BRIDGING DATA SILOS WITH SCENE-DEGRADATION-AWARE LEARNING FOR GENERALIZABLE REMOTE SENSING CLOUD REMOVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Cloud removal models often struggle to generalize across diverse land-surface scenes and cloud degradations. While combining multiple datasets can broaden the range of training conditions, their uneven and redundant distributions limit the benefit of direct data aggregation; meanwhile, the increased diversity introduces substantially different restoration patterns that are difficult to accommodate with a single shared-parameter model. To address this challenge, we approach it from two complementary perspectives. First, Scene-Degradation-Aware Data Distillation (SDA-DD) derives a scene-degradation-aware sampling distribution over the pooled cloudy-clear pairs, encouraging the selected training data to be more evenly distributed across diverse restoration conditions. Second, we develop OmniCR with a Local-Routing Mixture-of-Experts (LR-MoE) architecture that routes different windows to experts with complementary restoration preferences, allowing spatially varying regions within the same image to adopt different restoration behaviors. Soft-Scene Optimization further emphasizes scene-degradation relations with restoration losses, preventing easy-restored conditions from dominating training. Experiments on multiple CR benchmarks demonstrate consistent performance gains across diverse scenes and cloud-degradation conditions.

1 INTRODUCTION

Cloud removal (CR) is a fundamental task in optical remote sensing, restoring surface information obscured by clouds to support continuous and reliable Earth observation (Meraner et al., 2020; Ning et al., 2025; Ebel et al., 2023). Unlike conventional image restoration, cloud degradation in satellite imagery is neither uniform nor stationary: it ranges from translucent thin clouds to opaque formations and varies in morphology and spatial distribution across locations, seasons, and weather conditions (Ebel et al., 2022; Guo et al., 2025). The obscured content may correspond to cities, forests, or water bodies observed at different spatial resolutions and imaging conditions (Zhou et al., 2024a). CR must therefore learn not a single inverse mapping, but a family of restoration relations jointly determined by cloud type, terrain, and observation regime.

However, the problem of generalization across heterogeneous scene–degradation conditions remains insufficiently addressed, constrained by existing data and research paradigms. On the one hand, existing CR datasets usually exhibit different collection preferences during data acquisition, causing approaches developed on these datasets to be exposed to only a small portion of the underlying real-world distribution (Ebel et al., 2022; Wang et al., 2025d). Although direct pooling expands the theoretical support, it still preserves source-specific empirical frequencies (Torralba & Efros, 2011; Sorscher et al., 2022). On the other hand, conventional shared-parameter models must integrate a set of condition-dependent inverse relations into a single restoration pathway. Because the required restoration relations vary with terrain, cloud morphology, and even local regions within the same image, the lack of explicit conditional specialization forces these models to make trade-offs among heterogeneous restoration behaviors (Tian et al., 2025). As illustrated in Figure 1 (left), imbalanced data coverage and insufficient model specialization jointly hinder the translation of data diversity into robust cross-condition generalization (Rajagopalan et al., 2025; Li et al., 2025a).

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.

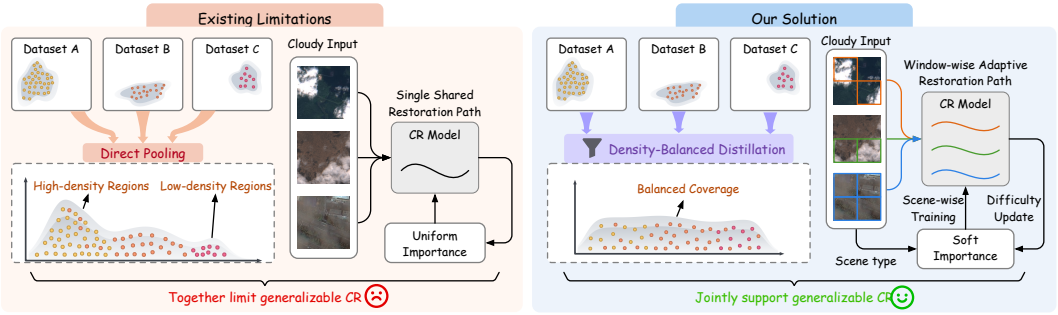


Figure 1: Motivation for jointly addressing data and model limitations in generalizable cloud removal. Left: Direct pooling preserves imbalanced scene–degradation distributions, while a shared restoration pathway and uniform training importance limit adaptation to heterogeneous conditions. Right: Our approach combines density-balanced data distillation with window-wise adaptive restoration and difficulty-aware scene reweighting, jointly promoting generalization across diverse scenes and cloud degradations.

We address these limitations from two complementary perspectives: data and model, as illustrated in Figure 1 (right). On the data side, the observation that representative subsets can approximate full-data training under limited budgets motivates us to define representativeness for CR (Mirzasoleiman et al., 2020). We argue that a degradation relation is primarily characterized by two factors: the terrain information obscured by clouds and the cloud type itself (Zhou et al., 2024a; Guo et al., 2025; He et al., 2010). To this end, we propose Scene-Degradation-Aware Data Distillation (SDA-DD). SDA-DD leverages a vision foundation model (VFM) with real-world visual priors (Siméoni et al., 2025) to construct representations of scene content and cloud degradation. Within this representation space, we measure pairwise sample similarities and assign scene-adaptive selection weights according to the density of similar representations. This gives greater coverage priority to sparsely represented scene–degradation relations while reducing the influence of redundant samples in dense regions. By optimizing a density-corrected coverage objective (Wei et al., 2015), we construct a balanced set spanning diverse terrains and cloud forms from over 300K pairs from 14 public CR datasets.

On the model side, the diversity of terrain and cloud-degradation conditions calls for multiple restoration preferences (Shazeer et al., 2017; Zamfir et al., 2025). At the same time, cloud degradation is typically spatially non-uniform, with cloud-obscured and relatively clear regions coexisting within an image (Meraner et al., 2020; Ning et al., 2025). Accordingly, we construct a Local-Routing Mixture-of-Experts (LR-MoE) network, allowing different experts to learn complementary restoration behaviors over local regions. Furthermore, we propose Soft-Scene Optimization to address the underemphasis of difficult scene relations under average-risk minimization (Sagawa et al., 2020). It reallocates learning emphasis according to restoration difficulty, promoting balanced learning across terrain–cloud relations. We refer to the complete framework as OmniCR; it complements SDA-DD's data coverage through local adaptive restoration and balanced learning across scenes.

In summary, our main contributions are as follows:

- We develop a scene-degradation-aware data distillation method that jointly characterizes scene content and cloud-degradation shifts to distill a compact training resource with broad coverage and reduced redundancy.
- We develop OmniCR, which uses LR-MoE for locally adaptive expert specialization and Soft-Scene Optimization to emphasize difficult scene–degradation relations, promoting balanced learning across heterogeneous conditions.
- Experiments on multiple CR datasets demonstrate that SDA-DD and OmniCR consistently improve performance across diverse scene and cloud-degradation conditions.

2 Related Work

2.1 Data Distillation for Image Restoration

Vision models continue to benefit from larger and more diverse training corpora, but the associated storage and optimization costs have also intensified interest in data efficiency (Sun et al., 2017; Gadre et al., 2023). Dataset distillation aims to condense a large training set into a compact training resource that preserves its learning utility while reducing data and optimization costs (Wang et al., 2018; 2025a; Bao et al., 2025; Tan et al., 2026). Its canonical formulations and benchmarks largely center on classification. Recent studies have begun to extend data distillation to image restoration (Li et al., 2025b; Dey et al., 2026). TripleD performs complexity-aware selection and feature-level distribution alignment for multiple restoration settings (Zheng et al., 2025). Cho et al. combine informative patch selection, semantic clustering, and diffusion-based generation for super-resolution data distillation (Cho et al., 2025), while Xu et al. develop personalized low-level dataset distillation for pixel-level mappings in medical image enhancement (Xu et al., 2026). UniRain retrieves real rainy references and employs VLMs to assess sample reliability for unified deraining (Yang et al., 2026). However, data-driven research in CR remains confined to isolated datasets, while direct pooling introduces redundancy and distribution imbalance. Compact data distillation for generalizable CR therefore remains underexplored.

2.2 Optical Remote Sensing Cloud Removal

General restoration models, including MPRNet (Mehri et al., 2021), Restormer (Zamir et al., 2022), AST (Zhou et al., 2024b), and MambaIR (Guo et al., 2024), provide representative CNN-, Transformer-, and state-space-based baselines for CR. Dedicated CR methods further incorporate cloud-specific assumptions. DFCFormer generates dynamic frequency-domain filters conditioned on cloud characteristics to accommodate diverse cloud types and aligns semantic features to preserve information for downstream tasks (Wang et al., 2025d). Generative CR methods instead model the restoration trajectory explicitly. Recent studies adapt conditional diffusion and progressive generation to optical imagery (Sui et al., 2024; Guo et al., 2025), while EMRDM establishes a mean-reverting process directly between cloudy and cloud-free images and redesigns its denoising and sampling procedures (Liu et al., 2025). GACR anchors residual flow to the cloudy observation and introduces VFM-based geo-contextual alignment, targeting both reconstruction fidelity and downstream semantic integrity (Wang et al., 2026). Despite these advances, existing CR methods primarily refine restoration architectures or generative processes within fixed data settings, rather than explicitly addressing generalization across diverse terrain and cloud-degradation conditions. MoE models suggest that specialized experts can accommodate heterogeneous degradations, but generally organize specialization around predefined tasks or image-level conditions (Zamfir et al., 2025; Wang et al., 2025b). Consequently, exploiting complementary experts for generalizable CR, while adapting their processing to spatially nonuniform cloud degradation, remains underexplored.

