Revealing the Invisible: Latent Structure Modeling for Semantically Consistent Cloud Removal

Jingwei Xin[1], Kai Guo[1], Jie Li[2], Nannan Wang[1]*

[1]School of Communication Engineering, Xidian University
[2]School of Electronic Engineering, Xidian University
jwxin@xidian.edu.cn, gkxdu@stu.xidian.edu.cn, leejie@mail.xidian.edu.cn, nnwang@xidian.edu.cn

Abstract

Cloud removal (CR) in remote sensing imagery is a critical yet challenging task due to complex cloud patterns and diverse underlying ground structures. Despite recent progress in generative models such as diffusion models, CR remains limited by their inadequate capability to perceive and reconstruct structured information beneath cloud-covered areas. In this work, we propose a Visibility-guided Semantic Estimation and Reconstruction network for cloud removal (VISER-CR), which reformulates CR as a structure-guided completion problem. Specifically, VISER-CR explicitly models cloud interference via spatial masking, encouraging the model to reason beyond pixel-level appearance and enhance scene-level structural understanding. Moreover, to further improve the representation of structural information, we introduce Patch Saliency Encoding, a self-guided mechanism that implicitly models structural alignment among patches, significantly enhancing clustering consistency and semantic separability in the latent space. This adaptive mechanism guides the network to focus on learning and reconstructing structurally important regions, thereby reducing redundancy and improving overall cloud removal performance. Extensive experiments on multiple benchmark datasets demonstrate the superior effectiveness of our method.

Introduction

Cloud coverage poses a persistent obstacle in the field of remote sensing, as it blocks direct observation of the Earth's surface. According to long-term satellite statistics, nearly 67% of the Earth's surface is affected by cloud cover at any given time, leading to considerable data gaps across both spatial and temporal dimensions. These missing observations significantly degrades the quality and reliability of downstream tasks such as land cover classification (Carranza-García, García-Gutiérrez, and Riquelme 2019; Tong et al. 2020), vegetation monitoring (Rußwurm and Korner 2017; Turkoglu et al. 2021), and ground target detection (Azimi et al. 2018; Wang et al. 2019). As a result, cloud removal has become an essential preprocessing step to enhance the usability of remote sensing imagery. However, this task remains highly challenging, primarily due to the high variability of cloud formations and the inherent uncertainty in reconstructing obscured regions from limited visible information.

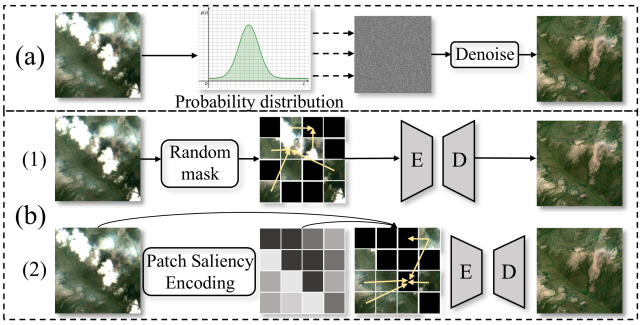

Figure 1: (a) A widely adopted probabilistic generative paradigm for cloud removal. (b) The proposed mask-guided reconstruction framework. Subfigures (1) and (2) depict variants employing random and adaptive masks, respectively. Yellow arrows indicate the flow of contextual information.

Generative adversarial networks (GANs) (Sarukkai et al. 2020; Ebel et al. 2022) were among the first models to tackle cloud removal by learning to synthesize plausible cloud-free images from multi-cloud inputs. These approaches improved visual realism but were often limited by unstable training and difficulty in preserving fine structural details. More recently, diffusion-based methods (Zou et al. 2024; Sui et al. 2024) have shown promising advancements in perceptual quality and training stability by modeling the progressive denoising process. However, despite their superior visual fidelity, these generative models frequently face challenges in maintaining structural fidelity and semantic consistency, especially under dense or irregular cloud coverage. This limitation stems from their reliance on learned statistical priors rather than leveraging visible cues present in the input images to guide the reconstruction process, as shown in Fig. 1(a). In addition, some studies have attempted to improve the cloud removal process by enhancing attention mechanisms (Huang et al. 2024) to better capture spatial dependencies. However, we argue that such designs are still insufficient for reliable context-aware reconstruction, as the inherent ambiguity and structural fragmentation in complex

*Corresponding author.

remote sensing imagery often disrupt semantic alignment.

To address this challenge, we redefine cloud removal (CR) as a structure-guided completion problem, where reasoning about visible regions is crucial for reconstructing occluded content. However, unlike traditional image reconstruction frameworks(Meraner et al. 2020; Ding, Zi, and Xie 2022), we design a model that is more aligned with the essence of the CR task. Specifically, we propose VISER-CR, a novel framework that leverages spatial masks to model realistic cloud interference(Fig. 1(b)). This framework utilizes partial visibility to optimize patch clustering in the latent space, achieving excellent semantic consistency. Additionally, it can utilize more diverse training sample forms beyond the input data (He et al. 2022), thereby accelerating convergence and enhancing its structural restoration capabilities.

However, traditional random occlusion strategies are still insufficient to fully utilize structural information. Inspired by recent research on self-guided mask autoencoders (SG-MAE) (Shin et al. 2024), which demonstrates that selectively retaining semantically and structurally important patches can significantly enhance latent representation learning performance, we believe that cloud removal models can also benefit from guided occlusion modeling. Existing random occlusion methods treat all spatial regions as equally important, leading to inefficient allocation of modeling capabilities toward easily predictable or redundant content while neglecting critical structural regions. To address this issue, we propose a Patch Saliency Encoding mechanism that leverages an implicitly captured spatial importance distribution, enabling the model to focus attention on structurally important regions. Unlike the binary-based object-centered information masking in SG-MAE, this mechanism demonstrates higher robustness when dealing with the complex and diverse features of objects in remote sensing images.

