MOSAIC: MULTI-SCALE OBJECT-CENTRIC SLOT ADAPTATION WITH IMAGE-GUIDED PROTOTYPE CONDITIONING FOR REMOTE SENSING IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained diffusion Transformers provide powerful generative priors for image restoration, yet adapting them to structure-sensitive tasks remains challenging. Existing methods mainly rely on dense low-resolution features as conditions, which often entangle salient objects with dominant backgrounds and provide limited abstraction for recurring structures across scales and samples. We propose MOSAIC, an object-centric representation learning framework that learns compact structural conditions for generative model adaptation. MOSAIC organizes degraded observations into structural slots that capture complementary object-level and layout-level evidence, integrates them across resolutions, and retrieves recurring priors from a shared prototype memory to complement incomplete image-specific information. The resulting representations are injected into selected layers of a frozen diffusion Transformer through lightweight residual modulation, preserving its pretrained generative prior. We instantiate MOSAIC for remote sensing image super-resolution, where small objects, repetitive structures, and large-scale variation provide a challenging testbed. Experiments on UCMerced, AID, and RSSCN7 show consistent PSNR improvements across $\times 2$ and $\times 4$ settings and favorable LPIPS results at $\times 4$. Qualitative results further demonstrate improved preservation of small objects, structural boundaries, and repetitive spatial patterns.

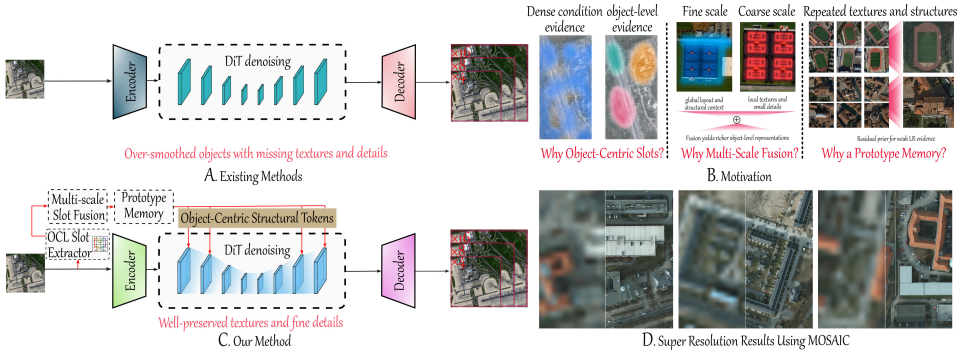

Figure 1: MOSAIC at a glance. (a) Conventional DiT-based SR relies on dense LR conditions. (b) Remote sensing scenes motivate object-centric, cross-scale, and shared structural modeling. (c) MOSAIC introduces compact multi-scale structural slots and shared prototype priors into a frozen DiT backbone. (d) Representative super-resolution results produced by MOSAIC.

1 INTRODUCTION

High-resolution remote sensing imagery is essential for fine-grained Earth observation and supports applications such as urban monitoring, environmental analysis, and disaster assessment. Recent surveys and real-world SR studies highlight both the rapid progress of deep-learning-based super-resolution and the remaining challenges under realistic degradations (Moser et al., 2023; Sun

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

& Chen, 2024). However, sensor limitations, complex imaging conditions, and transmission constraints often obscure small objects, structural boundaries, and repetitive textures; recent generative SR methods likewise emphasize the difficulty of maintaining fidelity under real degradations (Han et al., 2026; Wu et al., 2026). Remote sensing image super-resolution (RSISR) aims to recover high-resolution observations while preserving both large-scale layouts and fine local structures. Compared with natural images, remote sensing scenes exhibit substantial scale variation, dense spatial patterns, and small foreground objects embedded in large background regions. Therefore, effective RSISR requires not only plausible texture generation but also faithful reconstruction of object geometry and regular spatial arrangements.

Existing RSISR methods have evolved from convolutional and Transformer architectures to diffusion-based generative models, with recent DiT-based approaches benefiting from strong generative priors and long-range modeling (Chen et al., 2021b; Xiao et al., 2024; Huo et al., 2026). However, they typically inject LR observations as dense spatial conditions (Kim et al., 2026; Zheng et al., 2026). As illustrated in Figure 1(a), such dense conditions entangle foreground objects with dominant backgrounds, weakening small targets and structural boundaries during denoising. This reveals that effective RSISR depends not only on the generative capacity of the backbone, but also on how conditional evidence is organized.

Figure 1(b) highlights three characteristics of remote sensing scenes that motivate our design. (i) **Object-centric structural organization.** Foreground objects and structural regions occupy limited spatial areas but carry critical geometric information, which is easily diluted in dense representations dominated by backgrounds. (ii) **Cross-scale complementarity.** Coarser representations capture stable layouts and long-range structures, whereas finer ones preserve boundaries, textures, and small objects, making multi-scale cues inherently complementary. (iii) **Incomplete image-specific evidence.** Severe degradation may erase image-specific structural cues, while similar object layouts and textures frequently recur across remote sensing scenes, providing reusable structural knowledge.

