NATIVE-RESOLUTION READOUT AND SUPPORT-CENTERED GEOGRAPHIC CONDITIONING FOR DIFFUSION CLOUD REMOVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Auxiliary observations provide complementary information for cloud removal, but their utility depends on how spatial structure is represented at reconstruction. Spatial coarsening can suppress fine variation in continuous observation features; categorical maps require class composition and missing coverage to be represented distinctly. We view these as two forms of conditioning-representation mismatch and introduce native-resolution readout and support-centered geographic conditioning for diffusion-based cloud removal. Native-resolution readout exposes joint cloudy-optical and synthetic aperture radar (SAR) features directly to residual reconstruction, retaining spatial variation beyond an existing multiscale conditioning pathway. Geographic conditioning encodes land use and land cover (LULC) as target-grid class-area fractions with explicit coverage. A support-centered transformation removes the patch-average class composition while retaining within-patch spatial deviations; these features combine with SAR and elevation to update an RGB reconstruction embedding. Matched controls isolate representation effects in a 13-band pixel-space system on SEN12MS-CR and a latent-space RGB system on CloudRGB. With aligned auxiliary SAR channels, preserving the native field improves peak signal-to-noise ratio (PSNR) by 1.323 dB and reduces learned perceptual image patch similarity (LPIPS) by 0.0785 relative to spatially pooled conditioning. With geographic inputs fixed, support-centering improves full-image PSNR by 1.09 dB and reduces LPIPS by 0.0227 over uncentered area fractions in controlled CloudRGB gap reconstruction. These results support conditioning representations that preserve native spatial variation and explicitly encode valid support and local categorical deviations.

1 INTRODUCTION

Cloud removal seeks to reconstruct optical surface appearance from observations that are only partially informative. Synthetic aperture radar (SAR) provides complementary measurements when clouds obscure the optical signal, motivating a substantial body of optical–SAR restoration methods (Meraner et al., 2020; Xu et al., 2022; Ebel et al., 2023; Zhang et al., 2026). The complementary information is heterogeneous: joint optical–SAR features carry local spatial variation, while geographic maps describe terrain or categorical composition on grids that can differ from the output and have incomplete coverage. Effective reconstruction therefore requires more than making additional modalities available. It requires a representation that matches their spatial structure and valid support to the reconstruction interface.

Recent restoration models combine expressive backbones with increasingly rich conditioning mechanisms. Diffusion models provide flexible reconstruction priors, while spatial adapters and representation-based guidance offer different ways to exploit auxiliary information (Zou et al., 2024; Liu et al., 2025; Zhang et al., 2023; Mou et al., 2024; Yu et al., 2025). These developments motivate a common conditioning-representation question: how should a reconstruction interface retain the fine spatial variation of continuous observation features while explicitly representing categorical composition and valid support? For optical–SAR features, spatial coarsening can suppress output-aligned detail. For categorical maps, class proportions and missing coverage require distinct representations,

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.

and local composition can be expressed relative to the patch mixture. This separates the usefulness of an auxiliary observation from the usefulness of its representation.

We address this question through two complementary conditioning interfaces. For continuous optical–SAR joint features, we introduce native-resolution readout. Joint cloudy-optical and SAR feature extraction supplies both multiscale features to the existing restoration pathway and a native-resolution context to residual reconstruction. The additional context remains spatially aligned with the output grid and complements the existing multiscale conditioning pathway. We evaluate its role using matched full-native and pooled-native configurations with aligned auxiliary SAR channels, keeping the feature source, readout capacity, initialization, and adaptation budget fixed. Spatially pooling the context tests whether fine structure adds value beyond a coarse representation.

For categorical geographic information, the representation mismatch concerns class composition and valid support. Elevation and land-cover maps complement SAR backscatter with explicit geographic structure. Digital elevation models (DEMs) represent terrain elevation on a grid (Guth et al., 2021); land-use/land-cover (LULC) maps encode surface types such as vegetation, water, and built areas, with annually updated global products available online (Esri, 2024). LULC class identifiers are categorical, their grid may differ from the optical grid, and coverage can be incomplete. Interpolating label values is not a meaningful way to represent mixed classes, while ignoring coverage confounds missing data with valid observations. We instead aggregate class indicators into target-grid area fractions and retain their coverage explicitly. We then center each class channel over the valid support, separating within-patch spatial deviations from the patch-average class composition. The resulting geographic features combine with SAR and DEM to condition an additive update to the reconstruction embedding.

The two interfaces improve reconstruction in complementary diffusion restoration systems. On SEN12MS-CR, the native-field comparison evaluates output-resolution context at the reconstruction interface. Full native context improves peak signal-to-noise ratio (PSNR) by 1.323 dB over pooled context with the auxiliary SAR channels aligned. On CloudRGB, a compact geographic adapter augments RGB cloud-gap reconstruction. Combining elevation and land cover with SAR gains 3.11 dB PSNR over SAR alone, and support-centering gains 1.09 dB over uncentered area fractions. These gains demonstrate the value of native spatial context and support-centered categorical deviations for reconstruction.