3 Methodology

Our approach addresses generalizable CR through two complementary aspects: data construction and model learning. SDA-DD distills heterogeneous training data into a compact resource with balanced coverage of scene and cloud-degradation conditions. OmniCR combines LR-MoE with Soft-Scene Optimization to learn locally adaptive restoration preferences and balance learning across terrain–cloud relations. The overall pipeline is illustrated in Figure 2.

3.1 Scene-Degradation-Aware Data Distillation

Problem formulation. Given paired training data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^{N}$, where X_i is a cloudy observation and Y_i its clear reference, we seek $B \ll N$ real pairs that provide broad and balanced coverage of scene and cloud-degradation conditions. To this end, we propose SDA-DD, which jointly characterizes each pair by its scene content and cloud-degradation shift and performs density-balanced selection in the resulting joint space.

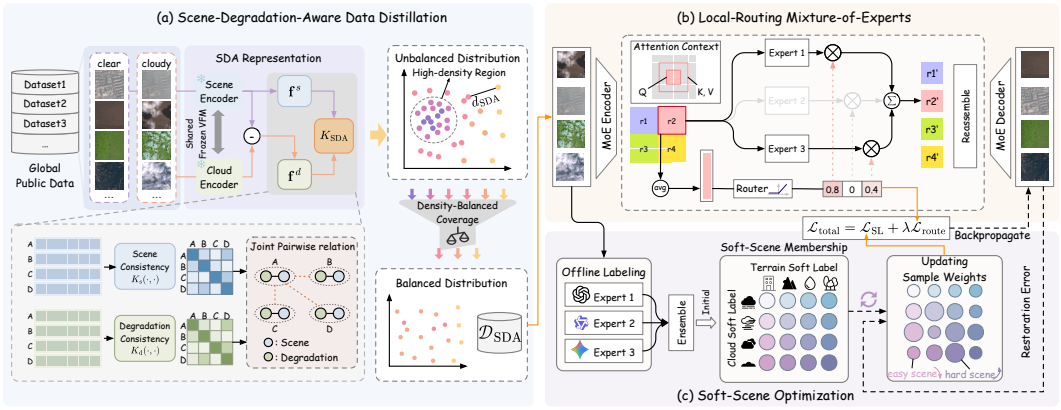


Figure 2: Overview of the proposed framework. (a) SDA-DD jointly models scene content and degradation shifts and distills the pooled data into a training resource with balanced scene–degradation coverage. (b) LR-MoE routes local regions and aggregates activated expert outputs, using overlapping neighborhoods as attention context. (c) Soft-Scene Optimization combines terrain and cloud soft labels into scene relations and updates learning weights according to restoration difficulty. OmniCR combines (b) and (c), with routing regularization during training.

SDA representation. We use a frozen VFM F as a shared high-level encoder for heterogeneous images, keeping distillation independent of an additional sample-value predictor or training phase. For the i-th pair, we extract the normalized clear and cloudy features as:

$$\mathbf{f}_i^s = \frac{F(Y_i)}{\|F(Y_i)\|_2}, \qquad \mathbf{f}_i^c = \frac{F(X_i)}{\|F(X_i)\|_2}, \tag{1}$$

where $\mathbf{f}_i^s$ serves as the scene representation of unobstructed content, while $\mathbf{f}_i^c$ additionally contains cloud effects. We further construct a degradation-shift representation as:

$$\mathbf{f}_i^d = \frac{\mathbf{f}_i^c - \mathbf{f}_i^s}{\|\mathbf{f}_i^c - \mathbf{f}_i^s\|_2}, \tag{2}$$

where the degradation-induced representation shift $\mathbf{f}_i^d$ describes the clear-to-cloudy transition in the latent space and reflects how cloud degradation affects the image representation.

SDA similarity. We first measure scene consistency and degradation-shift consistency using two nonnegative shifted-cosine similarities:

$$K_s(i,j) = \frac{1 + \langle \mathbf{f}_i^s, \mathbf{f}_j^s \rangle}{2}, \qquad K_d(i,j) = \frac{1 + \langle \mathbf{f}_i^d, \mathbf{f}_j^d \rangle}{2}. \tag{3}$$

We then couple the two components multiplicatively:

$$K_{\mathrm{SDA}}(i,j) = K_s(i,j) K_d(i,j). \tag{4}$$

Two pairs have high joint similarity only when both their scenes and degradation shifts agree. Accordingly, we express the distance d_{SDA} shown in Figure 2 (a) as $d_{\mathrm{SDA}}(i,j) = \sqrt{2 - 2K_{\mathrm{SDA}}(i,j)}$.

Density-balanced SDA distillation. The pooled data are generally distributed unevenly in the SDA space: dense regions contain many similar pairs, whereas sparse regions are represented by only a few samples, causing frequently observed regions to dominate the data distribution. We therefore seek a subset that preserves diverse scene–degradation patterns while providing more balanced coverage of the SDA space. Thus, we estimate the representational density around each pair by its average SDA similarity to the full collection as:

$$\rho_j = \frac{1}{N} \sum_{\ell=1}^{N} K_{\mathrm{SDA}}(j,\ell), \tag{5}$$

where a larger ρ_j indicates that pair j lies in a denser region. We then invert and normalize the densities to obtain the density-corrected coverage weight:

$$\omega_j = \frac{\rho_j^{-1}}{\sum_{\ell=1}^{N} \rho_\ell^{-1}}, \tag{6}$$

where ω_j suppresses the contribution of individual samples in dense regions while assigning greater coverage priority to sparsely represented regions. Then, let $\mathcal{I}$ denote a candidate subset of B real pairs. We score each original pair by its maximum SDA similarity and use the density-corrected weights ω_j defined above to maximize total coverage over the full collection:

$$\mathcal{I}^* = \arg\max_{|\mathcal{I}|=B} \sum_{j=1}^{N} \omega_j \max_{i\in\mathcal{I}} K_{\mathrm{SDA}}(i,j). \tag{7}$$

This optimization problem requires repeatedly evaluating marginal coverage gains, which becomes prohibitive over the full similarity matrix at large data scales. We therefore approximate the SDA similarity matrix using the Nyström method and construct the subset from the resulting approximation (Williams & Seeger, 2000). The detailed procedure is provided in Appendix A.

3.2 Local-Routing Mixture-of-Experts

The diversity of terrains, cloud morphologies, and occlusion levels in CR calls for complementary restoration preferences across experts. Since cloud degradation is spatially non-uniform, we perform expert processing over local regions, allowing specialization on degraded content while limiting unnecessary changes to clear regions. We term this architecture LR-MoE.

Local Routing and Expert Processing. LR-MoE treats local windows as its basic units of expert processing. Given a feature $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, we partition it into $N_w = HW/w^2$ non-overlapping windows of size $w \times w$ and denote the r-th window by $\mathbf{X}_r$. We describe each window by its mean feature to obtain the routing descriptor:

$$\mathbf{q}_r = \frac{1}{w^2} \sum_{t=1}^{w^2} \mathbf{X}_{r,t}, \tag{8}$$

where $\mathbf{X}_{r,t}$ is the t-th token in window r. At level ℓ, the router $\mathcal{R}_\ell$ maps $\mathbf{q}_r$ to expert scores, with the gate for expert e obtained through ReLU as:

$$g_{r,e}^{\ell} = [\mathcal{R}_\ell(\mathbf{q}_r)_e]_+ . \tag{9}$$

Expert e is activated for window r when $g_{r,e}^{\ell} > 0$. Each activated expert applies independently parameterized overlapping cross-attention (OCA) (Chen et al., 2023; 2024) to the entire window. The gated responses from the E_ℓ experts are then reassembled at their original spatial positions:

$$\mathcal{M}_\ell(\mathbf{x}) = \mathrm{Reassemble}\left(\left\{ \sum_{e=1}^{E_\ell} g_{r,e}^{\ell}\ \mathrm{OCA}_e^{\ell}(\mathbf{X}_r) \right\}_{r=1}^{N_w} \right). \tag{10}$$

A zero gate excludes an inactive expert from the weighted ensemble, and the resulting update $\mathcal{M}_\ell(\mathbf{x})$ is incorporated into the backbone through the block residual connection.

3.3 Soft-Scene and Routing Optimization

Soft-scene reweighting. Terrain–cloud relations differ in restoration difficulty, whereas the aggregate restoration loss weights all samples equally. To model variations across terrain–cloud relations, we precompute soft terrain and cloud-form labels for each image pair by averaging the category distributions predicted by three VLM experts, as shown in Figure 2 (c). For sample i, the corresponding probability vectors are denoted by $\mathbf{p}_i^s$ and $\mathbf{p}_i^c$, and their outer product defines its soft membership:

$$m_{i,s,c} = [\mathbf{p}_i^s]_s [\mathbf{p}_i^c]_c. \tag{11}$$

We initialize the scene importance as $a_{s,c}^{(0)} = 1$ and retain uniform weighting during warm-up. Thereafter, $a_{s,c}$ is updated at the end of each epoch according to relative restoration difficulty, assigning greater weight to more difficult scene relations. In implementation, we smooth the difficulty estimates with an exponential moving average; the detailed update is provided in Appendix B.