Motivated by these observations, we propose a **M**ulti-Scale **O**bject-Centric **S**lot **A**daptation framework with **I**mage-Guided Prototype **C**onditioning, termed **MOSAIC**, for RSISR. As illustrated in Figure 1(c), MOSAIC equips a frozen DiT4SR backbone with object-centric conditioning. It organizes dense LR features into compact structural slots, exploits cross-scale complementarity, and retrieves shared prototypes to supplement degraded image-specific evidence. The resulting conditions guide foreground-aware reconstruction while preserving the pretrained generative prior. As shown in Figure 1(d), MOSAIC recovers more coherent object structures, sharper boundaries, and finer textures.

Our main contributions are summarized as follows:

- We introduce an object-centric representation learning paradigm for diffusion Transformer adaptation, transforming dense LR features into compact structural slots while preserving the pretrained generative prior.
- We develop a multi-scale structural representation learning mechanism that fuses complementary slot representations across resolutions to capture both global layouts and fine-grained local structures.
- We incorporate image-guided retrieval from a shared prototype memory as residual context for fused slots, providing a complementary conditioning path alongside image-specific features.

2 Related Work

2.1 Remote Sensing Image Super-Resolution

Remote sensing image super-resolution reconstructs high-resolution images from low-resolution observations while preserving scene layouts and structures such as roads, buildings, fields, and small objects. Large-scale variation, repeated textures, and complex backgrounds make this challenging. CNN methods recover local details: SRCNN, EDSR, and RCAN use end-to-end convolution, deep residual learning, and channel attention (Dong et al., 2014; Lim et al., 2017; Zhang et al., 2018b), while RASAF combines residual aggregation and split attention to fuse multi-level remote-sensing

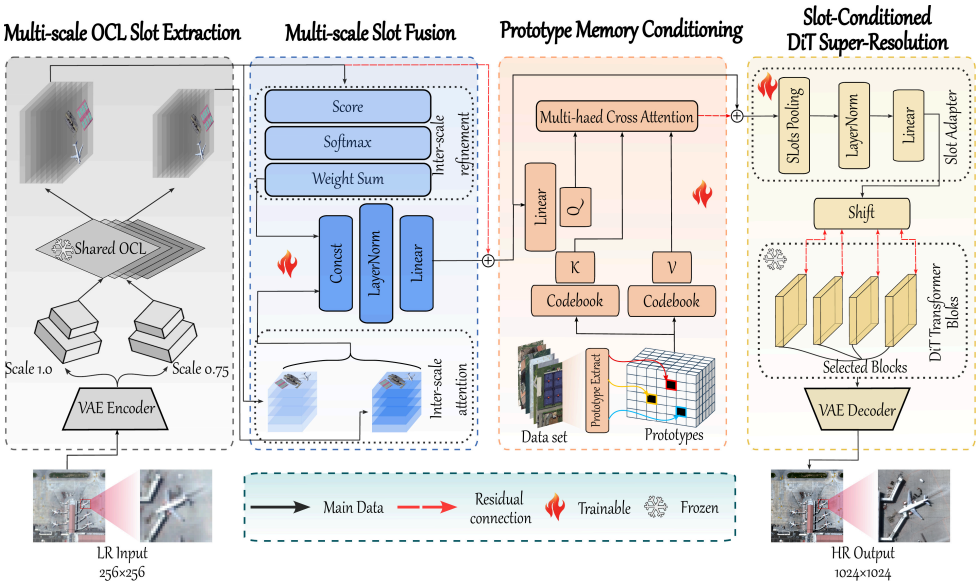

Figure 2: Overview of MOSAIC. The LR image is encoded into a control latent. OCL extracts object-centric slots. MSF fuses multi-scale slots. An offline prototype memory, constructed from training slots and used during adaptation and inference (shown as a codebook in the diagram), provides residual context. SlotFiLM injects the resulting condition into selected frozen DiT4SR blocks. The VAE decoder outputs the SR image.

features (Chen et al., 2021b). These methods are effective for texture recovery but limited in long-range structural modeling; early CNN designs therefore struggle to maintain globally consistent layouts over large scenes.

Transformer approaches extend the receptive field: IPT transfers knowledge across restoration tasks, SwinIR and HAT use window and cross-window attention (Chen et al., 2021a; Liang et al., 2021; Chen et al., 2023), and TTST selects informative tokens to suppress irrelevant regions and model long-range ground-object relations (Xiao et al., 2024). Although attention expands context, dense representations still mix foreground structures with redundant background content. Most approaches therefore lack compact object-level conditions, making structural evidence selection difficult in complex backgrounds.