Our contributions are threefold. (1) We develop native-resolution readout that exposes joint optical–SAR features to residual reconstruction at output resolution, preserving spatial structure beyond the existing multiscale conditioning pathway. (2) We formulate support-centered geographic conditioning that represents categorical composition with explicit coverage and separates patch-average composition from local spatial deviations. (3) We evaluate these interfaces in two complementary diffusion restoration systems, using matched controls to distinguish observation availability, spatial preservation, geographic complementarity, and representation effects.

2 Related work

Optical–SAR and diffusion-based cloud removal. DSen2-CR and GLF-CR fuse optical and SAR observations for cloud removal (Meraner et al., 2020; Xu et al., 2022). UnCRtainTS models reconstruction and uncertainty in optical time series, while ECRformer uses multimodal extraction with semantic-decoupled learning (Ebel et al., 2023; Zhang et al., 2026). DiffCR and EMRDM develop diffusion-based cloud restoration (Zou et al., 2024; Liu et al., 2025). Our native-resolution readout exposes output-aligned ECRformer features directly to residual reconstruction within the diffusion pipeline.

Spatial representations and conditioning interfaces. ControlNet and T2I-Adapter establish effective spatial conditioning paths (Zhang et al., 2023; Mou et al., 2024). REPA aligns diffusion features with clean-image representations during training (Yu et al., 2025). More closely, iREPA emphasizes spatial structure in alignment targets and combines spatial normalization with convolutional projection (Singh et al., 2026). Our native-resolution readout uses observed optical–SAR context at reconstruction. Our geographic transform acts on partially supported categorical area fractions: it subtracts support-weighted patch composition without variance normalization. This support-centered categorical representation is evaluated with coverage and adapter capacity held fixed.

LucidFlux routes degraded observations and a lightly restored proxy through timestep- and layer-dependent modulation, with proxy-derived semantic conditioning (Fei et al., 2026). Ada-RefSR regulates reference attention using implicit input–reference correlations (Wang et al., 2026). Our interventions preserve native joint-observation features beyond an existing multiscale pathway and encode categorical geography with valid coverage. They concern observed, spatially aligned conditions; these related systems focus on proxy-derived semantic conditioning or reference-image transfer.

3 METHOD

3.1 OVERVIEW

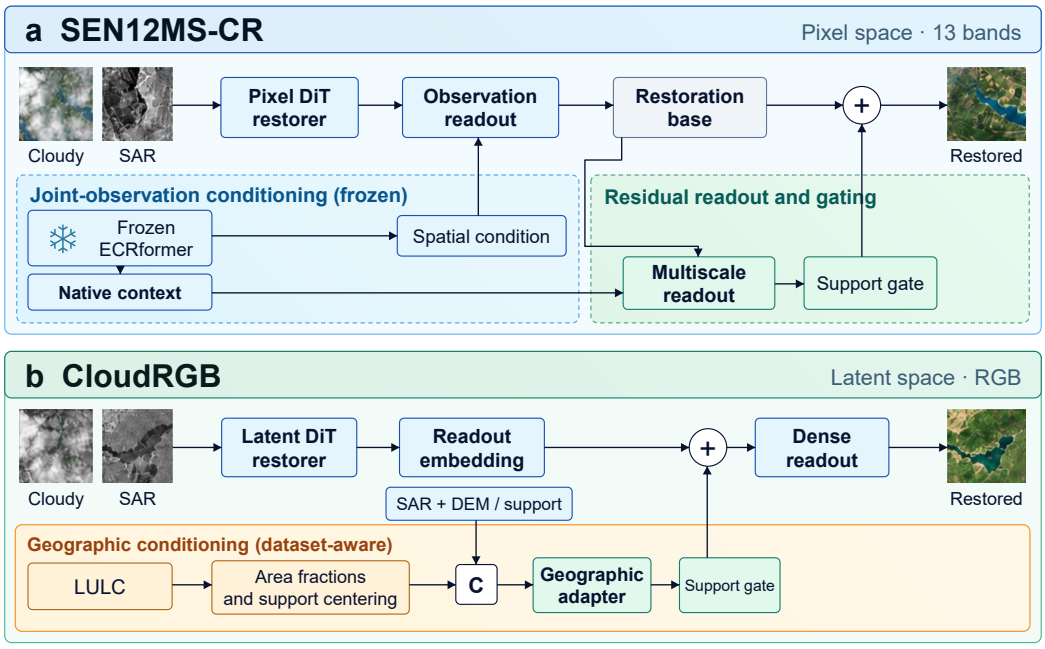


Figure 1: **Two dataset-specific conditioning pathways.** (a) SEN12MS-CR uses pixel-space diffusion to reconstruct 13 optical bands. A frozen optical–SAR extractor supplies spatial conditioning to the observation readout and native context to the residual readout. (b) CloudRGB uses latent-space diffusion with a geographic adapter that updates the RGB readout embedding from SAR, DEM, and support-centered land-cover features. Thumbnails are schematic.

