Interpretation-Oriented Cloud Removal via Observation-Anchored Residual Flow with Geo-Contextual Alignment

Ziyao Wang[1,*], Maonan Wang[1,2], Yucheng He[1], Xianping Ma[3,†], Ziyi Wang[1], Hongyang Zhang[1], Yirong Cheng[2], and Man-on Pun[1,†]

[1] School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen, China
[2] Shanghai Ai Lab, Shanghai, China
[3] Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu, China

Abstract. Cloud removal (CR) is essential for optical remote sensing, serving as a prerequisite for reliable downstream interpretation, such as semantic segmentation and change detection. However, existing CR approaches often prioritize visual realism while overlooking their impact on subsequent analytical tasks, leading to semantic drift and degraded downstream performance. To address this issue, we propose Geo-Anchored Cloud Removal (GACR), a unified framework that jointly ensures faithful reconstruction and robust interpretability. At its core, GACR incorporates Observation-Anchored Residual Flow (OAR-Flow), which reformulates CR as a physically grounded residual inversion process. By anchoring the generative trajectory to the cloudy observation rather than pure noise, OAR-Flow enables fast, stable, and faithful reconstruction. To further preserve semantic structures critical for downstream interpretation, GACR integrates Geo-Contextual Prior Alignment (GCPA) to constrain the reconstruction within a semantic manifold induced by a Vision Foundation Model (VFM). Consequently, GACR strictly maintains the spatial-semantic integrity of complex landscapes. Extensive experiments across six CR datasets and twelve downstream tasks demonstrate that GACR produces superior reconstruction quality while consistently improving downstream task accuracy. The code is available at `https://github.com/wzy6055/GACR`.

Keywords: Cloud Removal · Remote Sensing · Observation-Anchored Generative Modeling · Geo-Contextual Semantic Alignment

1 Introduction

Optical satellite imagery constitutes a primary data source for a broad spectrum of Earth observation applications, including urban development monitoring, resource management, and land-cover mapping [3,15,48,49]. The reliability of these

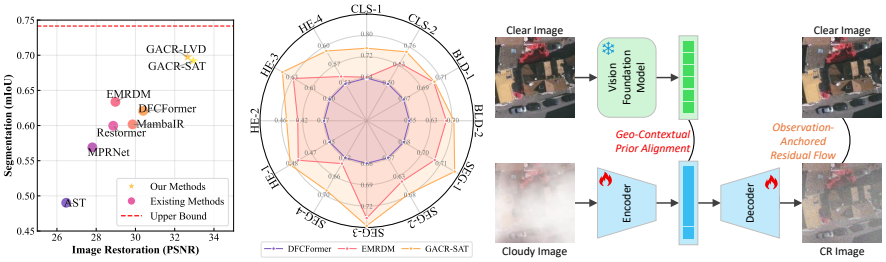
∗ First author. † Co-corresponding authors.

applications fundamentally depends on accurate surface representation and semantic consistency in the observed imagery. However, the presence of clouds severely limits the usability of optical imagery by obscuring surface information and introducing uncertainty in analyses [20,23]. As a result, cloud removal (CR) has evolved from a simple preprocessing operation to a critical component in the remote sensing interpretation pipeline, motivating extensive efforts toward generating cloud-free and semantically reliable imagery [8,20].

Recent advances in deep learning-based CR methods can be broadly categorized into denoising-based [32,46,47] and generative-based approaches [27,31,42, 51]. Denoising-based methods typically treat cloud occlusion as additive residual noise and learn a direct mapping to the underlying clear image. However, these methods are built upon the assumption that the residual noise follows a Gaussian prior distribution [44]. Under heavy occlusion, where surface information is largely obscured rather than merely perturbed, such assumptions often lead to structural ambiguity and over-smoothed reconstructions. In contrast, generative approaches, particularly diffusion-based models, demonstrate a strong capability in synthesizing visually plausible textures under severe cloud coverage. Yet, their inherent stochastic sampling process lacks explicit observation anchoring, frequently resulting in geographically inconsistent structures or semantic drift in heavily occluded regions. While visually realistic, such hallucinated details may contradict the true land-cover distribution, thereby undermining reliability in downstream interpretation tasks.

Beyond the lack of observation anchoring in generative modeling, a more fundamental challenge lies in the visual-fidelity-oriented optimization paradigm underlying most existing CR methods. Current approaches primarily minimize pixel-level discrepancies against cloud-free references, implicitly equating visual fidelity with semantic correctness. Such objectives encourage aggressive artifact removal to maximize conventional image quality metrics (e.g., PSNR, SSIM) [5,44], yet provide no explicit constraint on preserving task-relevant structural and categorical information. As a result, reconstructed regions may appear visually plausible while deviating from the true geographical context, leading to subtle but critical semantic inconsistencies. When deployed in downstream applications, such as semantic segmentation, these inconsistencies accumulate and translate into degraded representational reliability. This misalignment between low-level restoration objectives and high-level interpretative requirements limits the practical utility of current CR methods.

To address these challenges, we propose Geo-Anchored Cloud Removal (GACR), a unified framework built upon Observation-Anchored Residual Flow (OAR-Flow). Instead of initiating generation from pure noise, OAR-Flow starts from the cloudy observation and models CR as a physically grounded residual inversion process. By initializing the generative trajectory with structured perturbations around the observed image, the model adapts its behavior according to cloud opacity. In thin-cloud regions, where surface signals remain partially observable, the observation dominates the dynamics, guiding the model to perform physical inversion and preserve subtle yet authentic surface cues. In contrast,



(a) PSNR and mIoU comparison. **(b)** Downstream performance. **(c)** OAR-Flow and GCPA.

Fig. 1: (a) Comparison with existing methods in terms of PSNR and mIoU on Vaihingen-CR-Thick. (b) Performance across 12 downstream tasks, where the outermost ring denotes the upper bound. (c) GACR reconstructs cloud-free imagery from cloudy observations via OAR-Flow, while GCPA constrains the generative process within a geo-contextually consistent semantic manifold.

under thick cloud coverage where information is largely obscured, residual perturbations provide generative flexibility, enabling controlled semantic completion. This observation-anchored mechanism shortens the generative trajectory and suppresses unnecessary stochastic exploration, ensuring that reconstructed structures remain spatially aligned with the original observation while avoiding geographically implausible artifacts.