2.2 Diffusion and DiT-based Restoration

Diffusion models have progressed from multi-step denoising to efficient sampling and stronger architectures. DDPM formulates progressive noising and reverse denoising (Ho et al., 2020), DDIM accelerates sampling while preserving quality (Song et al., 2021), Latent Diffusion Models shift diffusion to a compressed latent space (Rombach et al., 2022), and DiT processes latent tokens with Transformers for long-range spatial modeling (Peebles & Xie, 2023). SinSR and TSD-SR distill multi-step diffusion into one-step super-resolution (Wang et al., 2024; Dong et al., 2025); these advances improve efficiency and fidelity but do not specify how structural conditions should be organized for restoration.

Iterative generative priors and global latent-space modeling support structural recovery and detail synthesis. SR3 demonstrated progressive high-frequency recovery (Saharia et al., 2023), and DiT4SR injects low-resolution control latents into DiT for real-world super-resolution (Duan et al., 2025). However, dense single-scale conditions retain excessive background information and can weaken small objects, regular textures, and local structures; MOSAIC instead supplies compact multi-scale object-level conditions to a frozen DiT backbone.

2.3 Object-Centric Representation and Prototype Priors

Object-centric learning compresses many input tokens into a small set of slots. MONet and IODINE learn scene decomposition and multi-object representations, while Slot Attention iteratively aggregates input features into exchangeable slots (Burgess et al., 2019; Greff et al., 2019; Locatello et al., 2020). Later work improves slot stability and multi-scale modeling through Smoothing Slot Attention and multi-scale object representation (Zhao et al., 2026; 2025).

However, image-specific slots do not explicitly summarize structures shared across samples. Prototype and codebook methods provide shared learned entries: VQ-VAE discretizes continuous features, VQ-VAE-2 expands capacity hierarchically, and VQGAN couples codebooks with adversarial synthesis (van den Oord et al., 2017; Razavi et al., 2019; Esser et al., 2021). Direct quantization or feature replacement can discard image-specific details. MOSAIC therefore retrieves prototype memory as residual context for multi-scale slots before SlotFiLM injects compact conditions into frozen DiT blocks.

3 Method

3.1 Overall Framework

Figure 2 summarizes MOSAIC. The LR observation I_{LR} is resized to the VAE operating resolution and encoded into a control latent z_c, while flow matching starts from noisy latent z_t. The shared OCL Slot Extractor forms compact slots at the original and resized scales; MSF fuses structural layouts with fine details. We first pretrain the OCL extractor and MSF, construct a frozen prototype memory from training slots, and then optimize the prototype cross-attention and SlotFiLM modules. SlotFiLM pools the enhanced slots and predicts layer-specific modulation for selected frozen DiT4SR blocks. At each sampling step, the conditioned DiT predicts the flow-matching target, the scheduler updates z_t, and the VAE Decoder produces I_{SR}.

3.2 Object-Centric Slot Adaptation

Remote sensing features mix multi-scale objects with complex backgrounds. Object-centric slot adaptation converts them into compact object-level conditions for frozen DiT denoising.

Step 1: Multi-scale slot extraction. Following the multi-scale OCL stage in Figure 2, the shared extractor processes z_c at scales 1.0 and 0.75 with shared parameters, producing:

$$S^a \in \mathbb{R}^{B\times K\times 256}, \tag{1}$$

where B, K, and 256 denote batch size, slot number, and slot dimension. The slots organize spatial evidence for local regions, textures, and layouts without masks or category labels.

Step 2: Prototype-enhanced slot conditioning. MSF fuses the two scales into S_{msf}, as described in the following subsection. Prototype retrieval then produces S_{pmc} (Section 3.4). As shown in Figure 3, Slot Pooling and LayerNorm convert these enhanced slots into a global condition z for selected DiT blocks; the subsequent FiLM modulation is channel-wise and shared over DiT tokens:

$$z = \mathrm{LayerNorm}(\mathrm{SlotPool}(S_{\mathrm{pmc}})). \tag{2}$$

Here, $\mathrm{SlotPool}$ denotes mean pooling across the slots. This permutation-invariant operation summarizes the enhanced object-level representations as a fixed-dimensional condition for channel-wise FiLM, with cost linear in the number of slots. Its uniform weights give every slot an equal direct contribution to the pooled condition and an equal direct gradient path, without privileging an arbitrary slot index.

Step 3: Layer-wise residual modulation. For selected DiT block l, a projection produces channel-wise scale and shift terms:

$$[\gamma_l, \beta_l] = W_l z + b_l, \tag{3}$$

and the gate g_l controls residual strength, giving the residual FiLM modulation

$$H_{l+1} = H_l^{\mathrm{base}} + g_l \left(H_l^{\mathrm{base}} \odot \gamma_l + \beta_l\right), \tag{4}$$

where H_l^{base} is the base hidden state, $\odot$ denotes channel-wise multiplication, and the terms are broadcast over tokens. The gate g_l is a learned scalar for block l, initialized to one. Zero-initialized

projections make the adapter initially close to identity; residual modulation preserves the DiT token length and generation path. Pseudocode and further implementation details of the object-centric slot adaptation procedure are provided in the *Technical Supplement*.