We match conditioning representations to spatial structure and valid support (Figure 1). On SEN12MS-CR, a pixel-space restorer predicts 13-band imagery. On CloudRGB, a latent-space restorer reconstructs RGB through a codec and a dense readout. The two implementations use dataset-specific weights and diffusion processes.

Let C, S, and b denote cloudy optical observations, SAR, and the restoration base. The multispectral system exposes joint observation features directly to residual reconstruction. In the RGB system, geographic features update the readout embedding e before RGB prediction. Target images supply training supervision and evaluation references.

3.2 NATIVE-RESOLUTION READOUT

Our pixel-space implementation uses an x_0-predicting DiT and receives projected multiscale features from a joint optical–SAR extractor (Zhang et al., 2026). We preserve this existing conditioning route and additionally expose the extractor's native-resolution intermediate features N to a residual readout. Its optical and joint-feature branches combine the restoration base, observations, and native context.

Fine, middle, and coarse heads predict corrections, align them to the output grid, and sum them:

$$\widehat{X} = b + \mathbf{1}[V \wedge p_S] \odot (R_f + R_m + R_c), \tag{1}$$

where V is valid image support and p_S indicates SAR availability. The gate confines the residual update to supported locations. The extractor supplies intermediate features to the residual branch.

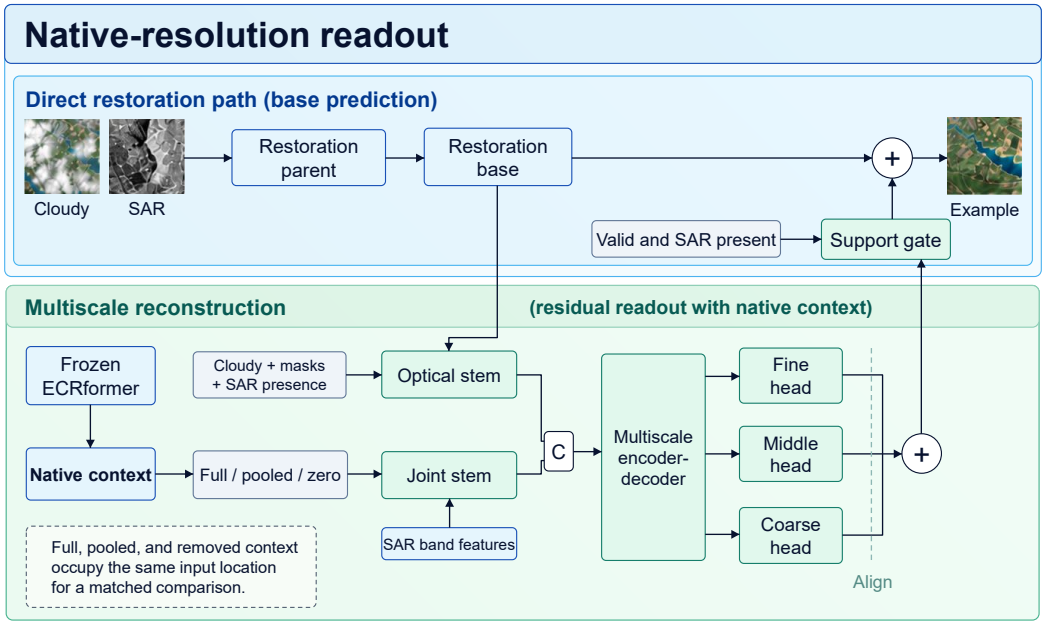


Figure 2: **Native-resolution residual readout.** The restoration base is combined with corrections from fine, middle, and coarse heads. The optical stem combines the restoration base with optical inputs, while the joint stem combines native context with SAR band features. The aligned residual sum is gated by valid image support and SAR availability. Full, pooled, and zero denote alternative native-context inputs. Thumbnails are schematic.

To evaluate spatial preservation, we compare full native context N and spatially pooled context PN under aligned auxiliary SAR conditioning (Figure 2). The pooled control retains a coarse spatial grid and restores its resolution by bilinear interpolation:

$$P = U_{H,W}^{\text{bilinear}} \circ \text{AvgPool}_{16}, \qquad N = PN + (I - P)N. \tag{2}$$

Full versus pooled context evaluates retaining the native field rather than its specified pooled representation, with feature source and trainable readout held fixed. The eight auxiliary SAR-derived channels, joint features, and other SAR paths remain spatially aligned in this comparison. Appendix B specifies the interfaces and the readout configurations used by the complete model and matched controls.