While physical anchoring stabilizes the generative dynamics, visual fidelity alone remains insufficient to guarantee semantic reliability for downstream interpretation. To bridge the gap between low-level restoration and high-level semantic preservation, we introduce Geo-Contextual Prior Alignment (GCPA), which leverages representational priors from a pretrained Vision Foundation Model (VFM) to guide cloud removal. Rather than optimizing solely for pixel-level similarity, GCPA aligns dense representations within a VFM-induced semantic manifold, encouraging reconstructed regions to remain coherent with their surrounding geographical context. This alignment is enforced through the proposed Geo-Contextual Integrity Loss (GCI Loss), which explicitly regularizes the generative process to preserve task-relevant structural patterns and category-specific semantic signatures. By jointly integrating OAR-Flow and GCPA, our framework produces cloud-free reconstructions that maintain both visual fidelity and semantic integrity. Extensive experiments across six CR datasets and twelve downstream tasks demonstrate that GACR consistently improves both reconstruction quality and downstream interpretation accuracy, with PSNR gains reaching 3.3 dB, semantic segmentation improvements of 3.1 mIoU, and approximately 5× faster convergence, effectively mitigating semantic distortion, as shown in Fig. 1.

The main contributions of this work are:

- We formulate cloud removal as an interpretation-oriented generative inversion problem and propose OAR-Flow, which grounds generation in the observed cloudy image to achieve stable and faithful reconstruction.
- We introduce GCPA, which constrains reconstruction within a semantic manifold induced by a VFM to preserve task-relevant structural patterns and category-specific information.
- Extensive experiments across six benchmarks and twelve downstream tasks demonstrate that GACR consistently improves reconstruction quality while yielding higher accuracy in downstream interpretation tasks.

2 Related Work

2.1 Cloud Removal Method

CR is a fundamental preprocessing step in Earth observation, aiming to recover cloud-free imagery for reliable surface analysis. With deep learning, data-driven CR methods have greatly improved visual quality and adaptability, and can be broadly grouped into residual-prediction-based and generation-based methods. The former estimates cloud residuals or clean reflectance using CNNs [22, 50], Transformers [18,44], or Mamba architectures [13,14,25,34]. However, they commonly rely on the cloud-as-residual assumption, which oversimplifies complex atmospheric scattering and limits recovery in regions severely occluded by thick clouds. Generation-based models instead synthesize cloud-free images through adversarial or diffusion mechanisms. Early GAN-based methods [4, 9, 12, 31, 39] improved perceptual realism but often suffered from instability and artifacts, whereas recent diffusion-based approaches [27, 37, 42, 51] achieve better fidelity and convergence. Nevertheless, most CR models are still mainly optimized for pixel-level similarity, offering limited semantic constraints for downstream interpretation tasks [5, 44]. In contrast, our framework introduces OAR-Flow and GCPA to jointly promote physically grounded reconstruction and semantic integrity.

2.2 Diffusion and Flow Model

Diffusion models have recently shown remarkable success in image generation [10, 17, 43, 45], evolving from DDPM [16, 33] and DDIM [40] to score-based SDE/ODE formulations [19, 41]. Through iterative denoising of Gaussian noise, they achieve higher fidelity and diversity than GAN-based [11] and VAE-based [21] paradigms. Early studies mainly used U-Net architectures [42, 51], while recent Diffusion Transformers (DiT) [35] improve scalability. However, many transformer-based diffusion models work in latent space [30, 35, 36], which limits pixel-level restoration. HDiT [6] therefore enables direct pixel-space reconstruction, with MRDM [29] and EMRDM [27] further adapting diffusion to CR scenarios. Nevertheless, diffusion models often require long stochastic sampling trajectories and high computational cost [16]. Flow matching methods [1,2,24,26]

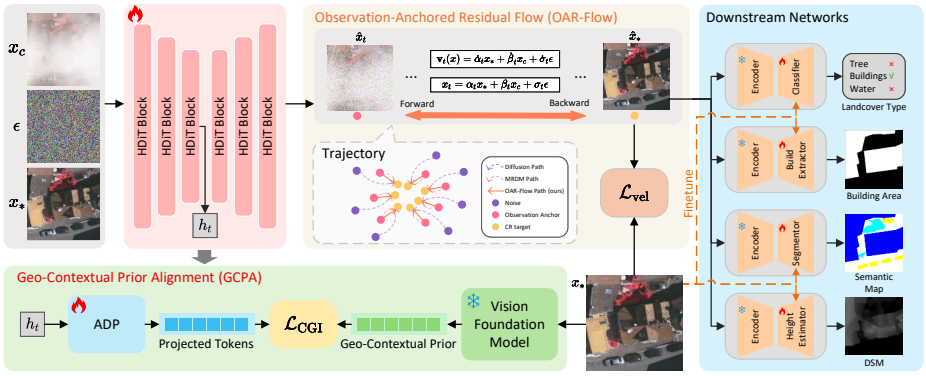


Fig. 2: Overview of the proposed GACR framework. (1) OAR-Flow reconstructs cloud-free imagery from cloudy observations via an observation-anchored residual trajectory, replacing pure noise initialization with a physically grounded anchor and enabling stable deterministic flow dynamics supervised by $\mathcal{L}_{vel}$. (2) GCPA leverages a pretrained Vision Foundation Model to extract geo-contextual representations from clean images, enforcing semantic consistency through the $\mathcal{L}_{GCI}$ and constraining the generative process within a coherent feature manifold. (3) Downstream networks trained on cloud-free data are used to evaluate the semantic fidelity and interpretation reliability of the reconstructed CR outputs.

reformulate diffusion as deterministic probability flow ODEs for more efficient trajectory learning. Different from existing flow-based CR methods, OAR-Flow introduces an observation-conditioned residual formulation that grounds generative dynamics in the cloudy input rather than unconditional noise.

