Dual-view Prompting for Cloud Removal

Ye Deng, *Member, IEEE,* Wenli Huang, ZiXin Tang and Jiang Duan

***Abstract*—Cloud cover significantly impedes the utilization of remote sensing data, limiting the effectiveness of satellite imagery in critical applications such as environmental monitoring and disaster response. While deep learning methods have advanced cloud removal, existing models predominantly focus on spatial-domain feature discrepancies, often overlooking distinctive spectral difference introduced by clouds. To address this gap, we propose a Dual-view Prompting Network (DVPNet) that integrates spatial and frequency information via prompt learning to generate robust guidance features. The core innovation, the Dual-view Prompting Block (DVPB), operates cascadedly: first, a spatial gating module refines features to capture contextual cues; these features are then transformed into the Fourier domain, where a frequency-gating structure and a learnable spectral prompt further calibrate and enhance representations. The holistically refined dual-view prompt is integrated into the decoder through an efficient windowed cross-attention mechanism, enabling precise cloud removal. Extensive experiments on benchmark datasets demonstrate that DVPNet achieves state-of-the-art performance. This work validates the critical role of frequency-domain modeling in cloud removal and establishes a new spatial-frequency collaborative paradigm for remote sensing image restoration. The code will be made available at https://github.com/huangwenwenlili/DVPNet.**

***Index Terms*—Cloud removal; Spectral–spatial analysis; Dual-view prompting network; Remote sensing image restoration.**

I. Introduction

APPROXIMATELY 60% of the Earth's surface is obscured by clouds daily, disrupting remote sensing observations and resulting in significant loss of ground object information in acquired imagery [1]. This limits the usability of data for applications such as environmental monitoring, disaster early warning, and agricultural yield estimation [2]. As a fundamental challenge in remote sensing, advancements in cloud removal technology are essential to improve data quality.

Recent progress in deep learning has significantly advanced image restoration techniques, particularly for cloud removal. Numerous end-to-end methods leveraging convolutional neural networks (CNNs) [3], generative adversarial networks (GANs) [4], and Transformer architectures [5] have been developed to learn the mapping from cloudy to cloud-free images. While these methods have shown notable improvements, they often share a key limitation: their designs primarily focus on spatial-domain information, neglecting the valuable frequency-domain data. Unlike spatial-domain processing, which emphasizes local pixel relationships, the differences between cloudy and cloud-free images are often more clearly defined in the frequency domain, where global spectral-structural characteristics are more evident. This limitation becomes especially apparent in complex cloud scenarios. For example, thin clouds typically cause low-frequency intensity modulation, while thick clouds introduce high-frequency structural disturbances. These variations are difficult to analyze effectively using spatial-domain models alone, often resulting in spectral distortion, blurred textures, or structural discontinuities in the restored images. Therefore, an in-depth investigation into the intrinsic correlation between cloud obscuration and frequency characteristics is crucial for a comprehensive understanding and effective resolution of the cloud removal problem.

In this paper, we first experimentally reveal the intricate relationship between cloud obscuration and the characteristics of amplitude and phase spectra of images in the frequency domain, as illustrated in Figure 1. Specifically, we transform spatial-domain images into the frequency domain using Fast Fourier Transform (FFT) to separate their amplitude and phase spectra. Subsequently, we swap the amplitude and phase spectra between cloudy images and their corresponding cloud-free counterparts, and then convert the modified spectral information back to the spatial domain using Inverse Fast Fourier Transform (IFFT). This allows us to observe the impact of different spectral components on cloud representation. From Figure 1, we can clearly observe the following phenomena: (1) For images obscured by both thin clouds and thick clouds, their amplitude spectra exhibit significant differences compared to their corresponding cloud-free images. This suggests that the amplitude spectrum carries substantial information regarding cloud intensity and distribution. (2) For thin cloud images, the difference in phase spectra compared to their cloud-free counterparts is relatively minor. However, for thick cloud images, a noticeable difference also exists in their phase spectra, implying that the phase spectrum is also critical for describing the structural characteristics of complex or dense clouds. Consequently, as shown by the results of the swapping experiments in Figure 1, simply swapping phase spectra has a limited visual impact on thin cloud images, further confirming that thin cloud information is predominantly concentrated in the amplitude spectrum. Conversely, for thick cloud images, the image degradation caused by clouds is diffused across both

Manuscript was received on June 8, 2025, date of current version June 8, 2025. This work was supported by the Fundamental Research Funds for the the Sichuan Science Foundation project under Grant 2024ZDZX0002 and Grant 2024NSFTD0054, and the Public Welfare Research Program of Ningbo City under Grant 2024S063. (Corresponding author: ZiXin Tang.)

Ye Deng, ZiXin Tang and Jiang Duan are with the Engineering Research Center of Intelligent Finance, Ministry of Education, School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Wenjiang, Chengdu, Sichuan, 611130, China (e-mail: dengye@swufe.edu.cn;sctangzixing123@163.com;duanj_t@swufe.edu.cn).

Wenli Huang is with the School of Electronic and Information Engineering, Ningbo University of Technology, Jiangbei, Ningbo, Zhejiang, 315211, China (e-mail: huangwenwenlili@stu.xjtu.edu.cn).

ZiXin Tang is also with Kash Institute of Electronics and Information Industry, Kash, China.

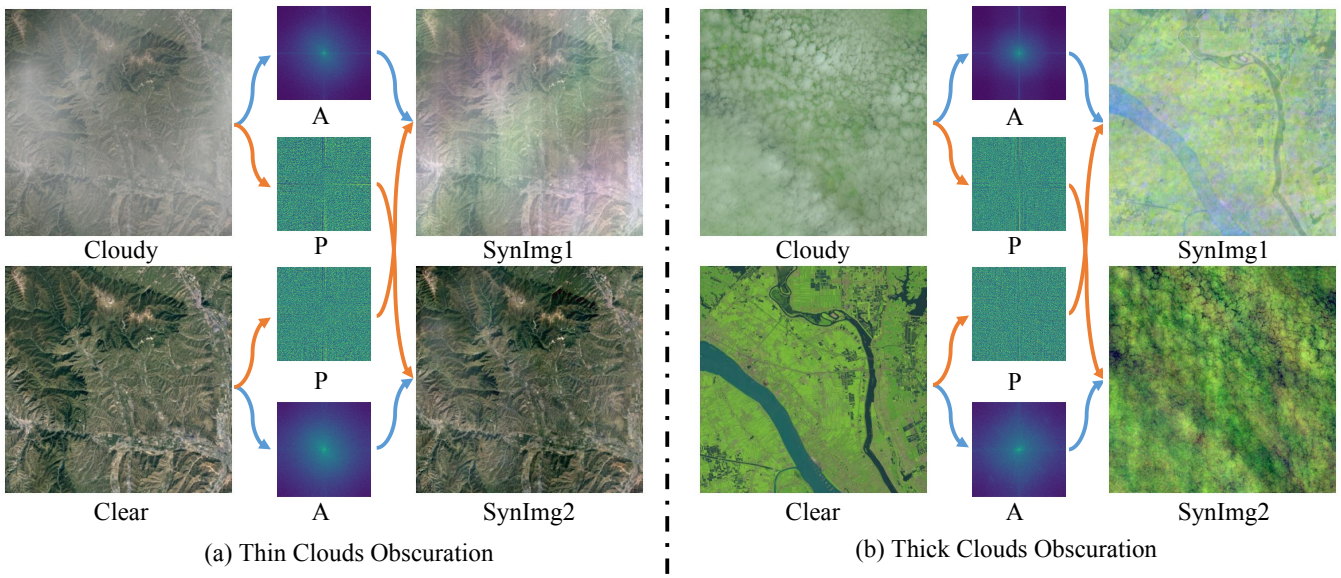


Fig. 1: Visualization of cloud obscuration effects on image frequency-domain characteristics. **Amplitude (A)** and **Phase (P)** spectra are shown, alongside Synthesized images reconstructed after phase spectrum swapping. (a) **Thin clouds**: Swapping phase spectra between thin-cloud and clean images produces minimal visual changes in the synthesized images. (b) **Thick clouds**: Swapping phase spectra between thick-cloud and clean images causes significant changes. The synthesized thick-cloud image recovers partial ground-feature details, while the synthesized clean image exhibits cloud-like artifacts.

amplitude and phase spectra. Swapping phase spectra can, to some extent, remove clouds but may also affect ground feature details. These observations collectively indicate that during the cloud removal process, effectively leveraging frequency-domain information—particularly through the synergistic utilization of both amplitude and phase—holds the promise for more comprehensive and precise cloud removal and restoration of underlying details.