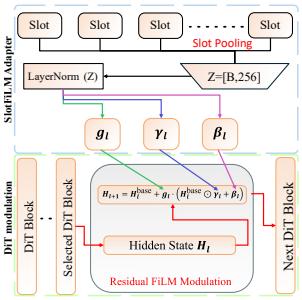

Figure 3: Object-organized global conditioning. The fused slots are pooled and normalized to form a global condition. SlotFiLM then generates layer-specific modulation parameters and injects a residual condition into selected frozen DiT blocks.

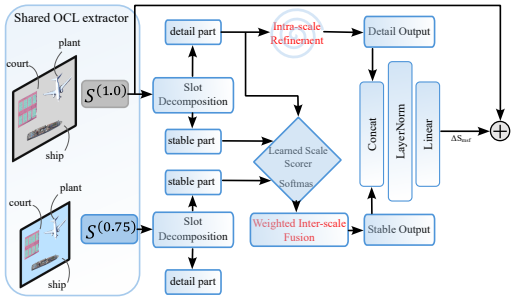

Figure 4: Multi-scale slot fusion. MSF selects stable information across scales and refines base-scale details. It produces a residual update for the original slots.

3.3 Multi-Scale Slot Fusion

Figure 4 summarizes Multi-Scale Slot Fusion (MSF). Coarse slots capture stable layouts and long-range relations, whereas original-scale slots retain textures, boundaries, and small objects. MSF fuses these cues in slot space and predicts a residual update for the original-scale slots.

Step 1: Slot decomposition. The shared OCL extractor outputs slots at both scales:

$$S^{1.0}, S^{0.75} \in \mathbb{R}^{B\times K\times 256}, \tag{5}$$

where shared parameters place both scales in a comparable feature space. MSF splits each 256-D slot into branches used for cross-scale structure selection and base-scale detail refinement:

$$S^a = [S^a_{\text{sta}}, S^a_{\text{det}}], \quad a \in \{1.0, 0.75\}, \tag{6}$$

where

$$S^a_{\text{sta}}, S^a_{\text{det}} \in \mathbb{R}^{B\times K\times 128}, \tag{7}$$

where S^a_{sta} is used for cross-scale structure selection, whereas S^a_{det} is used for current-resolution texture and boundary refinement. This architectural partition is functional rather than a claim that fixed channel subsets have inherent semantics.

Step 2: Adaptive inter-scale fusion for stable structures. For each slot, the Learned Scale Scorer predicts a score per scale from the stable components:

$$r_{k,a} = \mathbf{w}^\top S^a_{\text{sta},k} + b, \quad a \in \{1.0, 0.75\}, \tag{8}$$

and softmax over scales gives:

$$\alpha_{k,a} = \frac{\exp(r_{k,a})}{\sum_{a'\in\{1.0,0.75\}} \exp(r_{k,a'})}, \tag{9}$$

while the stable output is:

$$\hat{S}_{\text{sta},k} = \sum_{a\in\{1.0,0.75\}} \alpha_{k,a} S^a_{\text{sta},k}. \tag{10}$$

Thus, each slot selects reliable structural evidence instead of fixed averaging.

Step 3: Intra-scale refinement for local details. The detail branch refines the original-scale feature $S^{1.0}_{\text{det}}$ with self-attention among K slots:

$$\hat{S}_{\text{det}} = \text{MHA}\left(S^{1.0}_{\text{det}}, S^{1.0}_{\text{det}}, S^{1.0}_{\text{det}}\right), \tag{11}$$

which exchanges local context for boundaries, textures, and small objects; omitting coarse-scale details avoids over-smoothing.

Step 4: Residual fusion. The stable and detail outputs are concatenated, normalized, and projected:

$$\Delta S_{\text{msf}} = W_o \,\text{LayerNorm}\left(\left[\hat{S}_{\text{sta}}, \hat{S}_{\text{det}}\right]\right), \tag{12}$$

and is added to the original-scale slots:

$$S_{\text{msf}} = S^{1.0} + \Delta S_{\text{msf}}. \tag{13}$$

This residual preserves image-specific information while MSF learns complementary corrections. S_{msf} is passed to prototype memory conditioning; the resulting S_{pmc} is then passed to SlotFiLM.

3.4 Image-Guided Prototype Conditioning

The fused slots S_{msf} describe current-image structures, but severely degraded LR inputs can miss evidence for some objects and textures. Remote sensing scenes also contain recurring patterns. As shown in Figure 2, Image-Guided Prototype Conditioning constructs memory from multi-scale fused training slots and retrieves residual context for the current slots through cross-attention.

Step 1: Constructing the prototype memory. We flatten MSF slots from all training images and apply KMeans; the M centroids form the prototype memory:

$$P \in \mathbb{R}^{M\times 256}. \tag{14}$$

The centroids summarize recurring structures and spatial patterns in the training distribution as shared residual priors.