3 Methodology

This section presents the proposed GACR framework illustrated in Fig. 2. We start by reviewing the theoretical background in Sec. 3.1. We then introduce OAR-Flow in Sec. 3.2, followed by GCPA in Sec. 3.3. Finally, we describe the downstream evaluation protocols in Sec. 3.4.

3.1 Preliminaries

A diffusion process with an initial distribution $p_0(x)$ can be described by the following stochastic differential equation (SDE) [7, 28, 41] for $t \in [0, T]$:

$$dx = f(x,t)\,dt + g(t)\,d\mathbf{w}, \quad x(0) \sim p_0(x), \tag{1}$$

where $f(\cdot,\cdot)$ and $g(\cdot)$ denote the drift and diffusion coefficients, and $\mathbf{w}$ is a standard Wiener process. The corresponding reverse-time SDE is given by:

$$dx = \big[f(x,t) - g(t)^2 \nabla_x \log p_t(x)\big]dt + g(t)\,d\hat{\mathbf{w}}, \tag{2}$$

where $\hat{\mathbf{w}}$ is the reverse-time Wiener process and $\nabla_x \log p_t(x)$ denotes the score function of the marginal distribution $p_t(x)$. In image restoration problems, the objective is to model the conditional distribution between degraded and clean images by constructing a continuous trajectory that links the two states.

To better characterize structured degradation, several works introduced modified diffusion dynamics that interpolate between a reference state and the target distribution via mean-reverting formulations [27, 29]. A representative forward process can be written as

$$dx = \theta_t(\mu - x)\,dt + \sigma_t\,d\mathbf{w}, \tag{3}$$

where μ denotes a reference state and θ_t, σ_t control the drift strength and stochastic noise intensity, respectively. This formulation provides an interpretable trajectory that gradually transports samples toward a designated state, offering a structured alternative to purely noise-driven diffusion. The corresponding reverse process can be expressed in either SDE or ODE form depending on the parameterization [27, 29]. Such formulations provide a flexible basis for designing conditional generative dynamics, which we further specialize for CR in the following section.

3.2 Observation-Anchored Residual Flow

Forward Process. Unlike unconditional generation, CR aims to recover the clean surface image x_* from a structured degradation x_c (cloudy observation). In remote sensing imagery, cloud coverage is not random noise but a spatially varying atmospheric scattering layer that partially or fully obscures surface reflectance. To explicitly model this conditional relationship, OAR-Flow constructs a continuous trajectory that transports samples between the clean state x_* and the cloudy observation x_c under structured perturbation, as illustrated in Fig. 2.

We define the forward interpolant as:

$$x_t = \alpha_t x_* + \beta_t x_c + \sigma_t \epsilon, \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, I)$ denotes Gaussian noise, and α_t, β_t, and σ_t are time-dependent coefficients satisfying:

$$\begin{aligned} &\alpha_0 = 1, \quad \beta_0 = 0, \quad \sigma_0 = 0, \\ &\alpha_T = 0, \quad \beta_T > 0. \end{aligned} \tag{5}$$

Here, x_c acts as an observation anchor. In thin-cloud regions, where surface signals remain partially observable, the contribution of x_c preserves low-frequency and structural cues. In thick-cloud regions, the stochastic component ϵ provides generative flexibility for semantic completion. This formulation explicitly reflects the opacity-aware characteristics of cloud degradation.

In practice, we adopt a simple linear schedule:

$$\alpha_t = 1 - t, \quad \beta_t = \rho t, \quad \sigma_t = t, \tag{6}$$

where ρ controls the strength of the observation anchor relative to stochastic perturbations, ensuring that the trajectory remains grounded in the cloudy observation while retaining flexibility in severely occluded regions.

Backward Process. The marginal distribution $p_t(x)$ induced by Eq. (4) satisfies the transport equation [1, 30]:

$$\partial_t p_t(x) + \nabla_x \cdot \big(\mathbf{v}_t(x) p_t(x)\big) = 0, \tag{7}$$

where $\mathbf{v}_t(x)$ denotes the velocity field of the deterministic probability flow.

Following the flow matching formulation, the ideal velocity field can be expressed as:

$$v_t(x) = \mathbb{E}[\dot{\alpha}_t x_* + \dot{\beta}_t x_c + \dot{\sigma}_t \epsilon \mid x_t = x], \tag{8}$$

which explicitly decomposes the dynamics into three components: clean target guidance, observation anchoring, and stochastic perturbation. This residual decomposition differs from purely noise-driven flows by incorporating structured observational information into the trajectory. The correspondence between the marginal distribution $p_t(x)$ and the velocity formulation in Eq. (8) follows from the transport equation in Eq. (7), and the detailed derivation is provided in Appendix A.

We parameterize the velocity field using a neural network $\mathbf{u}_t$:

$$\mathbf{u}_t(x) = \mathrm{Net}_\theta(x_t, t, x_c), \tag{9}$$

where x_c is provided as a condition to preserve spatial alignment with the observed cloudy image.

During inference, the clean estimate is obtained by integrating the deterministic flow from $t = T$ to $t = 0$:

$$\hat{x}_{t-1} = \hat{x}_t + \Delta t\, \mathbf{u}_t(\hat{x}_t), \tag{10}$$

yielding the final reconstruction $\hat{x}_* = \hat{x}_0$. Because the trajectory remains anchored to x_c throughout integration, the reconstructed surface structures are spatially consistent with the original observation, effectively reducing geographically implausible artifacts.

3.3 Geo-Contextual Prior Alignment

To jointly ensure reconstruction fidelity and semantic integrity in CR, we optimize OAR-Flow under two complementary objectives: a velocity matching loss $\mathcal{L}_{\text{vel}}$ in pixel space and a geo-contextual consistency constraint loss $\mathcal{L}_{\text{GCI}}$ in the representation space.