At epoch t, the sample weight combines soft memberships with the corresponding scene importance:

$$w_i^{(t)} = \sum_{s,c} m_{i,s,c} a_{s,c}^{(t)}. \tag{12}$$

The resulting weight is then applied to the sample-wise restoration error:

$$\mathcal{L}_{\mathrm{SL}}^{(t)} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i^{(t)} \left\| \widehat{Y}_i - Y_i \right\|_1 . \tag{13}$$

Routing regularization. ReLU routing allows each window to activate a variable number of experts. We adopt adaptive routing regularization (Wang et al., 2025c), which couples the mean gate magnitude and activation frequency of each expert and averages them over all routed blocks:

$$\mathcal{L}_{\mathrm{route}} = \frac{\alpha_t}{|\mathcal{R}|} \sum_{\ell \in \mathcal{R}} E_\ell \sum_{e=1}^{E_\ell} \mathbb{E}_r \left[g_{r,e}^{\ell} \right] \mathbb{E}_r \left[\mathbb{I}(g_{r,e}^{\ell} > 0) \right] . \tag{14}$$

Here, $\mathcal{R}$ denotes the set of routed blocks, $\mathbb{E}_r$ averages over local windows in the current batch, and $\mathbb{I}$ is the indicator function. The internal coefficient α_t is adjusted after each optimizer step toward one active expert per window. With a weight λ, the final objective is:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{SL}} + \lambda \mathcal{L}_{\mathrm{route}}. \tag{15}$$

4 Experiments

4.1 Experimental Settings

Datasets and evaluation protocol. To build a training resource with broad coverage, we collect 14 publicly available remote-sensing CR datasets (Ding et al., 2022; Dai et al., 2020; Lin et al., 2019; Sarukkai et al., 2020; Ebel et al., 2020; Pan et al., 2024; Sui et al., 2024; Wang et al., 2025d; 2026; Xu et al., 2023). Their spatial resolutions range from $0.1\,\mathrm{m}$ to $30\,\mathrm{m}$, encompassing diverse terrain and cloud-degradation conditions. We follow the existing dataset splits and merge the designated training sets into a shared candidate pool. When no public split is available, all samples are included in the pool. For evaluation, we retain the original test sets of the ten datasets that provide official splits. Further dataset details are provided in Appendix C.

Implementation details. During data distillation, we use a frozen DINOv3 ViT-L/16 pretrained on SAT-493M as the VFM to construct scene representations and degradation shifts. OmniCR is trained for 100K steps on four NVIDIA RTX A6000 GPUs using AdamW, a constant learning rate of 2×10^{-4}, and a global batch size of 8. Data augmentation consists of random crops and flips. Evaluation uses exponential moving average weights with a decay of 0.999. Soft labels for Soft-Scene Optimization are produced by three API-based VLM experts: GPT-5-nano, Gemini-2.5-Flash-Lite, and Qwen3.5-Flash.

Evaluation metrics and compared methods. We compare with 7 representative methods, including MPRNet (Mehri et al., 2021), Restormer (Zamir et al., 2022), AST (Zhou et al., 2024b), MambaIR (Guo et al., 2024), DFCFormer (Wang et al., 2025d), EMRDM (Liu et al., 2025), and GACR (Wang et al., 2026). For GACR, we evaluate its LVD and SAT variants with $p = 2$. For a fair comparison, all methods are retrained on the distillation subset. We report peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

4.2 Comparative Experiments

To comprehensively evaluate the proposed approach, we conduct comparative experiments from both model and data perspectives. The former examines the restoration performance of OmniCR

Table 1: Quantitative comparison on 4 representative CR benchmarks and the overall test set.

Method	CUHK-CR1		CUHKCR-EXT-CS		T-Cloud		SEN12MS-CR		Overall	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
MPRNet	21.29	0.563	20.26	0.618	25.17	0.790	<u>31.62</u>	0.817	26.35	0.802
Restormer	21.75	0.574	20.95	0.623	<u>25.62</u>	0.783	31.30	0.833	26.65	0.826
AST	20.73	0.514	19.92	0.583	23.99	0.755	28.72	0.828	26.63	0.798
MambaIR	21.59	0.542	19.36	0.590	24.53	0.768	30.21	0.841	26.46	0.812
DFCFormer	<u>22.26</u>	<u>0.581</u>	<u>20.98</u>	<u>0.643</u>	25.30	0.790	30.80	0.838	27.47	0.817
EMRDM	21.39	0.574	20.09	0.632	25.11	0.755	31.26	0.852	27.73	0.822
GACR-LVD/2	21.74	0.575	20.50	0.610	25.55	0.804	30.64	0.869	28.17	0.839
GACR-SAT/2	22.03	0.575	20.74	0.615	25.59	<u>0.810</u>	31.12	<u>0.876</u>	<u>28.49</u>	<u>0.844</u>
OmniCR	**22.54**	**0.590**	**21.90**	**0.665**	**25.93**	**0.815**	**33.63**	**0.906**	**30.75**	**0.876**

across diverse data distributions and scene conditions, while the latter evaluates the effectiveness of SDA-DD under different distillation strategies.

Table 1 reports the quantitative results of all compared models on four representative datasets, namely CUHK-CR1, CUHKCR-EXT-CS, T-Cloud, and SEN12MS-CR, together with their average performance over all ten test sets. OmniCR ranks first in every reported PSNR and SSIM metric. On the ten-set average, it surpasses the strongest competing method, GACR-SAT/2, by 2.26 dB in PSNR; on SEN12MS-CR, it leads by 2.51 dB. This consistent advantage in both individual and aggregated evaluations indicates that OmniCR's improvement is not confined to a particular data distribution. We also observe that GACR-LVD/2 performs below the 24.230 dB PSNR reported under its original single-dataset setting. This further supports our hypothesis that a model governed by a single shared mapping struggles to accommodate multiple complex restoration relations, whereas OmniCR benefits from its LR-MoE design and exhibits stronger generalization across diverse scenes. Figure 3 further compares four representative scenes. In the densely built-up scene in the second row, MPRNet produces a pronounced color shift, while DFCFormer, EMRDM, and GACR-SAT/2 retain cloud residuals to varying degrees; in contrast, OmniCR recovers the building layout more faithfully to the clear target. Overall, these results further demonstrate OmniCR's stronger reconstruction capability across diverse scenes and spatially non-uniform cloud degradations. Additional visual results and edge cases are provided in Appendix E.

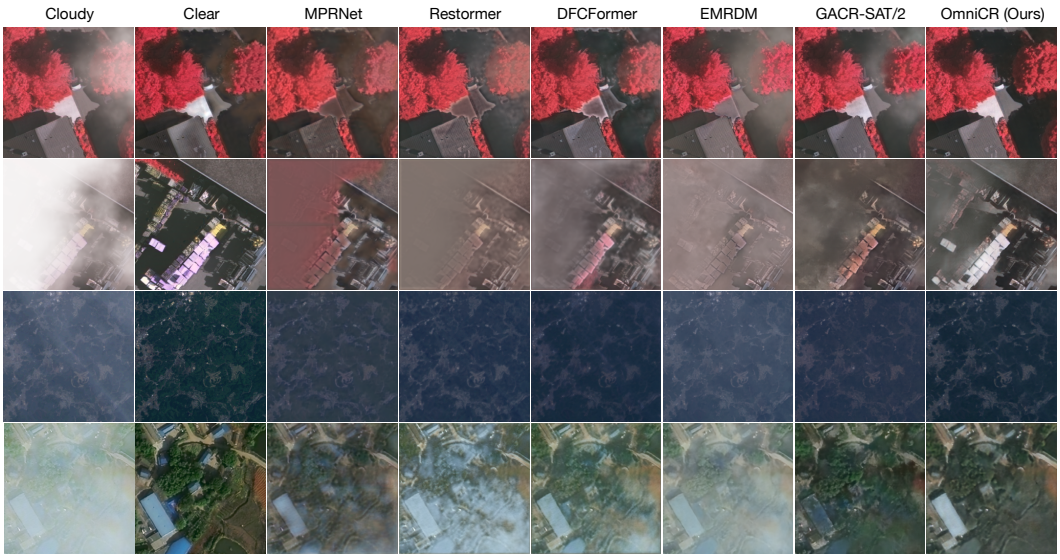
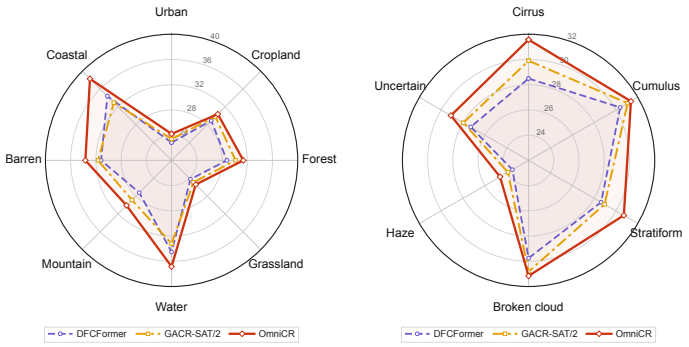


Figure 3: Visual comparison on 4 representative scenes. From left to right: cloudy inputs, clear targets, and restoration results from MPRNet, Restormer, DFCFormer, EMRDM, GACR-SAT/2, and OmniCR.



(a) Terrain categories (b) Cloud-form categories

Figure 4: Category-level PSNR comparison across terrain and cloud-form conditions. The two charts report results over eight terrain classes and six cloud-form classes, respectively.