3.3 Support-centered geographic conditioning

Categorical geographic conditions require explicit representations of class proportions and valid coverage. We aggregate class indicators into target-grid area fractions $a_{i,c}$, with coverage $w_i = \sum_c a_{i,c}$. A small class fraction and a missing observation thus have distinct representations. To describe local composition relative to each 256×256 patch, we center each class over its effective support:

$$\mu_c = \frac{\sum_i a_{i,c}}{\sum_i w_i}, \qquad \widetilde{a}_{i,c} = a_{i,c} - \mu_c w_i. \tag{3}$$

When no support is available, both terms are defined as zero. The centered channels encode local excess or deficit relative to the patch mixture, while the separate coverage channel records how much valid land cover is observed. Equivalently, for $w_i > 0$, $\widetilde{a}_{i,c} = w_i(a_{i,c}/w_i - \mu_c)$: valid coverage weights the local deviation from patch composition.

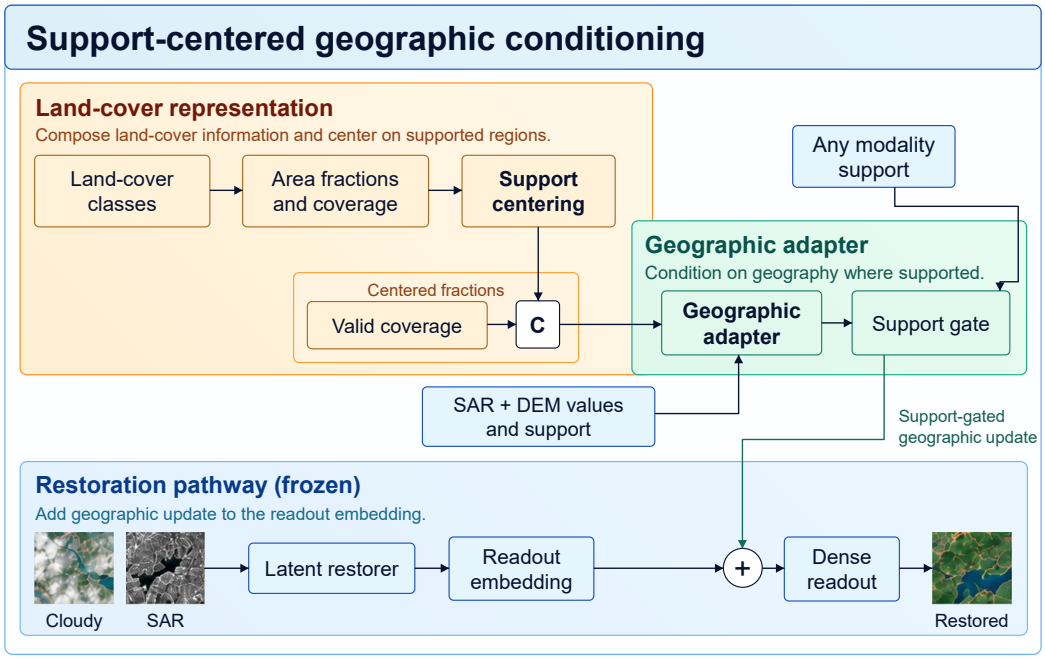


Figure 3: **Support-centered geographic conditioning.** Land-cover classes are aggregated into area fractions, then centered over their effective support. Centered composition and valid coverage are combined with supported SAR and DEM features in the geographic adapter. The union of modality supports gates its output, which updates the frozen restorer's readout embedding before dense RGB prediction. Thumbnails are schematic.

The adapter combines these channels with supported SAR and elevation values (Figure 3). A spatial transform with 1,031 trainable parameters predicts an additive update to the frozen RGB restorer's readout embedding. Its output is gated by the union of the available modality supports:

$$g = \mathbf{1}[v_S \vee v_D \vee (w > 0)] \odot \mathcal{A}_\phi(F_G), \qquad e' = e + g. \tag{4}$$

The dense readout then maps e' to RGB. Representation controls keep coverage and adapter capacity fixed while replacing centered fractions with uncentered fractions or nearest-class encoding. Appendix B gives the aggregation rule, channel layout, zero initialization, and centering identity.

4 Experiments

4.1 Experimental setup

Datasets. SEN12MS-CR pairs cloudy and clear 13-band optical observations with two-channel SAR (Ebel et al., 2021). CloudRGB evaluates controlled gap reconstruction on 256×256 HLSL30 v2.0 B4/B3/B2 reflectance patches, sampled at 30 m through Google Earth Engine (Masek et al., 2021). A fixed cloud mask removes pixels from each target to construct the observed RGB input. SAR, elevation, and annual land-cover maps provide auxiliary conditions. CloudRGB contains 131,538 patches, with 86,745/14,931/29,862 in TRAIN/VAL/TEST. Sample IDs, AOI IDs, and RGB source tuples are disjoint across splits; TEST covers 748 AOIs. Preprocessing and split details appear in Appendix A.