Velocity Matching Loss. The primary supervision for OAR-Flow is to match the analytical velocity defined in Eq. (8). The network $\mathbf{u}_t$ is trained to approximate the target velocity field:

$$\begin{aligned} \mathcal{L}_{\text{vel}} &= \mathbb{E}\Big[\|\mathbf{u}_t(x_t) - \mathbf{v}_t(x_t)\|^2\Big] \\ &= \mathbb{E}_{x_*,\epsilon,x_c}\Big[\Big\|\mathbf{u}_t(x_t) - \dot{\alpha}_t x_* - \dot{\beta}_t x_c - \dot{\sigma}_t \epsilon\Big\|^2\Big]. \end{aligned} \tag{11}$$

This objective ensures that the learned flow follows the observation-anchored residual trajectory defined in OAR-Flow. The network is implemented using an HDiT-based backbone [6], receiving the current state x_t, timestep t, and the cloudy observation x_c as condition, and predicting the deterministic velocity toward the clean state.

Geo-Contextual Integrity Loss. While velocity supervision guarantees pixel-level reconstruction consistency, it does not explicitly enforce preservation of land-cover semantics. In remote sensing imagery, geographical context exhibits strong structural continuity (e.g., forests form contiguous regions and urban layouts follow spatial regularity). To maintain such geo-contextual coherence, we introduce GCPA, implemented through the proposed GCI Loss.

Let $f_{\text{vfm}}(\cdot)$ denote a pretrained VFM encoder. The geo-contextual prior of the clear image is defined as

$$z_* = f_{\text{vfm}}(x_*) \in \mathbb{R}^{B\times L\times D}. \tag{12}$$

We extract the intermediate feature $h_t \in \mathbb{R}^{B\times C\times H\times W}$ from the bottleneck of the HDiT backbone and map it into the same representation space via an Adaptive Projector (ADP):

$$\begin{aligned} z_t &= \text{ADP}(h_t) \\ &= \text{MLP}\big(\text{RE}(\text{AP}(h_t))\big), \end{aligned} \tag{13}$$

where $\text{RE}(\cdot)$ and $\text{AP}(\cdot)$ denote rearrangement and adaptive pooling operations, respectively.

The GCI Loss is defined as a patch-wise cosine similarity:

$$\mathcal{L}_{\text{GCI}} = -\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\frac{\langle z_*^{[n]}, z_t^{[n]}\rangle}{\|z_*^{[n]}\|_2\,\|z_t^{[n]}\|_2}\right], \tag{14}$$

where n indexes spatial tokens. By aligning reconstructed features with the VFM-induced semantic manifold, this loss constrains the generative process to remain consistent with large-scale geographical structures, thereby reducing semantic drift in heavily occluded regions.

Unified Objective. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{vel}} + \lambda\mathcal{L}_{\text{GCI}}, \tag{15}$$

where λ balances reconstruction fidelity and geo-contextual integrity. Ablation analysis of the objective components is provided in Section 4.4.

3.4 Downstream Networks

We evaluate GACR on four remote sensing tasks: land-cover classification, building extraction, semantic segmentation, and height estimation as shown in Fig. 2. These tasks assess whether reconstructed CR results preserve task-relevant semantic and structural information.

We adopt a pretrained DINOv3 backbone [38] as a unified visual encoder for downstream evaluation. For each task, the backbone is frozen and paired with a lightweight decoder trained on clear images to establish reliable reference performance. Notably, the VFM used for GCPA and the downstream encoders are independent, ensuring that downstream evaluation does not share parameters or supervision signals with the upstream semantic guidance. The architectural details of the downstream networks are presented in Appendix B.

4 Experiments and Discussion

4.1 Implementation Details

Dataset. To enable a unified evaluation of both reconstruction fidelity and downstream performance, we construct a comprehensive benchmark for joint assessment. Specifically, we employ six CR datasets, including two publicly available datasets: CUHKCR-EXT-GZ and CUHKCR-EXT-CS [44], along with four synthetic cloud datasets constructed by ourselves: Potsdam-CR-thin, Potsdam-CR-thick, Vaihingen-CR-thin, and Vaihingen-CR-thick. These datasets cover both real and simulated cloud scenarios, encompassing thin and thick cloud conditions to reflect varying levels of opacity and structural occlusion.

Downstream Tasks. CUHKCR-EXT-GZ and CUHKCR-EXT-CS support land-cover classification (CLS-1 and CLS-2) and building extraction (BLD-1 and BLD-2), while the four synthetic datasets include semantic segmentation (SEG-1 to SEG-4) and height estimation (HE-1 to HE-4). These tasks evaluate whether reconstructed CR outputs preserve task-relevant semantic structures and spatial consistency. Further details on dataset construction and the cloud synthesis process are provided in Appendix C.

Model Settings. For GCPA, we employ two pretrained DINOv3 variants as the VFM module: ViT-L/16-SAT-300M and ViT-L/16-LVD-1689M. Model configurations are denoted as GACR-SAT/p or GACR-LVD/p, where "SAT" and "LVD" indicate the corresponding VFM pretraining variants, and "/p" specifies the patch size used in OAR-Flow. In our experiments, the upstream VFM for geo-contextual guidance and the downstream encoder for evaluation are initialized from different pretrained weights, ensuring no parameter sharing between guidance and evaluation stages. This separation avoids potential bias arising from shared representation priors and enables a fair assessment of downstream improvements. To further verify robustness with respect to backbone selection, we additionally conduct experiments using heterogeneous downstream encoders, with detailed results reported in Appendix D.

Evaluation Metrics. For CR evaluation, we adopt PSNR and SSIM to measure reconstruction fidelity from complementary perspectives of distortion, perceptual similarity, and error magnitude. For downstream evaluation, we report accuracy (Acc.) for CLS, IoU for BLD, mIoU for SEG, and RMSE for HE. More implementation details of metrics and training settings are provided in Appendix D.

Table 1: Quantitative comparison across six CR datasets. The best and second-best scores are marked in **bold** and underline, respectively.

