ISPRS Journal of Photogrammetry and Remote Sensing 237 (2026) 323–338

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs

ECRformer: An efficient cloud removal Transformer with semantic-decoupled learning for multimodal satellite imagery

Zaiyan Zhang [a], Jie Li [a,*], Yuanqi Liang [b], Jining Yan [d], Yi Xiao [e], Xin Su [c], Qiangqiang Yuan [a,*]

[a] School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, Hubei, China
[b] School of Remote Sensing Information and Engineering, Wuhan University, Wuhan 430079, Hubei, China
[c] School of Artificial Intelligence, Wuhan University, Wuhan 430072, Hubei, China
[d] School of Computer Science, China University of Geosciences, Wuhan 430074, Hubei, China
[e] School of Computer Science and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, Henan, China

ARTICLE INFO

Keywords:
Cloud removal
Image processing
Multimodal fusion
Synthetic aperture radar
Transformer

ABSTRACT

Remote sensing imagery is essential for global environmental monitoring, but frequent cloud cover severely limits the utility of optical images. Fusing cloud-prone optical images with cloud-penetrating Synthetic Aperture Radar (SAR) data offers a path to all-weather Earth observation. However, this task faces a dual challenge: the escalating computational cost of state-of-the-art methods and the inherent ill-posedness of the reconstruction under information loss, which complicates the learning process. To tackle this, we propose ECRformer (Efficient Cloud Removal Transformer). ECRformer pairs an efficient architecture with a principled learning paradigm to address both challenges through: (1) a suite of efficient attention mechanisms, including Cross-Covariance Attention (XCA) for computationally-aware multimodal feature fusion and Multi-Dilation Window Attention (MDWA) for capturing multi-scale spatial context with linear complexity; and (2) the Semantic-Decoupled Feature Learning (SDFL) paradigm, a novel training strategy that decomposes the ill-posed reconstruction task into two well-defined sub-problems: structure recovery and texture rendering. By applying asymmetric supervision (structural loss on the encoder, texture loss on the decoder), SDFL provides a more principled learning process. These improvements enhance reconstruction quality, training stability, and reliability, culminating in new state-of-the-art (SOTA) performance on both the SEN12MS-CR and LuojiaSET-OSFCR large-scale optical-SAR cloud removal datasets. Notably, ECRformer surpasses previous SOTA methods by 1.23/0.90 dB in PSNR, while requiring only 28.9% of the parameters and 24.5% of the FLOPs, providing a powerful, efficient, and reliable solution for multimodal cloud removal. The code is available at https://github.com/zzaiyan/ECRformer.

1. Introduction

Optical remote sensing provides invaluable data for a myriad of applications, from agricultural monitoring and urban planning to disaster response (Yuan et al., 2020; Liu et al., 2024). However, its effectiveness is fundamentally constrained by weather conditions (Shen et al., 2015). Globally, clouds obscure approximately 67% of the Earth's surface at any given time, with land areas experiencing about 55% coverage (King et al., 2013). This pervasive contamination leads to significant data loss, hindering continuous monitoring and analysis (Chen et al., 2025; Sun et al., 2025; Li et al., 2025a). Addressing this challenge, the task of cloud removal from optical remote sensing images has garnered substantial research interest, with methodologies evolving progressively to address inherent limitations.

Early efforts primarily focused on single-image cloud removal, treating the task as an image inpainting problem (Pathak et al., 2016; Criminisi et al., 2004). These methods attempt to reconstruct corrupted regions using contextual information from surrounding cloud-free pixels. While some success has been achieved with modern deep learning models like Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Transformers (Vaswani et al., 2017; Dosovitskiy et al., 2020), they face a fundamental bottleneck: when clouds are thick and opaque, the underlying ground information is irrevocably lost.

* Corresponding authors.
E-mail addresses: zzaiyan@whu.edu.cn (Z. Zhang), jli@sgg.whu.edu.cn (J. Li), yqliang@whu.edu.cn (Y. Liang), jnyan@cug.edu.cn (J. Yan), yixiao@zzu.edu.cn (Y. Xiao), xinsu.rs@whu.edu.cn (X. Su), qqyuan@sgg.whu.edu.cn (Q. Yuan).

https://doi.org/10.1016/j.isprsjprs.2026.04.009
Received 17 October 2025; Received in revised form 11 March 2026; Accepted 3 April 2026

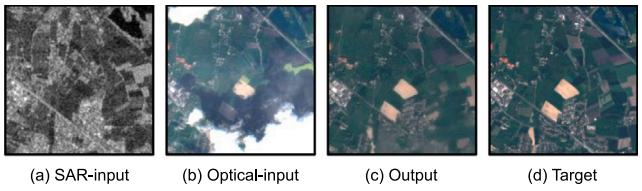

Fig. 1. Typical inputs and outputs of the optical-SAR multimodal cloud removal task, where SAR provides complementary structural information to aid optical image reconstruction.

This renders the problem severely ill-posed, often leading to blurry, structurally inconsistent, or factually incorrect reconstructions.

To overcome this information deficit, the community turned to multi-temporal approaches, which leverage a time-series of images of the same location (Gao and Gu, 2017). By sourcing clear pixels from cloud-free reference images captured at different times, these methods can restore information with high fidelity. However, this paradigm introduces its own stringent dependencies. Its success hinges on the availability of high-quality, precisely registered, cloud-free reference images from a similar season; when available, these methods can be advantageous. Performance degrades significantly in the face of rapid land-cover changes, registration errors, or in regions plagued by persistent cloud cover, limiting its reliability for on-demand analysis.

These limitations paved the way for multimodal fusion, particularly leveraging cloud-penetrating Synthetic Aperture Radar (SAR) imagery, which has emerged as a more robust and versatile solution (Fig. 1). Unaffected by weather and providing rich structural detail, SAR data offers an all-weather, temporally coincident information source, sidestepping the need for a "perfect" temporal reference. Within this promising multimodal paradigm, the field has witnessed its own rapid architectural evolution. Initial approaches utilized Convolutional Neural Networks (CNNs) (LeCun et al., 1989) to fuse local features (Meraner et al., 2020). However, their inherently limited receptive fields struggled to model global context. This motivated the shift to Vision Transformers (ViTs) (Vaswani et al., 2017; Dosovitskiy et al., 2020) to capture long-range dependencies (F. Xu et al., 2022; Gu et al., 2025). Most recently, the field has been dominated by conditional diffusion models, which set new benchmarks by generating highly realistic textures (Zou et al., 2024; Liu et al., 2025; Cai et al., 2025). This progress, however, has come at the cost of a dramatic increase in computational demand, rendering these models slow and resource-intensive. The field appears to have hit a computational wall, where further gains in accuracy seem to require unsustainable increases in cost.

Beyond this efficiency challenge, we argue there is a more fundamental issue at play. As observed in the literature, even state-of-the-art methods face an implicit trade-off between SAR-guided structural consistency and optical-style textural realism (Gu et al., 2025; Liu et al., 2025). In this context, we define *structure* as the low-frequency geometric information, such as object boundaries, road networks, and land shapes, which can be reliably captured by SAR sensors regardless of weather conditions. Conversely, *texture* refers to the high-frequency details, specifically the spectral information and color patterns characteristic of optical imagery, which are often corrupted by clouds. We contend that this is a symptom of a deeper problem: the conventional problem formulation itself. Multimodal cloud removal is an inherently **ill-posed problem**, where a single optical-SAR input pair can correspond to multiple plausible reconstructions. By treating it as a single end-to-end task, existing methods force a single network to simultaneously resolve structural ambiguities using SAR data and render fine-grained textures based on optical context. This convoluted learning objective often leads to suboptimal compromises. Regardless of their architectural power, from CNNs to diffusion models, existing approaches have not sufficiently addressed this foundational challenge in problem formulation.

This paper tackles these dual challenges of efficiency and ill-posedness head-on. We propose a solution that combines a new, principled learning paradigm with a highly efficient architecture. Our contributions are threefold:

- We propose ECRformer, a method for optical-SAR cloud removal that integrates theoretical innovation with an efficient architecture. The architecture features a suite of task-specific efficient attention mechanisms, specifically Cross-Covariance Attention (XCA) for computationally-aware channel-wise fusion and Multi-Dilation Window Attention (MDWA) for linear-complexity spatial modeling.
- We introduce the **Semantic-Decoupled Feature Learning (SDFL)** paradigm, a novel training strategy that reframes the ill-posed reconstruction task. By decomposing the task into two well-defined sub-problems, structure recovery and texture rendering, and applying targeted, asymmetric regularization, SDFL mitigates the risk of artifact generation and makes the learning process more stable, reliable, and interpretable.
- Extensive experiments on multiple large-scale datasets show that ECRformer achieves state-of-the-art reconstruction quality among single-temporal multimodal methods while being computationally efficient; notably, its lightweight variant, ECRformer-Light, attains competitive performance under extremely low computational load.

The organization of the remaining paper is delineated as follows: Section 2 reviews related work in multimodal cloud removal and efficient deep learning architectures. Section 3 details the ECRformer architecture and the SDFL training paradigm. Section 4 presents comprehensive experiments, including quantitative evaluations, qualitative analyses, and ablation studies. Finally, Section 5 concludes with a summary of findings and directions for future research.

2. Related work

Cloud removal from optical remote sensing images is a challenging image restoration problem. Methodologies are broadly categorized into single-image, multi-temporal, and multimodal (SAR-assisted) (Shen et al., 2016).

2.1. Single-image cloud removal

Single-image cloud removal relies solely on a single cloudy optical observation, thus the key is exploring spatial context. Early studies commonly formulated this task as image inpainting, using handcrafted priors such as interpolation-based gap filling (Zhang et al., 2007), PDE-driven inpainting (Bertalmio et al., 2000), and patch-level exemplar matching (Criminisi et al., 2004; He and Sun, 2014). With the advent of deep learning, reconstruction quality improved substantially: early CNN-based completion models (Malek et al., 2017; Pathak et al., 2016) and encoder–decoder designs (Zheng et al., 2020) provided stronger feature representations, while adversarial learning further enhanced perceptual realism in the restored regions (Yu et al., 2019; Sun et al., 2019; Shao et al., 2022). More recently, attention-based architectures have been introduced to better model long-range dependencies and global context, including Transformer-based image restoration backbones (Dosovitskiy et al., 2020; Liang et al., 2021) and task-specific designs for remote sensing cloud removal (M. Xu et al., 2022; Christopoulos et al., 2022; Dai et al., 2024; Wan et al., 2025; Li et al., 2025b). In parallel, diffusion-based formulations have gained traction by learning richer generative distributions and producing visually convincing details (Jing et al., 2023; Sui et al., 2024). However, for large, opaque clouds, the problem remains ill-posed, often leading to blurry or incorrect reconstructions.

2.2. Multi-temporal cloud removal

Multi-temporal cloud removal leverages complementary observations of the same area acquired at different dates. Early work typically relied on relatively strong assumptions such as limited temporal change, and employed hand-crafted fusion rules including interpolation and compositing (Chen et al., 2011; Gao and Gu, 2017), regression-style transfer (Zeng et al., 2013), and tensor decompositions (Liu et al., 2012). While these methods can be effective under mild changes, they often struggle with complex nonlinear appearance variations and temporal degradations. Deep learning has strengthened the ability to learn non-linear temporal correspondences and to model richer temporal dynamics, ranging from CNN-based restoration (Chen et al., 2019; Sarukkai et al., 2020) to 3D-CNN or recurrent formulations (Q. Zhang et al., 2018). More recently, attention-based designs have become a dominant trend by adaptively selecting informative frames and suppressing unreliable observations, enabling more robust aggregation under varying cloud coverage and observation quality (Ebel et al., 2023; Liu et al., 2023; Stucker et al., 2023; Zhang et al., 2025; Shu et al., 2025). The main drawback is the dependency on high-quality, cloud-free reference images, with performance suffering from registration errors, land cover changes, or persistent cloud cover.

2.3. Multimodal cloud removal

Fusing optical data with cloud-penetrating SAR imagery enables all-weather sensing. The dominant approach is deep learning-based fusion. Early works used CNNs to fuse multimodal data (Meraner et al., 2020; Ebel et al., 2020). Generative Adversarial Networks (GANs) were then employed to enhance realism (Grohnfeldt et al., 2018). More advanced hybrid architectures, combining CNNs and Transformers, were designed to better handle modality-specific characteristics (F. Xu et al., 2022; Ebel et al., 2023; He et al., 2023; Gu et al., 2025; Ma et al., 2024). In parallel, SAR-to-optical image translation methods were explored, leveraging GANs to directly synthesize cloud-free optical imagery from SAR data (Bermudez et al., 2018; Yang et al., 2022; Wang et al., 2025). Recently, conditional diffusion models have become state-of-the-art, framing cloud removal as a generative process guided by the cloudy optical and clear SAR images (Zou et al., 2024; Liu et al., 2025; Tu et al., 2025; Cai et al., 2025). Despite impressive results, challenges remain in fully exploiting complementary information, often leading to a trade-off between SAR-guided structural consistency and optical-style textural realism. Our work aims to develop a more effective cross-modal fusion strategy to enhance both aspects.