Step 2: Retrieving shared structural context. The memory P remains frozen. For the current image, S_{msf} provides queries and P keys and values:

$$H_p = \text{PMC}(Q = S_{\text{msf}}, K = P, V = P), \tag{15}$$

$$S_{\text{pmc}} = S_{\text{msf}} + \alpha H_p, \tag{16}$$

where PMC retrieves context relevant to the current slots.

Step 3: Residual conditioning. With $\alpha = 0.10$, the normalized memory output is added residually to S_{msf}, complementing image-specific evidence without replacing it. Only the PMC projections and SlotFiLM Adapter are updated; S_{pmc} then conditions frozen DiT blocks through SlotFiLM. The memory is constructed offline from training slots, remains frozen thereafter, and is used at both adaptation and inference.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate MOSAIC on UCMerced (Yang & Newsam, 2010), AID (Xia et al., 2017), and RSSCN7 (Zou et al., 2015) at $\times 2$ and $\times 4$, reporting PSNR, SSIM, LPIPS (Zhang et al., 2018a), MANIQA (Yang et al., 2022), and CLIP-IQA (Wang et al., 2023). The no-reference MANIQA and CLIP-IQA metrics are evaluated only at $\times 4$.

Implementation Details. Data are partitioned before LR generation, so held-out images are excluded from both optimization and the construction of the frozen prototype memory. Model selection and ablations use validation sets; main comparisons use held-out test sets. Detailed data splits, LR generation, and evaluation protocols are provided in the *Technical Supplement*. Implemented in PyTorch on 4 Tesla V100-SXM2-32GB GPUs, MOSAIC uses AdamW, FP16 mixed precision, a batch size of 1/GPU, and learning rates of 5×10^{-7}–5×10^{-6}. Unless noted, we use 8 256-D slots (3 iterations), MSF scales $[1.0, 0.75]$, 16 prototypes, 4-head prototype cross-attention (64-D heads), residual scale 0.1, SlotFiLM blocks $[2, 8, 12, 16]$, and 10 sampling steps.

4.2 Main Results

Quantitative results. Table 1 reports the full-reference results (PSNR and SSIM). MOSAIC achieves the best PSNR on all three datasets under both scales. Under the $\times 4$ setting, it improves PSNR over DiT4SR by 2.03 dB on UCMerced, 4.54 dB on AID, and 1.49 dB on RSSCN7; the corresponding $\times 2$ gains are 2.90 dB, 5.40 dB, and 2.25 dB. These comparisons establish the overall PSNR gains; the ablations below examine the contributions of the individual components.

Table 2 reports the perceptual-quality results (LPIPS, MANIQA, and CLIP-IQA). MOSAIC obtains the lowest LPIPS on both RSSCN7 and AID. The MANIQA and CLIP-IQA results vary across datasets while remaining competitive. Together, these comparisons indicate strong perceptual similarity alongside competitive no-reference image quality.

Method	×2		×4	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
UCMerced				
Real-ESRGAN	**26.8314**	**0.8023**	**23.3802**	0.6215
KDSR-GANV2	26.2297	0.7842	23.1964	0.6113
LDL	25.7230	0.8004	22.9224	**0.6339**
MM-RealSR	24.5995	0.7895	22.1259	**0.6255**
DiT4SR	26.4208	0.7922	22.1419	0.5579
MOSAIC	**29.3194**	**0.8473**	**24.1745**	0.6069
AID				
Real-ESRGAN	**27.6707**	**0.8104**	**23.8303**	0.6153
KDSR-GANV2	27.0080	0.7902	23.6919	0.6097
LDL	26.1695	0.8063	23.2775	**0.6273**
MM-RealSR	25.6580	0.8031	22.6745	0.6200
DiT4SR	27.0105	0.8016	22.6609	0.5534
MOSAIC	**32.4056**	**0.8910**	**27.1982**	**0.6908**
RSSCN7				
Real-ESRGAN	**27.7757**	0.7801	24.8375	0.6018
KDSR-GANV2	27.6795	0.7703	**24.8750**	0.5997
LDL	26.1831	0.7779	24.2620	**0.6089**
MM-RealSR	26.1496	**0.7827**	23.7773	**0.6073**
DiT4SR	27.3353	0.7740	23.8106	0.5430
MOSAIC	**29.5817**	**0.8249**	**25.2957**	0.5875

Table 1: PSNR and SSIM on the test sets of three datasets under $\times 2$ and $\times 4$. ■ best; ■ second best.

Method	LPIPS ↓	MANIQA ↑	CLIP-IQA ↑
RSSCN7			
Real-ESRGAN	**0.3199**	**0.3276**	0.4446
KDSR-GANV2	0.3554	0.3088	0.4336
LDL	0.3231	0.2911	0.4579
DiT4SR	0.3426	0.3260	**0.5391**
MOSAIC	**0.2737**	**0.3297**	**0.4674**
AID			
Real-ESRGAN	0.2803	**0.3243**	0.4663
KDSR-GANV2	0.3011	0.3154	0.4688
LDL	0.2847	0.2801	0.4679
DiT4SR	**0.2739**	**0.3207**	**0.5653**
MOSAIC	**0.2149**	0.2849	**0.4862**

Table 2: Perceptual quality on the RSSCN7 and AID test sets at $\times 4$. ■ best; ■ second best.