Model	CUHKCR-EXT-GZ		CUHKCR-EXT-CS		Potsdam-CR-thin		Potsdam-CR-thick		Vaihingen-CR-thin		Vaihingen-CR-thick	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
MPRNet [32]	23.454	0.712	23.365	0.695	28.349	0.960	26.418	0.902	31.209	0.978	27.805	0.935
Restormer [46]	25.839	0.743	23.632	0.710	31.413	0.970	28.831	0.922	31.210	0.970	28.867	0.922
AST [47]	25.482	0.735	23.365	0.695	28.924	0.957	25.886	0.890	29.877	0.973	26.471	0.916
MambaIR [14]	25.626	0.733	23.445	0.704	30.881	0.967	27.027	0.903	32.764	0.983	29.852	0.945
DFCFormer [44]	25.816	0.746	23.876	0.711	30.836	0.969	28.196	0.917	33.342	0.985	30.396	0.951
EMRDM [27]	25.862	**0.747**	23.736	0.712	30.335	0.972	27.199	0.923	33.620	0.988	28.979	0.951
GACR-SAT/2	25.964	0.736	24.230	0.709	33.141	0.975	30.578	0.934	36.293	0.990	33.018	0.964
GACR-LVD/2	25.899	0.735	24.220	0.707	32.880	0.975	30.308	0.932	35.749	0.989	32.656	0.963
GACR-SAT/1	**26.100**	0.744	**24.354**	0.713	**33.642**	**0.976**	**31.049**	**0.938**	**36.918**	**0.991**	**34.048**	**0.970**
GACR-LVD/1	26.059	0.744	24.331	**0.714**	33.305	0.975	30.586	0.935	36.590	0.991	33.808	0.969

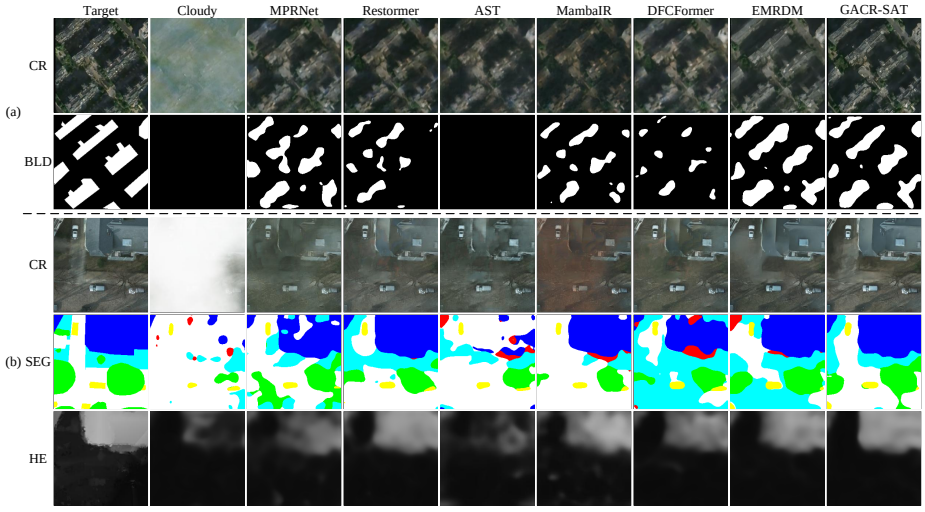


Fig. 3: Visualization of CR and downstream results. (a) The CR results on CUHKCR-EXT-GZ and the corresponding BLD results. (b) The CR results on Potsdam-CR-thick and the corresponding SEG and HE results.

4.2 CR Evaluation

This section evaluates the reconstruction fidelity of CR results. Quantitative comparisons are reported in Tab. 1. On the CUHKCR-EXT datasets, GACR achieves highly competitive performance across most metrics. Except for a slightly lower SSIM on CUHKCR-EXT-GZ, GACR consistently surpasses existing approaches in terms of PSNR and RMSE, indicating a favorable balance between pixel-level accuracy and perceptual consistency. For example, on CUHKCR-EXT-CS, GACR-SAT/1 achieves a PSNR of 24.354, improving upon the strongest baseline by approximately 0.5 dB. Notably, the improvements are not limited to a single metric but are consistently reflected across multiple evaluation metrics, demonstrating the robustness of the proposed reconstruction strategy.

Table 2: Quantitative comparison of different tasks on downstream networks with ViT-L/16-LVD-1689M weight, including classification (CLS), building extraction (BLD), semantic segmentation (SEG), and height estimation (HE). The best and second-best results are highlighted in **bold** and underline, respectively.

Model	CLS-1 Acc. ↑	CLS-2 Acc. ↑	BLD-1 IoU ↑	BLD-2 IoU ↑	SEG-1 mIoU ↑	SEG-2 mIoU ↑	SEG-3 mIoU ↑	SEG-4 mIoU ↑	HE-1 RMSE ↓	HE-2 RMSE ↓	HE-3 RMSE ↓	HE-4 RMSE ↓
Upper Bound	0.882	0.916	0.718	0.762	0.733	0.733	0.755	0.755	1.868	1.868	1.477	1.477
Lower Bound	0.746	0.739	0.669	0.596	0.657	0.490	0.677	0.550	2.144	2.703	1.672	2.095
End-to-End	0.833	0.876	0.691	0.698	0.709	0.643	0.737	0.684	2.000	2.245	1.583	1.830
MPRNet [32]	0.809	0.752	0.653	0.654	0.707	0.615	0.741	0.671	1.987	2.412	1.549	1.757
Restormer [46]	0.803	0.744	0.655	0.660	0.710	0.650	0.718	0.675	1.913	2.228	1.520	1.681
AST [47]	0.813	0.767	0.668	0.651	0.697	0.580	0.736	0.632	2.019	2.492	1.585	1.885
MambaIR [14]	0.759	0.778	0.660	0.648	0.713	0.620	0.736	0.671	1.948	2.350	1.552	1.697
DFCFormer [44]	0.763	0.771	0.659	0.657	0.713	0.630	0.722	0.671	1.948	2.317	1.526	1.687
EMRDM [27]	0.776	0.726	0.703	0.696	0.722	0.668	0.747	0.692	1.902	2.133	1.512	1.629
GACR-SAT/2	**0.833**	**0.781**	**0.704**	**0.710**	**0.727**	**0.699**	**0.750**	**0.737**	1.891	**2.014**	**1.482**	**1.554**
GACR-LVD/2	0.806	0.768	0.702	0.704	0.727	0.695	0.748	0.728	**1.889**	2.025	1.490	1.575

On the four synthetic datasets, GACR demonstrates clear advantages under both thin- and thick-cloud conditions. In thin-cloud scenarios, where surface structures remain partially observable, the observation-anchored formulation effectively preserves fine-grained textures and low-frequency spatial continuity, preventing unnecessary alterations to already reliable regions. This leads to the highest PSNR of 33.642 dB on Potsdam-CR-thin and 36.918 dB on Vaihingen-CR-thin with GACR-SAT/1. In thick-cloud settings, where large areas are severely occluded and lack direct visual cues, the residual generative component provides controlled semantic completion guided by the anchored trajectory. This mechanism enables the model to recover plausible yet geographically consistent structures, resulting in consistent improvements over prior methods across multiple quantitative metrics. The performance gains under heavy occlusion further validate the effectiveness of modeling CR as a physically grounded residual inversion process.