Comparisons and metrics. System-level comparisons evaluate reconstruction fidelity; matched controls isolate observation availability, native spatial detail, and geographic representation. All SEN12MS-CR comparison rows report local TEST evaluations. We report MAE, SAM, PSNR, SSIM, and LPIPS (Zhang et al., 2018) for system-level and matched comparisons. The multispectral metrics use all 13 bands; LPIPS uses the RGB view. CloudRGB metrics use RGB over the full valid image, including gaps and visible regions, and average finite per-image values with equal image weight.

PSNR uses a data range of one. Predictions are evaluated without clipping or visible-pixel copying. Qualitative comparisons use common display settings across methods. Appendix C.5 reports paired AOI-bootstrap intervals for the current-SAR and fixed-permutation native controls. Implementation settings and the model configuration map are provided in the supplementary material.

4.2 Overall Performance

4.2.1 Results on SEN12MS-CR

Table 1: **Quantitative comparison on the SEN12MS-CR TEST split.** All rows report local evaluations. MAE is in 10^{-3}, SAM in degrees, and PSNR in dB. Bold marks the best value in each column.

Method	MAE (10^{-3})↓	SAM↓	PSNR↑	SSIM↑	LPIPS↓
DSen2-CR (Meraner et al., 2020)	31.3	9.472	27.76	0.874	0.354
UnCRtainTS L2 (Ebel et al., 2023)	27.2	8.324	28.90	0.880	0.287
EMRDM (Liu et al., 2025)	17.9	5.267	32.14	0.924	**0.181**
ECRformer-Light (Zhang et al., 2026)	17.8	5.026	32.75	0.920	0.224
ECRformer (Zhang et al., 2026)	15.7	4.718	33.66	0.931	0.196
Ours	**12.3**	**3.921**	**35.24**	**0.938**	0.236

Table 1 compares our complete model with DSen2-CR, UnCRtainTS L2, EMRDM, ECRformer-Light, and ECRformer on the local TEST evaluation. Our complete model obtains 35.24 dB PSNR and 0.938 SSIM, with the lowest MAE and SAM and the highest PSNR and SSIM among the compared methods. Its PSNR exceeds the local ECRformer result by 1.58 dB. The matched spatial comparison in Section 4.3 evaluates the contribution of native-resolution context.

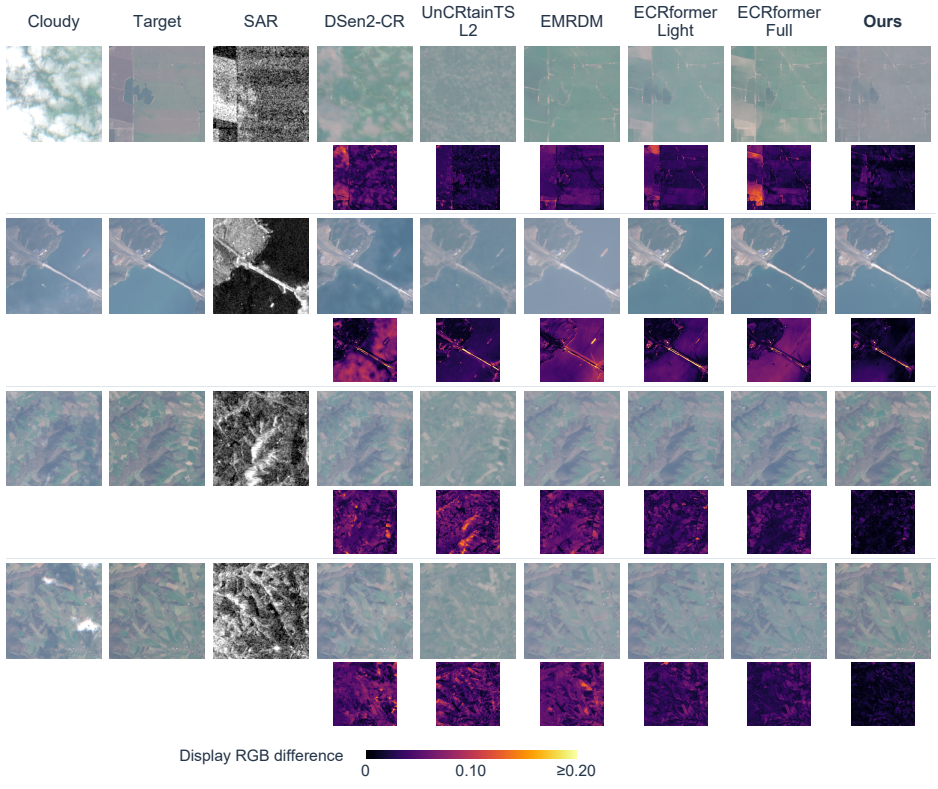


Figure 4: **SEN12MS-CR qualitative comparison.** Four selected scenes compare five reference methods with our complete model. The target is in the second column, with each model's RGB display error directly below its prediction. Errors are channel-mean absolute differences from the displayed target, using a common scale across both datasets. Further comparisons appear in Appendix D.