In summary, while SAR-assisted multimodal fusion is the leading paradigm for cloud removal, current methods face two major challenges. First, top-performing generative models are often too computationally expensive for practical use. Second, a fundamental conflict persists between preserving structural details from SAR data and generating realistic optical textures. This work addresses these issues by proposing an efficient architecture and a novel learning approach that explicitly disentangles structure and texture, aiming to improve both reconstruction quality and efficiency.

3. Methodology

Our proposed method consists of two main components: the ECRformer model, a hierarchical Transformer architecture, and the Semantic-Decoupled Feature Learning (SDFL) paradigm, a novel training strategy. Fig. 2 illustrates the overall architecture.

3.1. Problem formulation

Formally, let $X_{\text{opt}} \in \mathbb{R}^{H\times W\times C_{\text{opt}}}$ represent the cloudy optical image and $X_{\text{sar}} \in \mathbb{R}^{H\times W\times C_{\text{sar}}}$ be a co-registered SAR image of the same geographical area. Here, H and W denote the spatial height and width, while C_{opt} and C_{sar} are the number of channels for the optical and SAR images, respectively. The desired output is a cloud-free optical image $\hat{Y} \in \mathbb{R}^{H\times W\times C_{\text{opt}}}$ that is as close as possible to the ground-truth cloud-free image $Y \in \mathbb{R}^{H\times W\times C_{\text{opt}}}$.

The objective is to learn a mapping function f_θ, parameterized by θ, that takes the cloudy optical and SAR images as input and generates the estimated cloud-free image:

$$\hat{Y} = f_\theta\left(X_{\text{opt}}, X_{\text{sar}}\right). \tag{1}$$

The learning process seeks to find the optimal parameters θ by minimizing a loss function $\mathcal{L}\left(\hat{Y}, Y\right)$ that quantifies the discrepancy between the predicted image and the ground truth.

3.2. Overall architecture

To learn the mapping f_θ, we propose ECRformer, a U-shaped hierarchical Transformer model. As depicted in Fig. 2, its architecture comprises three primary components: a Shallow Feature Embedding module, a U-Shaped Backbone for deep feature extraction, and a final Refinement Network.

3.2.1. Shallow feature embedding

To preserve modality-specific information, the cloudy optical input X_{opt} and SAR input X_{sar} are first processed independently by two parallel 7×7 convolutional layers. This allows the network to learn low-level features for each modality before fusion. The resulting feature maps are then concatenated to form the input for the U-shaped backbone:

$$F_{\text{shallow}} = \text{Concat}\left(\text{Conv}^{\text{opt}}_{7\times7}\left(X_{\text{opt}}\right), \text{Conv}^{\text{sar}}_{7\times7}\left(X_{\text{sar}}\right)\right). \tag{2}$$

3.2.2. U-shaped backbone

The core of our model is a symmetric encoder–decoder backbone with a bottleneck, designed to capture and progressively refine multi-scale features.

The encoder progressively extracts more abstract, semantic features by reducing spatial resolution while increasing channel depth. It consists of K_{enc} encoder stages. Here, we define a *stage* as a sequence of ECRformer Blocks operating at a specific spatial resolution followed by a downsampling or upsampling operation. Let F^{D-1}_{enc} be the input to the Dth stage (with $F^{0}_{\text{enc}} = F_{\text{shallow}}$). This feature map is first processed by a stack of L_D ECRformer Blocks (where L_D denotes the depth of this stage). The resulting features are then downsampled to produce the stage's output, F^{D}_{enc}:

$$F^{D}_{\text{enc}} = \text{Downsample}\left(\text{Blocks}_{L_D}\left(F^{D-1}_{\text{enc}}\right)\right), \quad D = 1, \ldots, K_{\text{enc}} \tag{3}$$

where $\text{Blocks}_{L_D}(\cdot)$ denotes the sequential application of L_D ECRformer Blocks.

Positioned at the deepest part of the network, the bottleneck connects the encoder and decoder. It processes the final encoder output, $F^{K_{\text{enc}}}_{\text{enc}}$, with a series of L_B ECRformer Blocks to further transform the features at the lowest spatial resolution:

$$F_{\text{bottleneck}} = \text{Blocks}_{L_B}\left(F^{K_{\text{enc}}}_{\text{enc}}\right). \tag{4}$$

The decoder symmetrically restores the spatial resolution while refining features to reconstruct the image. It has K_{dec} decoder stages (i.e., we set $K_{\text{dec}} = K_{\text{enc}}$ to ensure size consistency). At each stage D, the output from the previous decoder stage, F^{D-1}_{dec} (with $F^{0}_{\text{dec}} = F_{\text{bottleneck}}$), is first upsampled. The result is then fused with the corresponding

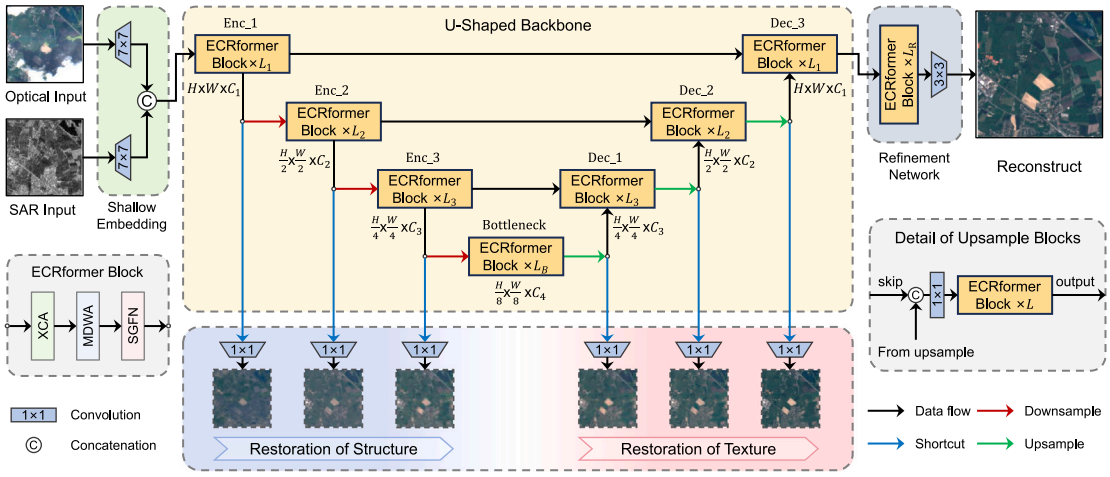

Fig. 2. The overall framework of our proposed method. The ECRformer model, a U-shaped backbone with efficient ECRformer Blocks, processes concatenated optical and SAR feature maps. The Semantic-Decoupled Feature Learning (SDFL) paradigm applies asymmetric supervision to intermediate layers: a structural loss on the encoder and a textural loss on the decoder, guiding a progressive reconstruction from structure to detail.

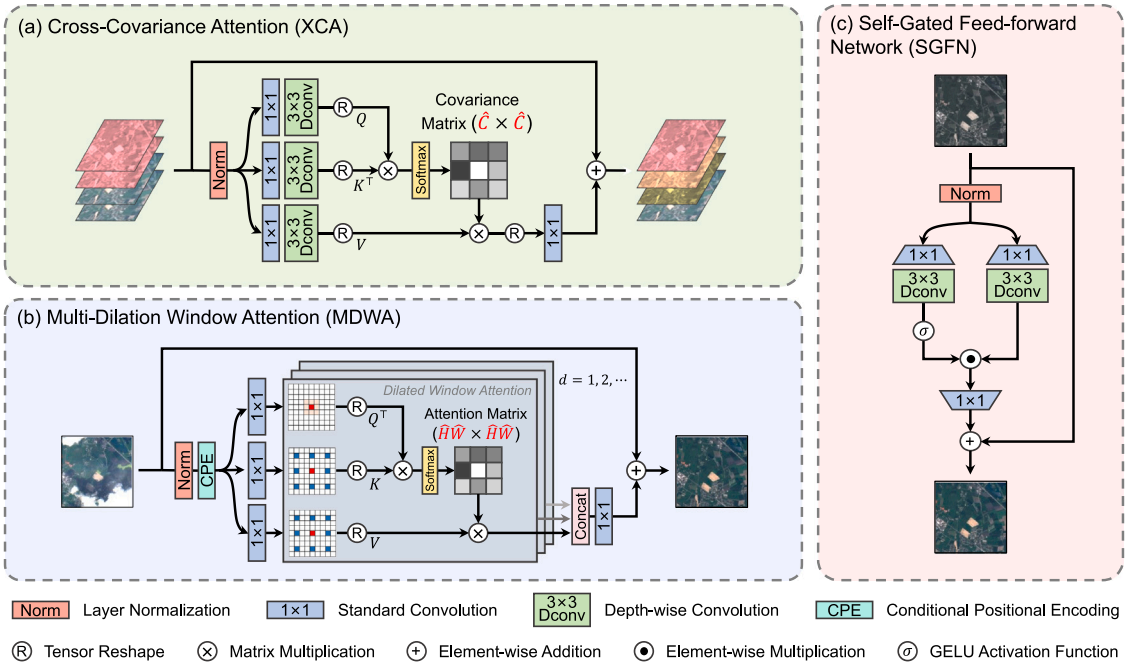

Fig. 3. The ECRformer Block architecture. It consists of three main components: (a) Cross-Covariance Attention (XCA) for efficient cross-modal channel fusion, (b) Multi-Dilation Window Attention (MDWA) for capturing multiscale spatial context, and (c) Self-Gated Feed-forward Network (SGFN) for enhanced non-linear feature transformation.

feature map from the encoder, $F_{\text{enc}}^{K_{\text{enc}}-D}$, via a skip connection. This fused feature map is then processed by L_D ECRformer Blocks to yield the intermediate output F_{dec}^{D}:

$$F_{\text{dec}}^{D} = \text{Blocks}_{L_D}\left(\text{Concat}\left(\text{Upsample}\left(F_{\text{dec}}^{D-1}\right), F_{\text{enc}}^{K_{\text{enc}}-D}\right)\right), \quad D = 1, \ldots, K_{\text{dec}}. \tag{5}$$

3.2.3. Refinement network

After the final decoder stage, the full-resolution feature map $F_{\text{dec}}^{K_{\text{enc}}}$ is fed into a refinement network. This network consists of a final stack of L_R ECRformer Blocks that further enhance fine-grained details and suppress potential artifacts. The final cloud-free prediction $\hat{Y}$ is then generated by a terminal convolutional layer:

$$F_{\text{refined}} = \text{Blocks}_{L_R}\left(F_{\text{dec}}^{K_{\text{enc}}}\right), \quad \hat{Y} = \text{Conv}_{\text{final}}\left(F_{\text{refined}}\right). \tag{6}$$

3.3. ECRformer Block

The backbone of ECRformer is constructed from a novel ECRformer Block, which serially integrates three specialized modules, each designed for efficiency and effectiveness in the context of remote sensing. Fig. 3 illustrates each component. For an input feature F_{in}, the block computes:

$$F_{\text{XCA}} = \text{XCA}\left(\text{LN}\left(F_{\text{in}}\right)\right) + F_{\text{in}}, \tag{7}$$

$$F_{\text{MDWA}} = \text{MDWA}\left(\text{LN}\left(F_{\text{XCA}}\right)\right) + F_{\text{XCA}}, \tag{8}$$

$$F_{out} = \text{SGFN}\left(\text{LN}\left(F_{\text{MDWA}}\right)\right) + F_{\text{MDWA}}, \tag{9}$$

where LN denotes Layer Normalization. The following sections detail the design of each component.

3.3.1. Cross-Covariance Attention (XCA)

Standard self-attention computes an attention map of size $N \times N$ (where $N = H \times W$), leading to a complexity of $O(N^2C)$, which is prohibitive for high-resolution images where $N \gg C$. We employ Cross-Covariance Attention (XCA) (Ali et al., 2021), which transposes the attention operation to the feature dimension. Given query Q, key K, and value V matrices (all of shape $\mathbb{R}^{N\times C}$), XCA is computed as:

$$\text{XCAttention}(Q,K,V) = V \cdot \text{Softmax}\left(\frac{K^\top Q}{\tau}\right), \tag{10}$$

where τ is a learnable temperature parameter that scales the attention logits.

The attention map $K^\top Q$ is of size $C \times C$, representing the cross-covariance between feature channels. This shifts the complexity to $O(NC^2)$, making it highly efficient.

To further enhance its capability for multimodal fusion, we introduce local spatial context into the Q, K, V projections by adding a lightweight depth-wise convolution (DWConv) after a point-wise convolution:

$$\begin{aligned} Q', K', V' &= \text{Split}\left(\text{DWConv}\left(\text{Conv}_{1\times1}\left(F_{\text{in}}\right)\right)\right), \\ Q, K, V &= \text{Reshape}\left(Q', K', V'\right), \end{aligned} \tag{11}$$

where Q, K, and V are of shape $\mathbb{R}^{N\times C}$, N is the number of spatial tokens.

This mechanism is particularly effective for multimodal fusion. After concatenating optical and SAR features along the channel axis, the $C\times C$ attention map in XCA explicitly models the inter-channel dependencies, including the crucial cross-modal correlations. It allows the network to learn how features from one modality (e.g., SAR structure) relate to features in another (e.g., optical color), and to adaptively up-weight or down-weight channels to achieve effective fusion.