We further evaluate two patch-size configurations to examine the trade-off between reconstruction granularity and computational cost. Results indicate that GACR-SAT/2 already surpasses existing methods by a clear margin, confirming that the proposed framework remains effective even under coarser spatial partitioning. A smaller patch size (GACR-SAT/1) further enhances reconstruction fidelity by enabling finer spatial modeling and more detailed residual refinement. However, considering computational efficiency and fairness in comparison with existing baselines, we adopt $p = 2$ in subsequent experiments unless otherwise specified. Qualitative comparisons are shown in Fig. 3, where GACR produces sharper structural boundaries and fewer texture inconsistencies. Additional visualizations are provided in Appendix E to further illustrate the stability of the reconstruction results across diverse cloud conditions.

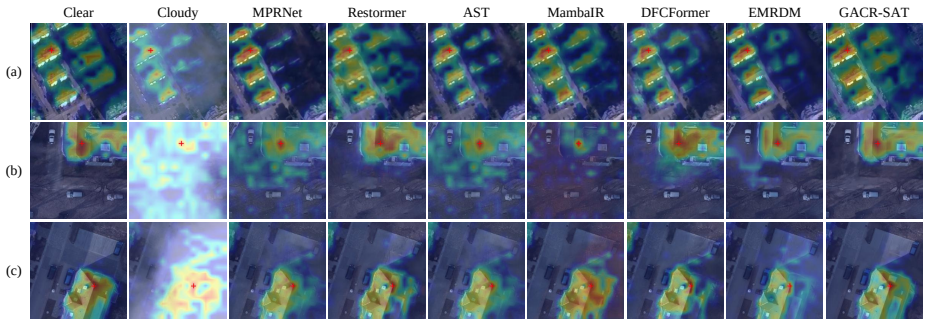


Fig. 4: Heatmaps of different CR obtained from the pretrained DINOv3 ViT-L/16-LVD-1689M. Regions with higher intensity indicate stronger similarity to the locations marked by red crosses.

4.3 Downstream Evaluation

This section evaluates downstream performance to examine whether GACR preserves task-relevant semantic structures beyond pixel-level fidelity. The quantitative results in Tab. 2 are obtained using a DINOv3-based downstream encoder initialized with the ViT-L/16-LVD-1689M weights. In the table, the upper bound denotes testing directly on clear images, the lower bound represents performance without CR preprocessing, and end-to-end refers to training and testing downstream models directly on cloudy images. Results obtained using the ViT-L/16-SAT-300M weights are provided in Appendix E.

Across all downstream tasks, GACR consistently achieves superior downstream performance compared with existing CR methods. In particular, GACR-SAT/2 attains the highest classification accuracies of 0.833 and 0.781 on CLS-1 and CLS-2, exceeding the strongest baseline (AST) by 2.0% and 1.4%, respectively. For height estimation, GACR-SAT/2 achieves the lowest RMSE values of 1.891 and 1.482 on HE-1 and HE-3, outperforming the second-best method by 0.1-0.2. These improvements indicate that GACR better preserves semantic consistency and spatial structure critical for downstream recognition.

For the CLS task, the performance gap between different CR methods is relatively small. We observe that the end-to-end approach yields results close to the upper bound and, in some cases, performs comparably to inputs preprocessed via CR. This phenomenon can be attributed to the global nature of classification, where coarse semantic representations are less sensitive to localized cloud occlusion. The t-SNE analysis in Appendix E further supports this observation: cloudy inputs already form distinguishable clusters similar to cloud-free samples, implying that CR only provides limited additional separability.

In contrast, for dense prediction tasks such as segmentation and height estimation, CR leads to substantial downstream improvements, with GACR consistently delivering greater gains than competing methods. As shown in Fig. 3, different CR approaches enhance land-cover discriminability to varying extents, yet GACR produces more distinct and semantically coherent object boundaries.

To further analyze this behavior, we present feature activation maps in Fig. 4. For example, in row (a), GACR enables clearer separation of cloud-affected buildings on the right side of the image, indicating improved semantic localization. Moreover, Fig. 5 visualizes feature distance distributions across methods. While cloudy inputs exhibit a noticeable distribution shift relative to clear references, GACR more effectively aligns the feature distribution with that of the cloud-free references. This observation is consistent with the geo-contextual alignment mechanism, which constrains reconstruction within a semantically coherent feature manifold.

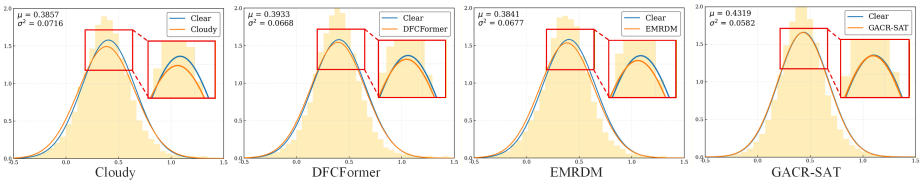


Fig. 5: Feature distance distribution comparison between CR result and corresponding cloud-free reference.