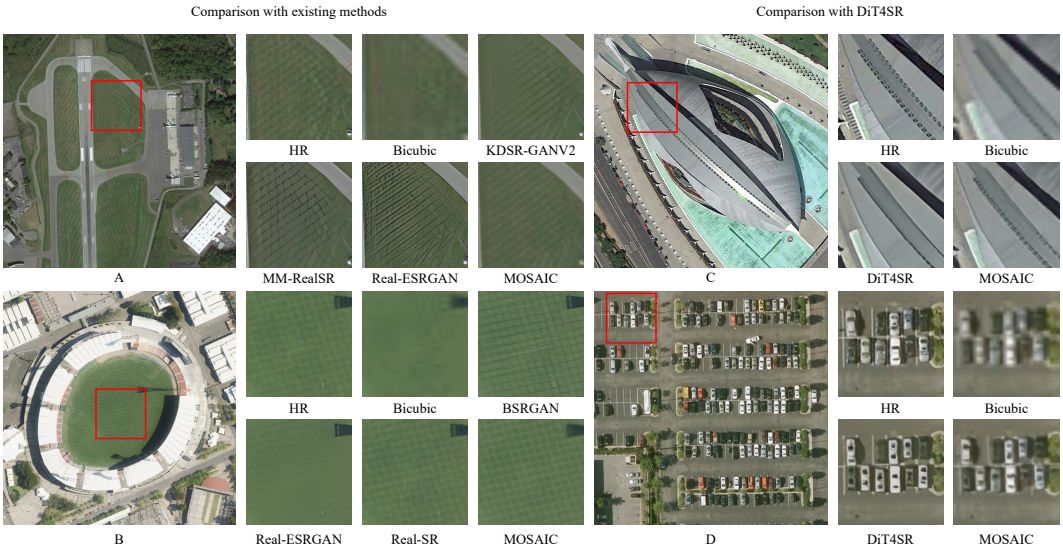

Figure 5: Qualitative super-resolution comparisons. Left: MOSAIC is compared with Bicubic and existing SR methods on Airport and Stadium scenes. Right: MOSAIC is compared directly with DiT4SR on Center and Parking scenes. Red boxes mark the enlarged regions.

Qualitative results. Figure 5 presents qualitative comparisons on Airport, Stadium, Center, and Parking scenes. Compared with existing SR methods, MOSAIC recovers more regular stripe and grid-like textures in Airport and Stadium while reducing local over-smoothing. In Center, DiT4SR largely flattens the serrated roof pattern into straight lines, whereas MOSAIC preserves non-straight roof details and structural variations. In Parking, DiT4SR smooths vehicle roofs into rectangular blocks, obscuring their geometry and depth cues; MOSAIC better preserves the boundaries, intensity variations, and spatial separation of individual roofs. The red boxes highlight these representative regions.

4.3 Ablation Studies

We study the main components and hyperparameters on the UCMerced $\times 4$ validation set. Tables 1 and 2 use test sets, so their absolute scores need not match the ablations. Within each ablation table, the split, seed, schedule, and checkpoint-selection rule are fixed; scores across exploratory tables are not directly compared. We examine OCL, MSF, prototype memory conditioning, slot number, SlotFiLM placement, MSF variants, and memory size. Further results are provided in the *Technical Supplement*.

Main-component ablation. Table 3 evaluates the cumulative contributions of OCL, MSF, and prototype memory conditioning (PMC). OCL accounts for the largest gain in this setting, raising PSNR by 1.8820 dB while improving SSIM and LPIPS. MSF adds 0.2820 dB PSNR, 0.0084 SSIM, and a 0.0102 reduction in LPIPS. PMC yields a smaller additional change under this protocol: 0.0079 dB PSNR, 0.0002 SSIM, and a 0.0033 reduction in LPIPS. Additional PMC controls are reported in the *Technical Supplement*.

OCL	MSF	PMC	PSNR ↑	SSIM ↑	LPIPS ↓
			22.2700	0.5670	0.2870
✓			24.1520	0.6274	0.2325
✓	✓		24.4340	0.6358	0.2223
✓	✓	✓	**24.4419**	**0.6360**	**0.2190**

Table 3: Main-component ablation on UCMerced $\times 4$ validation set. OCL denotes object-centric learning and PMC denotes prototype memory conditioning.

MSF variant	PSNR ↑	SSIM ↑	LPIPS ↓
Single-scale	24.1520	0.6274	0.2325
Multi-scale Avg	24.1620	0.6282	0.2323
Inter-scale Only	24.2440	0.6368	0.2215
Full MSF	**24.2510**	**0.6386**	**0.2210**

Table 4: MSF-variant ablation on UCMerced $\times 4$ validation set. Higher PSNR and SSIM are better; lower LPIPS is better.