3.3.2. Multi-Dilation Window Attention (MDWA)

To efficiently capture spatial context, we propose MDWA. It integrates the efficiency of window-based attention (Liu et al., 2021) with a dilated sampling mechanism inspired by atrous convolutions (Yu and Koltun, 2016). Before projecting to Q, K, V, we inject learnable relative positional information using a Conditional Positional Encoding (CPE) (Chu et al., 2023), implemented as a 3×3 DWConv:

$$Q, K, V = \text{Split}\left(\text{Conv}_{1\times1}\left(F_{\text{in}} + \text{DWConv}\left(F_{\text{in}}\right)\right)\right). \tag{12}$$

The core of MDWA lies in its dilated sampling. For each query pixel, instead of forming a window from its spatially contiguous neighbors, dilated sampling constructs a sparse window by selecting key and value elements at intervals. Specifically, for a dilation rate d, the sampling process skips $d-1$ pixels between window elements in both horizontal and vertical directions, effectively expanding the receptive field at no extra computational cost.

The attention output for this position is then computed by allowing the central query $q^{(h,w)}$ to attend to its context of sampled keys and aggregate the corresponding values:

$$\text{DilatedAttn}^{(h,w)} = \text{Softmax}\left(\frac{q^{(h,w)}K_{\text{dilated}}^\top}{\sqrt{d_k}}\right)V_{\text{dilated}}, \tag{13}$$

where K_{dilated} and V_{dilated} are matrices formed by stacking the sampled key and value vectors, respectively.

By running multiple parallel attention branches with different dilation rates $(d_1, d_2, \ldots, d_{N_d})$, MDWA aggregates multi-scale contextual information:

$$\text{MDWA}\left(F_{\text{in}}\right) = \text{Conv}_{\text{fuse}}\left(\begin{bmatrix} \text{DilatedAttn}_{(d_1)} \\ \vdots \\ \text{DilatedAttn}_{(d_{N_d})} \end{bmatrix}\right), \tag{14}$$

where $\text{DilatedAttn}_{(d)}$ denotes the attention output from the branch with dilation rate d, and $\text{Conv}_{\text{fuse}}$ is a 1×1 convolution that fuses the multi-branch outputs.

This structure enables MDWA to effectively model both local details and larger structures, such as object boundaries and cloud edges.

3.3.3. Self-Gated Feed-forward Network (SGFN)

To improve the non-linear transformation capability of the feed-forward network, we replace the standard MLP with a Self-Gated Feed-forward Network (SGFN). SGFN employs a gating mechanism to adaptively modulate features. To further enhance local feature perception with minimal overhead, we also incorporate a DWConv within the network (Zamir et al., 2022). Given the input feature F_{in}, the core computation of SGFN is:

$$X_1, X_2 = \text{Split}\left(\text{DWConv}\left(\text{Conv}_{1\times1}\left(F_{\text{in}}\right)\right)\right), \tag{15}$$

$$F_{\text{out}} = \text{Conv}_{1\times1}\left(\text{GELU}\left(X_1\right) \odot X_2\right), \tag{16}$$

where $\odot$ denotes element-wise multiplication, GELU denotes the Gaussian Error Linear Unit nonlinear activation (Hendrycks and Gimpel, 2016).

This combination of gating and depth-wise convolution improves the model's expressive power and local feature extraction capacity with negligible additional computational cost.

3.4. Semantic-Decoupled Feature Learning

A key conceptual contribution of our work is the Semantic-Decoupled Feature Learning (SDFL) paradigm. Traditional end-to-end training with a single loss at the final output can lead to undertrained shallow layers and fails to exploit the distinct functional roles of the encoder and decoder. SDFL addresses this by introducing targeted supervision throughout the network, which we conceptualize in two stages: Multiscale Feature Regularization and Semantic Decoupling.

3.4.1. Multiscale feature regularization

To ensure that features at all levels of the network are meaningful, we apply regularization constraints to the intermediate feature maps at different scales. Let $F_{enc}^{(k)}$ and $F_{dec}^{(k)}$ denote the feature maps from the kth stage of the encoder and decoder, respectively. We attach lightweight convolutional projection heads, $H_{enc}^{(k)}$ and $H_{dec}^{(k)}$, to these intermediate layers. Each head predicts a downscaled version of the cloud-free image:

$$\hat{Y}_{\text{enc}}^{(k)} = H_{enc}^{(k)}\left(F_{enc}^{(k)}\right), \quad \hat{Y}_{\text{dec}}^{(k)} = H_{dec}^{(k)}\left(F_{dec}^{(k)}\right). \tag{17}$$

The regularization term, $\mathcal{R}$, is the average of losses from all intermediate predictions:

$$\begin{aligned} \mathcal{R} = \mathcal{R}_{\text{enc}} + \mathcal{R}_{\text{dec}} &= \mathbb{E}_{k_e\sim\mathcal{U}(1,K_{\text{enc}})}\mathcal{L}\left(\hat{Y}_{\text{enc}}^{(k_e)}, Y^{(k_e)}\right) \\ &+ \mathbb{E}_{k_d\sim\mathcal{U}(1,K_{\text{dec}})}\mathcal{L}\left(\hat{Y}_{\text{dec}}^{(k_d)}, Y^{(k_d)}\right), \end{aligned} \tag{18}$$

where $\mathcal{L}$ is a reconstruction loss (e.g., L1), $Y^{(k)}$ is the correspondingly downsampled ground truth, and K_{enc} and K_{dec} are the number of stages in the encoder and decoder, respectively. This deep supervision strategy provides richer gradient signals throughout the network, improving feature space utilization, reducing redundancy, and accelerating convergence by enhancing the network's ability to process multiscale information.

3.4.2. Semantic decoupling

Our semantic decoupling strategy is motivated by the principle of simplifying a complex learning task through probabilistic factorization. The direct mapping from a cloudy optical input X_{opt} and a SAR input X_{sar} to a clean output Y requires learning a highly complex conditional probability distribution $p(Y|X_{\text{opt}}, X_{\text{sar}})$. This mapping is ill-posed due to the information loss under clouds, leading to high variance in the generation of plausible textures.

We decompose this problem by introducing a latent intermediate variable, S, representing the underlying **structure** of the scene. The joint probability can be factorized as:

$$p(Y|X_{\text{opt}}, X_{\text{sar}}) = \int p(Y|S)\cdot p(S|X_{\text{opt}}, X_{\text{sar}})\cdot dS. \tag{19}$$

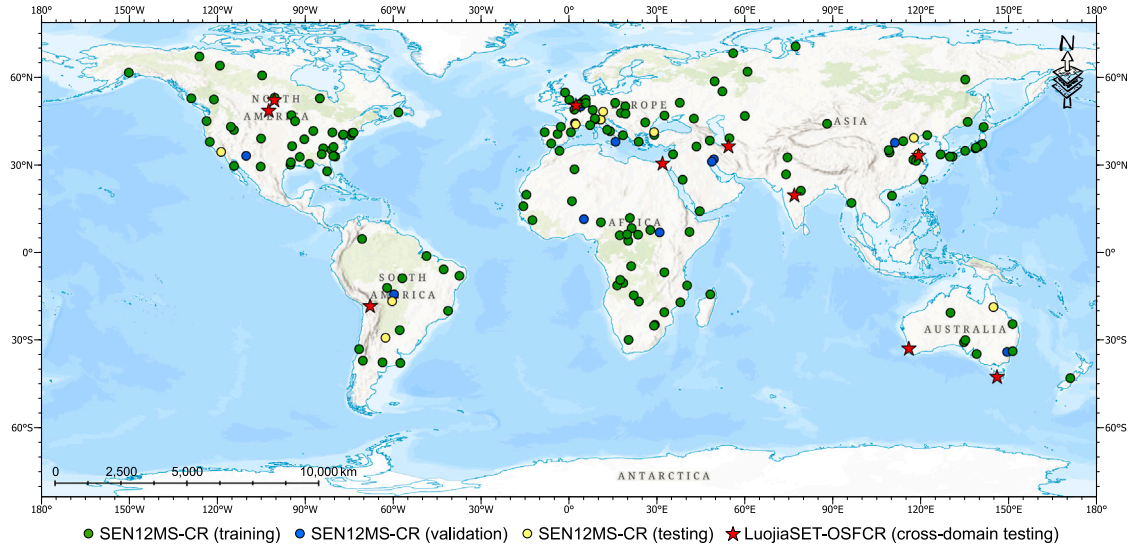

Fig. 4. Geographical distribution of Regions of Interest (ROIs) in the SEN12MS-CR and LuojiaSET-OSFCR datasets. The training set (green) covers diverse global locations and seasons, while the validation (blue), testing (yellow) and cross-domain testing (red star) sets are from distinct ROIs to assess generalization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

This simplifies the generation process to $p(S|X_{\text{opt}}, X_{\text{sar}})$ and $p(Y|S)$. Our model is thus designed to learn this two-stage process, enforced by asymmetric loss functions on the intermediate predictions:

1. **Encoder**: Models $p(S|X_{\text{opt}}, X_{\text{sar}})$ to recover the structural distribution. We enforce this with a **structural loss** using the Structural Similarity Index (SSIM). The expectation over the learned structural distribution is approximated using the intermediate predictions $\hat{Y}_{\text{enc}}^{(k)}$:
$$\mathcal{R}_{\text{enc}} = \mathbb{E}_{k\sim\mathcal{U}(1,K_{\text{enc}})}\left[1 - \text{SSIM}\left(\hat{Y}_{\text{enc}}^{(k)}, Y^{(k)}\right)\right]. \tag{20}$$
2. **Decoder**: Models $p(Y|S)$ to render high-fidelity textures. We enforce this with a **texture loss** using the L1 norm, which is well-suited for maximizing the likelihood of ground truth under a Laplacian distribution assumption. The expectation is approximated using the intermediate predictions $\hat{Y}_{\text{dec}}^{(k)}$:
$$\mathcal{R}_{\text{dec}} = \mathbb{E}_{k\sim\mathcal{U}(1,K_{\text{dec}})}\left[\left\|\hat{Y}_{\text{dec}}^{(k)} - Y^{(k)}\right\|_1\right], \tag{21}$$
where $\|\cdot\|_1$ denotes the L1 norm.

This factorization reduces a single, high-difficulty learning problem into two more constrained sub-problems, simplifying the feature learning path and leading to more stable and accurate reconstruction results.

3.4.3. Overall training objective

The overall training objective combines the final output loss with the intermediate regularization terms from the SDFL paradigm.

The final output loss, $\mathcal{L}_{\text{output}}$, is applied to the final prediction $\hat{Y}$ to balance pixel-level accuracy and perceptual quality. It is a weighted combination of the L1 loss and the SSIM loss:

$$\mathcal{L}_{\text{output}} = \alpha\left\|\hat{Y} - Y\right\|_1 + \beta\left(1 - \text{SSIM}\left(\hat{Y}, Y\right)\right), \tag{22}$$

where α and β are weighting factors.

The total loss $\mathcal{L}$ is then formulated by integrating the output loss with the semantic-decoupled regularization terms, $\mathcal{R}_{\text{enc}}$ and $\mathcal{R}_{\text{dec}}$:

$$\mathcal{L} = \mathcal{L}_{\text{output}} + \lambda_{\text{enc}}\mathcal{R}_{\text{enc}} + \lambda_{\text{dec}}\mathcal{R}_{\text{dec}}, \tag{23}$$

where λ_{enc} and λ_{dec} are hyperparameters that control the influence of the structural and textural regularization from the encoder and decoder, respectively.

4. Experiments

4.1. Settings

This section details the experimental setup, including the datasets, baseline methods, evaluation metrics, and training configurations used to validate our proposed model.

4.1.1. Dataset

To comprehensively evaluate our model, we employ two distinct datasets. We use SEN12MS-CR (Ebel et al., 2020) for training, validation, and initial testing. To further assess the model's generalization capabilities, we conduct cross-domain testing on the LuojiaSET-OSFCR (Pan et al., 2024) dataset. The geographical distribution of Regions of Interest (ROIs) for both datasets is illustrated in Fig. 4.

SEN12MS-CR. is a large-scale, public benchmark for multimodal cloud removal. It contains 122,218 triplets of co-registered, 256 × 256 pixel patches of cloudy Sentinel-2 optical images, cloud-free Sentinel-2 optical ground truth, and Sentinel-1 SAR data, covering diverse global locations and seasons. We follow the official data split of 155 ROIs for training, 10 for validation, and 10 for testing.

LuojiaSET-OSFCR. was created following the standards of SEN12MS-CR. It is also sourced from the Sentinel-1/2 satellites. The dataset comprises 10 distinct, non-overlapping ROIs. Data pre-processing for each ROI mirrors the methodology of SEN12MS-CR. Images are segmented into 256 × 256 pixel patches, totaling 20,000 patches. All these samples are used as testing sets to verify cross-domain generalization ability of the proposed model.