Motivated by these profound insights into frequency-domain characteristics, we propose a novel Dual-view Prompting Network (DVPNet), specifically designed for cloud removal in remote sensing imagery. The core idea of our method is to learn a powerful and adaptive prompting feature by jointly exploring information from both the spatial and frequency domains. This prompting feature is then used to guide and modulate the feature representations within the network, thereby more effectively addressing the cloud removal problem. To ensure that the learned prompting feature fully integrates dual information from both spatial and frequency domains, we meticulously designed a Dual-view Prompting Block (DVPB), which comprises three synergistic key components: First, the **Spatial Domain Modulator** employs a gated mechanism in the spatial domain to adaptively select and enhance crucial spatial contextual information vital for cloud removal from the features. Second, the **Frequency Domain Modulator** constructs a gated mechanism operating on frequency-domain features, aiming to intelligently screen and reinforce key frequency components effective for cloud removal from a spectral perspective. Finally, the **Feature Fusion Module**, leveraging the spatio-frequency joint prompting features learned by the preceding two modulators, employs an efficient windowed attention-based cross-attention mechanism to guide and refine features within the decoder for precise cloud removal. Benefiting from our in-depth analysis of frequency attributes and the meticulously constructed modular design, our DVPNet efficiently achieves state-of-the-art (SOTA) performance on multiple benchmark datasets.

The main contributions of this paper can be summarized as follows:

- We experimentally analyze and reveal the intrinsic correlation between cloud obscuration and the image amplitude and phase spectra, emphasizing the importance of comprehensively utilizing both frequency-domain and spatial-domain information in the cloud removal task. Based on this insight, we propose a novel Dual-view Prompting Network (DVPNet) that effectively addresses cloud obscuration in remote sensing images through spatio-frequency dual-view prompt learning.
- We design a core Dual-view Prompting Block (DVPB), which innovatively integrates a Spatial Domain Modulator and a Frequency Domain Modulator to learn adaptive prompts from spatial and frequency dimensions, respectively. It then utilizes a windowed attention-based cross-attention mechanism to effectively guide the image cloud removal process using the learned joint prompts.
- Extensive experimental results demonstrate that our proposed DVPNet achieves leading performance on multiple public remote sensing image cloud removal datasets, validating the effectiveness and superiority of the proposed method.

II. Related Works

This section reviews recent advancements in deep learning-based image cloud removal, with a focus on two key research

directions: visual prompt learning and frequency-based learning.

A. Deep Learning-based Cloud Removal Methods

Recent advancements in deep learning have significantly advanced cloud removal in remote sensing images, primarily through convolutional neural networks (CNNs), generative models, and Transformer architectures utilizing attention mechanisms for high-level feature extraction.

Early CNN-based approaches focused on improving cloud removal through multi-source data fusion and optimized network structures. For example, Meraner et al. [6] introduced the DSen2-CR model, which fuses SAR-optical data and uses a cloud-adaptive loss function for effective thick cloud removal and surface reconstruction in Sentinel-2 imagery. RSC-Net [7] employs an encoder-decoder structure for end-to-end thin cloud removal, while Ding et al. [8] used conditional variational autoencoders (CVAE) to overcome traditional point estimation limitations, enhancing generalization with real-world datasets. Other advanced techniques, such as KGSR [9], focus on learning the specific degradation kernel from the input image itself to guide the restoration process. Gong et al. [10] incorporated multiscale convolution and determinantal point process (DPP) priors to improve feature discriminability in hyperspectral image classification. Liu et al. [3] designed CMNet, which combines a local information memory module and a global information assistance module to improve cloud removal while preserving image details. However, CNNs are limited by their local feature extraction approach, which struggles to restore global structural information in complex cloud-covered environments.

The introduction of generative models has advanced cloud removal through generative modeling. Conditional GANs (cGANs) [11], [12] use auxiliary data like near-infrared (NIR) or SAR images to remove simulated clouds in datasets like WorldView-2 and Sentinel-2. Pan et al. [4] proposed SPA-GAN, which enhances image quality by incorporating a local-to-global spatial attention mechanism. CloudGAN [13] applied CycleGANs to unpaired data, but faced training instability in thick cloud scenarios. Zheng et al. [14] combined cloud segmentation with GAN generation to optimize restoration, and Wen et al. [15] used Wasserstein GAN to enhance color fidelity by modeling luminance and chrominance separately in the YUV color space. DADIGAN [16] explicitly disentangles shared and private features from SAR and optical data using a model-driven deep unfolding network and progressively fuses them with a dual-attention mechanism to improve interpretability and performance. Furthermore, MT_GAN [16], frame the task as a direct SAR-to-optical image translation problem, employing a multilayer translation generator within a cycle-consistent adversarial framework to generate cloud-free optical images without requiring the cloudy image as an input during inference. Recent diffusion-based approaches like ACDMSR [17] accelerate the restoration process by using pre-trained models to provide a strong initial condition for the iterative refinement. Despite progress, generative models are still limited by the diverse and complex nature of cloud formations, impacting their reliability and efficiency.

Transformer architectures, which leverage attention mechanisms, have gained prominence for their ability to model global features. SPA-GAN [4] incorporated a spatial attention mechanism into GANs, surpassing traditional models like cGANs and CycleGANs. Restormer [18] introduced a channel-level attention mechanism to reduce the computational complexity of global attention while enabling high-precision image restoration. ACA-Net [19] uses a contextual attention module to capture global information beneficial for cloud removal within a residual network framework. CR-former [5] improves global dependency modeling in high-resolution images using a Taylor series approximation of the Softmax attention mechanism. IDF-CR [20] combines Swin Transformer-based cloud removal with latent-space iterative noise diffusion, enhancing cloud removal in complex scenarios. These methods significantly improve the understanding and restoration of complex cloud structures through global attention modeling.

In summary, remote sensing cloud removal methods have evolved from CNN-based local feature extraction to GAN-based generative modeling, and more recently, to Transformer-based global attention modeling. While these advancements have led to significant progress, further research is needed to improve the robustness, computational efficiency, and adaptability of cloud removal methods, particularly in handling the diversity of cloud morphologies.

B. Visual Prompt Learning

With the rapid development of language prompt learning, Visual Prompt Learning (VPL) has emerged as a crucial paradigm in computer vision [21], demonstrating significant advances in image enhancement and restoration through fine-grained visual instructions and adaptive learning. Early works, such as VPT [22], pioneered the adaptation of prompt-based learning from NLP to vision by introducing lightweight, learnable parameters within Transformer layers while preserving pretrained backbones. This approach notably improved few-shot learning performance and enhanced local feature discrimination for fine-grained classification. Despite their success, textual prompts inherently suffer from semantic ambiguity and spatial imprecision, often resulting in visual hallucinations and modality biases. To address these challenges, contemporary VPL frameworks incorporate multi-modal prompt types—such as bounding boxes for regional localization, semantic markers for feature emphasis, pixel-level prompts for sub-visual detail, and soft prompts for continuous spatial modulation [21].