Our main contributions are summarized as follows:

- We propose a novel cloud removal framework, VISER-CR, by reformulating it as a structure-aware completion task. By aligning the training process with the physical characteristics of cloud interference, this design encourages the model to reason over semantically visible regions and enhances its ability to model latent scene structures beyond appearance-level cues.
- We design a Patch Saliency Encoding (PSE) mechanism that replaces random masking with a spatially guided occlusion strategy. By estimating region-level importance during training, PSE generates structured visibility patterns that simulate realistic cloud interference and concentrate reconstruction efforts on structurally informative regions. This improves latent feature separation and enhances restoration accuracy under challenging occlusion conditions.
- We conduct extensive experiments on several benchmark datasets, demonstrating that our approach consistently outperforms existing methods in terms of perceptual realism, structural accuracy, and generalization to complex cloud conditions.

Related Work

Cloud removal

Cloud removal (CR) remains a fundamental challenge in optical remote sensing, aiming to restore cloud-covered regions for improved data continuity and usability. Traditional approaches (Lin et al. 2012; Xu et al. 2015, 2019), such as temporal interpolation and spectral unmixing, provide interpretable but often inflexible solutions that struggle under varying cloud conditions. Recent deep learning methods (Goodfellow et al. 2014; Mirza and Osindero 2014; Ho, Jain, and Abbeel 2020) have surpassed these limitations by leveraging data-driven representations. Current learning-based CR approaches are mainly divided into mono-temporal (Bermudez et al. 2018; Grohnfeldt, Schmitt, and Zhu 2018; Li, Liu, and Li 2023) and multi-temporal (Li, Liu, and Li 2023; Ebel et al. 2022; Huang and Wu 2022) methods. Mono-temporal techniques reconstruct cloud-free images from single inputs using encoder-decoder architectures, often enhanced with attention mechanisms (Pan 2020), or transformer modules (Li, Liu, and Li 2023). While suitable when temporal data is scarce, these methods face difficulties recovering heavily occluded areas due to missing temporal cues. Diffusion models (Sui et al. 2024; Zou et al. 2024) have recently been introduced here, improving perceptual realism but typically requiring high computational cost during sampling. Multi-temporal methods exploit the redundancy in satellite time series by aggregating complementary observations at different timestamps. They utilize temporal alignment, fusion attention, or recurrent designs to recover missing content more effectively. Some also incorporate multi-modal data such as infrared (Li et al. 2021) or SAR (Meraner et al. 2020; Ebel et al. 2022) to tackle persistent cloud coverage. Nevertheless, these approaches rely on well-aligned, high-quality multi-temporal data, which is not always available in practice.

Masked Autoencoders

Masked Autoencoders (MAEs) (He et al. 2022) have recently shown remarkable performance in visual representation learning by reconstructing missing image patches from partially observed inputs. Originally proposed as a self-supervised pretraining framework for high-level tasks, MAEs employ random masking and an asymmetric encoder-decoder design to encourage semantic modeling with reduced computational cost. Their success has spurred extensions into low-level vision problems, including image inpainting (Liu et al. 2023), denoising (Chen et al. 2023), and super-resolution (Kim et al. 2024). Despite their promising generative capability, conventional MAEs are not directly tailored for structured image reconstruction tasks involving complex, spatially correlated degradations—such as cloud occlusion in remote sensing imagery. The use of random masking, while effective for general-purpose representation learning, does not reflect the spatial and semantic properties of natural occlusions. This gap has motivated recent efforts to adapt the masking mechanism to better align with task-specific priors, such as preserving structural continuity or focusing reconstruction on semantically critical regions

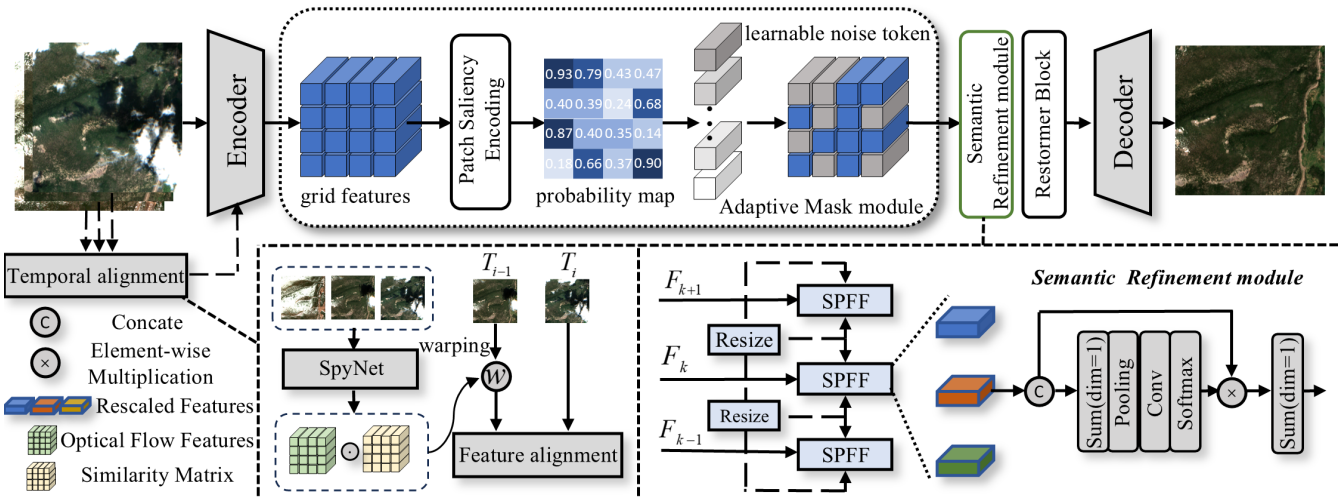

Figure 2: The core restoration stream, guided by an adaptive spatial mask, performs structure-aware completion by selectively reconstructing occluded regions based on visible contextual cues. The SPFF module enhances semantic coherence across occluded and visible regions by promoting feature-level complementarity and suppressing semantic ambiguity. The temporal alignment module, guided by optical flow and structural similarity matrices, ensures robust feature alignment across frames, improving the reconstruction quality.