MSF-variant ablation. Table 4 compares different ways to combine multi-scale slots. Direct averaging provides only a small improvement over the single-scale baseline. Inter-scale fusion gives a larger gain, demonstrating the benefit of learned cross-scale interaction over fixed averaging. Full MSF reaches 24.2510 dB PSNR, 0.6386 SSIM, and 0.2210 LPIPS. Compared with inter-scale-only fusion, it further improves all three metrics, supporting the complementarity of cross-scale fusion and original-scale detail refinement.

Slot-number ablation. Figure 6 evaluates the number of object-centric slots. PSNR varies non-monotonically across the tested values and reaches its maximum of 24.0658 dB at $K = 8$; it decreases at $K = 10$. This result motivates our choice of $K = 8$ for the final model.

Adapter-placement ablation. Figure 7 evaluates different SlotFiLM injection locations. The best result is obtained when SlotFiLM is injected into early and middle DiT blocks. This configuration, using blocks $[2, 8, 12, 16]$, achieves the best PSNR of 24.2439 dB. Early blocks receive the structural condition before the denoising features become highly specialized. Middle blocks further refine the recovered content. This placement provides a better balance between structural control and texture reconstruction than using only a limited part of the denoising network.

Prototype-memory-size ablation. Figure 8 evaluates the number of prototypes in the memory. Performance reaches its peak at $M = 16$, with a PSNR of 24.1156 dB. Other memory sizes yield lower performance under this ablation protocol; we therefore use 16 prototypes in the final model.

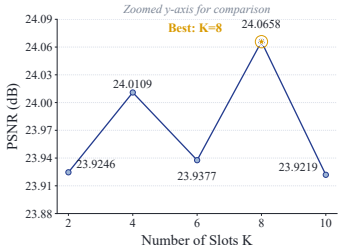

Figure 6: Slot-number ablation (UCMerced $\times 4$ validation set).

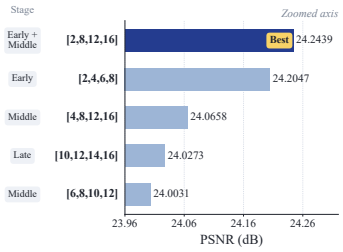

Figure 7: Adapter-layer ablation (UCMerced $\times 4$ validation set).

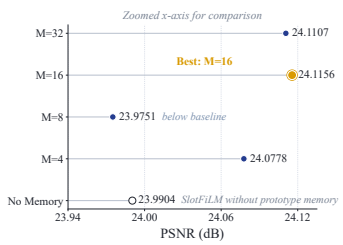

Figure 8: Prototype-memory-size ablation (UCMerced $\times 4$ validation set).

4.4 Slot Attention Visualization

Each example in Figure 9 contains a real-world input, its learned slot response, and the corresponding localized response. Different slots attend to complementary structural regions. These localized responses illustrate representation-level specialization before slot pooling and provide qualitative evidence that the OCL extractor can distinguish salient structures from redundant background patterns.

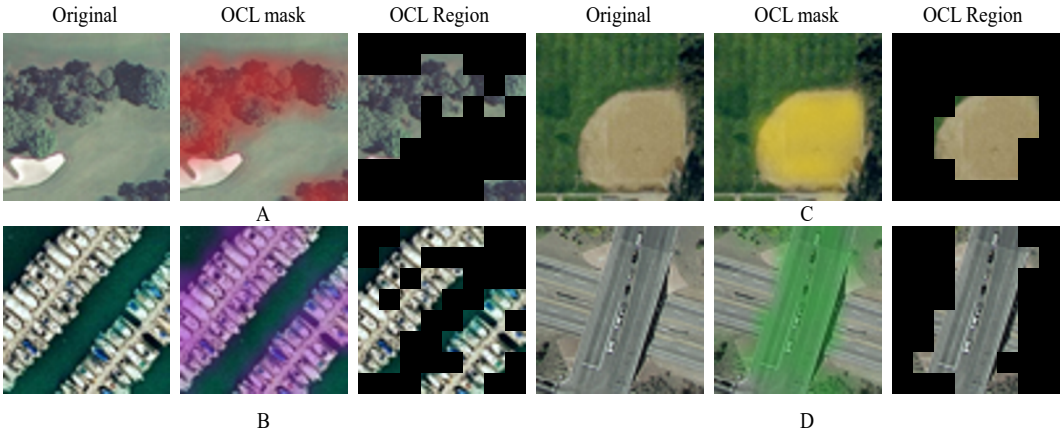

Figure 9: Slot attention responses. Each example shows an input, its learned slot response, and a localized response; slots attend to complementary structural regions.

5 Conclusion

In this paper, we presented MOSAIC, an object-centric adaptation framework for DiT-based remote sensing image super-resolution. MOSAIC organizes dense LR conditions into compact structural slots, combines cues across two scales, and retrieves shared prototypes to guide a frozen DiT backbone while preserving its generative prior. Experiments on three remote sensing datasets under $\times 2$ and $\times 4$ settings show consistent PSNR gains and favorable LPIPS results on the evaluated $\times 4$ datasets. The qualitative examples illustrate more coherent local structures and finer details in the shown scenes.