Data preprocessing: For the original Sentinel-1/2 images, we use the data preprocessing consistent with (Ebel et al., 2020). For radiometric normalization, Sentinel-1 SAR images were clipped (VV: $[-25, 0]$ dB, VH: $[-35, 0]$ dB) to remove outliers, then both channels were independently scaled to $[0, 1]$. Sentinel-2 optical images were clipped to an effective reflectance range of $[0, 10000]$ and then uniformly scaled to $[0, 1]$. For resolution alignment, all Sentinel-2 bands at 20-m and 60-m were upsampled to 10-m using bilinear interpolation, creating a unified data cube where all bands are represented as 256 × 256 patches.

Table 1
Quantitative comparison against state-of-the-art methods on the SEN12MS-CR and LuojiaSET-OSFCR testing sets. Best results are highlighted in **bold**, and second-best are underlined. ↑ indicates higher is better and ↓ indicates lower is better.

Method	Venue	SEN12MS-CR					LuojiaSET-OSFCR				
		MAE ↓	SAM ↓	PSNR ↑	SSIM ↑	LPIPS ↓	MAE ↓	SAM ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Optical-only (Single Image CR)											
McGAN	CVPRW'17	0.0475	15.676	25.14	0.744	0.528	0.0503	16.112	24.95	0.731	0.552
SpA GAN	ArXiv'20	0.0446	18.085	24.78	0.754	0.451	0.0479	18.521	24.53	0.740	0.489
CloudRuler	RSE'25	0.0314	11.512	27.95	0.882	0.336	0.0328	12.015	27.34	0.871	0.354
SAR-only (SAR-to-Optical Translation)											
SAR2OPT	ISPRS'18	0.0418	14.788	25.87	0.793	0.393	0.0425	15.103	25.68	0.781	0.426
ICGAN	PR'22	0.0387	11.184	27.15	0.844	0.369	0.0403	11.467	26.89	0.836	0.384
MT-GAN	ISPRS'25	0.0325	9.752	27.88	0.871	0.343	0.0341	10.125	27.73	0.872	0.351
Multimodal (Optical-SAR Fusion)											
SAR-Opt-cGAN	IGARSS'18	0.0431	15.494	25.59	0.764	0.476	0.0457	15.953	25.31	0.752	0.498
DSen2-CR	ISPRS'20	0.0313	9.472	27.76	0.874	0.354	0.0317	9.511	27.68	0.873	0.359
GLF-CR	ISPRS'22	0.0280	8.981	28.64	0.885	0.321	0.0284	9.039	28.57	0.884	0.327
UnCRtainTS	CVPRW'23	0.0272	8.324	28.90	0.880	0.287	0.0299	8.495	28.05	0.878	0.294
DiffCR	TGRS'24	0.0191	5.821	31.77	0.902	0.244	0.0194	5.886	31.71	0.900	0.263
HPN-CR	TGRS'25	0.0242	7.637	30.23	0.898	0.275	0.0246	7.692	30.17	0.897	0.299
EMRDM	CVPR'25	0.0179	5.267	32.14	0.924	**0.181**	0.0182	5.338	32.15	0.921	0.201
ECRformer-Light	*Ours*	0.0178	5.026	32.75	0.920	0.224	0.0182	5.185	32.41	0.918	0.235
ECRformer	*Ours*	**0.0164**	**4.693**	**33.37**	**0.932**	0.188	**0.0167**	**4.751**	**33.05**	**0.929**	**0.196**

4.1.2. Baseline methods

We compare ECRformer against a comprehensive set of state-of-the-art methods representing different architectural paradigms: CNN-based (DSen2-CR (Meraner et al., 2020)), GAN-based (McGAN (Enomoto et al., 2017), SAR-Opt-cGAN (Grohnfeldt et al., 2018), SAR2OPT (Bermudez et al., 2018), SpA GAN (Pan, 2020), ICGAN (Yang et al., 2022), MT-GAN (Wang et al., 2025)), Transformer-based (GLF-CR (F. Xu et al., 2022), UnCRtainTS (Ebel et al., 2023), CloudRuler (Li et al., 2025b), HPN-CR (Gu et al., 2025)), and diffusion-based (DiffCR (Zou et al., 2024), EMRDM (Liu et al., 2025)).

Among them, McGAN, SpA GAN, and CloudRuler only accept optical (multispectral) input, SAR2OPT, ICGAN, and MT-GAN only accept SAR input, and the remaining methods use both optical and SAR inputs. For fair comparison, we use official pre-trained weights (if available) or retrain on the same dataset according to the original paper settings.

4.1.3. Implementation details

All experiments were conducted on an Ubuntu 22.04 server equipped with two AMD EPYC 7K62 48-core CPUs, 512 GB RAM, and four NVIDIA RTX 4090 GPUs (24 GB graphic memory each). Our model was implemented using PyTorch and PyTorch Lightning libraries. The code for the proposed method is available at https://github.com/zzaiyan/ECRformer.

Training settings: ECRformer was trained with AdamW (Kingma, 2015; Loshchilov and Hutter, 2017) ($\beta_1 = 0.9$, $\beta_2 = 0.999$), batch size 16, for up to 200 epochs. The initial learning rate was 4×10^{-4}, reduced by 0.1 if validation loss did not improve for 5 epochs. Early stopping (patience 10) was used. Data augmentation included random cropping (128×128), flips, and rotations (90, 180, 270 degrees). The best model was selected based on validation performance.

Inference settings: During inference, for methods that cannot handle dynamic resolution (e.g., HPN-CR), we used a sliding window approach with a window size of 128×128 and a 50% overlap to handle larger images. The final output was averaged over the overlapping regions to ensure smooth transitions. Other methods used full resolution for inference. No test-time augmentation was used.

Model variants: We instantiate the proposed model at two capacity levels: ECRformer and ECRformer-Light. Both share the same building blocks and the SDFL training paradigm, and only differ in width/depth hyperparameters, enabling a controllable accuracy-efficiency trade-off.

4.1.4. Metrics

To quantitatively evaluate the effectiveness of cloud removal algorithms, we adopt five widely-used image quality assessment metrics to measure the differences between the reconstructed image ($\hat{Y}$) and the ground truth (Y).

Peak Signal-to-Noise Ratio (PSNR) measures pixel-level fidelity based on Mean Squared Error (MSE). A higher PSNR value (in dB) indicates lower distortion.

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}_I^2}{\mathrm{MSE}}\right), \tag{24}$$

$$\mathrm{MSE} = \mathbb{E}\left[\left\|Y - \hat{Y}\right\|_2^2\right], \tag{25}$$

where MAX_I is the maximum pixel value.

Structural Similarity Index Measure (SSIM) (Wang et al., 2004) evaluates the similarity in terms of structure, luminance, and contrast by operating on local windows, which is more consistent with human visual perception. The final score is the mean of SSIM values across all windows. Its value ranges from −1 to 1, with 1 indicating identical images.

$$\mathrm{SSIM}(x, y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{26}$$

where x and y are corresponding local image patches, μ and σ represent the patch's mean and variance (weighted by a Gaussian kernel), and C_1, C_2 are small constants.

Mean Absolute Error (MAE) calculates the average absolute pixel-wise difference. A lower MAE value signifies a more accurate reconstruction.

$$\mathrm{MAE} = \mathbb{E}\left[\left\|Y - \hat{Y}\right\|_1\right]. \tag{27}$$

Spectral Angle Mapper (SAM) (Kruse et al., 1993) assesses spectral fidelity by computing the average angle between spectral vectors of corresponding pixels. A smaller SAM value (in degrees) indicates less spectral distortion.

$$\mathrm{SAM} = \mathbb{E}\left[\arccos\left(\frac{Y \cdot \hat{Y}}{\|Y\|_2 \cdot \|\hat{Y}\|_2}\right)\right]. \tag{28}$$

Learned Perceptual Image Patch Similarity (LPIPS) (R. Zhang et al., 2018) evaluates image quality from a human-perception perspective by comparing high-level semantics such as texture and shape. Since

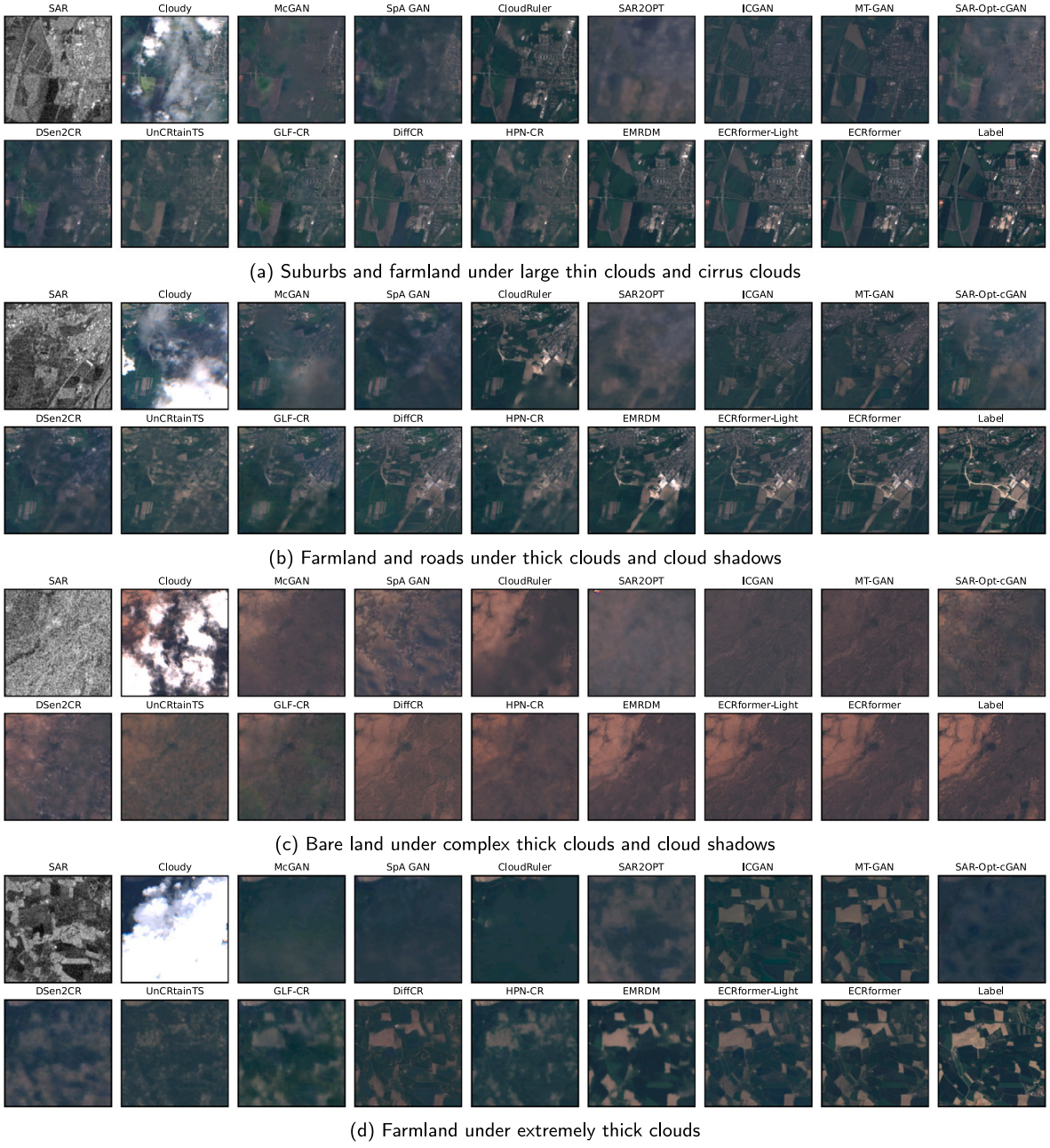

Fig. 5. Qualitative comparison on four challenging scenarios from the SEN12MS-CR testing set. ECRformer excels under thick cloud cover, maintaining fine structures and recovering realistic textures, outperforming all baseline methods.

the feature network used by LPIPS is pre-trained on natural images[1], we compute LPIPS using Sentinel-2 bands B4, B3, and B2 as the RGB channels.

$$\text{LPIPS} = \mathbb{E}\left[\sum_l w_l \left\|\phi_l(Y) - \phi_l(\hat{Y})\right\|_2^2\right], \tag{29}$$

where ϕ_l denotes the feature representation at layer l of pre-trained feature network, and w_l are learned weights.

For PSNR and SSIM, higher values are better. For MAE, SAM, and LPIPS, lower values are better.