(Shin et al. 2024). These developments highlight the potential of MAE-style formulations beyond pretraining, particularly as a controllable and interpretable paradigm for learning structure-aware restoration under partial observability.

Method

In this section, we first introduce the overall architecture of VISER-CR, as shown in Fig. 2. Then, we describe the adaptive mask module, the semantic refinement module, and the encoder–decoder process of VISER-CR. Finally, we define the loss functions used for training.

Overall Architecture

As illustrated in Fig. 2, the proposed VISER-CR adopts a unified architecture tailored for cloud removal in remote sensing imagery. The overall network follows a symmetric encoder-decoder paradigm, where both encoder and decoder operate over three spatial resolutions to model multi-level context and support progressive reconstruction. Given a cloud-covered image $\mathbf{I}_{cloudy} \in \mathbb{R}^{C\times H\times W}$, the network first projects the input into a latent space and then progressively refines representations using Restormer blocks (Zamir et al. 2022) along with residual down/up-sampling modules. Notably, the Adaptive Mask (AM) module leverages the spatial structures and visibility patterns of remote sensing imagery to generate content-aware spatial priors, guiding the network to accurately distinguish and focus on occluded regions. These priors are injected into both the encoding and decoding pathways. To further improve restoration quality under heavy cloud cover, we introduce a Semantic Refinement (SR) stream after adaptive masking. This stream aggregates high-level contextual clues from visible regions to enhance semantic consistency and compensate for large-scale structural missing information. In combination with AM, SR helps to maintain both local texture and global structure, facilitating the restoration of missing regions with both perceptual accuracy and structural coherence. Features from different branches are adaptively fused at multiple scales and finally reconstructed through an up-sampling decoder.

Adaptive Mask Module

Despite the strong representation capacity of deep networks, their ability to focus on structurally relevant regions under cloud occlusion remains limited. Motivated by the observation that regions with ambiguous textures or occlusions (e.g., clouds) exhibit inconsistent feature activations, we propose an adaptive mask module to guide the network towards learning spatially discriminative reconstruction priors.

Traditional masking strategies, such as random or uniform dropout, are agnostic to the content and structure of the input, often resulting in sub-optimal feature learning. In contrast, AM dynamically adapts to the content of the input image, allowing it to effectively handle cloud occlusion in remote sensing imagery. AM generates a spatial attention mask conditioned on the input features, focusing on occluded regions and structural boundaries that require enhanced modeling.

Formally, given an input feature map $X \in \mathbb{R}^{C\times H\times W}$, AM generates a spatial occlusion score map through a lightweight projection network $F_{proj}(\cdot)$:

$$S = F_{proj}(X), S \in \mathbb{R}^{1\times H\times W} \tag{1}$$

The score map S is spatially aggregated and flattened to obtain a per-patch occlusion confidence vector. During training, we select the top-k uncertain patches according to

Algorithm 1: Adaptive Masking and Semantic Refinement (AM + SR)

Input: Multi-scale features $\{F_j\}_{j=1}^{H}$; Mask predictors $\{\mathcal{M}_j\}$; Tokens $\{\mathcal{T}_j\}$
Output: Refined features $\{F_j^{\text{out}}\}_{j=1}^{H}$
Set H=3
for $j = 1$ *to* H **do**
  Predict mask: $M_j \leftarrow \mathcal{M}_j(F_j)$
  Expand tokens: $\tilde{\mathcal{T}}_j \leftarrow \text{Expand}(\mathcal{T}_j)$
  Replace: $\hat{F}_j \leftarrow M_j \cdot \tilde{\mathcal{T}}_j + (1 - M_j) \cdot F_j$
  Restore: $\tilde{F}_j \leftarrow \text{Reshape}(\hat{F}_j)$
for $j = 1$ *to* H **do**
  for $k = 1$ *to* H **do**
    $F_{j \leftarrow k} \leftarrow \text{Resize}(\tilde{F}_k, \text{scale} = j)$
  $F_j^{\text{out}} \leftarrow \text{SPFF}_j(\{F_{j \leftarrow k}\}_{k=1}^{H})$
return $\{F_j^{\text{out}}\}_{j=1}^{H}$

these scores to construct a binary occlusion mask $M \in \{0,1\}^{B \times N}$, where $N = H \cdot W$ and the masking ratio is controlled by a hyperparameter ρ:

$$M_i = 1\{i \in TopK(S, \rho)\}, i \in [1, N] \tag{2}$$

Instead of simply discarding masked features, we replace them with learnable token embeddings $\mathcal{T} \in \mathbb{R}^{N \times C}$, serving as semantic priors for plausible reconstruction. To stabilize training and encourage structural completion in severely occluded regions, these tokens are initialized with either zero vectors or controlled Gaussian noise. This initialization not only regularizes the early learning stage but also enables the network to distinguish between observed and missing regions, facilitating structure-aware decoding and enhancing perceptual consistency:

$$X^{masked} = M \odot \mathcal{T} + (1 - M) \odot X \tag{3}$$

From an information-theoretic perspective, the adaptive mask performs spatial importance sampling by prioritizing cloud-occluded regions exhibiting high predictive uncertainty. Let $p(x)$ denote the clean data distribution and $U(x)$ the uncertainty measure derived from the network's occlusion confidence. The optimal mask M^* solves

$$\begin{aligned} M^* = \arg\max_{M} \ & \mathbb{E}_{x \sim p(x)} \left[U(x) \cdot M(x) \right], \\ \text{s.t.} \quad & \|M\|_0 = \rho HW \end{aligned} \tag{4}$$

with masking ratio ρ, this formulation directs representational capacity toward structurally ambiguous, cloud-covered patches, enabling efficient allocation of modeling resources and improved reconstruction fidelity under heavy occlusion.

Notably, the AM module is deployed at multiple levels throughout the encoder, enabling multi-scale structural reasoning and progressive token injection across feature hierarchies. The learnable tokens are shared across samples and spatially broadcasted to align with masked locations, effectively acting as implicit shape priors in regions lacking reliable evidence.

Semantic Refinement Module

While low-level features are essential for preserving textures and local details, the recovery of structurally coherent content under severe cloud occlusion relies heavily on global semantic priors. To effectively enhance such high-level semantic guidance across multiple depths and resolutions, we introduce a semantic refinement module that enables dense, multi-scale interaction and progressive fusion through an iterative grid refinement process. SR module enhances the model's ability to complete occluded regions by reasoning over high-level semantics.

At each refinement step i, SR aggregates hierarchical features $F_k{}_{k=0}^{H-1}$ from all scales using a resolution-adaptive alignment mechanism. Specifically, for each target scale j, we spatially resample the features from all other scales k via upsampling or downsampling operations to match the resolution of j:

$$F_{j \leftarrow k} = \begin{cases} \text{Up}(F_k,\, s = 2^{|j-k|}), & k < j \\ F_k, & k = j \\ \text{Down}(F_k,\, s = 2^{|j-k|}), & k > j \end{cases} \tag{5}$$

where $F_{j \leftarrow k}$ denotes the resolution-aligned feature from stage k to stage j. The set $F_{j \leftarrow k}{}_{k=0}^{H-1}$ is then fused via our proposed Semantic-Prioritized Feature Fusion (SPFF) block:

$$\tilde{F} = SPFF_j(\{F_{j \leftarrow k}\}) \tag{6}$$

which assigns greater importance to semantically informative components based on inter-scale relevance. Unlike traditional attention modules that treat all scales uniformly, SPFF emphasizes semantically rich responses and filters out redundant or noisy features from less relevant depths.

This refinement process is repeated for multiple iterations across the horizontal axis of the feature grid, allowing semantic priors to flow bidirectionally across depth and resolution. The result is a set of refined features $\tilde{F}_j$ that are not only spatially consistent but also semantically aligned, enhancing the model's capacity to complete cloud-covered regions with structurally and perceptually faithful content.

By leveraging semantic-aware, cross-scale reasoning, the Semantic Refinement Module plays a critical role in bridging hierarchical representations and reinforces cloud removal under large-area occlusion.

Multi-temporal flow

Although VISER-CR is designed mainly for mono-temporal cloud removal, it can naturally extend to multi-temporal settings when additional observations are available. To improve temporal consistency and leverage complementary cues from adjacent frames, we adopt a classical optical flow-based feature alignment strategy widely used in video super-resolution tasks (Chan et al. 2022). This motion-aware alignment mechanism enables the model to compensate for inter-

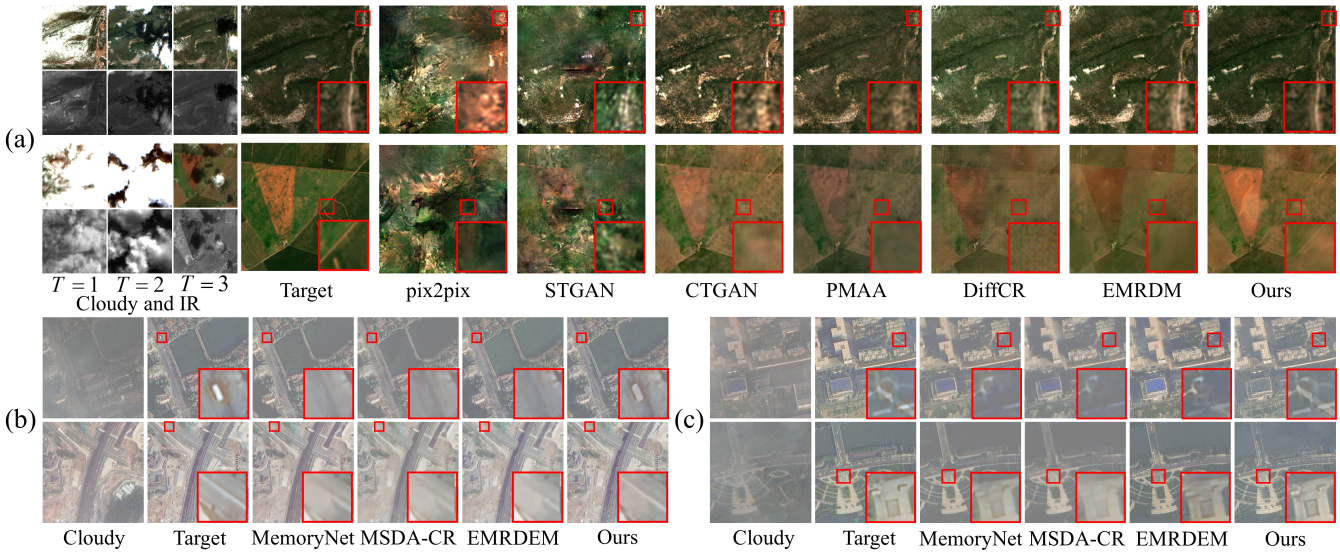

Figure 3: (a) Cloudy, Infrared (IR) and predicted images from the Sen2_MTC_New dataset. Results are obtained by processing images at a resolution of 256 × 256. (b) and (c) RGB channel results on CUHK-CR1 and CUHK-CR2 datasets, respectively. Red boxes highlight key regions of interest.