Visual prompt generation methods can be categorized into engineered and automated strategies. Manual prompt engineering employs explicit visual cues (e.g., annotated regions of interest) to guide model attention, whereas automated pipelines leverage segmentation models (e.g., SAM [23], SegFormer [24]) and object detectors (e.g., RCNN variants [25]) to delineate contextual regions, enabling modular multi-step reasoning [21]. Learnable soft prompts further facilitate task-adaptive encoding by optimizing parameters in pixel space, promoting cross-task knowledge transfer. Representative methods include CVP [26], which employs convolutional prompts to enhance out-of-distribution robustness via architec-

tural regularization; ProVP [27], featuring progressive inter-layer propagation with attenuation-controlled feature fusion and contrastive reconstruction to maintain alignment with CLIP's pretrained feature space; and MaPLe [28], which improves vision-language alignment by coupling depthwise prompts across Transformer layers. In biomedical imaging, BiomedCoOp [29] integrates domain-specific BiomedCLIP with large language model-generated prompts, achieving state-of-the-art pathological slide classification under sparse annotations through statistical noise pruning.

In image restoration, VPL enables precise processing by jointly encoding degradation characteristics and semantic guidance. ProRes [30] unifies multi-task restoration—including denoising, deblurring, and low-light enhancement—via degradation-aware prompt embeddings. The PIP framework [31] further demonstrates task adaptability by jointly optimizing degradation-specific and restorative prompts, achieving robust performance under complex and low-light conditions. SeeSR [32] utilizes dual-channel prompt extraction, combining hard structural cues with soft semantic information to guide super-resolution, especially in satellite imagery requiring texture fidelity. SUPIR [33] integrates textual annotations with LLaVA for diffusion-based directional restoration, enabling precise user-targeted edits. Recent innovations such as DPPD [34] enhance feature discriminability through dynamic prototype allocation and distribution learning, while Multi-dimensional Visual Prompt [35] balances computational efficiency and long-range dependency modeling via Mamba-Transformer hybrids. Similarly, CSCT [36] introduces a frequency-gated feed-forward network with a learnable filter to adaptively enhance high-frequency details for remote sensing image super-resolution.

Collectively, these developments mark a paradigm shift from explicit spatial constraints toward implicit feature modulation, fostering solutions characterized by high precision, strong generalization, and reduced reliance on dense annotations.

C. Frequency-based learning Methods

Frequency-domain representations have garnered increasing interest in deep learning due to their unique properties that complement spatial-domain analysis. Researchers have developed novel architectures leveraging frequency characteristics to address challenges that are difficult to resolve in the spatial domain. For instance, FcaNet [37] introduced frequency channel attention for image classification, while GFNet [38] utilized discrete Fourier transforms to capture long-range spatial dependencies. FDIT [39] preserved structural and detailed features by decomposing images into high- and low-frequency components and applying constraints in both pixel and spectral domains. Zhang et al. [40] integrated frequency enhancement modules within CNNs for object detection, and Tan et al. [41] employed convolution on phase and amplitude spectra from FFT for deepfake detection.

In image restoration and generation, frequency-based methods have also demonstrated efficacy. LaMa [42] applied fast Fourier convolutions for large-scale masked image inpainting, and Jeong et al. [43] designed FrepGAN to detect frequency-level artifacts and improve generalization. Mao et al. [44] introduced residual FFT-ReLU blocks capturing frequency-spatial dual-domain features for image deblurring, further extended by Loformer [45] using local channel-wise self-attention in the frequency domain. FourierUp [46] proposed a Fourier-domain up-sampling operator, while DDCN [47] combined discrete cosine transform (DCT) and pixel-domain convolutional networks to reduce compression artifacts. Other notable works include dynamic utilization of frequency features for super-resolution [48] and frequency-assisted architectures for remote sensing super-resolution incorporating frequency selection modules [49].

Within cloud removal, frequency-domain analysis reveals a natural separation: clouds predominantly manifest in low-frequency components centered in the frequency spectrogram, whereas background details are concentrated in high-frequency regions [44], [50]. This distinction has motivated frequency-based methods to enhance cloud removal performance. Early efforts such as Shen et al. [51] applied homomorphic filtering to suppress low-frequency cloud components, preserving clearer pixels. Li et al. [52] demonstrated the value of frequency information for cloud detection and removal. Jiang et al. [50] introduced frequency-domain residual and attention modules to fuse multi-temporal features for thick cloud elimination. Guo et al. [53] developed CP-FFCN, which combines frequency spatial attention with fast Fourier convolutions to model long-range cloud distributions and selectively extract cloud features. Zi et al. [54] integrated wavelet-based CNNs to capture multi-scale frequency components for remote sensing cloud removal, with Jiang et al. [55] advancing this line by employing two-stage training to leverage diverse frequency representations.

Despite the growing number of image restoration methods leveraging frequency-domain analysis, our DVPNet offers distinct and novel contributions that set it apart. In particular:

(1) A New Paradigm of Prompt Guidance. Unlike most existing methods that use frequency analysis for direct feature enhancement or filtering (e.g., FcaNet, LoFormer), we introduce a novel **prompt learning paradigm**. Our DVPNet learns a distinct, adaptive **dual-view prompt** that serves as a guiding signal for the restoration process. This shifts the focus from merely enhancing features to generating an explicit "instruction" on how to perform cloud removal.

(2) A Cascaded Refinement Prompt. Our Dual-view Prompting Block (DVPB) employs a unique **cascaded design**. It first refines features in the spatial domain (via SDM) to capture context, and then transforms these refined features into the frequency domain (via FDM) for further calibration with a learnable spectral prompt. This sequential, "spatial-to-frequency" refinement process for generating the prompt is a key architectural novelty compared to methods that use parallel processing streams.

III. Methodology

In this section, we systematically elucidate the construction details of our proposed Dual-view Prompting Network

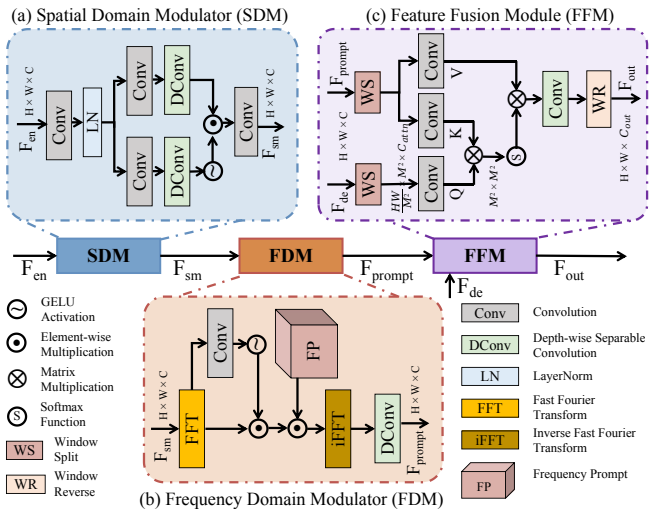


Fig. 2: Architecture of the Dual-view Prompt Block, comprising a Spatial Domain Modulator (SDM), a Frequency Domain Modulator (FDM), and a Feature Fusion Module (FFM). The SDM and FDM utilize gated mechanisms to capture critical prompt features from spatial and frequency domain perspectives, respectively. The FFM then enables these prompt features to guide the corresponding decoder features (F_{de}) via a window-based cross-attention mechanism.

(DVPNet). To present our approach with clarity, we will follow a bottom-up organizational structure. First, we will provide a detailed exposition of the core innovative component of DVPNet—the Dual-view Prompting Block (DVPB)—elucidating how its internal mechanisms learn and apply prompting features from both spatial and frequency dimensions. Second, we will describe the overall network architecture of DVPNet, illustrating how the DVPB is integrated within an encoder-decoder framework to achieve end-to-end cloud removal. Finally, we will define the loss function employed for training and optimizing our network.