[1] https://github.com/richzhang/PerceptualSimilarity

4.2. Evaluation results

4.2.1. Quantitative comparison

As shown in Table 1, ECRformer establishes a new state-of-the-art on both the SEN12MS-CR and cross-domain LuojiaSET-OSFCR benchmarks. On SEN12MS-CR, our model surpasses all baselines across all four conventional reconstruction metrics (MAE, SAM, PSNR, and SSIM). Compared with the previous best diffusion model, EMRDM, ECRformer achieves a 1.23 dB higher PSNR, 8.4% lower MAE, and 10.9% lower SAM. For the perceptual metric LPIPS, ECRformer is competitive with the diffusion-based EMRDM on SEN12MS-CR (0.188 vs. 0.181) and surpasses it on LuojiaSET-OSFCR (0.196 vs. 0.201), demonstrating strong perceptual quality despite requiring only a single forward pass rather than iterative refinement. These advantages are also maintained

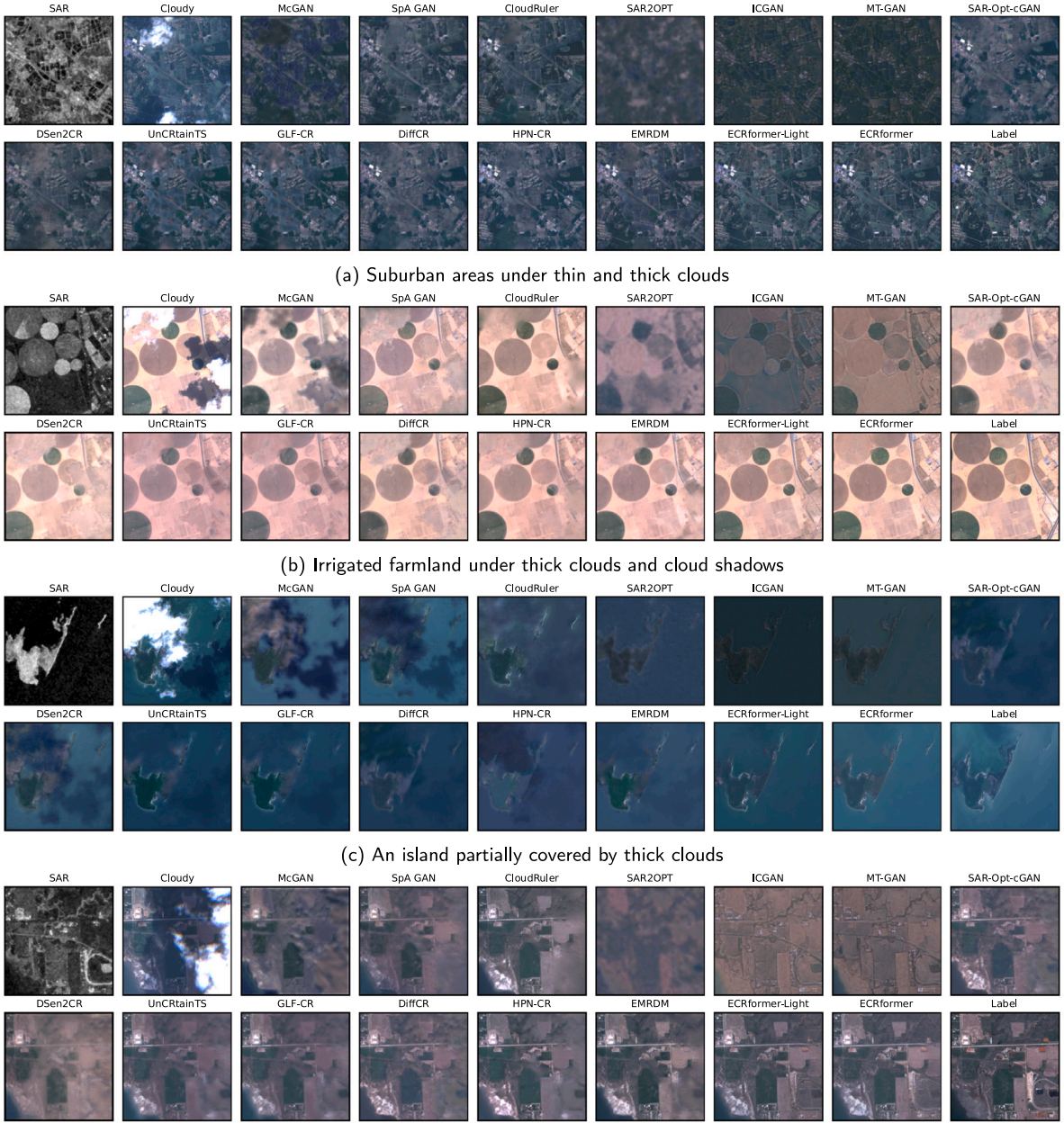

Fig. 6. Cross-domain qualitative comparison on four challenging scenarios from the LuojiaSET-OSFCR dataset. Our ECRformer consistently outperforms all baseline methods, effectively recovering fine structures and realistic textures.

on LuojiaSET-OSFCR across all five metrics, confirming strong cross-domain generalization.

Among methods that use only a single modality, the Transformer-based CloudRuler achieves the best performance in the optical-only category, while the recent MT-GAN leads the SAR-only category. However, both remain substantially below the multimodal fusion methods, confirming the necessity of leveraging complementary optical–SAR information for high-quality cloud removal.

Furthermore, our lightweight version, ECRformer-Light, also g demonstrates exceptional performance. It surpasses most baselines and outperforms the much larger EMRDM model in PSNR, MAE, and SAM, while using only a fraction of the computational resources. This underscores the remarkable efficiency and scalability of our proposed architecture.

4.2.2. *Qualitative comparison*

The superior quantitative performance of ECRformer is mirrored in its qualitative results, as illustrated in Figs. 5 and 6. Our model consistently generates reconstructions with high structural fidelity and realistic textures across different geographical regions.

In scenes with moderate to heavy cloud cover (Fig. 5), while some methods exhibit blurriness or artifacts, ECRformer accurately reconstructs intricate details such as road networks and field boundaries. SAR-only methods such as MT-GAN recover the approximate scene layout but introduce noticeable spectral distortion due to the absence of optical input, whereas the optical-only CloudRuler improves upon earlier single-image methods yet still produces blurred textures under heavy cloud cover. Even under extremely heavy cloud cover (Scenes (c) and (d)), where competing methods produce large-scale artifacts

Table 2
Quantitative comparison of different methods under different cloud coverage conditions (PSNR ↑/SSIM ↑). Best results are highlighted in **bold**, and second-best are underlined.

Method	Venue	Cloud Coverage (%) 0–20	20–40	40–60	60–80	80–100	Avg.
Optical-only							
McGAN	CVPRW'17	27.65 / 0.795	25.81 / 0.761	25.02 / 0.743	23.79 / 0.722	23.43 / 0.699	25.14 / 0.744
SpA GAN	ArXiv'20	27.21 / 0.805	25.43 / 0.772	24.69 / 0.753	23.55 / 0.731	23.02 / 0.709	24.78 / 0.754
CloudRuler	RSE'25	31.26 / 0.924	28.41 / 0.893	27.64 / 0.878	26.58 / 0.869	25.86 / 0.846	27.95 / 0.882
SAR-only							
SAR2OPT	ISPRS'18	26.22 / 0.808	25.95 / 0.799	26.05 / 0.802	25.78 / 0.791	25.35 / 0.765	25.87 / 0.793
ICGAN	PR'22	27.74 / 0.861	27.18 / 0.848	27.56 / 0.856	27.05 / 0.837	26.22 / 0.818	27.15 / 0.844
MT-GAN	ISPRS'25	28.36 / 0.889	27.95 / 0.874	28.22 / 0.883	27.74 / 0.867	27.13 / 0.842	27.88 / 0.871
Multimodal							
SAR-Opt-cGAN	IGARSS'18	28.18 / 0.817	26.23 / 0.781	25.41 / 0.763	24.28 / 0.745	23.85 / 0.714	25.59 / 0.764
DSen2-CR	ISPRS'20	30.82 / 0.915	28.45 / 0.887	27.53 / 0.873	26.31 / 0.858	25.69 / 0.837	27.76 / 0.874
GLF-CR	ISPRS'22	31.78 / 0.924	29.36 / 0.898	28.41 / 0.884	27.18 / 0.871	26.47 / 0.848	28.64 / 0.885
UnCRtainTS	CVPRW'23	32.11 / 0.920	29.69 / 0.894	28.65 / 0.879	27.50 / 0.865	26.55 / 0.842	28.90 / 0.880
DiffCR	TGRS'24	34.95 / 0.938	32.51 / 0.914	31.62 / 0.901	30.15 / 0.887	29.62 / 0.870	31.77 / 0.902
HPN-CR	TGRS'25	33.35 / 0.934	30.93 / 0.910	30.04 / 0.897	28.81 / 0.883	28.02 / 0.866	30.23 / 0.898
EMRDM	CVPR'25	35.43 / 0.954	32.89 / 0.936	31.91 / 0.923	30.58 / 0.911	29.89 / **0.896**	32.14 / 0.924
ECRformer-Light	*Ours*	35.99 / 0.947	33.05 / 0.931	33.08 / 0.925	32.02 / 0.914	29.51 / 0.881	32.75 / 0.920
ECRformer	*Ours*	**36.75** / **0.960**	**33.73** / **0.943**	**33.66** / **0.937**	**32.73** / **0.927**	**30.09** / 0.893	**33.37** / **0.932**

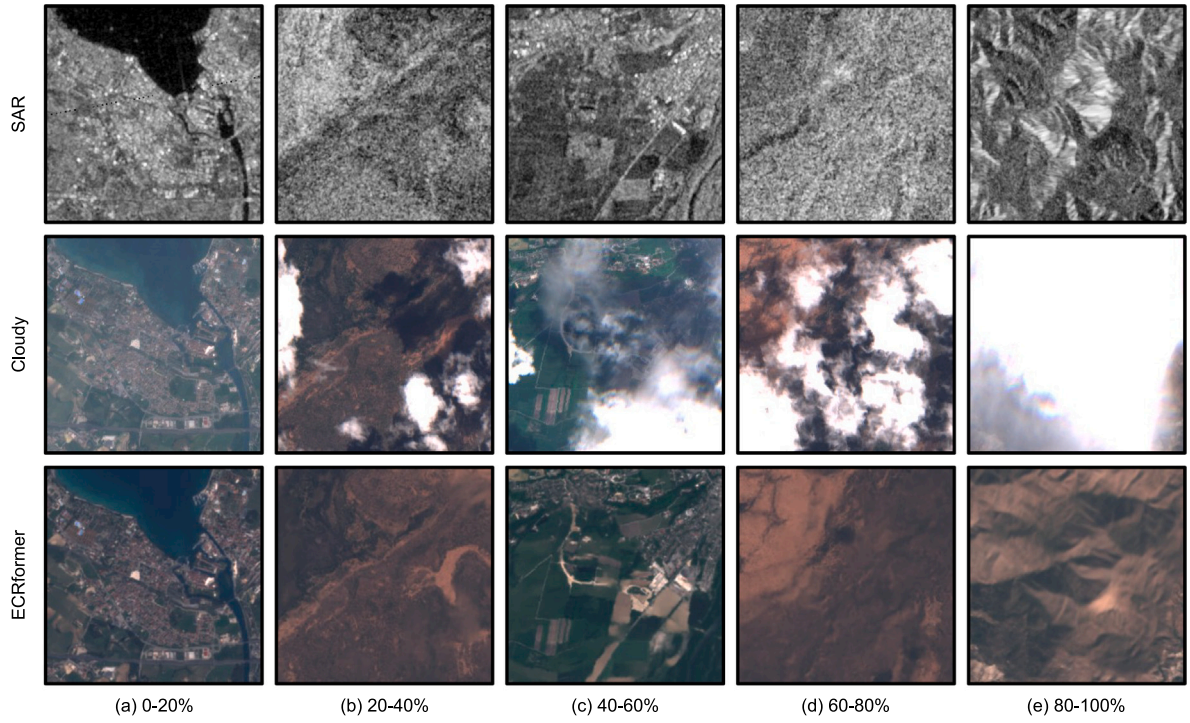

Fig. 7. Visualization of cloud removal results of ECRformer under different cloud cover conditions. Each column corresponds to a specific cloud cover range, showing the input SAR image (row 1), the input cloudy image (row 2), and the reconstructed cloud-free image (row 3). ECRformer consistently reconstructs detailed and realistic textures, even under extreme cloud cover.

or overly smooth regions, ECRformer restores a coherent and detailed landscape.

The cross-domain results on LuojiaSET-OSFCR (Fig. 6) further underscore the model's generalization capability. For instance, in challenging out-of-distribution scenes with thick clouds, such as the island in Scene (c), earlier GAN-based methods such as McGAN suffer from severe artifacts and mode collapse, while single-modality methods (e.g., CloudRuler and MT-GAN) also exhibit clear quality degradation in this unseen domain. In contrast, ECRformer demonstrates strong generalization by effectively handling diverse unseen landscapes and delivering clear, detailed reconstructions. This validates its potential for real-world applications across different geographical contexts.

4.2.3. Evaluation under different cloud coverage

To assess our model's performance across various levels of cloud obscuration, we conduct a comprehensive evaluation under different cloud coverage conditions. As summarized in Table 2 and visualized in Fig. 7, we analyze both quantitative metrics and qualitative results.

