	74.71	92.45	79.57	85.52				
	Low Vegetation	123,326,441	776,545,358	93,064,537	27,209,864	88.29	50.63	56.99	81.92	67.22				
	Tree	67,620,247	877,077,662	28,573,315	54,056,560	91.96	45.01	70.30	55.57	62.07				
	Vehicles	16,034,227	1,006,368,865	2,083,977	2,840,715	99.52	76.50	88.50	84.95	86.69				
	Clutter	16,153,963	948,734,652	6,930,123	55,509,046	93.92	20.55	69.98	22.54	34.10				
SRResNet	Impervious surface	298,518,050	577,606,026	79,800,851	71,402,857	85.28	66.38	78.91	80.70	79.79	72.84	52.84	58.80	61.61
	Building	233,807,382	703,431,023	29,240,947	60,848,432	91.23	72.19	88.88	79.35	83.85				
	Low Vegetation	105,176,132	772,879,065	103,912,414	45,360,173	85.47	41.33	50.30	69.87	58.49				
	Tree	71,629,093	857,850,653	47,800,324	50,047,714	90.48	42.26	59.98	58.87	59.42				
	Vehicles	16,347,217	1,003,254,932	5,197,910	2,527,725	99.25	67.91	75.87	86.61	80.89				
	Clutter	22,830,041	942,597,352	13,067,423	48,832,968	93.97	26.94	63.60	31.86	42.45				
MSCA	Impervious surface	332,786,952	567,098,089	90,308,788	37,133,955	87.59	72.31	78.66	89.96	83.93	78.75	58.58	65.51	68.82
	Building	246,458,471	713,878,508	18,793,462	48,197,343	93.48	78.63	92.91	83.64	88.04				
	Low Vegetation	116,413,414	813,847,251	62,944,228	34,122,891	90.55	54.53	64.91	77.33	70.58				
	Tree	83,128,170	875,697,297	29,953,680	38,548,637	93.33	54.82	73.51	68.32	70.82				
	Vehicles	17,838,649	1,003,981,639	4,471,203	1,036,293	99.46	76.41	79.96	94.51	86.63				
	Clutter	12,353,200	943,787,208	11,877,567	59,309,809	93.07	14.79	50.98	17.24	25.76				
MAPSRNet	Impervious surface	324,010,078	595,752,609	61,654,268	45,910,829	**89.53**	**75.08**	**84.01**	87.59	**85.76**	80.60	62.54	68.34	71.51
	Building	256,170,757	708,574,727	24,097,243	38,485,057	93.01	**80.37**	91.40	**86.94**	**89.11**				
	Low Vegetation	116,433,182	812,862,501	63,928,978	34,103,123	90.46	54.29	64.56	**77.35**	70.37				
	Tree	84,206,249	874,173,623	31,477,354	37,470,558	93.29	54.98	72.90	**69.20**	70.95				
	Vehicles	17,927,506	1,003,806,046	4,646,796	947,436	99.46	76.22	79.42	**94.98**	86.50				
	Clutter	29,232,265	942,121,667	13,543,108	42,430,744	**94.55**	**34.31**	**68.34**	**40.79**	**51.09**				

*Note: The underlined values refer to results worse than using LR images. The bolded values refer to the best results. All values except of overall Kappa in the table are percentages.

6. Discussion

Recent advances in computer vision have led to a wide range of SR methods that achieve strong performance on natural image benchmarks. In this study, although MAPSRNet achieved quite good performance compared with other typical SR networks, several limitations should be acknowledged.

First, the current evaluation is conducted primarily on the Potsdam dataset, which, although widely used, represents a specific urban scenario with high spatial resolution and limited variability in sensor characteristics. As such, the generalizability of MAPSRNet to other datasets—particularly those with different spatial resolutions, noise patterns, or acquisition conditions—remains to be systematically validated. Second, the model operates in a task-aware but not fully joint optimization framework, meaning that SR and segmentation are not optimized end-to-end. This may limit the model’s adaptability to other tasks or domains where task-specific features differ significantly.

Third, the computational efficiency of MAPSRNet has not been comprehensively analyzed. Important indicators, including parameter count, multiply–accumulate operations (MACs), floating-point operations (FLOPs), peak GPU memory consumption, and inference latency, are not yet reported. This makes it difficult to assess the feasibility of deploying the model in large-scale or real-time remote sensing applications, where computational constraints are often critical. In addition, while the current study focuses on RGB aerial imagery, the effectiveness of MAPSRNet on satellite data and multi-spectral imagery remains unclear. Variations in spectral characteristics, radiometric properties, and spatial resolution may affect model performance and require architectural adaptation.

More broadly, the role of SR in remote sensing pipelines is still not fully understood. While this study focuses on land-cover segmentation, the impact of SR—both MAPSRNet and existing methods—on other tasks, such as object detection, change detection, and instance segmentation, remains insufficiently explored. It is therefore unclear whether improvements observed in one task generalize to others, or whether task-specific SR strategies are required.