A. Dual-view Prompting Block

As illustrated in Figure 2, our proposed Dual-view Prompting Block comprises three core components designed to learn adaptive prompting features: a Spatial Domain Modulator, a Frequency Domain Modulator, and a Feature Fusion Module. These modules work synergistically to achieve effective feature correction and fusion. Each module is detailed below.

1) Spatial Domain Modulator: Cloud obscuration significantly distorts pixel values in remote sensing imagery, leading to blurred visual content or loss of information. To address this, we first meticulously adjust the encoder's output features from a spatial perspective. Considering the complexity and diversity of interactions between clouds and ground features across different images, we design an adaptive feature selection module based on a gating mechanism to dynamically filter and enhance critical spatial information.

Specifically, features $F_{\text{en}} \in \mathbb{R}^{H\times W\times C}$ from an encoder stage first undergo channel transformation via a 1×1 convolutional layer, followed by Layer Normalization (LayerNorm), yielding projected features $\hat{F}_{\text{en}}$:

$$\hat{F}_{\text{en}} = \text{LayerNorm}(\text{Conv}_{1\times 1}(F_{\text{en}})). \tag{1}$$

Subsequently, $\hat{F}_{\text{en}} \in \mathbb{R}^{H\times W\times C}$ is fed into a gating mechanism layer. This mechanism consists of two parallel branches: a gated branch and a content branch. The output of the gated branch serves as dynamic weights to modulate the feature response of the content branch. Both branches are initially processed by a 1×1 convolutional layer followed by a 3×3 Depth-wise Separable Convolutional (DConv) layer. Then, the output of the gated branch is passed through a GELU activation function $\sigma(\cdot)$ to generate weights, which are element-wise multiplied (Hadamard product) with the output of the content branch, enabling adaptive feature interaction. Finally, the interacted features are consolidated and outputted through another 1×1 convolutional layer. This entire gating process can be formulated as:

$$F_g = \sigma(\text{DConv}_{3\times 3}(\text{Conv}_{1\times 1}(\hat{F}_{\text{en}}))), \tag{2}$$

$$F_f = \text{DConv}_{3\times 3}(\text{Conv}_{1\times 1}(\hat{F}_{\text{en}})), \tag{3}$$

$$F_{\text{sm}} = \text{Conv}_{1\times 1}(F_g \odot F_f), \tag{4}$$

where $F_{\text{sm}} \in \mathbb{R}^{H\times W\times C}$ represents the spatially modulated features, which will be utilized for subsequent frequency-domain processing.

2) Frequency Domain Modulator: To capture and leverage the unique patterns of cloud influence from a frequency-domain perspective, we transform the spatially modulated features F_{sm} into the frequency domain. This transformation is achieved using the Fast Fourier Transform (FFT). Analogous to spatial modulation, we also employ a gating mechanism in the frequency domain to selectively enhance information-rich frequency components:

$$\hat{F}_{\text{f}} = \mathcal{F}(F_{\text{sm}}) \odot \sigma(\text{Conv}_{1\times 1}(\mathcal{F}(F_{\text{sm}}))), \tag{5}$$

where $\hat{F}_{\text{f}} \in \mathbb{R}^{H\times(\frac{W}{2}+1)\times 2C}$ are the gated complex spectral features, $\sigma(\cdot)$ refers the GELU activation, and $\mathcal{F}(\cdot)$ denotes the FFT operation. Subsequently, we introduce a learnable frequency prompt $P_f \in \mathbb{R}^{H\times(\frac{W}{2}+1)\times 2C}$, which is element-wise multiplied with $\hat{F}_{\text{f}}$ to calibrate the frequency features. The calibrated features are then transformed back to the spatial domain via the Inverse Fast Fourier Transform (IFFT) $\mathcal{F}^{-1}(\cdot)$, generating the frequency-prompted features F_{fm}:

$$F_{\text{fm}} = \mathcal{F}^{-1}(\hat{F}_{\text{f}} \odot P_f), \tag{6}$$

where $F_{\text{fm}} \in \mathbb{R}^{H\times W\times C}$. As discussed in the introduction, clouds and image content exhibit distinct characteristics and correlations in the frequency domain compared to the spatial domain. Therefore, incorporating supplementary information from the frequency domain, especially learned prompts, can robustly enhance the network's capability to model complex cloud conditions.

According to the Convolution Theorem [56], the Hadamard product of two signals in the Fourier domain is equivalent to the Fourier transform of their convolution in the original

spatial domain. Consequently, by combining Eq. (5) and Eq. (6), the generation of F_{prompt} can be rewritten as:

$$\begin{aligned} F_{\text{fm}} &= \mathcal{F}^{-1}\left(\mathcal{F}(F_{\text{sm}}) \odot \sigma(\text{Conv}_{1\times1}(\mathcal{F}(F_{\text{sm}}))) \odot P_f\right) \\ &= F_{\text{sm}} * \mathcal{F}^{-1}\left(\sigma(\text{Conv}_{1\times1}(\mathcal{F}(F_{\text{sm}}))) \odot P_f\right), \end{aligned} \tag{7}$$

where '$*$' denotes the convolution operation. The term $\mathcal{F}^{-1}(\sigma(\text{Conv}_{1\times1}(\mathcal{F}(F_{\text{sm}}))) \odot P_f)$ can be interpreted as a dynamic convolutional kernel. Finally, F_{fm} undergoes reshaping and is further refined by a 3×3 depth-wise separable convolutional (DConv) layer to produce the final frequency-modulated output features $F_{\text{prompt}} \in \mathbb{R}^{H\times W\times C}$:

$$F_{\text{prompt}} = \text{DConv}_{3\times3}(F_{\text{fm}}). \tag{8}$$

F_{prompt} serve as a guiding prompt for the Feature Fusion Module.

3) Feature Fusion Module: In this module, the guiding prompt features F_{prompt}, generated by the preceding Frequency Domain Modulator, are fused with the input features $F_{\text{de}} \in \mathbb{R}^{H\times W\times C}$ from the corresponding decoder layer via a cross-attention mechanism. Specifically, we project F_{de} to generate Query (Q) and project F_{prompt} to generate Key (K) and Value (V). Considering that high-resolution images contain a large number of tokens, rendering standard dot-product self-attention computationally expensive, we employ Window Attention [57] for efficient feature interaction. This process can be described as:

$$Q = W_q\, \text{R}(F_{\text{de}}), \tag{9}$$

$$K = W_k\, \text{R}(F_{\text{prompt}}), \tag{10}$$

$$V = W_v\, \text{R}(F_{\text{prompt}}), \tag{11}$$

where $W_{(\cdot)}$ represents linear projection matrices, and $R(\cdot)$ denotes the operation of partitioning features into non-overlapping windows, following the strategy in [57]. After window partitioning, $Q, K, V \in \mathbb{R}^{\frac{HW}{M^2}\times M^2\times C_{attn}}$, where M^2 is the number of tokens per window (window size), and C_{attn} is the feature dimension for attention. The attention is then computed as:

$$F_{\text{attn}} = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{12}$$

where d_k is the dimension of the Key (i.e., C_{attn}), used for scaling the dot product, following [57]. Finally, the attention output F_{attn} is reshaped back to its original feature map configuration via a window reverse operation, yielding the module's final output $F_{out} \in \mathbb{R}^{H\times W\times C_{out}}$, which represents the features guided and fused by the dual-view prompts.

B. Network Architecture

As depicted in Figure 3, our proposed Dual-view Prompting Network (DVPNet) adopts a U-Net variant architecture based on Transformers, comprising a symmetric Encoder and Decoder. The Transformer blocks within our network draw inspiration from Restormer [18]. Each Transformer block consists of two core sub-layers: the first is a Multi-Head Self-Attention (MHSA) mechanism, responsible for capturing long-range dependencies, and the second is a Feed-Forward Network (FFN) for feature transformation. Furthermore, residual connections [58] are applied after each sub-layer to facilitate gradient flow and stabilize training. Beyond these Transformer modules, a key innovation of DVPNet lies in the decoder stages, where we integrate our designed core module—the Dual-view Prompting Block (DVPB)—into the feature restoration process at each decoder level to enable precise feature guidance. The specific workflows of the encoder and decoder are detailed below.