frame displacement and effectively retrieve cloud-free content from neighboring observations, which is especially beneficial for reconstructing severely occluded regions. Building upon this, we design a more robust alignment strategy by introducing a similarity matrix to mitigate the impact of cloud occlusion on other regions. Instead of solely relying on optical flow to guide feature propagation, we compute a similarity matrix between feature maps of adjacent frames, which quantifies the structural consistency between frames. This matrix allows the model to focus on regions with higher similarity, ensuring more reliable alignment even in the presence of occlusions.

Specifically, we first use SpyNet (Ranjan and Black 2017) to compute the optical flow feature $F_{flow,i-1\to i}$ between frames T_{i-1} and T_i, capturing pixel-level displacement information across consecutive frames. We also compute a similarity matrix $S_{i-1,i}$ to evaluate the consistency of features between frames. Then, by combining the optical flow with the similarity matrix, we provide a more informative alignment mechanism, which reduces the negative impact of occlusions on the alignment process:

$$F_{i-1}^{warped} = \mathcal{W}(F_{i-1}, F_{flow,i-1\to i} \odot S_{i-1,i}) \tag{7}$$

where $\mathcal{W}$ denotes the spatial warping operation. Finally, the warped feature map is aligned with the feature map F_i of the adjacent frame T_i. The temporally aligned multi-frame maps are concatenated and fed into the backbone network for cloud removal.

Triple Learning Objective

To enhance the spatial fidelity, frequency consistency, and perceptual quality of the reconstructed cloud-free images, we formulate a composite loss function that jointly supervises the output in spatial, spectral, and semantic feature domains. Specifically, the loss comprises three complementary components:

$$\mathcal{L}_{L1} = \frac{1}{P}\|\hat{X} - X\|_1 \tag{8}$$

$$\mathcal{L}_{FFT} = \frac{1}{P}\|\mathcal{F}(\hat{X}) - \mathcal{F}(X)\|_1 \tag{9}$$

$$\mathcal{L}_{perceptual} = \frac{1}{P}\|\phi(\hat{X}) - \phi(X)\|_1 \tag{10}$$

Here, $\hat{X}$ and X denote the predicted image and ground truth, respectively; P is the total number of pixels used for normalization; $\mathcal{F}(\cdot)$ denotes the 2D fast Fourier transform, applied to compute differences in both amplitude and phase spectra; and $\phi(\cdot)$ represents deep feature embeddings extracted by a pre-trained VGG network, which captures high-level perceptual differences between images.

By simultaneously optimizing pixel-wise accuracy, spectral consistency, and perceptual realism, this compound loss guides the model toward producing visually faithful reconstructions even under heavy cloud occlusion. The overall loss is given by:

$$\mathcal{L}_{total} = \mathcal{L}_{L1} + \lambda_1\mathcal{L}_{FFT} + \lambda_2\mathcal{L}_{perceptual} \tag{11}$$

where λ_1 and λ_2 are weighting coefficients used to balance the contributions of each component. In our experiments, we empirically set $\lambda_1 = 0.05$ and $\lambda_2 = 0.1$.

Experimental Results

In this section, we evaluate our proposed VISER-CR on three datasets: CUHK-CR1 (Sui et al. 2024) and CUHK-CR2 (Sui et al. 2024) for mono-temporal CR tasks; and Sen2_MTC_New (Huang and Wu 2022) for multi-temporal

Method	CUHK-CR1		CUHK-CR2	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
SpA-GAN (Pan 2020)	20.999	0.5162	19.680	0.3952
AMGAN-CR (Xu et al. 2022)	20.867	0.4986	20.172	0.4900
CVAE (Ding, Zi, and Xie 2022)	24.252	0.7252	22.631	0.6302
MemoryNet	26.073	0.7741	24.224	0.6838
(Zhang, Gu, and Zhu 2023)	–	–	–	–
MSDA-CR	25.435	0.7483	23.755	0.6661
(Yu, Zhang, and Pun 2022)	–	–	–	–
DE-MemoryNet (Sui et al. 2024)	26.183	0.7746	24.348	0.6843
DE-MSDA (Sui et al. 2024)	25.739	0.7592	23.968	0.6737
EMRDM (Liu et al. 2025)	27.281	0.8007	24.594	0.6951
Ours (VISER-CR)	**31.105**	**0.8721**	**26.565**	**0.8078**

Method	Sen2_MTC_New		
	PSNR↑	SSIM↑	LPIPS↓
McGAN (Enomoto et al. 2017)	17.448	0.513	0.447
Pix2Pix	16.985	0.455	0.535
(Isola et al. 2017)	–	–	–
AE (Sintarasirikulchai et al. 2018)	15.100	0.441	0.602
STNet (Chen et al. 2020)	16.206	0.427	0.503
DSen2-CR	16.827	0.534	0.446
(Meraner et al. 2020)	–	–	–
STGAN (Sarukkai et al. 2020)	18.152	0.587	0.513
CTGAN (Huang and Wu 2022)	18.308	0.609	0.384
SEN12MS-CR-TS Net	18.585	0.615	0.342
(Ebel et al. 2022)	–	–	–
PMAA (Zou et al. 2023)	18.369	0.614	0.392
UnCRtainTS (Ebel et al. 2023)	18.770	0.631	0.333
DDPM-CR (Jing et al. 2023)	18.742	0.614	0.329
DiffCR (Zou et al. 2024)	19.150	0.671	0.291
EMRDM (Liu et al. 2025)	20.067	0.709	0.255
Ours (VISER-CR)	**20.541**	**0.733**	**0.231**

Table 1: Quantitative comparison on multiple cloud removal benchmarks.