Figure 4 places our predictions beside five reference methods on four selected scenes. Further comparisons appear in Appendix D.

4.2.2 RESULTS ON CLOUDRGB

Table 2: **Quantitative comparison on CloudRGB.** MAE is in 10^{-3}, SAM in degrees, and PSNR in dB. All methods use observed RGB; auxiliary inputs are listed per row. Ours uses SAR and DEM together with support-centered LULC area fractions (D: DEM; L: LULC). Metrics cover full valid images. These rows compare complete systems; matched modality and representation controls appear in Section 4.4. Bold marks the best MAE, PSNR, SSIM, and LPIPS values.

Method	Aux.	MAE (10^{-3})↓	SAM† ↓	PSNR↑	SSIM↑	LPIPS↓
UnCRtainTS L2 (Ebel et al., 2023)	SAR	6.907	2.183	40.16	0.9505	0.2133
DSen2-CR (Meraner et al., 2020)	SAR	6.808	1.755	39.78	0.9466	0.2097
DiffCR (Zou et al., 2024)	SAR	6.043	2.019	41.13	0.9523	0.2172
ECRformer (Zhang et al., 2026)	SAR	**4.976**	1.554	42.41	0.9621	0.1802
Ours	SAR+D+L	5.975	1.244	**44.11**	**0.9697**	**0.1780**

† SAM uses metric-specific finite image sets; Ours has 22,495 finite values. Its other four metrics each use 29,862 images.

Table 2 compares the geographic readout with existing models evaluated on CloudRGB. With SAR, DEM, and LULC available, the proposed complete geographic configuration obtains 44.11 dB PSNR and 0.9697 SSIM. It achieves the best full-image PSNR, SSIM, and LPIPS means among the compared systems, with 1.70 dB higher PSNR than the RGB–SAR ECRformer configuration. All baselines are retrained with observed RGB and SAR and evaluated at 100k updates; Table 2 lists their input conditions. The matched controls in Section 4.4 test how the geographic representation improves the same parent restorer.

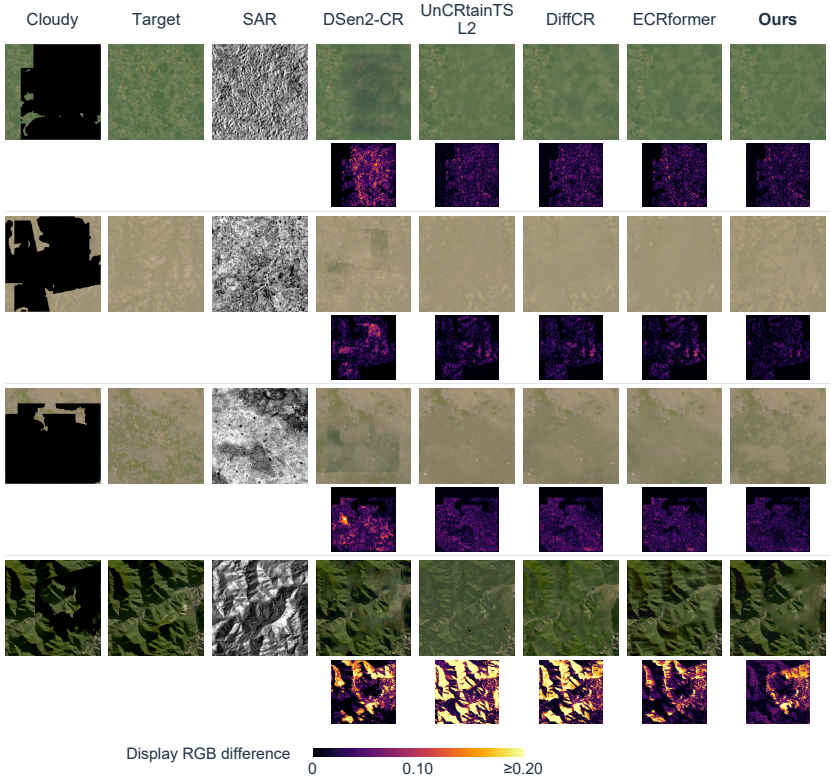


Figure 5: **CloudRGB qualitative comparison.** Four selected scenes use common display settings, with the target in the second column and RGB display error directly below each model's prediction. Error maps use the same scale as Figure 4. Black input regions are fixed gaps applied to the target.

Figure 5 illustrates reconstructions on four selected scenes with the same input and target.