To further examine performance under different terrain and cloud-form conditions, we group the test samples and evaluate the CR performance of each model over eight terrain categories and six cloud-form categories. Figure 4 reports PSNR across eight terrain and six cloud-form categories. Restoration difficulty varies substantially: urban scenes are the most challenging terrain type because dense structures provide limited contextual cues for recovering occluded details, while haze is the most difficult cloud form. Despite these variations, OmniCR achieves the highest PSNR in all 14 categories. Relative to the stronger result of DFCFormer and GACR-SAT/2 in each category, OmniCR yields average gains of 1.57 dB across terrain groups and 0.99 dB across cloud-form groups, demonstrating the adaptability of LR-MoE to diverse terrain–cloud relations.

Comparison of data distillation. Furthermore, to verify the effectiveness of SDA-DD, Table 2 reports the PSNR and SSIM of three models using different representation strategies at a 20% distillation scale. All three models achieve their best results with the training resources distilled by SDA-DD. Compared with random selection, the single-representation variants, Scene only and Degradation only, provide improvements to varying degrees, while SDA-DD outperforms all three alternatives. Specifically, SDA-DD improves the PSNR of MPRNet, Restormer, and OmniCR over random selection by 2.08, 2.35, and 2.87 dB, respectively. These results indicate that both SDA-DD data distillation and the OmniCR model design contribute to improved CR performance. Additional experimental analyses of SDA-DD are provided in Appendix E.

Table 2: Comparison of different representation strategies for data distillation.

Method	Metric	Random	Scene only	Degradation only	SDA-DD
MPRNet	PSNR	24.27	25.64	25.93	26.35
	SSIM	0.776	0.790	0.794	0.802
Restormer	PSNR	24.30	25.01	26.18	26.65
	SSIM	0.781	0.798	0.807	0.826
OmniCR	PSNR	27.88	29.11	30.12	30.75
	SSIM	0.816	0.843	0.868	0.876

4.3 Ablation Studies

Effect of distillation scale. Figure 5 compares SDA-DD with random sampling across different data budgets. At the 50% scale, SDA-DD enables Restormer and OmniCR to outperform full-data training, indicating that balanced coverage can be more effective than a larger but imbalanced pool. Under more aggressive compression, SDA-DD remains more stable than random sampling, demonstrating its robustness with limited training data.

Component analysis. Table 3 evaluates the contribution of each OmniCR component. The progressive improvements from Variants 1–3 verify the respective contributions of MoE, window-level routing, and ReLU routing, while Variant 4 shows that Soft-Scene Optimization is independently effective. Combining all components achieves the best PSNR of 30.75 dB, confirming the complementary contributions of LR-MoE and Soft-Scene Optimization.

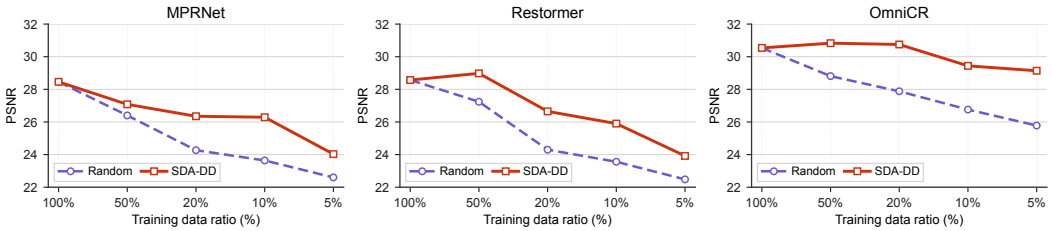


Figure 5: Comparison between random sampling and SDA-DD across different distillation ratios.

Table 3: Component ablation. Without MoE, Local, or ReLU, the model uses a shared-parameter architecture, image-level routing, or softmax gating, respectively.

Variant	MoE	Local	ReLU	$\mathcal{L}_{\mathrm{SL}}$	PSNR	SSIM
Baseline					28.92	0.843
Variant 1	✓				29.29	0.855
Variant 2	✓	✓			29.34	0.858
Variant 3	✓	✓	✓		29.52	0.861
Variant 4				✓	29.25	0.853
OmniCR	✓	✓	✓	✓	**30.75**	**0.876**

Table 4: Effect of routing coefficient λ. Experts denotes the number of expert activations per forward.

λ	PSNR	SSIM	Experts	FLOPs
0.1	30.56	0.865	1588.72	367.61
0.2	30.64	0.870	1368.16	367.29
0.5	30.66	**0.878**	1016.81	366.06
1	**30.75**	0.876	890.66	365.59
2	30.51	0.873	1025.35	366.23
5	30.45	0.869	1000.66	365.91

Table 5: Comparison of model complexity.

Metric	DFCFormer	EMRDM	GACR/2	OmniCR
FLOPs (G)	138.45	166.72	56.050	365.59
Params (M)	29.98	39.13	73.81	32.96

Table 6: GCPA variant results.

Metric	Original	w/ SAT	w/ LVD
PSNR	30.75	**30.86**	30.80
SSIM	0.876	0.877	**0.878**

Routing regularization. Table 4 shows that, as λ increases, the average number of activated experts decreases at $\lambda = 1$ and then rises. Layer-wise statistics suggest that stronger regularization suppresses shallow-layer activations, causing deeper layers to activate more experts to compensate for limited context. Since $\lambda = 1$ achieves the best balance between restoration performance and expert-activation efficiency, we adopt it as the default setting.

Computational complexity. Table 5 compares the computational complexity of different CR models. OmniCR contains 32.96 M parameters, fewer than EMRDM and GACR. Although OmniCR requires 365.59 GFLOPs per forward pass, iterative generative methods such as EMRDM and GACR typically require at least 4 sampling steps during inference. Accounting for this cumulative sampling cost, OmniCR has comparable inference complexity to these methods.

GCPA variants. In Table 6, we incorporate the GCPA module introduced by GACR (Wang et al., 2026) into OmniCR and evaluate its LVD and SAT variants. GCPA improves the restoration performance, but the gains are smaller than those reported for GACR. One possible explanation is that VFM features provide a stronger guidance prior for diffusion representations in latent space, whereas OmniCR directly processes the observed input rather than starting from randomly sampled latent variables, leaving less room for such guidance to improve performance.

5 Conclusion

We study generalizable remote-sensing CR through complementary data and model designs. SDA-DD constructs a compact training resource with broad scene–degradation coverage by jointly modeling scene content and degradation shifts. OmniCR combines LR-MoE with Soft-Scene Optimization to learn differentiated restoration behaviors and balance learning across scene relations. Experiments across multiple CR datasets demonstrate consistent gains in restoration and generalization. Ablations further highlight the importance of balanced coverage and adaptive modeling of heterogeneous conditions over simply increasing the training-data scale.

AI USE STATEMENT

In this work, we used generative AI tools for polishing and proofreading the final manuscript and for performing debugging checks during software development. We have not used generative AI tools for drafting the original manuscript or developing the core code, including the model structure, data distillation method, data preprocessing pipeline, and training framework, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for developing tools to organize and schedule experiment queues. We have reviewed all AI-assisted work. All AI-assisted textual revisions and utility code were manually checked by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Youneng Bao, Yiping Liu, Zhuo Chen, Yongsheng Liang, Mu Li, and Kede Ma. Dataset distillation as data compression: A rate-utility perspective. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 519–529. IEEE, 2025.

Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22367–22377. IEEE, 2023.

Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Sunwoo Cho, Yejin Jung, Nam Ik Cho, and Jae Woong Soh. Dataset distillation for super-resolution without class labels and pre-trained models. *IEEE Signal Processing Letters*, 32:3700–3704, 2025.

Peiyu Dai, Shunping Ji, and Yongjun Zhang. Gated convolutional networks for cloud removal from bi-temporal remote sensing images. *Remote Sensing*, 12(20):3427, 2020.

Priyam Dey, Aditya Sahdev, Sunny Bhati, Konda Reddy Mopuri, and Venkatesh Babu Radhakrishnan. Rethinking dataset distillation: Hard truths about soft labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 178–187, 2026.

Haidong Ding, Yue Zi, and Fengying Xie. Uncertainty-based thin cloud removal network via conditional variational autoencoders. In *Asian Conference on Computer Vision*, pp. 52–68. Springer, 2022.

Patrick Ebel, Andrea Meraner, Michael Schmitt, and Xiao Xiang Zhu. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 59(7):5866–5878, 2020.

Patrick Ebel, Yajin Xu, Michael Schmitt, and Xiao Xiang Zhu. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–14, 2022.

Patrick Ebel, Vivien Sainte Fare Garnot, Michael Schmitt, Jan Dirk Wegner, and Xiao Xiang Zhu. UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2086–2096. IEEE, 2023.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36:27092–27112, 2023.

Hang Guo, Jinmin Li, Tao Dai, Zhihao Ouyang, Xudong Ren, and Shu-Tao Xia. MambaIR: A simple baseline for image restoration with state-space model. In *European Conference on Computer Vision*, pp. 222–241. Springer, 2024.

Yujun Guo, Wei He, Ting Hu, Siti Khairunniza Bejo, and Hongyan Zhang. Cloud meets diffusion: Progressive cloud removal for optical remote sensing images. In *IGARSS 2025-2025 IEEE International Geoscience and Remote Sensing Symposium*, pp. 6123–6127. IEEE, 2025.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010.

Hao Li, Xiang Chen, Jiangxin Dong, Jinhui Tang, and Jinshan Pan. Foundir: Unleashing million-scale training data to advance foundation models for image restoration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12626–12636. IEEE, 2025a.

Hongcheng Li, Yucan Zhou, Xiaoyan Gu, Bo Li, and Weiping Wang. Diversity-enhanced distribution alignment for dataset distillation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3747–3756. IEEE, 2025b.