CR tasks with L = 3. MAE, PSNR, SSIM, and LPIPS are used as evaluation metrics.

Implementation Details

The proposed model is implemented using PyTorch with the Adam optimizer. We adopt a composite loss function that jointly supervises the reconstruction in spatial, frequency, and perceptual domains. The initial learning rate is set to 0.0001 and gradually decayed to 0.00005 during training. The embedding layer projects features to a default channel dimension of $C = 64$, while the final convolutional layer restores the output to match the ground truth channels of each dataset. For data preprocessing, we resize all images by a factor of 0.5, extract training patches of size 128×128, and apply random horizontal flipping with a probability of 0.5. For the NIR or IR channels provided by the datasets, we concatenate them with the RGB input and feed the result into the network, allowing the model to leverage complementary spectral information. In the course of the entire series of experiments, we utilized NVIDIA RTX 3090 GPU.

Comparisons

We compare the proposed ViSER-CR with a broad range of state-of-the-art cloud removal methods on three challenging benchmarks. As shown in Table 1, ViSER-CR consistently outperforms existing approaches across all evaluation metrics. On CUHK-CR1 and CUHK-CR2, which contain complex cloud patterns and rely solely on single-frame RGB inputs, ViSER-CR achieves a substantial margin over all competing methods in both PSNR and SSIM. Notably, it improves PSNR by 3.8 dB and 2.0 dB over the best prior method (EMRDM) on CUHK-CR1 and CUHK-CR2, respectively. These improvements indicate that our structure-aware representation and mask-driven modeling can effectively recover spatial details even under heavily degraded input conditions, where no temporal or auxiliary modality is available. On the more challenging Sen2_MTC_New dataset, which features severe cloud occlusion and complex multi-temporal dynamics, ViSER-CR demonstrates strong generalization capability. It surpasses all baselines in terms of both PSNR and LPIPS, suggesting that the reconstructed images are not only pixel-wise accurate but also perceptually more realistic. Taken together, the results highlight ViSER-CR's ability to recover both fine spatial details and perceptually realistic content under challenging conditions.

As shown in Fig. 3, our ViSER-CR exhibits strong generalization to real-world cloud-covered scenes with diverse cloud patterns and surface types. Compared to other methods, ViSER-CR more effectively restores fine structures such as vegetation textures and building edges, and better preserves spectral consistency. We use red boxes to highlight regions where structural details are difficult to recover. On CUHK-CR1 and CUHK-CR2, ViSER-CR reconstructs more complete and coherent textures, while in Sen2_MTC_New, where cloud occlusion is more severe, our method is still able to recover plausible color and structural information that closely resemble the ground truth. These results confirm the robustness of our method under both moderate and extreme cloud conditions.

Ablation Studies

We conduct comprehensive ablation studies to validate the effectiveness of key modules in ViSER-CR and to analyze the influence of temporal sequence length. All experiments are conducted on the CUHK-CR1 and Sen2_MTC_New datasets using the same training protocol to ensure fair comparison.

Effectiveness of Structural Modules. We perform ablation studies on three key components in ViSER-CR: patch saliency encoding, masking, and the SR module. As shown in Table 2, the baseline model without any of these components yields the lowest performance, with a PSNR of 25.62 dB, SSIM of 0.751, and MAE of 0.043. Adding the SR module alone results in a modest improvement, indicating its role in enhancing geometric regularity and perceptual sharpness. In contrast, incorporating the mask brings a more substantial performance boost—PSNR increases by over 3 dB and MAE drops to 0.032—highlighting the importance of mask-based guidance in restoring cloud-covered content. When

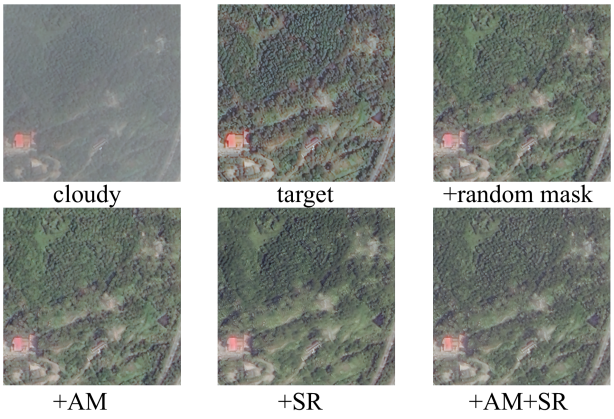

Figure 4: We conducted an ablation study on the CUHK-CR1 dataset to evaluate our method by incrementally adding modules.

	Pos Enc	Mask	SR	PSNR↑	SSIM↑	MAE↓
(a)				25.62	0.751	0.043
(b)			✓	26.95	0.762	0.039
(c)		✓		28.75	0.804	0.032
(d)	✓	✓		30.25	0.844	0.027
(e)	✓	✓	✓	**31.11**	**0.872**	**0.025**

Table 2: Ablation study of our modules on the CUHK-CR1 dataset.

both patch saliency encoding and masking are applied, the model achieves further gains, demonstrating the synergy between structural priors and adaptive occlusion modeling. The full model, with all three modules enabled, achieves the best overall performance with a PSNR of 31.11, SSIM of 0.872, and MAE of 0.025, confirming the effectiveness of our design in leveraging multi-scale features and structural semantics for robust cloud removal.