4.3 Analysis of native-resolution readout

Table 3: **Matched current-SAR and native-field controls.** MAE is in 10^{-3}, SAM in degrees, and PSNR in dB. Each block uses its own matched readout; bold marks the best value within that block. Native-field controls use aligned auxiliary SAR channels and 2,048 updates per arm.

Method / condition	MAE (10^{-3})↓	SAM↓	PSNR↑	SSIM↑	LPIPS↓
Current SAR availability					
Cloudy only	32.933	10.047	27.65	0.8486	0.4047
Current SAR	**12.386**	**3.951**	**35.19**	**0.9376**	**0.2108**
Native spatial condition (aligned auxiliary SAR)					
Pooled native	15.240	5.117	33.81	0.9265	0.2888
Full native	**12.518**	**4.003**	**35.14**	**0.9373**	**0.2103**

Table 3 separates current SAR availability from native-field preservation under their matched readouts. Providing SAR raises PSNR from 27.65 to 35.19 dB and reduces LPIPS from 0.4047 to 0.2108. The availability comparison uses the same SAR-trained parent and changes the current SAR input throughout the restoration path.

The native-field comparison keeps all auxiliary SAR channels aligned and the projected joint-condition path active, changing the native context seen by the residual head. With the same parent and 2,048 updates per arm, full native context gains 1.323 dB PSNR and reduces LPIPS by 0.0785 relative to pooled context; MAE, SAM, and SSIM also improve. The native field improves reconstruction over the specified pooled representation with aligned auxiliary observations and unchanged readout

capacity. Additional full, pooled, and removed-context controls under a fixed auxiliary-channel permutation appear in Appendix C.4.

4.4 Analysis of geographic conditioning

4.4.1 Elevation and land cover provide complementary gains

Table 4: **Matched geographic modality controls on CloudRGB.** MAE is in 10^{-3}, SAM in degrees, and PSNR in dB. Bold marks the best MAE, PSNR, SSIM, and LPIPS values. The comparison fixes the parent, adapter capacity, and 2,048-update adapter budget.

Method / condition	MAE (10^{-3})↓	SAM† ↓	PSNR↑	SSIM↑	LPIPS↓
SAR only	5.999	1.894	41.00	0.9486	0.2543
SAR + DEM	6.007	1.705	42.99	0.9585	0.2044
SAR + LULC	**5.949**	1.723	43.13	0.9500	0.2275
SAR + DEM + LULC	5.975	1.244	**44.11**	**0.9697**	**0.1780**

† SAM means use metric-specific finite image sets, as defined in Appendix A.

Table 4 isolates the contribution of each geographic modality under the same RGB parent. Adding LULC to SAR increases PSNR by 2.13 dB and reduces LPIPS by 0.0268; MAE and SSIM also improve. Adding elevation to SAR improves PSNR, SSIM, and LPIPS. Combining all three modalities improves PSNR by 3.11 dB and reduces LPIPS by 0.0763 over SAR only. The combined configuration gives the best PSNR, SSIM, and LPIPS among the modality controls.

4.4.2 Support-centering improves over uncentered area fractions

Table 5: **Matched LULC representation controls.** MAE is in 10^{-3}, SAM in degrees, and PSNR in dB. Bold marks the best MAE, PSNR, SSIM, and LPIPS values. All rows use SAR + DEM + LULC and the same area-derived coverage.

Method / condition	MAE (10^{-3})↓	SAM† ↓	PSNR↑	SSIM↑	LPIPS↓
Nearest one-hot	6.018	1.734	43.02	0.9594	0.2004
Uncentered area	6.018	1.733	43.02	0.9593	0.2007
Support-centered area	**5.975**	1.244	**44.11**	**0.9697**	**0.1780**

† SAM means use metric-specific finite image sets, as defined in Appendix A.

With the parent, capacity, 2,048-update budget, and coverage held fixed, support-centering improves PSNR from 43.02 to 44.11 dB and reduces LPIPS from 0.2007 to 0.1780 (Table 5). The corresponding differences are 1.09 dB and 0.0227; MAE and SSIM also improve. The centered representation improves MAE, PSNR, SSIM, and LPIPS over both uncentered area fractions and nearest one-hot encoding at the same adapter capacity.

The centered channels encode local class excess or deficit relative to patch composition. The zero-sum identity in Appendix B.4 specifies this patch-relative encoding; the matched representation comparison measures its reconstruction benefit.

5 Conclusion

We developed native-resolution readout and support-centered geographic conditioning for diffusion-based cloud removal. On SEN12MS-CR, retaining native joint features improves reconstruction over spatially pooled conditioning with aligned auxiliary SAR channels. On CloudRGB, elevation and land cover provide complementary gains, with their combination improving full-image PSNR, SSIM, and LPIPS over either modality paired with SAR. Support-centering further improves full-image fidelity and perceptual similarity over uncentered area fractions. Together, these results support matching conditioning interfaces to the spatial structure and valid support of each observation: native joint context for optical–SAR reconstruction and coverage-aware local composition for categorical maps.