Quantitative analysis shows that increasing cloud cover affects different method families in distinct ways. Optical-only and multimodal methods generally deteriorate as cloud density rises, whereas SAR-only methods remain comparatively stable because their inputs are unaffected by optical corruption, although their absolute performance remains clearly below that of strong multimodal approaches. Against this backdrop, ECRformer consistently delivers the best results across all cloud-coverage levels. In particular, under the most challenging

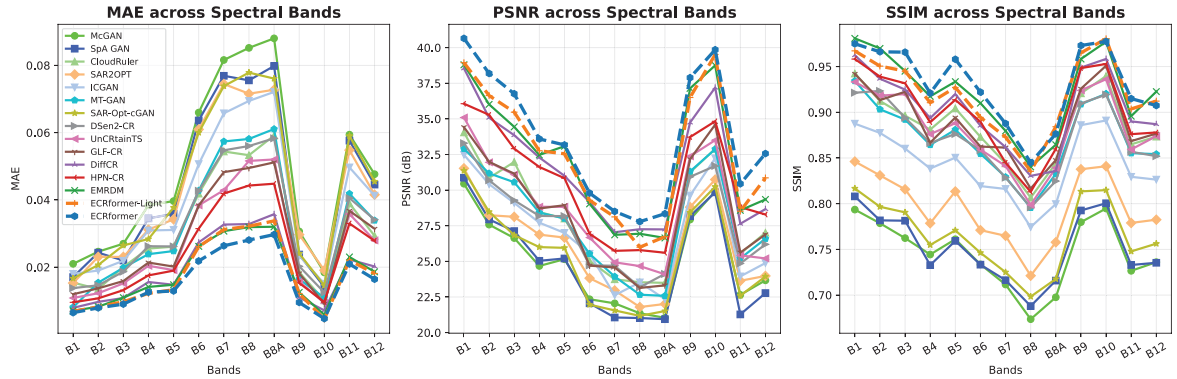

Fig. 8. Per-band quantitative comparison on the SEN12MS-CR testing set. ECRformer consistently outperforms all baselines across all 13 spectral bands, demonstrating superior reconstruction accuracy at the individual band level.

Table 3
Ablation study of ECRformer's core modules. Each component is enabled (✓) or disabled (–).

Configuration	XCA	MDWA	SGFN	PSNR ↑	SSIM ↑
Baseline	–	–	–	31.60	0.901
+ XCA	✓	–	–	32.56	0.919
+ MDWA	–	✓	–	32.19	0.912
+ SGFN	–	–	✓	31.95	0.908
Full (Ours)	✓	✓	✓	**33.37**	**0.932**

Table 4
Ablation study on the Semantic-Decoupled Feature Learning (SDFL) strategy.

Description	Regularization	PSNR ↑	SSIM ↑
No Reg.	None	32.51	0.923
Simple Reg.	Texture → Texture	32.88	0.927
Simple Reg.	Struct. → Struct.	32.95	0.929
SDFL (Reversed)	Texture → Struct.	32.90	0.928
SDFL (Ours)	**Struct. → Texture**	**33.37**	**0.932**

80%–100% cloud-cover bracket, it achieves a PSNR of 30.09 dB, outperforming the next-best method, EMRDM, by 0.20 dB. These results indicate that ECRformer can exploit SAR information effectively when optical observations are heavily degraded, supporting its robustness in diverse and challenging real-world scenarios.

4.2.4. Full spectral comparison results

To validate that our model's superior performance is consistent across the entire spectrum — a critical factor for downstream quantitative analysis — we conduct a per-band evaluation using metrics such as PSNR, SSIM, and MAE. Fig. 8 plots the per-band comparison over the 13 spectral bands of the Sentinel-2 sensor.

The results clearly demonstrate that ECRformer consistently outperforms all baseline methods across every single spectral band. This holds true for both the visible and near-infrared (VNIR) bands, which are critical for visual interpretation, and the short-wave infrared (SWIR) bands, which are vital for applications like vegetation health and soil moisture analysis. This consistent, band-level superiority in reconstruction accuracy is the foundation of the model's excellent overall spectral fidelity, as reflected in the aggregate SAM score presented in Table 1. This comprehensive spectral accuracy is a key advantage of our model, ensuring that the reconstructed data is reliable for scientific research and other quantitative tasks.

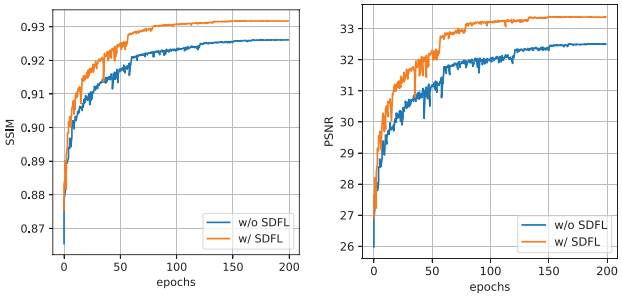

Fig. 9. Convergence curves comparing ECRformer with (blue) and without (orange) SDFL strategy, which accelerates training and leads to better final performance.

4.3. Ablation studies

4.3.1. ECRformer Block

To validate the efficacy of our proposed ECRformer Block, we conducted an ablation study on its core components: Cross-Covariance Attention (XCA), Multi-Dilation Window Attention (MDWA), and Self-Gated Feed-forward Network (SGFN). In our experiments, we used a simple model based on standard window-based attention (WA) and a standard feed-forward neural network (FFN) as the baseline. These standard WA modules were stacked with either XCA or MDWA, respectively, while the standard FFN was replaced by the SGFN, resulting in baseline model variants.

As detailed in Table 3, starting from a baseline model (31.60 dB PSNR), each module brings a significant performance boost when added individually. XCA provides the largest gain (+0.96 dB), underscoring the importance of effective cross-modal fusion. MDWA (+0.59 dB) and SGFN (+0.35 dB) also prove effective in capturing diverse spatial contexts and refining features. The full model, integrating all three, achieves the best performance (33.37 dB PSNR), demonstrating that the components are complementary and collectively enhance reconstruction quality.

4.3.2. Semantic-Decoupled Feature Learning

To validate our proposed Semantic-Decoupled Feature Learning (SDFL) strategy, we conducted an ablation study comparing different intermediate supervision configurations, as shown in Table 4. The baseline model without any regularization already performs well (32.51 dB PSNR). Applying symmetric regularization (either texture or structure loss on both encoder and decoder) improves performance, with structure loss being more effective.

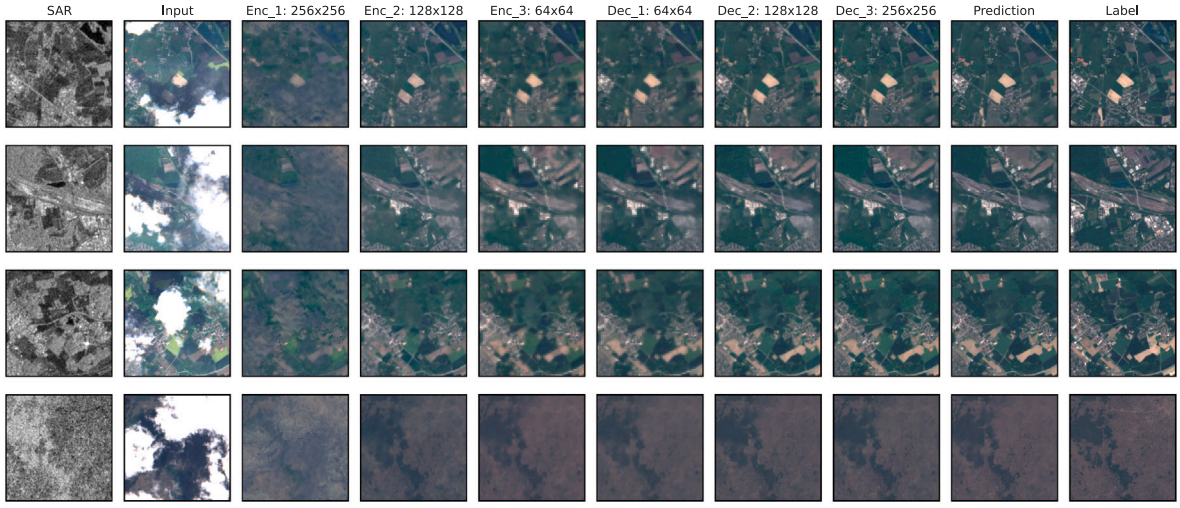

Fig. 10. Visualization of intermediate outputs from different stages of ECRformer. The titles indicate the network stage (e.g. Enc_1, Dec_2), and spatial resolution of each intermediate result. The encoder rapidly reconstructs the overall structure, while the decoder progressively refines textural details.

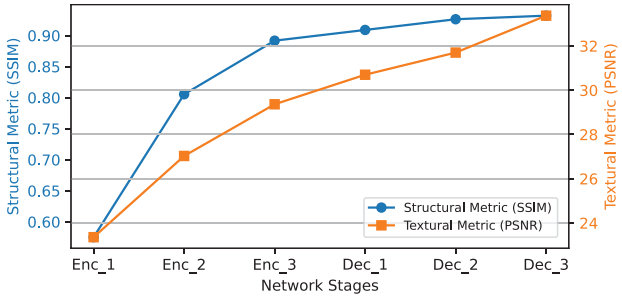

Fig. 11. The structure metric (SSIM) and texture metric (PSNR) of the intermediate results change with the network stage. The encoder quickly establishes structural integrity, while the decoder gradually adds fine textural details.

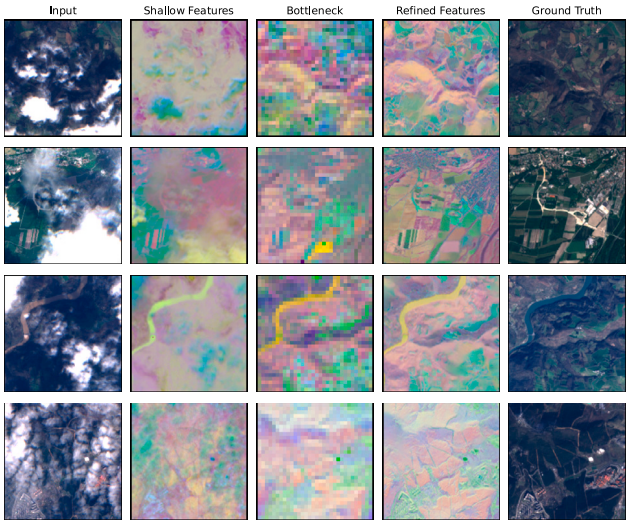

Fig. 12. Visualization of key feature maps. We show features from shallow features (F_{shallow}), bottleneck ($F_{\text{bottleneck}}$), and refined features (F_{refined}). The features evolve from severely missing information to structural recovery, finally reconstructing the fine texture details close to ground truth.

Our proposed asymmetric SDFL strategy, which applies structure loss to the encoder and texture loss to the decoder, achieves the best results, boosting PSNR by 0.86 dB over the baseline. This confirms our hypothesis that tailoring supervision to the semantic role of each network stage is crucial. Reversing the strategy (texture on encoder, structure on decoder) degrades performance, reinforcing that the encoder's primary role is to capture structure. Furthermore, Fig. 9 shows that SDFL not only improves final accuracy but also accelerates convergence. By providing clearer semantic guidance, SDFL enables more efficient learning, enhancing both performance and training speed.

4.4. Further analysis

4.4.1. Progressive refinement

To analyze how ECRformer progressively refines the image, we visualize intermediate outputs from both the encoder and decoder stages. As depicted in Fig. 10, the encoder's initial stages rapidly reconstruct the main structural outlines of the landscape, such as field boundaries and road networks, even from heavily obscured inputs. As the process moves to the decoder, subsequent stages progressively introduce finer details, enhancing textural realism. For instance, the texture of vegetation and subtle variations in terrain are gradually filled in, leading to a final image that is both structurally accurate and visually plausible.

This process is quantitatively validated by the line chart in Fig. 11. The data shows that in the encoder stage (left half of the chart), the curve for the structural metric, SSIM, rises sharply and then plateaus, indicating that the model quickly establishes the overall structural integrity. In contrast, the growth of the texture-oriented metric, PSNR, is more gradual but continuous. Upon entering the decoder stage (right half of the chart), the SSIM shows little change, while the PSNR continues to increase steadily, peaking at the final output. This quantitative evidence strongly supports our qualitative observations and validates the effectiveness of the SDFL strategy, where the network prioritizes learning coarse structure before refining fine-grained textures, leading to efficient and effective cloud removal.