Daoyu Lin, Guangluan Xu, Xiaoke Wang, Yang Wang, Xian Sun, and Kun Fu. A remote sensing image dataset for cloud removal. *arXiv preprint arXiv:1901.00600*, 2019.

Yi Liu, Wengen Li, Jihong Guan, Shuigeng Zhou, and Yichao Zhang. Effective cloud removal for remote sensing images by an improved mean-reverting denoising model with elucidated design space. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17851–17861. IEEE, 2025.

Armin Mehri, Parichehr B Ardakani, and Angel D Sappa. MPRNet: Multi-path residual network for lightweight image super resolution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2704–2713, 2021.

Andrea Meraner, Patrick Ebel, Xiao Xiang Zhu, and Michael Schmitt. Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 166:333–346, 2020.

Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pp. 6950–6960. PMLR, 2020.

Jin Ning, Lianbin Xie, Jie Yin, and Yiguang Liu. Cloud removal advances: A comprehensive review and analysis for optical remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:15914–15930, 2025.

Jun Pan, Jiangong Xu, Xiaoyu Yu, Guo Ye, Mi Wang, Yumin Chen, and Jianshen Ma. Hdrsa-net: Hybrid dynamic residual self-attention network for sar-assisted optical image cloud and shadow removal. *ISPRS Journal of Photogrammetry and Remote Sensing*, 218:258–275, 2024.

Sudarshan Rajagopalan, Nithin Gopalakrishnan Nair, Jay N Paranjape, and Vishal M Patel. Gendeg: Diffusion-based degradation synthesis for generalizable all-in-one image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28144–28154. IEEE, 2025.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Vishnu Sarukkai, Anirudh Jain, Burak Uzkent, and Stefano Ermon. Cloud removal in satellite images using spatiotemporal generative networks. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1785–1794. IEEE, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Jialu Sui, Yiyang Ma, Wenhan Yang, Xiaokang Zhang, Man-On Pun, and Jiaying Liu. Diffusion enhancement for cloud removal in ultra-resolution remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 843–852. IEEE, 2017.

Haoru Tan, Wang Wang, Sitong Wu, Xiuzhe Wu, Yang-Tian Sun, Chirui Chang, Shaofeng Zhang, and Xiaojuan Qi. Dataset distillation by influence matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19654–19664, 2026.

Xiangpeng Tian, Xiangyu Liao, Xiao Liu, Meng Li, and Chao Ren. Degradation-aware feature perturbation for all-in-one image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28165–28175. IEEE, 2025.

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pp. 1521–1528. IEEE, 2011.

Shaobo Wang, Yicun Yang, Zhiyuan Liu, Chenghao Sun, Xuming Hu, Conghui He, and Linfeng Zhang. Dataset distillation with neural characteristic function: A minmax perspective. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25570–25580. IEEE, 2025a.

Tao Wang, Peiwen Xia, Bo Li, Peng-Tao Jiang, Zhe Kong, Kaihao Zhang, Tong Lu, and Wenhan Luo. Moerl: When mixture-of-experts meet reinforcement learning for adverse weather image restoration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13673–13683. IEEE, 2025b.

Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A. Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018.

Ziteng Wang, Jun Zhu, and Jianfei Chen. ReMoE: Fully differentiable mixture-of-experts with ReLU routing. In *International Conference on Learning Representations*, volume 2025, pp. 59486–59507, 2025c.

Ziyao Wang, Xianping Ma, and Man-On Pun. Downstream task-aware cloud removal for very-high-resolution remote sensing images: An information loss perspective. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18:24531–24545, 2025d. ISSN 2151-1535. doi: 10.1109/jstars.2025.3610641.

Ziyao Wang, Maonan Wang, Yucheng He, Xianping Ma, Ziyi Wang, Hongyang Zhang, Yirong Cheng, and Man-on Pun. Interpretation-oriented cloud removal via observation-anchored residual flow with geo-contextual alignment. In *European Conference on Computer Vision*, pp. 248–265. Springer, 2026.

Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *International conference on machine learning*, pp. 1954–1963. PMLR, 2015.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Fang Xu, Yilei Shi, Patrick Ebel, Wen Yang, and Xiao Xiang Zhu. Multimodal and multiresolution data fusion for high-resolution cloud removal: A novel baseline and benchmark. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–15, 2023.

Fengzhi Xu, Ziyuan Yang, Mengyu Sun, Joey Tianyi Zhou, and Yi Zhang. Low-level dataset distillation for medical image enhancement. In *European Conference on Computer Vision*, 2026.

Qianfeng Yang, Qiyuan Guan, Xiang Chen, Jiyu Jin, Guiyue Jin, and Jiangxin Dong. Unirain: Unified image deraining with rag-based dataset distillation and multi-objective reweighted optimization. *CVPR*, 2026.

Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12753–12763. IEEE, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5728–5739, 2022.

Zhuoran Zheng, Xin Su, Chen Wu, and Xiuyi Jia. Distribution-aware dataset distillation for efficient image restoration. *arXiv preprint arXiv:2504.14826*, 2025.

Hangyu Zhou, Chia-Hsiang Kao, Cheng Perng Phoo, Utkarsh Mall, Bharath Hariharan, and Kavita Bala. Allclear: A comprehensive dataset and benchmark for cloud removal in satellite imagery. *Advances in Neural Information Processing Systems*, 37:53571–53597, 2024a.

Shihao Zhou, Duosheng Chen, Jinshan Pan, Jinglei Shi, and Jufeng Yang. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2952–2963, 2024b.

APPENDIX

A SCALABLE IMPLEMENTATION OF SDA-DD

The density-weighted coverage objective in the main paper is defined on the full SDA kernel, but explicitly constructing and optimizing an $N \times N$ similarity matrix is computationally infeasible at our data scale. We therefore adopt a scalable approximate solver that preserves the central design of the selection objective while keeping computation and memory manageable.

A.1 NYSTRÖM APPROXIMATION OF GLOBAL DENSITY

The SDA kernel defined in the main paper is symmetric and positive semidefinite, which allows it to be approximated using the Nyström method. Exact optimization requires computing and storing the SDA similarities for all N^2 sample pairs. To avoid this cost, we sample $m \ll N$ reference samples, termed landmarks, from the full collection and compute only two sets of similarities: those among the landmarks and those between every sample and the landmarks. Let $\mathcal{L} = \{\ell_1, \ldots, \ell_m\}$ denote the landmarks. The corresponding similarity matrices are:

$$W_{ab} = K_{\mathrm{SDA}}(\ell_a, \ell_b), \qquad C_{ia} = K_{\mathrm{SDA}}(i, \ell_a). \tag{16}$$

The matrix $\mathbf{W}$ summarizes the similarity structure among the landmarks. We first symmetrize it and then compute $\mathbf{W} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, retaining the numerically stable positive eigenvalues and their corresponding components $\mathbf{U}_r$ and $\mathbf{\Lambda}_r$. Each sample's similarities to the landmarks can then be converted into an explicit map:

$$\boldsymbol{\phi}_i = \mathbf{C}_{i,:}\mathbf{U}_r\mathbf{\Lambda}_r^{-1/2}, \qquad K_{\mathrm{SDA}}(i, j) \approx \langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle. \tag{17}$$

The global similarity structure of the original SDA kernel is thus approximated by inner products in the mapped space, without explicitly storing all pairwise similarities.

We next use this explicit map to approximate the global density. The exact density in the main paper averages the similarity between each sample and the full collection. Under the Nyström map, let $\overline{\boldsymbol{\phi}} = N^{-1}\sum_{j=1}^{N}\boldsymbol{\phi}_j$. The global summation can then be reduced to an inner product between the sample map and the mean map:

$$\widetilde{\rho}_i = \frac{1}{N}\sum_{j=1}^{N}\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j \rangle = \langle \boldsymbol{\phi}_i, \overline{\boldsymbol{\phi}} \rangle. \tag{18}$$

Because the exact SDA kernel is nonnegative and satisfies $K_{\mathrm{SDA}}(i, i) = 1$, its density is lower-bounded by $1/N$. To preserve this property and prevent numerical undershoot from the finite-rank approximation, we impose the same lower bound before applying the inverse-density normalization from the main paper:

$$\widehat{\rho}_i = \max(\widetilde{\rho}_i, 1/N), \qquad \widetilde{\omega}_i = \frac{\widehat{\rho}_i^{-1}}{\sum_{j=1}^{N}\widehat{\rho}_j^{-1}}. \tag{19}$$

The resulting $\widetilde{\omega}_i$ approximates the density-corrected coverage weight in the main paper and is used only to determine the distilled subset.

A.2 SUBSET SELECTION

Using the approximate densities and mapped similarities, we progressively construct a selection ordering for the density-weighted coverage objective. For a current subset $\mathcal{I}$, let $c_j(\mathcal{I}) = \max_{s\in\mathcal{I}}\widetilde{K}(s, j)$ denote the coverage already received by sample j, where $\widetilde{K}$ is a fixed nonnegative similarity approximation constructed from the Nyström map. The marginal gain of adding candidate i is:

$$\Delta(i \mid \mathcal{I}) = \sum_{j=1}^{N}\widetilde{\omega}_j \max\Big(0, \widetilde{K}(i, j) - c_j(\mathcal{I})\Big). \tag{20}$$

The coverage objective exhibits diminishing returns: once a scene–degradation pattern is already well represented, selecting another similar sample provides only a small coverage gain, whereas a

sample representing an underrepresented pattern contributes more. Consequently, iteratively adding the sample with the largest current coverage gain achieves at least $(1-1/e)$ of the optimal coverage under the same subset budget.