4.4 Ablation Study

Convergence Speed. Fig. 6 presents the PSNR progression over training steps for EMRDM, OAR-Flow without GCPA, and the complete GACR model. Compared with EMRDM, OAR-Flow converges substantially faster, reaching the high-PSNR regime with significantly fewer iterations. Specifically, OAR-Flow achieves a comparable PSNR level using approximately one-third of the training steps required by EMRDM, corresponding to about a 3× acceleration in convergence. This improvement highlights the optimization efficiency introduced by the observation-anchored residual trajectory. Furthermore, incorporating GCPA further accelerates convergence. The complete GACR not only attains higher PSNR but also converges approximately 5× faster than EMRDM.

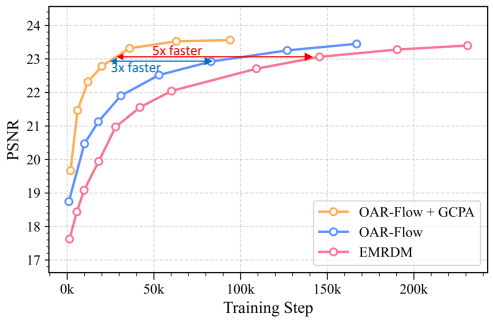


Fig. 6: The introduction of OAR-Flow and GCPA significantly accelerates the convergence of training.

Effectiveness of Model Choice. To investigate the effectiveness of each module, we perform an ablation study on the CUHKCR-EXT-CS dataset, as reported in Tab. 3. Four configurations are compared to disentangle the effects

Table 3: Comparison of different model settings on CUHKCR-EXT-CS.

Method	PSNR ↑	SSIM ↑	CLS ↑	BLD ↑	GFLOPs
DiT + MRDM	<u>24.037</u>	0.707	0.670	0.693	697.20
HDiT + MRDM	23.736	**0.712**	<u>0.726</u>	0.696	166.72
HDiT + OAR-Flow	23.986	0.698	0.690	<u>0.700</u>	56.05
HDiT + OAR-Flow + GCPA	**24.230**	<u>0.709</u>	**0.781**	**0.710**	56.05

of generative formulation and geo-contextual alignment. The diffusion-based baseline corresponding to EMRDM serves as a reference, while the complete GACR represents the final configuration. Although DiT-based modeling produces competitive visual results, it incurs substantially higher computational cost (GFLOPs = 697.20). Replacing diffusion dynamics with OAR-Flow consistently improves reconstruction metrics while maintaining efficiency. Moreover, incorporating GCPA further enhances both CR quality and downstream performance, indicating that geo-contextual alignment effectively mitigates semantic drift and strengthens interpretation reliability.

Effectiveness of λ and patch size p. We further conduct ablation studies on key hyperparameters in Tab. 4, including the balancing coefficient λ for GCI and the patch size p. When $\lambda = 0.5$, the model achieves the best trade-off between reconstruction fidelity and downstream performance. This indicates that moderate regularization enhances semantic consistency without over-constraining pixel-level refinement, whereas overly small or large λ values degrade either reconstruction quality or downstream metrics. Regarding patch size, smaller values lead to improved visual quality due to finer spatial modeling; however, setting $p = 1$ significantly increases computational cost in terms of FLOPs. Therefore, $p = 2$ is adopted as the default configuration in our experiments to balance reconstruction quality and computational efficiency while maintaining fair comparison with existing methods.

Table 4: Hyperparameter analysis on Vaihingen-CR-Thick.

	PSNR ↑	SSIM ↑	SEG ↑	HE ↓	GFLOPs
λ	$p = 2$				
0.2	**33.073**	**0.965**	0.728	1.578	-
0.5	33.018	0.964	**0.737**	**1.554**	-
1.0	32.838	0.964	0.730	1.563	-
2.0	32.720	0.963	0.729	1.576	-
5.0	32.615	0.962	0.727	1.596	-
p	$\lambda = 0.5$				
4	30.303	0.949	0.693	1.616	14.15
2	33.018	0.964	**0.737**	1.554	56.05
1	**33.799**	**0.969**	0.720	**1.548**	223.60

5 Conclusion

In this paper, we present GACR, an interpretation-oriented framework that jointly improves reconstruction fidelity and downstream reliability. With OAR-Flow, CR is reformulated as a physically grounded residual inversion process

anchored to cloudy observations, reducing geographically implausible artifacts. Meanwhile, GCPA constrains reconstruction within a VFM-induced semantic manifold to preserve task-relevant structures and category-specific information. Extensive evaluations on six datasets and twelve downstream tasks show that GACR consistently enhances both reconstruction quality and task accuracy across diverse cloud conditions and remote sensing scenarios. These results suggest that coupling observation-anchored physical priors with semantic constraints enables CR to move beyond visual enhancement toward a reliable, semantics-preserving foundation for Earth observation.

Acknowledgements

This work was supported by the Guangdong Science and Technology Department (Grant No. 2025A0505000062) and the Hong Kong Research Grants Council through the General Research Fund (Grant No. 17617024).

References

1. Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E.: Stochastic interpolants: A unifying framework for flows and diffusions. arXiv preprint arXiv:2303.08797 (2023)
2. Albergo, M.S., Vanden-Eijnden, E.: Building normalizing flows with stochastic interpolants. In: The Eleventh International Conference on Learning Representations (2023)
3. Astruc, G., Gonthier, N., Mallet, C., Landrieu, L.: Omnisat: Self-supervised modality fusion for earth observation. In: European Conference on Computer Vision. pp. 409–427. Springer (2024)
4. Bermudez, J.D., Happ, P.N., Oliveira, D.A.B., Feitosa, R.Q.: SAR to optical image synthesis for cloud removal with generative adversarial networks. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **4**, 5–11 (2018)
5. Chen, I., Chen, W.T., Liu, Y.W., Chiang, Y.C., Kuo, S.Y., Yang, M.H., et al.: Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17969–17979 (2025)
6. Crowson, K., Baumann, S.A., Birch, A., Abraham, T.M., Kaplan, D.Z., Shippole, E.: Scalable high-resolution pixel-space image synthesis with hourglass diffusion transformers. In: International Conference on Machine Learning (2024)
7. De Bortoli, V., Mathieu, E., Hutchinson, M., Thornton, J., Teh, Y.W., Doucet, A.: Riemannian score-based generative modelling. Advances in neural information processing systems **35**, 2406–2422 (2022)
8. Ebel, P., Xu, Y., Schmitt, M., Zhu, X.X.: SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–14 (2022)
9. Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., Kawaguchi, N.: Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 48–56 (2017)