REPRODUCIBILITY STATEMENT

The supplementary material specifies CloudRGB optical acquisition, quality filtering, input construction and splits, train-only normalization, model interfaces, matched controls, full-image metric domains, and AOI estimands.

AI USE STATEMENT

AI tools assisted with manuscript writing, figure preparation, and review.

REFERENCES

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, 2021. doi: 10.1109/TGRS.2020.3024744. URL https://doi.org/10.1109/TGRS.2020.3024744.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2086–2096, 2023. doi: 10.1109/CVPRW59228.2023.00202. URL https://openaccess.thecvf.com/content/CVPR2023W/EarthVision/papers/Ebel_UnCRtainTS_Uncertainty_Quantification_for_Cloud_Removal_in_Optical_Satellite_Time_CVPRW_2023_paper.pdf.

Esri. Esri releases latest land cover map with updated Sentinel-2 satellite data, 2024. URL https://www.esri.com/about/newsroom/announcements/esri-releases-latest-land-cover-map-with-updated-sentinel-2-satellite-data. April 30, 2024. Accessed September 24, 2026.

Song Fei, Tian Ye, Lujia Wang, and Lei Zhu. LucidFlux: Caption-free universal image restoration via a large-scale diffusion transformer. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 86233–86260, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8ac80f6cdd1fffc43a4ba4ba49f40186-Paper-Conference.pdf.

Peter L. Guth, Adriaan Van Niekerk, Carlos H. Grohmann, Jan-Peter Muller, Laurence Hawker, Igor V. Florinsky, Dean Gesch, Hannes I. Reuter, Virginia Herrera-Cruz, Serge Riazanoff, Carlos López-Vázquez, Claudia C. Carabajal, Clément Albinet, and Peter Strobl. Digital elevation models: Terminology and definitions. *Remote Sensing*, 13(18):3581, 2021. doi: 10.3390/rs13183581. URL https://doi.org/10.3390/rs13183581.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17851–17861, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Liu_Effective_Cloud_Removal_for_Remote_Sensing_Images_by_an_Improved_CVPR_2025_paper.pdf.

J. Masek, J. Ju, J. Roger, S. Skakun, E. Vermote, M. Claverie, J. Dungan, Z. Yin, B. Freitag, and C. Justice. HLS operational land imager surface reflectance and TOA brightness daily global 30m v2.0, 2021. URL https://doi.org/10.5067/HLS/HLSL30.002. Data set. Product metadata checked 25 September 2026.

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. doi: 10.1016/j.isprsjprs.2020.05.013. URL https://mediatum.ub.tum.de/doc/1575235/document.pdf.

Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. doi: 10.1609/aaai.v38i5.28226. URL https://ojs.aaai.org/index.php/AAAI/article/view/28226.

Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. What matters for representation alignment: Global information or spatial structure? In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 33807–33845, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/3929a7785bd56f57edcff0152ab41289-Paper-Conference.pdf.

Yuan Wang, Yuhao Wan, Siming Zheng, Bo Li, Qibin Hou, and Peng-Tao Jiang. Trust but verify: Adaptive conditioning for reference-based diffusion super-resolution via implicit reference correlation modeling. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 96929–96955, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/9d0947107ea92d6ce369dce7749180dd-Paper-Conference.pdf.

Fang Xu, Yilei Shi, Patrick Ebel, Lei Yu, Gui-Song Xia, Wen Yang, and Xiao Xiang Zhu. GLF-CR: SAR-enhanced cloud removal with global–local fusion. *ISPRS Journal of Photogrammetry and Remote Sensing*, 192:268–278, 2022. doi: 10.1016/j.isprsjprs.2022.08.002. URL https://arxiv.org/abs/2206.02850.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87400–87442, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d9e42b4d7163931f3689d6d6fbaa11d0-Paper-Conference.pdf.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Adding_Conditional_Control_to_Text-to-Image_Diffusion_Models_ICCV_2023_paper.html.

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 Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Zaiyan Zhang, Jie Li, Yuanqi Liang, Jining Yan, Yi Xiao, Xin Su, and Qiangqiang Yuan. ECRformer: An efficient cloud removal transformer with semantic-decoupled learning for multimodal satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 237:323–338, 2026. doi: 10.1016/j.isprsjprs.2026.04.009. URL https://zzaiyan.github.io/assets/pubs/ecrformer.pdf.

Xuechao Zou, Kai Li, Junliang Xing, Yu Zhang, Shiying Wang, Lei Jin, and Pin Tao. DiffCR: A fast conditional diffusion framework for cloud removal from optical satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024. doi: 10.1109/TGRS.2024.3365806. URL https://arxiv.org/abs/2308.04417.