A.3 Computational Complexity

Directly constructing the full SDA similarity matrix requires $O(N^2)$ memory. With m landmarks and a retained Nyström rank of r, we store only the $N \times r$ sample map and the $m \times m$ landmark kernel, reducing the memory complexity to $O(Nr+m^2)$. The global densities are obtained from the mean mapped feature in linear time $O(Nr)$, and the subsequent subset selection operates directly on the approximate similarities without materializing the full matrix. Therefore, when $m, r \ll N$, the implementation avoids the principal quadratic storage bottleneck as the dataset grows.

B Soft-Scene Optimization Details

B.1 VLM-Based Soft-Label Generation

For each cloudy–clear pair, the land-cover annotation is generated from the clear image, whereas the cloud-form annotation is generated from the cloudy image. Each VLM expert uses a fixed task-specific prompt to predict a probability vector whose entries represent the probabilities that the image belongs to the corresponding categories:

$$\begin{aligned} \mathcal{S} &= \{\text{urban}, \text{cropland}, \text{forest}, \text{grassland}, \text{water}, \text{mountain}, \text{barren}, \text{coastal}\}, \\ \mathcal{C} &= \{\text{cirrus}, \text{cumulus}, \text{stratiform}, \text{broken-cloud}, \text{haze}, \text{uncertain}\}. \end{aligned} \tag{21}$$

Let $\mathbf{p}_{i,e}^{x}$ denote the normalized probability vector predicted by valid expert e for sample i and task $x \in \{s, c\}$, and let E be the number of valid experts. The final soft label is obtained by equally averaging their predictions:

$$\mathbf{p}_i^x = \frac{1}{E} \sum_{e=1}^{E} \mathbf{p}_{i,e}^x, \qquad x \in \{s, c\}. \tag{22}$$

B.2 Scene-Difficulty Update

At the end of epoch t, let $\ell_i = \|\widehat{Y}_i - Y_i\|_1$ denote the reconstruction error of sample i. Using the soft memberships defined in the main text, the restoration difficulty of each scene relation is obtained by a weighted average of the associated sample errors:

$$D_{s,c}^{(t)} = \frac{\sum_{i \in \mathcal{D}_t} m_{i,s,c} \ell_i}{\sum_{i \in \mathcal{D}_t} m_{i,s,c}}. \tag{23}$$

Consequently, each sample contributes to multiple scene relations according to its soft memberships, while a larger $D_{s,c}^{(t)}$ indicates a higher restoration error for that relation at the current training stage. We smooth this estimate across epochs to reduce fluctuations and denote the result by $\overline{D}_{s,c}^{(t)}$.

B.3 Scene-Importance Update

After warm-up, we compare the smoothed difficulty of each scene relation with the overall mean difficulty in the log domain. Let $\pi_{s,c}^{(t)}$ denote the normalized soft sample mass of scene relation (s, c). Its relative difficulty is defined as:

$$r_{s,c}^{(t)} = \log \overline{D}_{s,c}^{(t)} - \sum_{u,v} \pi_{u,v}^{(t)} \log \overline{D}_{u,v}^{(t)}. \tag{24}$$

By comparing each scene relation with the overall mean difficulty, this calculation retains only the relative differences among scene relations: $r_{s,c}^{(t)} > 0$ indicates a relation that is more difficult than the current overall level, whereas $r_{s,c}^{(t)} < 0$ indicates a relatively easier relation. Finally, we convert

the relative difficulty into scene importance through an exponential mapping, where β controls the emphasis on difficult scene relations:

$$a_{s,c}^{(t)} = \frac{\exp\left(\beta r_{s,c}^{(t)}\right)}{\sum_{u,v} \pi_{u,v}^{(t)} \exp\left(\beta r_{u,v}^{(t)}\right)}. \tag{25}$$

This mapping assigns greater importance to scene relations with higher relative difficulty and reduces the contribution of relatively easier relations. Meanwhile, the normalization satisfies $\sum_{s,c} \pi_{s,c}^{(t)} a_{s,c}^{(t)} = 1$, so it redistributes learning emphasis across scene relations without changing the overall loss scale.

C Dataset Details

C.1 Dataset Overview

Table 7 summarizes the 14 public datasets used to construct the candidate pool, together with their training sizes, image dimensions, spatial resolutions, and evaluation usage. We follow the existing splits and merge only the training portions; the ten datasets with official test splits are also used for the overall evaluation.

Table 7: Overview of the 14 public cloud-removal datasets. Spatial resolution denotes the approximate ground sampling distance.

Dataset	Training pairs	Image size	Spatial resolution	Test pairs	Used for test
CUHK-CR1	534	512×512	0.5 m	134	✓
CUHK-CR2	448	512×512	0.5 m	111	✓
CUHKCR-EXT-CS	2,799	256×256	0.5 m	700	✓
CUHKCR-EXT-GZ	3,686	256×256	0.5 m	922	✓
LuojiaSET-OSFCR	20,000	256×256	10 m	–	×
M3R-CR	53,469	300×300	3 m	3,000	✓
Potsdam-CR	7,356	256×256	$\approx$ 0.1 m	1,840	✓
Vaihingen-CR	3,606	256×256	$\approx$ 0.1 m	902	✓
RICE-I	500	512×512	$\approx$ 30 m	–	×
RICE-II	736	512×512	$\approx$ 30 m	–	×
SEN12MS-CR	107,143	256×256	10 m	7,899	✓
STGAN-Single	97,640	256×256	10 m	–	×
T-Cloud	2,351	256×256	30 m	588	✓
WHUCloudDataset	680	512×512	30 m	16	✓
Total	300,948	–	0.1–30 m	16,112	-

C.2 Scene and Cloud-Form Composition

To summarize semantic composition across data sources, Tables 8 and 9 report both Hard and Soft statistics in each cell using the format Hard|Soft. Hard is the sample count obtained by assigning each sample to the highest-probability category in its VLM-ensemble prediction, whereas Soft is the average probability assigned to that category. Thus, the former provides a discrete category distribution, while the latter preserves the probabilistic information in the soft labels. In addition to the full candidate pool, both tables separately report the subset obtained by SDA-DD(20%).

Table 8: Land-cover composition in the format Hard count|Soft percentage.

Dataset / subset	Urban	Crop.	Forest	Grass.	Water	Mount.	Barren	Coastal	Total
CUHK-CR1	381|53.9%	2|3.7%	133|23.4%	1|5.1%	17|4.6%	0|1.2%	0|5.0%	0|3.1%	534
CUHK-CR2	422|72.4%	2|2.8%	15|7.8%	0|3.5%	9|3.9%	0|1.1%	0|4.7%	0|3.8%	448
CUHKCR-EXT-CS	1,380|35.5%	462|16.2%	811|24.9%	22|7.2%	102|5.7%	2|2.0%	20|5.8%	0|2.6%	2,799
CUHKCR-EXT-GZ	2,966|54.8%	86|5.8%	420|17.7%	3|4.1%	178|6.6%	1|1.8%	32|5.9%	0|3.3%	3,686
LuojiaSET-OSFCR	1,151|9.4%	9,813|34.7%	3,572|18.1%	361|8.0%	3,113|14.0%	845|5.7%	1,140|8.0%	5|2.1%	20,000
M3R-CR	8,210|13.6%	11,721|17.1%	17,964|26.8%	300|7.1%	3,690|7.0%	3,774|9.3%	7,768|15.8%	42|3.3%	53,469
Potsdam-CR	6,484|59.5%	110|5.1%	306|8.6%	370|9.6%	40|2.1%	0|1.5%	46|9.3%	0|4.2%	7,356
RICE-I	19|5.2%	39|7.9%	82|14.2%	4|11.6%	110|17.1%	147|23.1%	99|18.7%	0|2.3%	500
RICE-II	4|2.6%	82|9.9%	74|9.3%	29|7.8%	265|30.1%	89|11.2%	193|25.9%	0|3.0%	736
SEN12MS-CR	19,968|14.9%	22,844|17.5%	34,674|27.0%	385|6.5%	15,442|12.9%	7,878|9.2%	5,914|9.4%	38|2.6%	107,143
STGAN-Single	2,107|5.0%	22,249|19.4%	35,434|28.1%	3,968|10.4%	12,616|12.9%	8,677|9.4%	12,541|12.9%	48|2.0%	97,640
T-Cloud	219|10.0%	489|16.2%	302|13.4%	35|8.9%	145|6.7%	898|26.8%	263|16.2%	0|1.8%	2,351
Vaihingen-CR	2,968|56.5%	90|5.9%	538|19.4%	4|4.5%	4|1.7%	0|1.5%	2|6.8%	0|3.7%	3,606
WHUCloudDataset	97|15.5%	224|23.7%	157|19.8%	5|7.8%	20|6.9%	177|20.4%	0|4.4%	0|1.5%	680
Total	46,376|13.4%	68,213|18.5%	94,482|25.8%	5,487|8.0%	35,751|11.3%	22,488|8.8%	28,018|11.6%	133|2.5%	300,948
SDA-DD (20%)	7,040|10.8%	9,835|13.6%	13,853|20.3%	636|7.1%	10,439|14.5%	6,387|11.1%	11,927|19.6%	72|3.0%	60,189

Table 9: Cloud-form composition in the format Hard count|Soft percentage.