1) Encoder: Given a cloudy input image $I_{in} \in \mathbb{R}^{H\times W\times 3}$, the encoder first projects it into a high-dimensional feature space using a 3×3 convolutional layer, yielding initial feature maps $F_{eni} \in \mathbb{R}^{H\times W\times C}$. Subsequently, these feature maps are fed into a three-stage encoder backbone. Each stage is composed of a stack of several Transformer blocks. As the encoding progresses, the network generates hierarchical feature maps at multiple scales, denoted as $F_{en_j} \in \mathbb{R}^{H/2^{(j-1)}\times W/2^{(j-1)}\times 2^{(j-1)}C}$ for $j = 1, 2, 3$. The final stage outputs $F_{en_3} \in \mathbb{R}^{H/4\times W/4\times 4C}$, which serves as the bottleneck representation. Between adjacent stages, a downsampling layer is employed to reduce the spatial resolution of the feature maps while increasing the channel dimensionality. This downsampling layer consists of a 3×3 convolutional layer followed by a PixelUnshuffle operation.

2) Decoder: The decoder takes the deepest feature maps F_{en_3} from the encoder as its initial input, with the objective of progressively reconstructing the cloud-free image. The decoder also comprises three stages, symmetrically structured to the encoder, with each level again formed by stacking multiple Transformer blocks. Between adjacent decoder stages, an upsampling layer is used to increase the feature map resolution, specifically implemented as a 3×3 convolutional layer followed by a PixelShuffle operation. Similar to standard U-Net architectures, we utilize skip connections to concatenate the features F_{en_j} extracted by the encoder at corresponding levels with the upsampled features from the decoder at the same level. The concatenated features are then passed through a 1×1 convolutional layer for channel dimensionality adjustment (halving it) to integrate information and match the dimensionality for subsequent processing.

The core innovation of our architecture is the introduction of the Dual-view Prompting Block (DVPB) within each decoder stage. Specifically, following the features are processed by the Transformer blocks of that stage, but before the upsampling and skip-connection fusion, the DVPB is employed. It utilizes its learned dual-view prompts from both spatial and frequency domains to guide and rectify the fused features, thereby enhancing the cloud removal efficacy. Finally, after the output of the last decoder stage, a 3×3 convolutional layer is applied to map the high-dimensional features back to a three-channel image representation. This is then added residually to the original input image I_{in} to produce the final cloud-free image $I_{out} \in \mathbb{R}^{H\times W\times 3}$.

C. Loss Function

To train the proposed Dual-view Prompting Network, we adopt a composite loss function based on mean squared error (MSE), designed to enforce consistency between the predicted

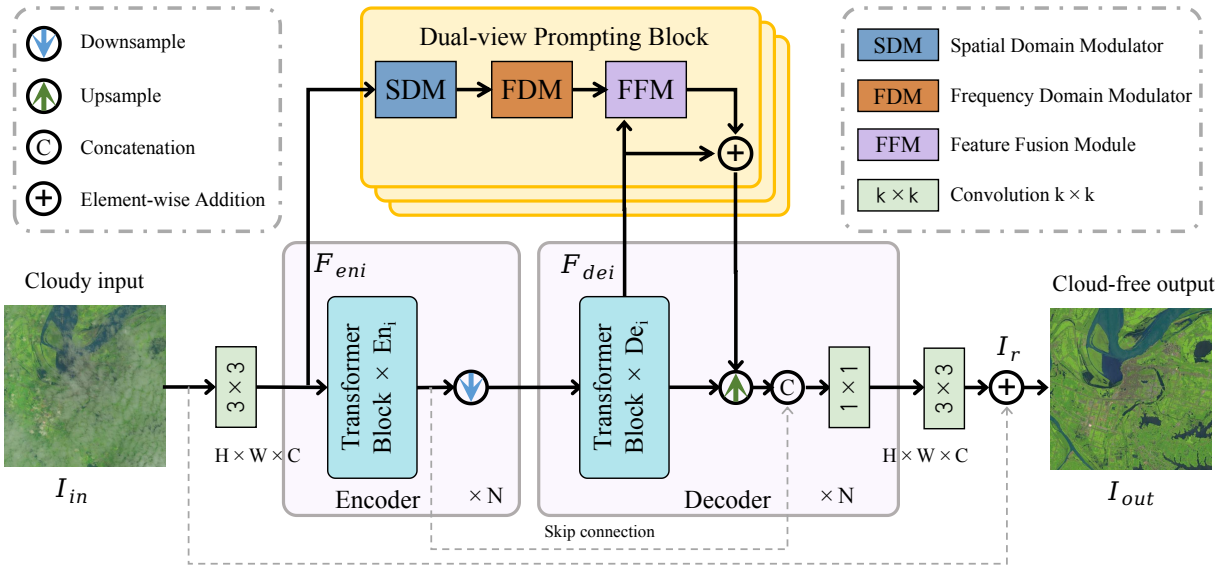


Fig. 3: Illustration of the Dual-view Prompting Network (DVPNet) architecture. Building upon a classic U-Net-style cloud removal network, the proposed Dual-view Prompt Block (DVPB) is incorporated to learn prompt features from dual perspectives (spatial and frequency domains), guiding the cloud removal process.

and reference cloud-free images in both the spatial (RGB) and frequency domains. The total loss function L is formulated as follows:

$$L = w_1 \|I_{out} - I_{gt}\|^2 + w_2 \|\mathcal{F}(I_{out}) - \mathcal{F}(I_{gt})\|^2, \quad (13)$$

where I_{out} denotes the predicted cloud-free image, I_{gt} is the ground truth image, $\mathcal{F}(\cdot)$ represents the FFT operation, and w_1, w_2 are weighting factors that balance the contributions of the spatial and frequency-domain losses, respectively.

IV. Experiment

This section presents comprehensive experimental evaluations, including performance benchmarking and ablation studies, to rigorously validate the effectiveness of the proposed DVPNet in the image cloud removal task.

A. Experiments Details

1) Datasets and Metrics: To rigorously evaluate the proposed cloud removal algorithm, we conducted experiments on four widely recognized remote sensing benchmarks: RICE-I, RICE-II, T-CLOUD, and SEN12MS-CR [6]. These datasets collectively encompass diverse operational challenges, including varying cloud morphologies (thin and thick) and heterogeneous geographic contexts (natural and urban landscapes).

The RICE series [59], developed by the Chinese Academy of Sciences, offers spatially registered cloudy and cloud-free image pairs with complementary annotation granularity. RICE-I consists of 500 spatially aligned 512×512 pixel image pairs from Google Earth, covering representative land cover types such as farmland, forest, and urban areas. RICE-II extends this dataset with 736 triplet sets derived from Landsat 8 OLI/TIRS imagery, each comprising cloud-contaminated observations, temporally proximate ($\leq$15 days) cloud-free references, and semantically enriched cloud masks extracted from Landsat Level-1 quality bands. These masks facilitate precise delineation of cloud layers, shadows, and cirrus clouds, thereby enhancing the representation of complex atmospheric interference. T-CLOUD [8], a state-of-the-art benchmark introduced at ACCV 2022, contains 2,939 spatially consistent 256×256 pixel image pairs captured under real-world conditions. By incorporating variations in geography, seasonality, and cloud optical thickness, T-CLOUD effectively mitigates synthetic-to-real domain gaps, enabling robust assessment of model generalization. The SEN12MS-CR [6] is a multimodal cloud removal dataset with about 110,000 samples from 169 regions, representing different seasonal conditions. Each sample contains paired Sentinel-2 multispectral images (13 bands) in both cloudy and cloud-free versions, along with corresponding Sentinel-1 SAR data, which penetrates cloud cover to reveal sub-cloud surface details.