Broader Implications and Limitations. These results highlight object-centric conditioning as a promising way to adapt pretrained generative models to structure-sensitive remote sensing restoration. One limitation is that the prototype memory is constructed offline from the training distribution, which may limit adaptation to unseen sensors and degradations; developing dynamic and more transferable structural priors is left for future work.

AI USE STATEMENT

Generative AI tools were used to assist with language and grammar polishing of the manuscript. The authors take responsibility for the final paper.

REFERENCES

Christopher P. Burgess, Loic Matthey, Nicholas Watters, et al. MONet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Hanting Chen, Yunhe Wang, Tianyu Guo, et al. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021a.

Liang Chen, Hu Liu, Ming Yang, Yi Qian, Zhen Xiao, and Xiaochun Zhong. Remote sensing image super-resolution via residual aggregation and split attentional fusion network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:9546–9556, 2021b.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*, 2014.

Li Dong, Qihang Fan, Yuchen Guo, Zhenyu Wang, Qian Zhang, Jiyang Chen, Yihang Luo, and Chang Zou. TSD-SR: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Zhi-Peng Duan, Jie Zhang, Xiaosong Jin, et al. DiT4SR: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Klaus Greff, Raphael Lopez Kaufman, Rishabh Kabra, et al. Multi-object representation learning with iterative variational inference. In *International Conference on Machine Learning*, 2019.

Shuhao Han, Wenjie Liao, Hayden Vance, Hang Dong, Rui Zhang, Chun-Le Guo, and Chongyi Li. DNF-SR: Dual-input and negative-aware feature fine-tuning for real-world image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38176–38186, 2026.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

Jia-Bin Huang, Abhishek Singh, and Narendra Ahuja. Single image super-resolution from transformed self-exemplars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5197–5206, 2015.

Yajun Huo, Shun Gang, Xin Xiao, et al. FATSNet: Transformer-based skip network with frequency attention for remote sensing image super-resolution. *Complex & Intelligent Systems*, 12:131, 2026. doi: 10.1007/s40747-026-02245-z.

Hyeonjae Kim, Dongjin Kim, Eugene Jin, and Tae Hyun Kim. Continuous degradation modeling via latent flow matching for real-world super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5665–5672, 2026. doi: 10.1609/aaai.v40i7.37486.

Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. SwinIR: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, et al. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Brian B. Moser, Federico Raue, Stanislav Frolov, Sebastian Palacio, Jörn Hees, and Andreas Dengel. Hitchhiker's guide to super-resolution: Introduction and recent advances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9862–9882, 2023. doi: 10.1109/TPAMI.2023.3243794.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, 2019.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, 2023.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

Wanjie Sun and Zhenzhong Chen. Learning many-to-many mapping for unpaired real-world image super-resolution and downscaling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9874–9889, 2024. doi: 10.1109/TPAMI.2024.3428546.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

Jianyi Wang, Kelvin C. K. Chan, Hui Zeng, Ke Yu, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 2555–2563, 2023.

Yuzhi Wang, Wei Yang, Xinyi Chen, Yiyang Wang, Liang Guo, Lap-Pui Chau, Zhen Liu, Yu Qiao, Alex C. Kot, and Bihan Wen. SinSR: Diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Rongyuan Wu, Lingchen Sun, Zhengqiang Zhang, Xiangtao Kong, Jixin Zhao, Shihao Wang, and Lei Zhang. VOSR: A vision-only generative model for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16311–16321, 2026.

Gui-Song Xia, Jingwen Hu, Fan Hu, et al. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(7):3965–3981, 2017.

Yanting Xiao, Qiangqiang Yuan, Kun Jiang, Junjun He, Chia-Wen Lin, and Liang Zhang. TTST: A top-k token selective transformer for remote sensing image super-resolution. *IEEE Transactions on Image Processing*, 33:738–752, 2024.

Sidi Yang, Tianhe Wu, Shuwei Shi, et al. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, 2010.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018a.

Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *European Conference on Computer Vision*, 2018b.

Rongzhen Zhao, Vivienne Huiling Wang, Juho Kannala, and Joni Pajarinen. Multi-scale fusion for object representation. In *International Conference on Learning Representations*, 2025.

Rongzhen Zhao, Wenyan Yang, Juho Kannala, and Joni Pajarinen. Smoothing slot attention iterations and recurrences. In *Proceedings of the International Conference on Machine Learning*, 2026.

Wenli Zheng, Huiyuan Fu, Zekai Xu, Xin Wang, and Huadong Ma. Selective diffusion distillation for real-world high-scale image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 13467–13475, 2026. doi: 10.1609/aaai.v40i16.38351.

Qin Zou, Li Ni, Tong Zhang, and Qian Wang. Deep learning based feature selection for remote sensing scene classification. *IEEE Geoscience and Remote Sensing Letters*, 12(11):2321–2325, 2015.