Dataset / subset	Cirrus	Cumulus	Stratiform	Broken cloud	Haze	Uncertain	Total
CUHK-CR1	2|4.3%	2|5.1%	36|18.7%	2|10.3%	393|47.4%	99|14.2%	534
CUHK-CR2	0|4.7%	0|5.1%	23|18.5%	1|11.3%	400|51.1%	24|9.3%	448
CUHKCR-EXT-CS	4|5.2%	26|9.2%	282|20.9%	7|10.4%	1,371|32.6%	1,109|21.7%	2,799
CUHKCR-EXT-GZ	1|4.3%	2|7.1%	214|19.0%	10|11.9%	1,355|28.3%	2,104|29.4%	3,686
LuojiaSET-OSFCR	176|8.0%	13,845|38.5%	2,810|18.5%	535|16.4%	1,474|10.2%	1,160|8.5%	20,000
M3R-CR	703|7.6%	6,498|11.9%	24,624|31.1%	1,425|14.6%	9,188|18.7%	11,031|16.1%	53,469
Potsdam-CR	7|4.9%	3|6.7%	365|19.0%	2|11.6%	6,510|48.1%	469|9.7%	7,356
RICE-I	2|5.9%	1|4.8%	151|27.3%	0|11.0%	247|34.0%	99|17.0%	500
RICE-II	42|10.2%	348|28.4%	168|23.2%	31|15.3%	112|14.5%	35|8.5%	736
SEN12MS-CR	1,193|7.9%	49,092|26.9%	24,990|22.2%	4,850|16.9%	10,650|12.7%	16,368|13.5%	107,143
STGAN-Single	1,575|9.3%	63,966|37.7%	19,501|19.9%	10,788|20.1%	908|6.6%	902|6.4%	97,640
T-Cloud	213|12.6%	16|7.9%	767|26.6%	20|13.8%	1,146|30.4%	189|8.6%	2,351
Vaihingen-CR	2|4.9%	49|9.1%	164|17.7%	11|12.8%	2,830|42.0%	550|13.5%	3,606
WHUCloudDataset	35|8.5%	120|15.6%	84|17.7%	65|16.6%	164|20.1%	212|21.4%	680
Total	3,955|8.1%	133,968|27.1%	74,179|22.6%	17,747|17.1%	36,748|13.5%	34,351|11.5%	300,948
SDA-DD (20%)	1,116|7.7%	11,395|15.0%	17,946|25.2%	2,023|14.7%	11,946|19.0%	15,763|18.4%	60,189

The per-dataset profiles reveal clear and complementary acquisition biases. CUHK-CR1/2, CUHKCR-EXT-GZ, Potsdam-CR, and Vaihingen-CR are urban-focused; LuojiaSET-OSFCR favors cropland; M3R-CR, SEN12MS-CR, and STGAN-Single emphasize forest and cropland; and RICE-I/II and T-Cloud add water, mountain, and barren scenes. Cloud-form labels show analogous specialization: CUHK-CR1/2, Potsdam-CR, and Vaihingen-CR favor haze, LuojiaSET-OSFCR, SEN12MS-CR, and STGAN-Single favor cumulus, M3R-CR favors stratiform clouds, and CUHKCR-EXT-GZ contains more uncertain clouds. Pooling these complementary sources therefore broadens semantic coverage beyond any single dataset.

This complementarity does not remove the size imbalance: SEN12MS-CR and STGAN-Single account for 68.0% of the Total candidate pool, rising to 85.8% with M3R-CR. Accordingly, the Total Soft distributions concentrate on forest (25.8%) and cropland (18.5%), and on cumulus (27.1%) and stratiform clouds (22.6%), while coastal scenes (2.5%) and cirrus clouds (8.1%) remain less represented. The complete pool is broad, but still dominated by frequent sources and their characteristic scenes.

At a 20% budget, SDA-DD redistributes this composition instead of reproducing the original proportions. In the Soft statistics, forest, cropland, and urban decrease by 5.5, 4.9, and 2.6 percentage points, while barren, water, and mountain increase by 8.0, 3.2, and 2.3 points; cumulus decreases by 12.1 points, while haze, uncertain clouds, and stratiform clouds increase by 5.5, 6.9, and 2.6 points. The Hard statistics show that all eight land-cover and six cloud-form categories remain represented. With only one fifth of the data, SDA-DD therefore reduces the dominance of frequent patterns and improves the relative coverage of underrepresented scenes and cloud forms, yielding a more balanced and diverse training subset.

D Complete Experimental Settings

D.1 OmniCR Configuration

Table 10 summarizes the architecture and training configuration of OmniCR in the main experiments. Unless an ablation specifically studies an individual component or data budget, all OmniCR experiments use the same settings, with only the factor under study being changed.

Table 10: Architecture and training configuration of OmniCR used in the main experiments.

Parameter	Configuration
Architecture	
input channel	3
output channel	3
dim	48
blocks	$[2, 4, 4, 6]$
refine blocks	4
channel heads	$[1, 2, 4, 8]$
spatial heads	$[2, 2, 3, 4]$
head dim	16
experts	$[3, 2, 1]$
window size	8×8
overlap ratio	$[0.5, 0.5, 0.5, 0.5]$
ffn ratio	2.66
routing	ReLU
Training	
steps	100K
patch size	128×128
batch size	8
optimizer	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$
lr	2×10^{-4}
weight decay	0
warmup epochs	10
difficulty ema / beta	$0.9/1.0$
weight ema decay	0.999
precision	FP16
seed	42
devices	$4\times$ NVIDIA RTX A6000

D.2 Compared Methods

We compare seven representative methods. MPRNet (Mehri et al., 2021) is a lightweight convolutional network that aggregates multi-level features through multi-path residual connections. Restormer (Zamir et al., 2022) adopts an encoder–decoder Transformer with channel-wise attention and gated depth-wise-convolutional feed-forward blocks. AST (Zhou et al., 2024b) uses adaptive sparse–dense attention and feature-refinement feed-forward modules to suppress irrelevant interactions and redundant channel features. MambaIR (Guo et al., 2024) adapts the Mamba state-space model to image restoration with local-enhancement and channel-attention modules. DFC-Former (Wang et al., 2025d) generates dynamic frequency-domain filters conditioned on cloud characteristics and aligns semantic features to preserve information relevant to downstream interpretation. EMRDM (Liu et al., 2025) constructs a mean-reverting diffusion process directly between the cloudy observation and cloud-free target, and redesigns the denoising network and sampling process to improve restoration efficiency. GACR (Wang et al., 2026) combines OAR-Flow anchored to the cloudy observation with GCPA based on vision-foundation-model priors; following the main setting, we report its LVD and SAT configurations with $p = 2$. All methods are retrained on the same distilled subset as described in the main paper.

E Additional Experimental Results

E.1 Additional Analysis of SDA-DD

To further analyze the selection behavior of SDA-DD, we examine coverage, sample retention across density groups, distributional changes in the representation space, and local neighborhood relations. Figures 6–8 report coverage and retention under different data budgets, while Figures 9 and 10 visualize the global distribution after selection and the local neighbors of selected representatives, respectively.

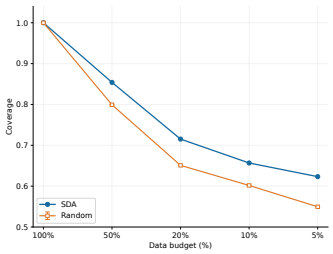


Figure 6: Coverage retained by SDA-DD and random sampling under different data budgets.

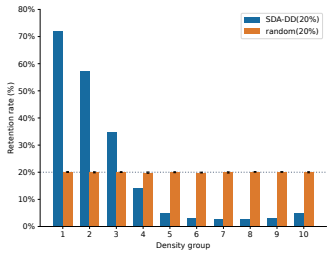


Figure 7: Retention rates of density-ranked groups at the 20% data budget.

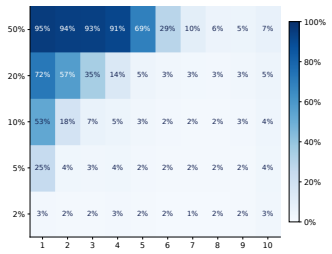


Figure 8: Retention rates across density groups and different data budgets.

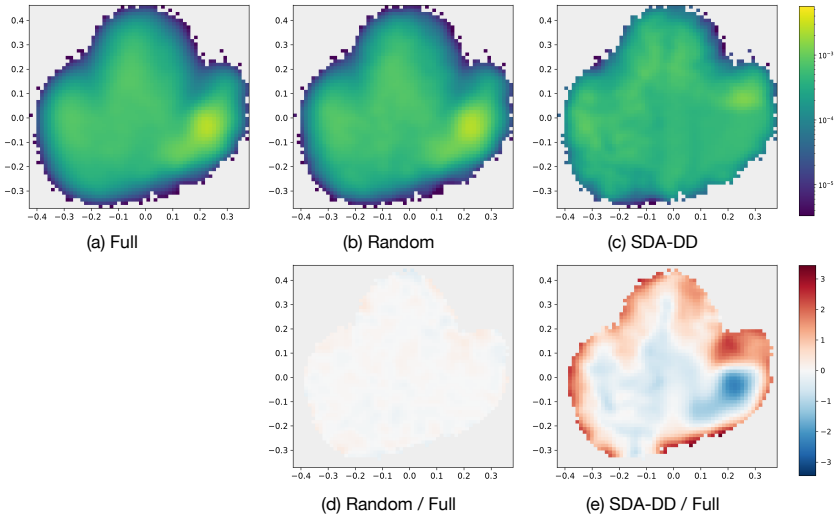


Figure 9: Distributions of the full data, random subset, and SDA-DD subset in the representation space, together with their relative density changes with respect to the full data.