In alignment with established cloud removal evaluation protocols [5], [6], we adopted a comprehensive set of quantitative metrics. Mean Absolute Error (MAE) [60] quantifies pixel-level reconstruction accuracy, while Peak Signal-to-Noise Ratio (PSNR) [61] and Structural Similarity Index (SSIM) [62] assess perceptual quality and structural integrity. Furthermore, Spectral Angle Mapping (SAM) [63] evaluates spectral consistency across multispectral bands. Collectively, these metrics enable systematic evaluation of spatial detail restoration, structural preservation, and spectral fidelity.

2) Implementation Details: All training and validation experiments were conducted on one NVIDIA RTX A100 GPU using the PyTorch framework. To ensure stable convergence, the AdamW optimizer was employed with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and an L2 weight decay of 1×10^{-4}. Training was performed with a batch size of 16 over 1000 epochs. Learning rates were adjusted

TABLE I: Detailed configuration of the network architecture for our cloud removal model.

Stage	Operator	Param & Output
Input	-	H×W× C_{in}
Stem	Conv($3 \times C_{in} \times C$)	H×W×C
En-Stage 1	[TransBlock] ×1	$R = 3$, $N_h = 2$, H×W×C
Downsample	Conv, PixelUnshuffle(2)	H/2×W/2× $2C$
En-Stage 2	[TransBlock] ×2	$R = 3$, $N_h = 4$, H/2×W/2×$2C$
Downsample	Conv, PixelUnshuffle(2)	H/4×W/4×$4C$
En-Stage 3	[TransBlock] ×4	$R = 3$, $N_h = 8$, H/4×W/4×$4C$
De-Stage 3	[TransBlock] ×4	$R = 3$, $N_h = 8$, H/4×W/4×$4C$
Prompting	DVPB	$P_f = [64 \times 33 \times 2C]$, H/4×W/4×$4C$
Upsample	Conv, PixelShuffle(2), Cat	H/2×W/2× $2C$
De-Stage 2	[TransBlock] ×2	$R = 3$, $N_h = 4$, H/2×W/2× $2C$
Prompting	DVPB	$P_f = [128 \times 65 \times 2C]$, H/2×W/2×$2C$
Upsample	Conv, PixelShuffle(2), Cat	H×W× C
De-Stage 1	[TransBlock] ×1	$R = 3$, $N_h = 2$, H×W× C
Prompting	DVPB	$P_f = [256 \times 129 \times 2C]$, H×W×$C$
Refine	Conv:($3 \times C \times C_{in}$)	H×W× C_{in}
Output	Add Input and Refine	H×W×C_{in}

C is the base number of channels. A convolutional filter $Conv(k \times C_1 \times C_2)$ is defined by k as the kernel size, C_1 and C_2 as the number of input and output channels. The feed-forward network (FFN) expansion factor is R, and the number of attention heads is N_h. The frequency prompt is P_f.

based on dataset characteristics, an initial rate of 4×10^{-4} was used for RICE-I and RICE-II, while a higher rate of 8×10^{-4} was adopted for the more challenging thin-cloud scenarios in T-CLOUD to expedite feature learning. A cosine annealing scheduler was applied to balance training efficiency and generalization, implementing learning rate decay by a factor of 0.1 at 30% and 70% of the total training progress. The network's base channel size C was set to 48; detailed architectural configurations are presented in Table I. During preprocessing, random cropping extracted 256×256 pixel patches from training images as input. The total loss function combined the RGB and frequency domains with weights of 1.0 and 0.1, respectively.

B. Performance of Our Cloud Removal Network

1) Comparison Baselines: To assess the effectiveness of our proposed cloud removal network, we conducted extensive quantitative and qualitative evaluations against seven state-of-the-art methods representing diverse architectural frameworks and published across leading venues. The comparative benchmark includes: (1) Pix2Pix [64] (CVPR 2017), a foundational image-to-image translation model; (2) SPA-GAN [4], which integrates spatial attention mechanisms; (3) CVAE [8] (ACCV 2022), exemplifying variational autoencoder architectures; (4) Restormer [18] (CVPR 2022), leveraging transformer-based restoration modules; (5) CMNet [3] (TGRS 2024), utilizing cascaded memory networks; (6) ACA-CRNet [19] (TGRS 2024), incorporating attention-guided context aggregation; (7) CR-former [5] (TGRS 2024), featuring linear attention mechanisms; (8) DSen2-CR [6] (ISPRS 2020), using residual connection networks; (9) GLF-CR [65] (ISPRS 2022), fusing global-local SAR and optical textures; and (10) UnCRtainTS [66] (CVPR 2023), leveraging uncertainty quantification with attention-based temporal encoding. This selection spans a period from 2017 to 2024, covering key advancements in computer vision and remote sensing, thereby providing a comprehensive validation of our approach against prominent restoration paradigms.

2) Quantitative Comparison with Baselines: This study performs systematic comparisons of the proposed cloud removal method against state-of-the-art models across four remote sensing benchmark datasets. The evaluation uses four accuracy metrics, as detailed in Table II. Our method consis-

TABLE II: Quantitative comparison results on the four datasets. Metrics marked with ↓ are better when lower, and those with ↑ are better when higher. Best results are shown in bold.

Datasets	Models	MAE↓	SAM↓	PSNR↑	SSIM↑
RICE-I	pix2pix [64]	0.0253	4.11	31.97	0.9161
	SPA-GAN [4]	0.0372	2.25	28.89	0.9144
	CVAE [8]	0.0198	1.11	33.70	0.9562
	Restoremer [18]	0.0176	1.01	35.42	0.9617
	CMNet [3]	0.0144	0.98	36.26	0.9625
	ACA-CRNet [19]	0.0151	0.98	36.05	0.9628
	CR-former [5]	0.0138	0.92	36.75	0.9627
	Ours	**0.0138**	**0.88**	**36.88**	**0.9648**
RICE-II	pix2pix [64]	0.0327	5.08	29.45	0.8473
	SPA-GAN [4]	0.0403	3.36	27.51	0.8177
	CVAE [8]	0.0216	1.69	33.62	0.9079
	Restoremer [18]	0.0163	1.27	36.05	0.9155
	CMNet [3]	0.0167	1.29	35.82	0.9151
	ACA-CRNet [19]	0.0164	1.29	35.65	0.9126
	CR-former [5]	0.0163	1.26	36.24	0.9161
	Ours	**0.0152**	**1.18**	**36.71**	**0.9201**
T-CLOUD	pix2pix [64]	0.0449	10.01	25.64	0.7563
	SPA-GAN [4]	0.0419	4.09	26.14	0.7954
	CVAE [8]	0.0342	2.95	28.19	0.8613
	Restoremer [18]	0.0248	2.46	30.49	0.8851
	CMNet [3]	0.0251	2.49	30.33	0.8829
	ACA-CRNet [19]	0.0234	2.35	30.85	0.8885
	CR-former [5]	0.0238	2.43	30.82	0.8867
	Ours	**0.0237**	**2.39**	**30.96**	**0.8945**
SEN12MS-CR	pix2pix [64]	0.0310	10.78	27.60	0.8640
	SPA-GAN [4]	0.0450	18.09	24.78	0.7540
	DSen2-CR [6]	0.0310	9.47	27.76	0.8740
	GLF-CR [65]	0.0280	8.98	28.64	0.8850
	UnCRtainTS [66]	0.0270	8.32	28.90	0.8800
	ACA-CRNet [19]	0.0250	7.77	29.78	0.8960
	Ours	**0.0247**	**7.68**	**29.90**	**0.9020**

TABLE III: Efficiency comparison across Params, FLOPs, Inference time, and GPU memory usage.