As shown in Fig. 4, adding AM significantly improves texture and color consistency, effectively restoring cloud-covered regions and enhancing perceptual quality. On the other hand, SR greatly improves the structural coherence of the image, but seems to have limited impact on fine details, as evidenced by the relatively uniform texture restoration. In contrast, the baseline model with random masking exhibits slight performance degradation in both overall color and detail fidelity compared to our VISER-CR, underscoring the importance of adaptive occlusion modeling in achieving robust cloud removal.

Effectiveness of Masking Ratio. The performance of the adaptive masking strategy is highly dependent on the masking ratio ρ. As shown in Table 3, we observe a convex trend: performance improves as ρ increases from 0 to 0.25, peaking at $\rho = 0.25$ with a PSNR gain of over 4 dB compared to the unmasked baseline. This confirms that moderate masking encourages the model to focus on semantically ambiguous regions, leading to better structural recovery. However, excessive masking (*e.g.*, $\rho \geq 0.5$) leads to performance degradation, indicating that overly sparse observations limit the model's reconstruction capacity. These results suggest that the masking strategy requires a carefully chosen ratio to be effective, and that cloud removal tasks do not tolerate overly aggressive masking.

Masking Ratio	PSNR↑	SSIM↑	MAE↓
$\rho = 0.00$	26.95	0.762	0.039
$\rho = 0.15$	28.47	0.815	0.031
$\rho = 0.25$	**31.11**	**0.872**	**0.025**
$\rho = 0.35$	29.03	0.841	0.029
$\rho = 0.50$	26.82	0.765	0.041
$\rho = 0.75$	24.56	0.721	0.048

Table 3: Analysis of masking ratio ρ on the CUHK-CR1 dataset.

Sequence Length	PSNR↑	SSIM↑	MAE↓	LPIPS↓
$L = 1$	16.17	0.503	0.142	0.424
$L = 2$	18.32	0.647	0.098	0.326
$L = 3$	**20.54**	**0.733**	**0.080**	**0.231**

Table 4: Analysis of sequence length L on the Sen2_MTC_New dataset.

Effectiveness of Temporal Sequence Length. We further investigate the effect of input temporal length L on the Sen2_MTC_New dataset, which poses challenges due to heavy cloud coverage and severe structural occlusion. As shown in Table 4, using only a single-frame input leads to significantly lower performance, reflecting the limitations of relying solely on spatial cues. As the temporal length increases, the model consistently improves, indicating that even limited historical context is beneficial for cloud disambiguation. With longer input sequences, ViSER-CR achieves the best overall results, highlighting its effectiveness in leveraging multi-temporal information.

Conclusion

In this work, we proposed VISER-CR, a structure-aware and semantics-guided framework for cloud removal in remote sensing imagery. By reformulating cloud removal as a structure-guided completion task, VISER-CR leverages partial visibility and adaptive spatial priors to infer occluded content more effectively. The proposed Adaptive Mask module enables content-aware masking that prioritizes structurally critical regions, while the Semantic Refinement module facilitates multi-scale context aggregation for enhanced semantic consistency. Additionally, we introduced a temporal fusion mechanism and a triple-domain loss to jointly optimize spatial, frequency, and perceptual objectives. These improvements enable VISER-CR to achieve outstanding performance in both mono-temporal and multi-temporal tasks.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62206211 and Grant U22A2096, in part by Scientific and Technological Innovation Teams in Shaanxi Province under grant 2025RS-CXTD-011, in part by the Shaanxi Province Core Technology Research and Development Project under grant 2024QY2-GJHX-11, in part by the Young Talent Fund of Association for Science and Technology in Shaanxi China under Grant 20240140, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23042 and Grant KYFZ25001, in part by the Innovation Fund of Xidian University under Grant YJSJ25007.

References

Azimi, S. M.; Vig, E.; Bahmanyar, R.; Körner, M.; and Reinartz, P. 2018. Towards multi-class object detection in unconstrained remote sensing imagery. In *Asian conference on computer vision*, 150–165. Springer.

Bermudez, J. D.; Happ, P. N.; Oliveira, D. A. B.; and Feitosa, R. Q. 2018. 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.

Carranza-García, M.; García-Gutiérrez, J.; and Riquelme, J. C. 2019. A framework for evaluating land use and land cover classification using convolutional neural networks. *Remote Sensing*, 11(3): 274.

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5972–5981.

Chen, H.; Gu, J.; Liu, Y.; Magid, S. A.; Dong, C.; Wang, Q.; Pfister, H.; and Zhu, L. 2023. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1692–1703.

Chen, Y.; Weng, Q.; Tang, L.; Zhang, X.; Bilal, M.; and Li, Q. 2020. Thick clouds removing from multitemporal Landsat images using spatiotemporal neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Ding, H.; Zi, Y.; and Xie, F. 2022. Uncertainty-based thin cloud removal network via conditional variational autoencoders. In *Proceedings of the Asian Conference on Computer Vision*, 469–485.

Ebel, P.; Garnot, V. S. F.; Schmitt, M.; Wegner, J. D.; and Zhu, X. X. 2023. UnCRtainTS: Uncertainty quantification for cloud removal in optical satellite time series. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2086–2096.

Ebel, P.; Xu, Y.; Schmitt, M.; and Zhu, X. X. 2022. SEN12MS-CR-TS: A remote-sensing data set for multimodal multitemporal cloud removal. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–14.