Future work will focus on addressing these limitations by conducting cross-dataset evaluations and incorporating domain adaptation techniques to enhance generalizability. Joint optimization frameworks that integrate SR with downstream tasks will also be explored to better align reconstruction with task objectives. Furthermore, a detailed analysis of computational efficiency, along with potential model compression and acceleration strategies, will be essential for practical deployment. Finally, extending the framework to satellite, multi-spectral, and multi-temporal data, as well as systematically evaluating SR methods across diverse remote sensing tasks, will provide a more comprehensive understanding of task-oriented super-resolution in real-world scenarios.

7. Conclusions

In this study, we applied MAPSRNet to aerial imagery working on remote sensing tasks. Experiments on the ISPRS Potsdam dataset show that MAPSRNet achieves state-of-the-art reconstruction quality, with PSNR of 32.92 dB and SSIM of 0.87, while maintaining competitive inference speed. Qualitative results indicate that MAPSRNet produces outputs visually closest to high-resolution imagery, preserving structural details and improving clarity for small objects such as vehicles and clutters. More importantly, MAPSRNet significantly enhances downstream land cover segmentation, achieving an overall accuracy of 80.60%, mean IoU of 62.54%, and FwIoU of 68.34%, outperforming low-resolution inputs and even surpassing high-resolution performance in certain categories such as impervious surface and clutter. These improvements confirm that perceptual and structural fidelity, rather than pixel-level similarity, drives superior performance in real-world urban mapping scenarios. Overall, MAPSRNet not only improves image reconstruction but also delivers measurable gains in segmentation accuracy, especially for complex and small-scale features. This makes it a practical solution in scenarios where high-resolution imagery is limited or unavailable. These advantages further highlight its strong potential for large-scale remote sensing applications.

References

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(4), 834–848.

Chen, P., Huang, H., Liu, J., Wang, J., Liu, C., Zhang, N., Su, M., Zhang, D., 2023a. Leveraging Chinese GaoFen-7 imagery for high-resolution building height estimation in multiple cities. *Remote Sens. Environ.*, 298, 113802.

Chen, X., Wang, X., Zhou, J., Qiao, Y., Dong, C., 2023b. Activating more pixels in image super-resolution transformer. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 22367–22377.

Dong, C., Loy, C. C., Tang, X., 2016. Accelerating the super-resolution convolutional neural network. *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Springer, 391–407.

He, D., Shi, Q., Liu, X., Zhong, Y., Xia, G., Zhang, L., 2022a. Generating annual high resolution land cover products for 28 metropolises in China based on a deep super-resolution mapping network using Landsat imagery. *GIScience & Remote Sensing*, 59(1), 2036–2067.

He, H., Gao, K., Tan, W., Wang, L., Chen, N., Ma, L., Li, J., 2022b. Super-resolving and composing building dataset using a momentum spatial-channel attention residual feature aggregation network. *Int. J. Appl. Earth Obs. Geoinf.*, 111, 102826.

Hore, A., Ziou, D., 2010. Image quality metrics: Psnr vs. ssim. *Proc. Int. Conf. Pattern Recognit.*, IEEE, 2366–2369.

Khan, J. A., Khan, M. A., Al-Khalidi, M., AlHammadi, D. A., Alasiry, A., Marzougui, M., Zhang, Y., Khan, F., 2025. Design of Super Resolution and Fuzzy Deep Learning Architecture for the Classification of Land Cover and Landsliding Using Aerial Remote Sensing Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, 337–351.

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z. et al., 2017. Photo-realistic single image super-resolution using a generative adversarial network. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 4681–4690.

Lenczner, G., Le Saux, B., Luminari, N., Chan-Hon-Tong, A., Le Besnerais, G., 2020. Disir: Deep image segmentation with interactive refinement. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2020, 877–884.

Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., Zhang, T., Chen, Q., 2021. Involution: Inverting the inherence of convolution for visual recognition. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 12321–12330.

Mao, Z., Abdi, O., Uusitalo, J., Laamanen, V., Kivinen, V.-P., 2025. Super-resolution supporting individual tree detection and canopy stratification using half-meter aerial data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 224, 251–271.

Muhammad, U., Laaksonen, J., 2025. Hybrid Deep Learning for Hyperspectral Single Image Super-Resolution. *IEEE Geoscience and Remote Sensing Letters*.

Pang, R., Tan, H., Yang, Y., Xu, X., Liu, N., Zhang, P., 2024. A novel segnet model for crack image semantic segmentation in bridge inspection. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 344–355.

Razzak, M. T., Mateo-Garcia, G., Lecuyer, G., Gómez-Chova, L., Gal, Y., Kalaitzis, F., 2023. Multi-spectral multi-image super-resolution of Sentinel-2 with radiometric consistency losses and its effect on building delineation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195, 1–13.

Sherrah, J., 2016. Fully convolutional networks for dense semantic labelling of high-resolution aerial imagery. *arXiv preprint arXiv:1606.02585*.

Tai, Y., Yang, J., Liu, X., 2017. Image super-resolution via deep recursive residual network. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 3147–3155.

Wambugu, N., Chen, Y., Xiao, Z., Wei, M., Bello, S. A., Marcato Junior, J., Li, J., 2021. A hybrid deep convolutional neural network for accurate land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 103, 102515.

Yin, Z., Li, X., Wu, P., Lu, J., Ling, F., 2025. CSSF: Collaborative spatial-spectral fusion for generating fine-resolution land cover maps from coarse-resolution multi-spectral remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 226, 33–53.