Models	Params(M)	FLOPs(G)	Times(ms)	GPU Memory(MB)
pix2pix [64]	54.41	6.1	4.6	3357
SPA-GAN [4]	0.21	15.2	26.0	3749
CVAE [8]	15.42	37.1	15.1	1097
Restoremer [18]	26.13	155.0	112.3	1629
CMNet [3]	16.51	236.0	197.0	1607
ACA-CRNet [19]	20.39	1422.0	223.7	6417
CR-former [5]	26.15	155.0	117.3	1601
Ours	9.98	50.3	38.6	1475

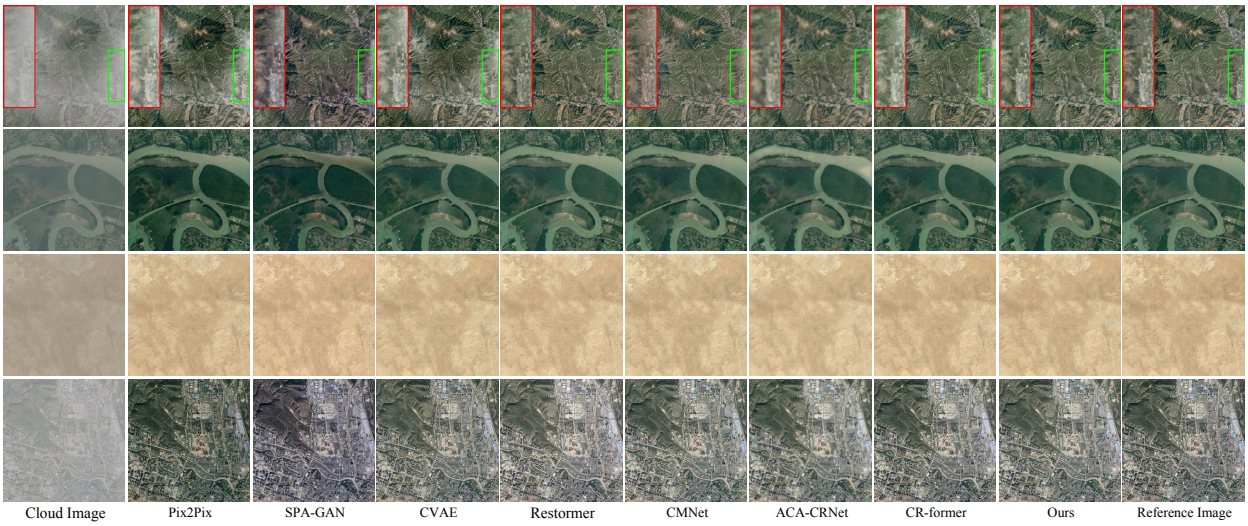


Fig. 4: Cloud removal results visualized on the RICE-I dataset. Key local details, highlighted in red boxes, are best viewed when zoomed in.

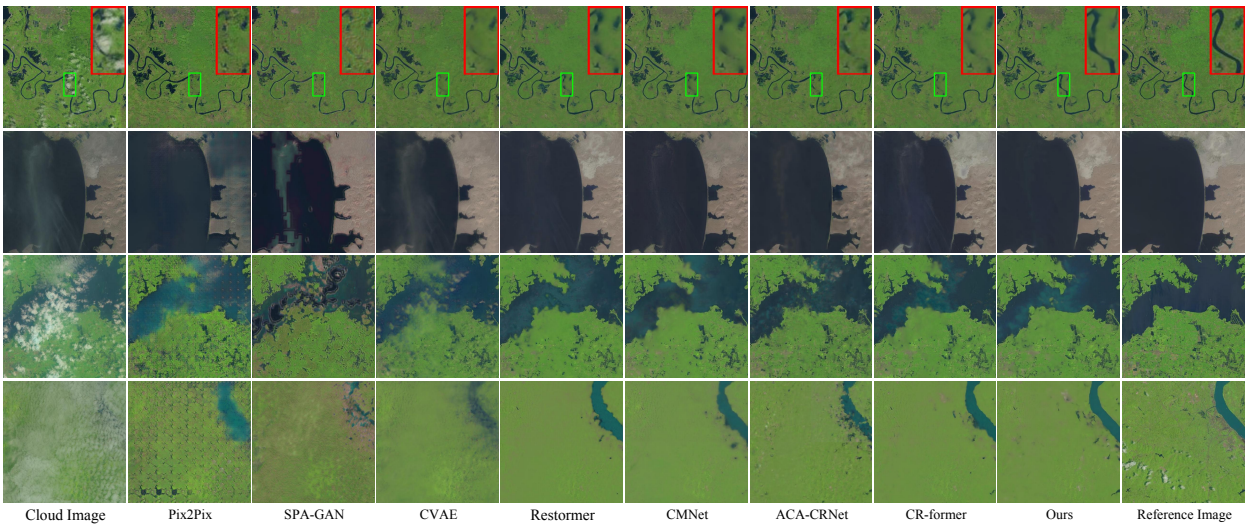


Fig. 5: Cloud removal results visualized on the RICE-II dataset. Key local details, highlighted in red boxes, are best viewed when zoomed in.

tently outperforms competitors, particularly in the challenging cloud-contaminated RICE-II dataset, the thin-cloud T-CLOUD dataset, and the complex multimodal SEN12MS-CR dataset. Compared to the leading baseline, CR-former [5], our method achieves average reductions of 2.55% in MAE and 4.33% in SAM, with improvements of 0.70% in PSNR and 0.51% in SSIM across the RICE-I, RICE-II, and T-CLOUD datasets. On the SEN12MS-CR dataset, our method yields the best results, with a 0.12 dB gain over ACA-CRNet [19].

3) Efficiency Analysis: The efficiency evaluation focuses on four key metrics: model complexity (Params), computational cost (FLOPs), inference time, and GPU memory usage. All experiments were conducted under identical conditions (NVIDIA RTX 3090 GPU, 256×256 input resolution, batch size = 1) for fair comparison, with results presented in Table III. Our proposed model shows significant advantages in key metrics while maintaining competitive performance. It has 9.98M parameters, a 61.8% reduction compared to CR-former (26.15M) and a 51.2% reduction compared to CVAE (15.42M), making it one of the most parameter-efficient models. In terms of computational cost, it requires 50.3 GFLOPs, representing a 67.5% reduction from CR-former (155 GFLOPs), 78.7% from CMNet (236 GFLOPs), and 96.5% from ACA-CRNet (1422 GFLOPs), demonstrating superior operational efficiency. For inference speed, our model processes in 38.6ms, outperforming ACA-CRNet (223.7ms) by $5.8\times$, CMNet (197.0ms) by $5.1\times$, and CR-former (117.3ms) by $3.0\times$. Although slightly slower than CVAE (15.1ms), it offers better reconstruction quality, balancing speed and effectiveness. In terms of memory efficiency, the model uses 1475MB of GPU memory, 7.9% less than CR-former (1601MB) and the least among modern methods (SPA-GAN: 3749MB, ACA-CRNet: 6417MB).