10. Feng, C., Chen, Z., Holynski, A., Efros, A.A., Owens, A.: GPS as a control signal for image generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2766–2778 (2025)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
12. Grohnfeldt, C., Schmitt, M., Zhu, X.: A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from Sentinel-2 images. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. pp. 1726–1729. IEEE (2018)
13. Gu, Y., Meng, Y., Ji, J., Sun, X.: ACL: Activating capability of linear attention for image restoration. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17913–17923 (2025)
14. Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., Xia, S.T.: MambaIR: A simple baseline for image restoration with state-space model. In: European Conference on Computer Vision. pp. 222–241. Springer (2024)
15. Guo, X., Lao, J., Dang, B., Zhang, Y., Yu, L., Ru, L., Zhong, L., Huang, Z., Wu, K., Hu, D., et al.: Skysense: A multi-modal remote sensing foundation model towards universal interpretation for earth observation imagery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27672–27683 (2024)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Jeong, J., Han, S., Kim, J., Kim, S.J.: Latent space super-resolution for higher-resolution image generation with diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2355–2365 (2025)
18. Jin, X., He, J., Xiao, Y., Lihe, Z., Liao, X., Li, J., Yuan, Q.: RFE-VCR: Reference-enhanced transformer for remote sensing video cloud removal. ISPRS Journal of Photogrammetry and Remote Sensing **214**, 179–192 (2024)
19. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in neural information processing systems **35**, 26565–26577 (2022)
20. King, M.D., Platnick, S., Menzel, W.P., Ackerman, S.A., Hubanks, P.A.: Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites. IEEE transactions on geoscience and remote sensing **51**(7), 3826–3852 (2013)
21. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
22. Li, W., Li, Y., Chan, J.C.W.: Thick cloud removal with optical and SAR imagery via convolutional-mapping-deconvolutional network. IEEE Transactions on Geoscience and Remote Sensing **58**(4), 2865–2879 (2019)
23. Li, Z., Shen, H., Weng, Q., Zhang, Y., Dou, P., Zhang, L.: Cloud and cloud shadow detection for optical satellite imagery: Features, algorithms, validation, and prospects. ISPRS Journal of Photogrammetry and Remote Sensing **188**, 89–108 (2022)
24. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
25. Liu, J., Pan, B., Shi, Z.: CR-Famba: A frequency-domain assisted mamba for thin cloud removal in optical remote sensing imagery. IEEE Transactions on Multimedia (2025)

26. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023)
27. Liu, Y., Li, W., Guan, J., Zhou, S., Zhang, Y.: Effective cloud removal for remote sensing images by an improved mean-reverting denoising model with elucidated design space. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17851–17861 (2025)
28. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps. Advances in neural information processing systems **35**, 5775–5787 (2022)
29. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. International Conference on Machine Learning (2023)
30. Ma, N., Goldstein, M., Albergo, M.S., Boffi, N.M., Vanden-Eijnden, E., Xie, S.: SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In: European Conference on Computer Vision. pp. 23–40. Springer (2024)
31. Ma, X., Huang, Y., Zhang, X., Pun, M.O., Huang, B.: Cloud-EGAN: Rethinking cyclegan from a feature enhancement perspective for cloud removal by combining cnn and transformer. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **16**, 4999–5012 (2023)
32. Mehri, A., Ardakani, P.B., Sappa, A.D.: 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)
33. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International conference on machine learning. pp. 8162–8171. PMLR (2021)
34. Pan, L., Song, X., Xie, F., Zhang, X., Ji, H., Shi, Z.: M3-CR: Multi-scale multi-branch Mamba for SAR-assisted optical image thick cloud removal. IEEE Transactions on Geoscience and Remote Sensing (2025)
35. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
36. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
37. Silva, L.H.F.P., Mari, J.F., Escarpinati, M.C., Backes, A.R.: Cloud removal with compact diffusion models: A residual block-based approach. Remote Sensing Applications: Society and Environment p. 101680 (2025)
38. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: DINOv3 (2025)
39. Singh, P., Komodakis, N.: Cloud-Gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks. In: IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium. pp. 1772–1775. IEEE (2018)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv:2010.02502 (October 2020)
41. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021)
42. Sui, J., Ma, Y., Yang, W., Zhang, X., Pun, M.O., Liu, J.: Diffusion enhancement for cloud removal in ultra-resolution remote sensing imagery. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)

43. Wang, C., Guo, L., Fu, Z., Yang, S., Cheng, H., Kot, A.C., Wen, B.: Reconciling stochastic and deterministic strategies for zero-shot image restoration using diffusion model in dual. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23207–23216 (2025)
44. Wang, Z., Ma, X., Pun, M.O.: 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 (2025)
45. Yang, H., Bulat, A., Hadji, I., Pham, H.X., Zhu, X., Tzimiropoulos, G., Martinez, B.: FAM diffusion: Frequency and attention modulation for high-resolution image generation with stable diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2459–2468 (2025)
46. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: 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)
47. Zhou, S., Chen, D., Pan, J., Shi, J., Yang, J.: 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 (2024)
48. Zhu, Q., Lao, J., Ji, D., Luo, J., Wu, K., Zhang, Y., Ru, L., Wang, J., Chen, J., Yang, M., et al.: Skysense-O: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14733–14744 (2025)
49. Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F.: Deep learning in remote sensing: A comprehensive review and list of resources. IEEE geoscience and remote sensing magazine **5**(4), 8–36 (2017)
50. Zi, Y., Xie, F., Zhang, N., Jiang, Z., Zhu, W., Zhang, H.: Thin cloud removal for multispectral remote sensing images using convolutional neural networks combined with an imaging model. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **14**, 3811–3823 (2021)
51. Zou, X., Li, K., Xing, J., Zhang, Y., Wang, S., Jin, L., Tao, P.: DiffCR: A fast conditional diffusion framework for cloud removal from optical satellite images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–14 (2024)