For each data budget, we compute coverage using the density-corrected objective in equation 7: for every sample in the full candidate pool, we take its maximum SDA similarity to the selected subset and aggregate these maxima using the inverse-density weights; the random result is averaged over five independent selections. In Figure 6, both curves decline as the budget decreases, but SDA-DD remains consistently above random sampling. At the 20% budget used in our experiments, its coverage is 6.4 percentage points higher than random sampling, and the advantage remains 7.3 percentage points when the budget is further reduced to 5%. This persistent advantage indicates that SDA-DD allocates the limited budget to complementary scene–degradation patterns rather than repeatedly selecting samples from regions that are already well covered.

Groups 1–10 in Figures 7 and 8 are fixed equal-sized groups ordered from low to high SDA density in the full candidate pool, and each value denotes the fraction retained within that group. At the 20%

budget, random sampling retains approximately 20% from every group, whereas SDA-DD retains about 72% from the lowest-density group but only around 5% or less from the denser half, showing that it reallocates the budget away from regions containing many similar samples. Figure 8 further shows that, under the less restrictive 50% budget, more than 90% of the samples in the first four low-density groups are retained; as the budget tightens, this high-retention region progressively contracts toward the lowest-density groups, while a small number of representatives remain in the other groups. SDA-DD therefore does not simply exclude common patterns, but removes redundancy from dense regions while maintaining basic coverage across the density range.

In Figure 9, random sampling largely reproduces the density profile of the full data, whereas SDA-DD reduces the concentration of high-density regions and increases the relative presence of distributional boundaries and other low-density areas. The selected subset therefore preserves the overall support of the original data while reducing the dominance of frequent patterns in the training resource.

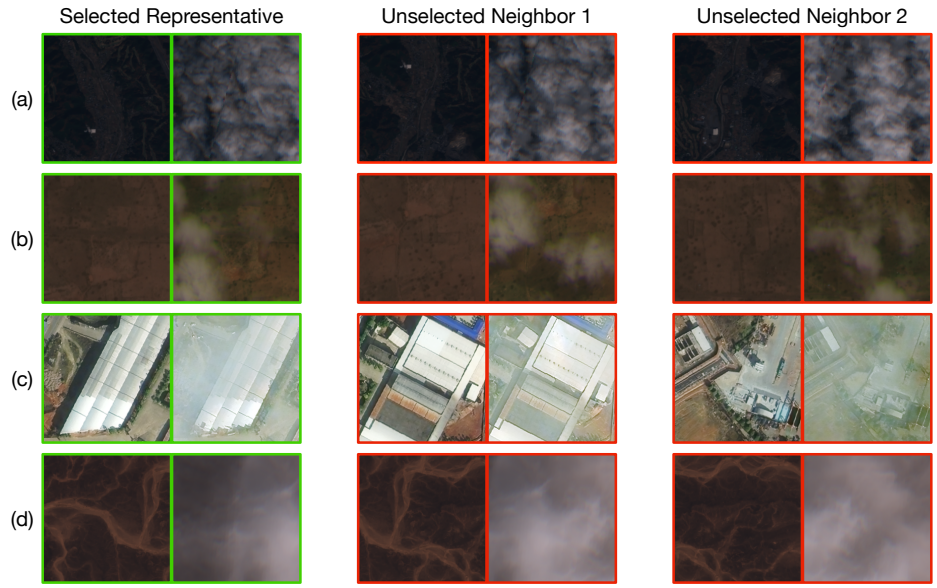


Figure 10: Selected representatives and their nearest unselected neighbors in the SDA representation space. Each pair shows the clear and cloudy images of one sample.

Figure 10 presents selected samples together with two of their nearest unselected neighbors. Samples within each row exhibit similar surface structures and cloud appearances, indicating that one selected sample can represent multiple nearby examples, whereas different rows cover distinct scene and degradation conditions. This visual evidence illustrates how SDA-DD reduces local redundancy while retaining diversity across conditions.

E.2 Additional Visual Comparisons

To further complement the quantitative comparisons in the main paper, we provide a richer collection of visual comparisons together with the corresponding analyses. Figures 11, 12, and 13 respectively report results on low-resolution scenes with limited surface detail, high-resolution scenes containing fine-grained structures, and challenging edge cases on which all compared methods struggle. These three groups examine how different methods preserve limited surface information, recover fine local structures, and handle rare and particularly difficult samples.

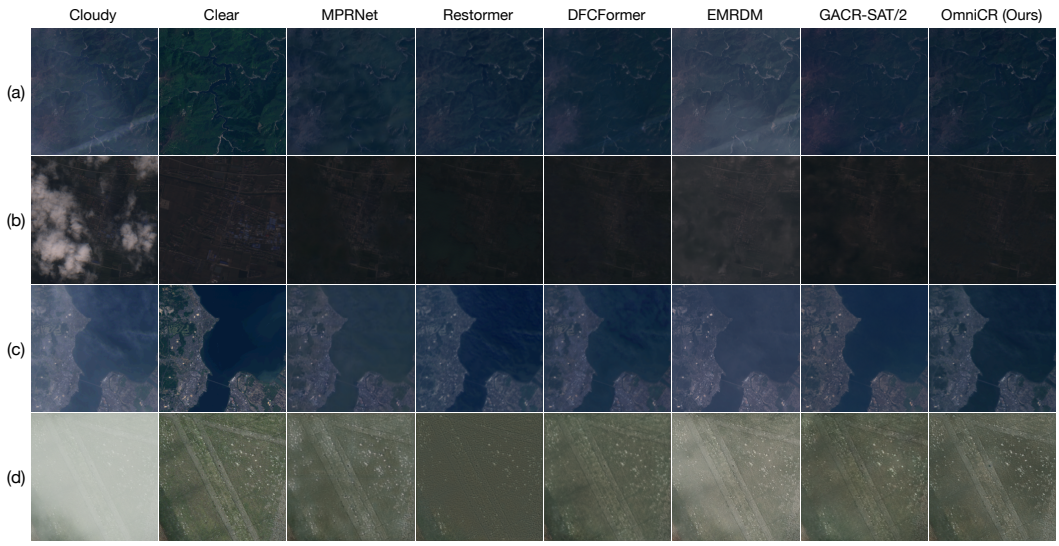


Figure 11: Visual comparisons on low-resolution scenes with limited spatial detail.

Low-resolution inputs contain weak textures and few distinctive structures, making it important to remove cloud contamination without further suppressing the limited surface information. As shown in Figure 11, several competing methods retain a dim veil or further smooth the already sparse texture cues. OmniCR more consistently recovers the global land-cover layout and major boundaries while maintaining brightness and color closer to the clear reference, with particularly visible improvements in the coastline and agricultural examples.

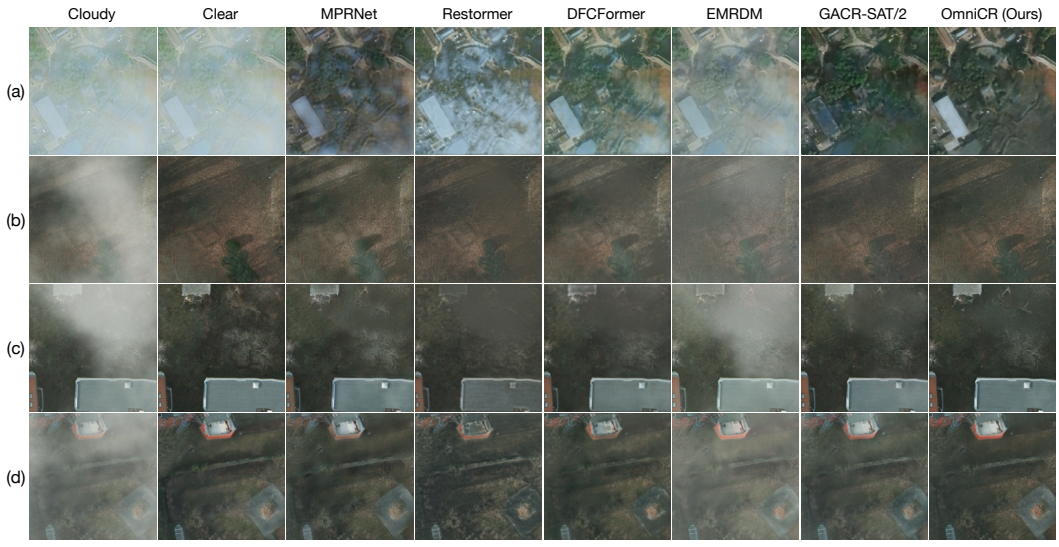


Figure 12: Visual comparisons on high-resolution scenes with fine-grained details.

High-resolution scenes further test whether cloud removal can preserve fine-grained structures. In Figure 12, most methods attenuate the dominant cloud regions but tend to blur roof contours, road patterns, and field boundaries, or introduce noticeable local color shifts. OmniCR provides clearer structural separation and more coherent colors, producing results closer to the corresponding clear references for both natural and man-made surfaces.



Figure 13: Visual comparisons on challenging edge cases where all methods struggle under severe cloud contamination.

Figure 13 presents challenging edge cases in which unusually heavy or spatially complex cloud contamination makes restoration extremely difficult. All compared methods fail to fully recover the obscured surface and commonly exhibit residual clouds, color shifts, or substantial structure loss. Nevertheless, OmniCR produces comparatively cleaner outputs and retains more recognizable road, building, and parcel contours, demonstrating a clearer relative advantage even on these difficult samples.