4.4.2. Semantic-decoupled representation

To provide qualitative evidence for the proposed structure-texture decoupling, we visualize intermediate representations produced by different stages of ECRformer. Specifically, we extract (i) shallow features (F_{shallow}) from the Shallow Feature Embedding, (ii) bottleneck features ($F_{\text{bottleneck}}$), and (iii) refined features (F_{refined}) from the Refinement Network. Since these representations are high-dimensional feature tensors, we project each feature map to a three-channel image using PCA

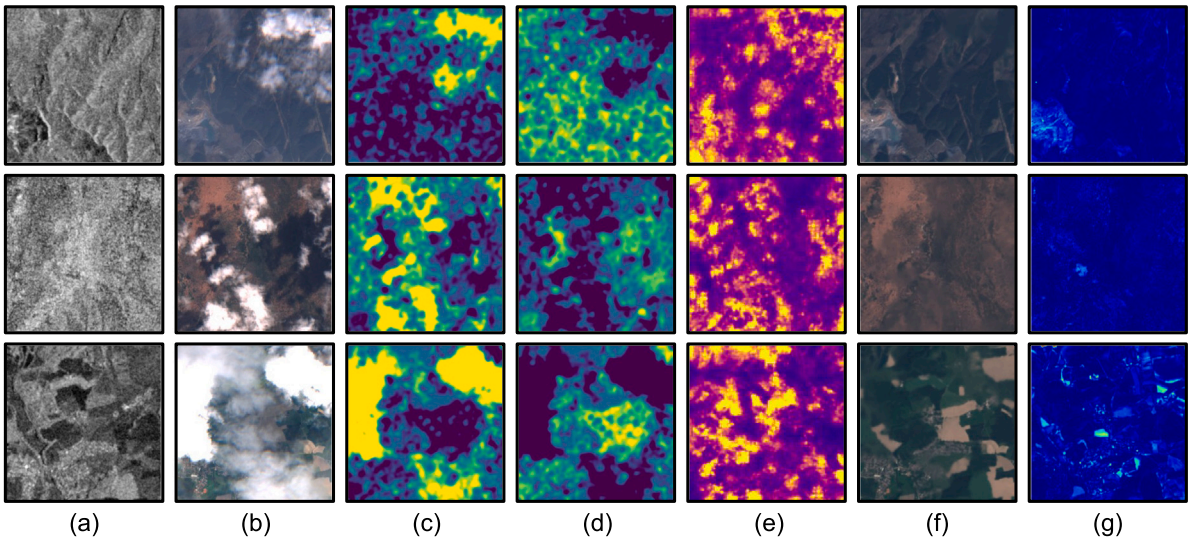

Fig. 13. Visualization of attention attribution maps for the first XCA and MDWA modules. (a) SAR inputs, (b) Optical inputs, (c) Contribution of SAR, (d) Contribution of optical, (e) Selection of receptive field, (f) Refined outputs, and (g) Error maps.

along the channel dimension and visualize the first three principal components as RGB.

As shown in Fig. 12, the shallow features still preserve the cloud/shadow degradations inherited from the cloudy input. In contrast, the bottleneck features emphasize the coarse structural layout (e.g., large-scale geometry and boundaries) but exhibit noticeably reduced high-frequency textures, consistent with a more structure-oriented encoding. Finally, the refined features progressively recover realistic textures and spectral appearance, yielding representations that are visually closer to the ground truth. Although PCA visualization is inherently qualitative, these observations support our design rationale that the encoder tends to prioritize structural content while the refinement stage is responsible for texture restoration.

4.4.3. Modality contribution

To analyze the contribution of the optical and SAR modalities, we use a gradient-based attribution method (Selvaraju et al., 2017; Wang et al., 2024) to visualize the saliency of each input. As shown in Fig. 13(c)–(d), where brighter areas indicate a greater contribution from that modality, the results reveal a clear and intelligent pattern of modal fusion. In regions completely obscured by thick clouds, where optical information is absent, the model's output is almost entirely dominated by contributions from the SAR data, which provides crucial all-weather structural information. Conversely, in clear or cloud-free areas, the model relies primarily on the high-quality optical input. For intermediate zones, such as areas with thin clouds or at cloud edges, both modalities contribute synergistically. The model leverages the residual textural information from the optical data while using SAR to supplement and correct structural details. This demonstrates that our XCA module can accurately identify the local degradation level and flexibly orchestrate the appropriate fusion of modalities for optimal image restoration.

4.4.4. Adaptive receptive field selection

We further analyze the effectiveness of the adaptive receptive field selection mechanism in MDWA. Fig. 13(e) shows the receptive field selection heatmap, where brighter areas indicate a preference for a larger receptive field (i.e., a larger dilation rate). The map reveals that MDWA adaptively adjusts its focus based on the image content and cloud characteristics. In regions with thick, homogeneous clouds or complex, large-scale ground structures, it tends to select a larger receptive field to capture a wider range of contextual information for robust reconstruction. In contrast, for clear areas with fine-grained details or simple textures, it narrows the receptive field to focus on capturing local, high-frequency information. This dynamic, content-aware behavior confirms the model's ability to efficiently allocate computational resources and effectively handle the diverse spatial frequencies present in various cloud and landscape conditions.

Table 5
Performance vs. Efficiency. ECRformer models offer superior accuracy with significantly lower computational cost and faster training convergence than diffusion-based SOTA methods.

Method	Performance ↑ PSNR/SSIM	Complexity ↓ Params/FLOPs	Training time ↓ (GPU hours)
DSen2-CR	27.76/0.874	18.95M/1241.18G	212.9
GLF-CR	28.64/0.885	14.83M/249.71G	142.4
UnCRtainTS	28.90/0.880	0.52M/28.56G	89.5
DiffCR	31.77/0.902	22.91M/45.86G	396.0
HPN-CR	30.23/0.898	3.69M/19.61G	130.4
EMRDM	32.14/0.924	39.13M/417.85G	231.7
ECRformer-Light	32.75/0.920	3.70M/35.78G	78.4
ECRformer	**33.37/0.932**	11.29M/102.47G	142.1

4.4.5. Efficiency analysis

A central claim of our work is efficiency, which is substantiated in Table 5 and visually represented in Fig. 14. ECRformer achieves the highest accuracy while being significantly more efficient than the top-performing diffusion models. Computational complexity is evaluated by floating point operations (FLOPs), specifically estimated using the "fvcore"[2] library based on processing a single 256 × 256 image. For diffusion-based models, we report the total FLOPs across all sampling steps to ensure a fair comparison. It uses only 28.9% of the parameters and 24.5% of the FLOPs of EMRDM, yet delivers a 1.23 dB higher PSNR. In terms of training cost, ECRformer converges in 142.1 GPU hours, approximately 61% of EMRDM's 231.7 GPU hours and 36% of DiffCR's 396.0 GPU hours, while ECRformer-Light requires only 78.4 GPU hours.

The efficiency of our architecture is further highlighted by ECRformer-Light. It achieves a PSNR of 32.75 dB, surpassing EMRDM, with merely 9.5% of its parameters and 8.6% of its FLOPs. Compared to HPN-CR, a competitive baseline with a comparable parameter

[2] https://github.com/facebookresearch/fvcore

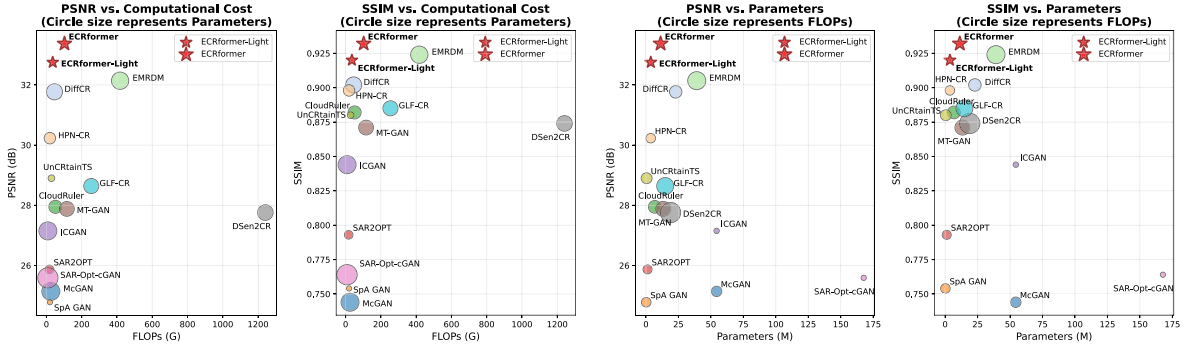

Fig. 14. Trade-off analysis between performance and computational cost for different cloud removal methods (top left is better). The plots compare PSNR/SSIM against model parameters/FLOPs. Our ECRformer method (red star) demonstrates superior efficiency, achieving high accuracy with competitive computational cost.

Table 6
Key hyperparameter settings and statistics information of ECRformer and ECRformer-Light.

Configuration	ECRformer	ECRformer-Light
Network Architecture		
Input Channels (C_{in})	13 + 2	13 + 2
Output Channels (C_{out})	13	13
Embedding Dimension (C_{embed})	48	32
Encoder Stages (K_{enc})	3	3
Decoder Stages (K_{dec})	3	3
Depth per Stage (L)	[2, 3, 2]	[2, 2, 1]
Blocks in Bottleneck (L_B)	2	1
Refinement Depth (L_R)	4	2
Core Components		
XCA Attention Heads	4	4
XCA Channel per Head (C_{head})	24	16
MDWA Window Size (M)	3	3
MDWA Dilation Rates (d)	[1, 2, 3, 4]	[1, 2, 3, 4]
SGFN Hidden Expansion	2.0	2.0
Training Hyperparameters		
α (Output Loss)	0.9	0.9
β (Output Loss)	0.1	0.1
λ_{enc} (SDFL)	0.05	0.05
λ_{dec} (SDFL)	0.05	0.05
Model Complexity		
Parameters (M) ↓	11.29	3.70
FLOPs (G) ↓	102.47	35.78

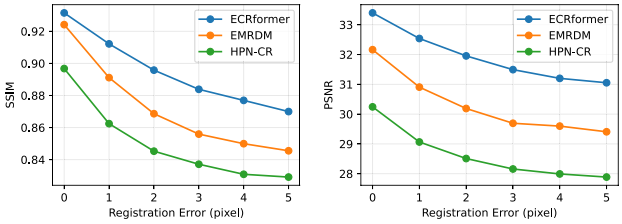

Fig. 15. Robustness analysis of representative multimodal models under different levels of SAR-optical registration error.

count, ECRformer-Light offers a dramatic 2.52 dB PSNR improvement with only a modest increase in FLOPs. This clearly demonstrates that our ECRformer architecture establishes a new Pareto frontier for the accuracy-efficiency trade-off in cloud removal, making high-quality reconstruction more practical for real-world deployment.

4.4.6. Robustness of registration error

The ideal scenario assumes perfect co-registration between SAR and optical images. However, in real-world applications, slight misalignments may occur due to differences in sensor geometry, acquisition times, or atmospheric conditions. To assess robustness against registration errors, we introduce controlled random translations of up to ±5 pixels to the SAR input during inference — without any corresponding augmentation during training — and compare three representative multimodal models: ECRformer, EMRDM, and HPN-CR.

As shown in Fig. 15, PSNR and SSIM of all three models decrease monotonically as misalignment increases, yet ECRformer consistently achieves the highest scores at every error level and exhibits the smallest relative degradation. This resilience can be attributed to XCA's channel-wise fusion mechanism and MDWA's multi-scale spatial modeling, which together enable effective cross-modal feature alignment even under imperfect registration. These results confirm that ECRformer possesses a reasonable tolerance to the registration errors encountered in practical deployment.

4.4.7. Hyperparameters

The key hyperparameters for the proposed ECRformer and ECRformer-Light are detailed in Table 6. These settings are optimized to balance model capacity, computational efficiency, and task-specific requirements.

ECRformer's superior performance stems from a larger **embedding dimension** (48 vs. 32) and greater **network depth** (9 vs. 6 blocks), which provide a richer feature space and stronger modeling capabilities. This increased capacity results in approximately three times the parameters and FLOPs but yields significant gains across all metrics. The **core components** are designed for efficiency: MDWA uses a fixed window with multi-level dilation to capture diverse spatial contexts without extra cost, while other parameters scale with the embedding dimension or follow standard practices.

In our **training strategy**, the loss function weights ($\alpha = 0.9, \beta = 0.1$) prioritize pixel-level accuracy (L1 loss) while preserving structural similarity (SSIM loss). The small SDFL loss weights ($\lambda_{\text{enc}} = \lambda_{\text{dec}} = 0.05$) provide effective semantic guidance without excessive regularization.

Overall, ECRformer-Light offers an excellent performance-efficiency trade-off for resource-constrained scenarios. The scalability of the architecture is proven by ECRformer, which achieves a higher performance ceiling with more resources. This validates that our design is effective and adaptable across different model capacities.

5. Conclusion

In this paper, we addressed the critical challenge of balancing accuracy and efficiency in multimodal remote sensing cloud removal. We proposed ECRformer, a solution that pairs a highly efficient architecture with a principled learning paradigm and surpasses prior state-of-the-art methods in both reconstruction quality and computational efficiency. Our key contributions, which are the Semantic-Decoupled Feature Learning (SDFL) paradigm and a suite of efficient attention

modules (XCA and MDWA), enable the model to achieve superior results with significantly fewer parameters and FLOPs than recent, computationally intensive diffusion models. The empirical results on the SEN12MS-CR and LuojiaSET-OSFCR datasets validate our approach, establishing ECRformer as a new, powerful, and practical baseline for cloud removal.

Although our method demonstrates strong performance and efficiency, there are still areas for future improvement. First, while SDFL effectively decomposes the learning task, further exploration into more sophisticated loss functions or multi-task learning strategies could enhance the disentanglement of structure and texture features. Second, although our attention mechanisms are efficient, investigating alternative architectures or hybrid models that combine the strengths of CNNs and Transformers may yield additional gains in both speed and accuracy. Third, the recent emergence of large-scale visual foundation models presents a promising direction; their pre-trained representations could potentially serve as powerful priors for image reconstruction tasks, including cloud removal, and their integration with task-specific architectures like ours warrants future investigation. Finally, extending our approach to handle other types of remote sensing data or environmental conditions could broaden its applicability and robustness.

CRediT authorship contribution statement

Zaiyan Zhang: Writing – original draft, Visualization, Software, Methodology, Data curation. **Jie Li:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Yuanqi Liang:** Validation, Software, Investigation. **Jining Yan:** Validation, Software, Investigation. **Yi Xiao:** Validation, Software, Investigation. **Xin Su:** Validation, Software, Investigation. **Qiangqiang Yuan:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM910 and the National Natural Science Foundation of China under Grants 42471504 and 42230108.