Enomoto, K.; Sakurada, K.; Wang, W.; Fukui, H.; Matsuoka, M.; Nakamura, R.; and Kawaguchi, N. 2017. 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*, 48–56.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Grohnfeldt, C.; Schmitt, M.; and Zhu, X. 2018. 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*, 1726–1729. IEEE.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Huang, G.-L.; and Wu, P.-Y. 2022. Ctgan: Cloud transformer generative adversarial network. In *2022 IEEE International Conference on Image Processing (ICIP)*, 511–515. IEEE.

Huang, W.; Deng, Y.; Wu, Y.; and Wang, J. 2024. Attentive Contextual Attention for Cloud Removal. *IEEE Transactions on Geoscience and Remote Sensing*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jing, R.; Duan, F.; Lu, F.; Zhang, M.; and Zhao, W. 2023. Denoising diffusion probabilistic feature-based network for cloud removal in Sentinel-2 imagery. *Remote Sensing*, 15(9): 2217.

Kim, S.-B.; Lee, S.-H.; Choi, H.-Y.; and Lee, S.-W. 2024. Audio super-resolution with robust speech representation learning of masked autoencoder. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32: 1012–1022.

Li, C.; Liu, X.; and Li, S. 2023. Transformer meets GAN: Cloud-free multispectral image reconstruction via multisensor data fusion in satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–13.

Li, J.; Wu, Z.; Hu, Z.; Li, Z.; Wang, Y.; and Molinier, M. 2021. Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for Sentinel-2A imagery. *Remote Sensing*, 13(1): 157.

Lin, C.-H.; Tsai, P.-H.; Lai, K.-H.; and Chen, J.-Y. 2012. Cloud removal from multitemporal satellite images using information cloning. *IEEE transactions on geoscience and remote sensing*, 51(1): 232–241.

Liu, J.; Huang, X.; Zheng, J.; Liu, Y.; and Li, H. 2023. MixMAE: Mixed and masked autoencoder for efficient pretraining of hierarchical vision transformers. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6252–6261.

Liu, Y.; Li, W.; Guan, J.; Zhou, S.; and Zhang, Y. 2025. 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*, 17851–17861.

Meraner, A.; Ebel, P.; Zhu, X. X.; and Schmitt, M. 2020. 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.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Pan, H. 2020. Cloud removal for remote sensing imagery via spatial attention generative adversarial network. *arXiv preprint arXiv:2009.13015*.

Ranjan, A.; and Black, M. J. 2017. Optical flow estimation using a spatial pyramid network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4161–4170.

Rußwurm, M.; and Korner, M. 2017. Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 11–19.

Sarukkai, V.; Jain, A.; Uzkent, B.; and Ermon, S. 2020. Cloud removal from satellite images using spatiotemporal generator networks. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1796–1805.

Shin, J.; Lee, I.; Lee, J.; and Lee, J. 2024. Self-guided masked autoencoder. *Advances in Neural Information Processing Systems*, 37: 58929–58954.

Sintarasirikulchai, W.; Kasetkasem, T.; Isshiki, T.; Chanwimaluang, T.; and Rakwatin, P. 2018. A multi-temporal convolutional autoencoder neural network for cloud removal in remote sensing images. In *2018 15th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, 360–363. IEEE.

Sui, J.; Ma, Y.; Yang, W.; Zhang, X.; Pun, M.-O.; and Liu, J. 2024. Diffusion enhancement for cloud removal in ultra-resolution remote sensing imagery. *arXiv preprint arXiv:2401.15105*.

Tong, X.-Y.; Xia, G.-S.; Lu, Q.; Shen, H.; Li, S.; You, S.; and Zhang, L. 2020. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237: 111322.

Turkoglu, M. O.; D'Aronco, S.; Perich, G.; Liebisch, F.; Streit, C.; Schindler, K.; and Wegner, J. D. 2021. Crop mapping from image time series: Deep learning with multi-scale label hierarchies. *Remote Sensing of Environment*, 264: 112603.

Wang, P.; Sun, X.; Diao, W.; and Fu, K. 2019. FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5): 3377–3390.

Xu, M.; Deng, F.; Jia, S.; Jia, X.; and Plaza, A. J. 2022. Attention mechanism-based generative adversarial networks for cloud removal in Landsat images. *Remote sensing of environment*, 271: 112902.

Xu, M.; Jia, X.; Pickering, M.; and Jia, S. 2019. Thin cloud removal from optical remote sensing images using the noise-adjusted principal components transform. *ISPRS Journal of Photogrammetry and Remote Sensing*, 149: 215–225.

Xu, M.; Pickering, M.; Plaza, A. J.; and Jia, X. 2015. Thin cloud removal based on signal transmission principles and spectral mixture analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 54(3): 1659–1669.

Yu, W.; Zhang, X.; and Pun, M.-O. 2022. Cloud removal in optical remote sensing imagery using multiscale distortion-aware networks. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5728–5739.

Zhang, X. F.; Gu, C. C.; and Zhu, S. Y. 2023. Memory augment is all you need for image restoration. *arXiv preprint arXiv:2309.01377*.

Zou, X.; Li, K.; Xing, J.; Tao, P.; and Cui, Y. 2023. Pmaa: A progressive multi-scale attention autoencoder model for high-performance cloud removal from multi-temporal satellite imagery. In *ECAI 2023*, 3165–3172. IOS Press.

Zou, X.; Li, K.; Xing, J.; Zhang, Y.; Wang, S.; Jin, L.; and Tao, P. 2024. Diffcr: A fast conditional diffusion framework for cloud removal from optical satellite images. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–14.