References

Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., et al., 2021. XCiT: Cross-covariance image transformers. Adv. Neural Inf. Process. Syst. 34, 20014–20027.

Bermudez, J., Happ, P., Oliveira, D., Feitosa, R., 2018. SAR to optical image synthesis for cloud removal with generative adversarial networks. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. 4, 5–11.

Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C., 2000. Image inpainting. In: Proceedings of the ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques. pp. 417–424.

Cai, J., Huang, B., Liu, H., 2025. Fusing sentinel-1 and sentinel-2 data with diffusion models for cloud removal. Remote Sens. Environ. 331, 115049.

Chen, Y., Tang, L., Yang, X., Fan, R., Bilal, M., Li, Q., 2019. Thick clouds removal from multitemporal ZY-3 satellite images using deep learning. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 13, 143–153.

Chen, Y., Yuan, Q., Xie, H., Tang, Y., Xiao, Y., He, J., Guan, R., Liu, X., Zhang, L., 2025. Hyperspectral video tracking with spectral-spatial fusion and memory enhancement. IEEE Trans. Image Process..

Chen, J., Zhu, X., Vogelmann, J.E., Gao, F., Jin, S., 2011. A simple and effective method for filling gaps in landsat ETM+ SLC-off images. Remote Sens. Environ. 115 (4), 1053–1064.

Christopoulos, D., Ntouskos, V., Karantzalos, K., 2022. Cloudtran: Cloud removal from multitemporal satellite images using axial transformer networks. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 43, 1125–1132.

Chu, X., Tian, Z., Zhang, B., Wang, X., Shen, C., 2023. Conditional positional encodings for vision transformers. In: International Conference on Learning Representations. pp. 1–19, URL: https://openreview.net/forum?id=3KWnuT-R1bh.

Criminisi, A., Perez, P., Toyama, K., 2004. Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. 13 (9), 1200–1212. http://dx.doi.org/10.1109/TIP.2004.833105.

Dai, J., Shi, N., Zhang, T., Xu, W., 2024. TCME: Thin cloud removal network for optical remote sensing images based on multi-dimensional features enhancement. IEEE Trans. Geosci. Remote Sens..

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations. pp. 1–22.

Ebel, P., Garnot, V.S.F., Schmitt, M., Wegner, J.D., 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. pp. 2086–2096.

Ebel, P., Meraner, A., Schmitt, M., Zhu, X.X., 2020. Multisensor data fusion for cloud removal in global and all-season sentinel-2 imagery. IEEE Trans. Geosci. Remote Sens. 59 (7), 5866–5878.

Enomoto, K., Sakurada, K., Wang, W., Fukui, H., Matsuoka, M., Nakamura, R., 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. pp. 48–56.

Gao, G., Gu, Y., 2017. Multitemporal landsat missing data recovery based on tempo-spectral angle model. IEEE Trans. Geosci. Remote Sens. 55 (7), 3656–3668.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Grohnfeldt, C., Schmitt, M., Zhu, X., 2018. A conditional generative adversarial network to fuse SAR and multispectral optical data for cloud removal from sentinel-2 images. In: Proceedings of the IEEE International Geoscience and Remote Sensing Symposium. IEEE, pp. 1726–1729.

Gu, P., Liu, W., Feng, S., Wei, T., Wang, J., Chen, H., 2025. HPN-CR: Heterogeneous parallel network for SAR-optical data fusion cloud removal. IEEE Trans. Geosci. Remote Sens..

He, K., Sun, J., 2014. Image completion approaches using the statistics of similar patches. IEEE Trans. Pattern Anal. Mach. Intell. 36 (12), 2423–2435. http://dx.doi.org/10.1109/TPAMI.2014.2330611.

He, J., Yuan, Q., Li, J., Xiao, Y., Zhang, L., 2023. A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection. ISPRS J. Photogramm. Remote Sens. 204, 131–144.

Hendrycks, D., Gimpel, K., 2016. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415.

Jing, R., Duan, F., Lu, F., Zhang, M., Zhao, W., 2023. Denoising diffusion probabilistic feature-based network for cloud removal in sentinel-2 imagery. Remote. Sens. 15 (9), 2217.

King, M.D., Platnick, S., Menzel, W.P., Ackerman, S.A., Hubanks, P.A., 2013. Spatial and temporal distribution of clouds observed by MODIS onboard the terra and aqua satellites. IEEE Trans. Geosci. Remote Sens. 51 (7), 3826–3852.

Kingma, D., 2015. Adam: a method for stochastic optimization. In: International Conference on Learning Representations. pp. 1–15.

Kruse, F.A., Lefkoff, A., Boardman, y.J., Heidebrecht, K., Shapiro, A., Barloon, P., Goetz, A., 1993. The spectral image processing system (SIPS)—interactive visualization and analysis of imaging spectrometer data. Remote Sens. Environ. 44 (2–3), 145–163.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1989. Handwritten digit recognition with a back-propagation network. Adv. Neural Inf. Process. Syst. 2.

Li, J., Shi, S., Lin, L., Yuan, Q., Shen, H., Zhang, L., 2025a. A multi-task learning framework for dual-polarization SAR imagery despeckling in temporal change detection scenarios. ISPRS J. Photogramm. Remote Sens. 221, 155–178.

Li, J., Wang, Y., Sheng, Q., Wu, Z., Wang, B., Ling, X., Liu, X., Du, Y., Gao, F., Camps-Valls, G., et al., 2025b. CloudRuler: Rule-based transformer for cloud removal in landsat images. Remote Sens. Environ. 328, 114913.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. SwinIR: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 1833–1844.

Liu, H., Huang, B., Cai, J., 2023. Thick cloud removal under land cover changes using multisource satellite imagery and a spatiotemporal attention network. IEEE Trans. Geosci. Remote Sens. 61, 1–18.

Liu, Y., Li, W., Guan, J., Zhou, S., Zhang, Y., 2025. Effective cloud removal for remote sensing images by an improved mean-reverting denoising model with elucidated design space. arXiv preprint arXiv:2503.23717.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, R., Luo, T., Huang, S., Wu, Y., Jiang, Z., Zhang, H., 2024. CrossMatch: Cross-view matching for semi-supervised remote sensing image segmentation. IEEE Trans. Geosci. Remote Sens. 62, 1–15. http://dx.doi.org/10.1109/TGRS.2024.3507050.

Liu, J., Musialski, P., Wonka, P., Ye, J., 2012. Tensor completion for estimating missing values in visual data. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 208–220.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. In: International Conference on Learning Representations. pp. 1–19.

Ma, J., Chen, Y., Pan, J., Xu, J., Li, Z., Xu, R., Chen, R., 2024. SCT-CR: A synergistic convolution-transformer modeling method using SAR-optical data fusion for cloud removal. Int. J. Appl. Earth Obs. Geoinf. 130, 103909.

Malek, S., Melgani, F., Bazi, Y., Alajlan, N., 2017. Reconstructing cloud-contaminated multispectral images with contextualized autoencoder neural networks. IEEE Trans. Geosci. Remote Sens. 56 (4), 2270–2282.

Meraner, A., Ebel, P., Zhu, X.X., Schmitt, M., 2020. Cloud removal in sentinel-2 imagery using a deep residual neural network and SAR-optical data fusion. ISPRS J. Photogramm. Remote Sens. 166, 333–346.

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

Pan, J., Xu, J., Yu, X., Ye, G., Wang, M., Chen, Y., Ma, J., 2024. HDRSA-Net: Hybrid dynamic residual self-attention network for SAR-assisted optical image cloud and shadow removal. ISPRS J. Photogramm. Remote Sens. 218, 258–275.

Pathak, D., Krähenbühl, P., Donahue, J., Darrell, T., Efros, A.A., 2016. Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2536–2544. http://dx.doi.org/10.1109/CVPR.2016.278.

Sarukkai, V., Jain, A., Uzkent, B., 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. pp. 1796–1805.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626.

Shao, M., Wang, C., Zuo, W., Meng, D., 2022. Efficient pyramidal GAN for versatile missing data reconstruction in remote sensing images. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li, H., Zhang, L., 2015. Missing information reconstruction of remote sensing data: A technical review. IEEE Geosci. Remote. Sens. Mag. 3 (3), 61–85.

Shen, H., Meng, X., Zhang, L., 2016. An integrated framework for the spatio-temporal-spectral fusion of remote sensing images. IEEE Trans. Geosci. Remote Sens. 54 (12), 7135–7148. http://dx.doi.org/10.1109/TGRS.2016.2596290.

Shu, Q., Zhu, X., Xu, S., Wang, Y., Liu, D., 2025. RESTORE-DiT: Reliable satellite image time series reconstruction by multimodal sequential diffusion transformer. Remote Sens. Environ. 328, 114872.

Stucker, C., Garnot, V.S.F., Schindler, K., 2023. U-TILISE: A sequence-to-sequence model for cloud removal in optical satellite time series. IEEE Trans. Geosci. Remote Sens. 61, 1–16.

Sui, J., Ma, Y., Yang, W., Zhang, X., Pun, M.-O., Liu, J., 2024. Diffusion enhancement for cloud removal in ultra-resolution remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 62, 1–14.

Sun, L., Jin, M., Yan, J., He, H., Cao, L., 2025. Semantic-TemporalNet: A novel urban block change detection method based on semantic coherence analysis. IEEE Trans. Geosci. Remote Sens..

Sun, L., Zhang, Y., Chang, X., Wang, Y., Xu, J., 2019. Cloud-aware generative network: Removing cloud from optical remote sensing images. IEEE Geosci. Remote. Sens. Lett. 17 (4), 691–695.

Tu, K., Yang, C., Qing, Y., Qi, K., Chen, N., Gong, J., 2025. Cloud removal with optical and SAR imagery via multimodal similarity attention. ISPRS J. Photogramm. Remote Sens. 226, 116–126.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 30.

Wan, Y., Li, J., Lin, L., Yuan, Q., Shen, H., 2025. Collaboration of dehazing and object detection tasks: A multi-task learning framework for foggy image. IEEE Trans. Geosci. Remote Sens..

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Wang, P., Chen, Y., Huang, B., Zhu, D., Lu, T., Dalla Mura, M., Chanussot, J., 2025. MT_GAN: A SAR-to-optical image translation method for cloud removal. ISPRS J. Photogramm. Remote Sens. 225, 180–195.

Wang, Y., Zhang, T., Guo, X., Shen, Z., 2024. Gradient based feature attribution in explainable ai: A technical review. arXiv preprint arXiv:2403.10415.

Xu, M., Deng, F., Jia, S., Jia, X., Plaza, A.J., 2022. Attention mechanism-based generative adversarial networks for cloud removal in landsat images. Remote Sens. Environ. 271, 112902.

Xu, F., Shi, Y., Ebel, P., Yu, L., Xia, G.-S., Yang, W., Zhu, X.X., 2022. GLF-CR: SAR-enhanced cloud removal with global–local fusion. ISPRS J. Photogramm. Remote Sens. 192, 268–278.

Yang, X., Zhao, J., Wei, Z., Wang, N., Gao, X., 2022. SAR-to-optical image translation based on improved CGAN. Pattern Recognit. 121, 108208.

Yu, F., Koltun, V., 2016. Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations. pp. 1–13.

Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2019. Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4471–4480.

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., et al., 2020. Deep learning in environmental remote sensing: Achievements and challenges. Remote Sens. Environ. 241, 111716.

Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., 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. pp. 5728–5739.

Zeng, C., Shen, H., Zhang, L., 2013. Recovering missing pixels for landsat ETM+ SLC-off imagery using multi-temporal regression analysis and a regularization method. Remote Sens. Environ. 131, 182–194.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 586–595.

Zhang, C., Li, W., Travis, D., 2007. Gaps-fill of SLC-off landsat ETM+ satellite image using a geostatistical approach. Int. J. Remote Sens. 28, 5103–5122. http://dx.doi.org/10.1080/01431160701250416.

Zhang, Z., Yan, J., Liang, Y., Feng, J., He, H., Cao, L., 2025. Multiscale restoration of missing data in optical time-series images with masked spatial-temporal attention network. IEEE Trans. Geosci. Remote Sens. 63, 1–15. http://dx.doi.org/10.1109/TGRS.2025.3574799.

Zhang, Q., Yuan, Q., Zeng, C., Li, X., Wei, Y., 2018. Missing data reconstruction in remote sensing image with a unified spatial–temporal–spectral deep convolutional neural network. IEEE Trans. Geosci. Remote Sens. 56 (8), 4274–4288. http://dx.doi.org/10.1109/TGRS.2018.2810208.

Zheng, J., Liu, X.-Y., Wang, X., 2020. Single image cloud removal using U-Net and generative adversarial networks. IEEE Trans. Geosci. Remote Sens. 59 (8), 6371–6385.

Zou, X., Li, K., Xing, J., Zhang, Y., Wang, S., Jin, L., Tao, P., 2024. DiffCR: A fast conditional diffusion framework for cloud removal from optical satellite images. IEEE Trans. Geosci. Remote Sens. 62, 1–14.