4) Qualitative Comparison with Baselines: Figures 4-7 present visual comparisons across four benchmarks (RICE-

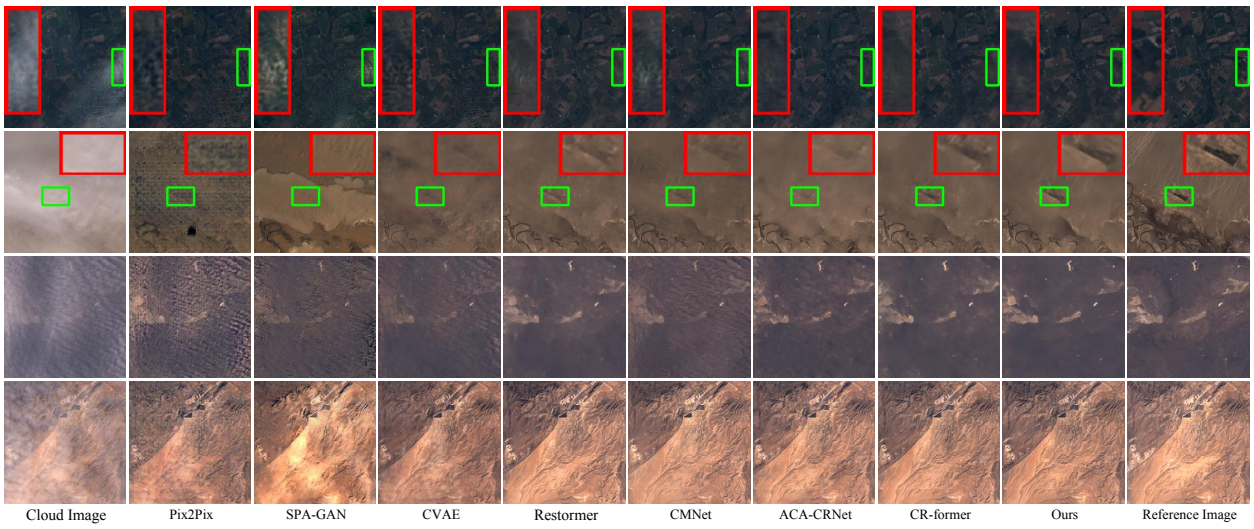


Fig. 6: Cloud removal results visualized on the T-CLOUD dataset. Key local details, highlighted in red boxes, are best viewed when zoomed in.

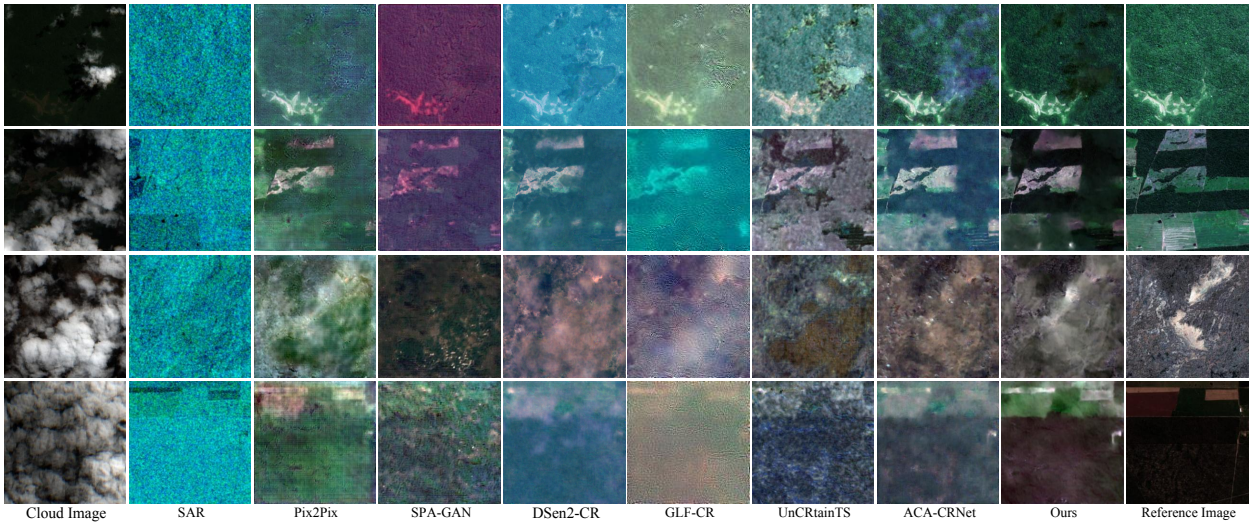


Fig. 7: Cloud removal results visualized on the SEN12MS-CR dataset. Key local details, highlighted in red boxes, are best viewed when zoomed in.

I, RICE-II, T-CLOUD, SEN12MS-CR), demonstrating consistent superiority of the proposed method in cloud removal fidelity and structural preservation under diverse cloud conditions.

On RICE-I (thin clouds, Figure 4), our method and CR-former lead in color/texture preservation. Unlike SPA-GAN's color shifts (Rows 1, 4), ours matches references with clearer details (e.g., Row 4 architecture). On RICE-II (thick clouds, Figure 5), ours preserves structural continuity: coherent river flow (Row 1) vs. fragmented baselines, and sharper grassland textures (Row 4). On T-CLOUD (mixed clouds, Figure 6), ours generalizes best, outperforming pix2pix/SPA-GAN/CVAE by eliminating artifacts and residual clouds. It preserves ground textures, with mountain structures (Rows 3-4) matching references. On SEN12MS-CR (multimodal optical and SAR data, Figure 7), ours excels in restoring thin/thick cloud-contaminated images. SAR-derived subsurface info enables near-ground-truth recovery in heavy cloud scenarios (Row 4).

5) Analysis of Performance Under Different Cloud Conditions: To rigorously evaluate the robustness of our proposed DVPNet, we specifically analyzed its performance across different cloud types, using the RICE-I dataset as a benchmark for thin cloud scenarios and the RICE-II dataset for more challenging thick cloud conditions.

For thin cloud removal on the **RICE-I dataset**, as show in Table II our method achieves the best performance across all metrics, including a **PSNR of 36.88 and an SSIM of 0.9648**, confirming its ability to restore fine details with high fidelity.

The performance gap becomes more pronounced in the challenging thick cloud scenarios of the **RICE-II dataset**. Our robustness is quantitatively confirmed in Table II, where our model again ranks first across all metrics, achieving a **PSNR of 36.71 and an SSIM of 0.9201**.

This robust performance across different cloud types is

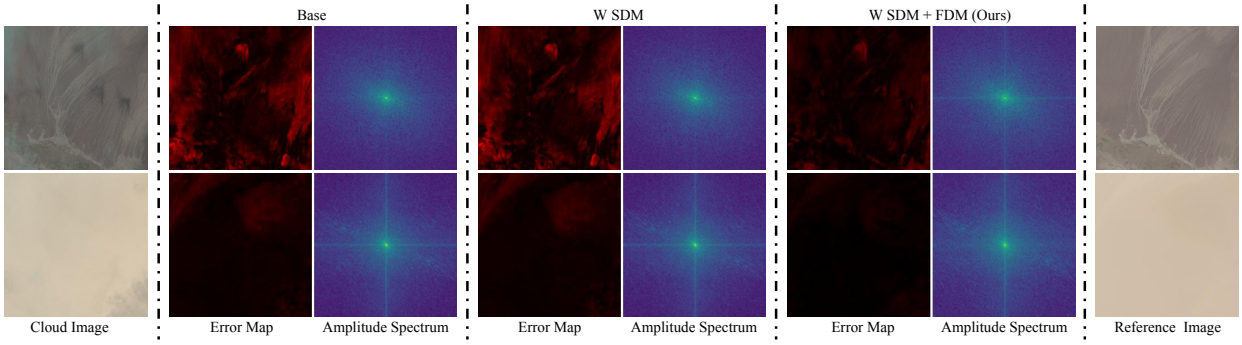


Fig. 8: Amplitude spectrum analysis of error maps with incremental addition of SDM and FDM to the base model. Error maps show pixel-wise differences between restored and reference images, with a color scale from black (low error, better performance) to red (high error).

TABLE IV: Ablation studies of the proposed method on the RICE-I and RICE-II dataset. Metrics marked with ↓ are better when lower, and those with ↑ are better when higher. Best results are shown in bold.

Datasets	Models	MAE↓	SAM↓	PSNR↑	SSIM↑
RICE-I	Base	0.0144	0.933	36.50	0.9640
	w/ SDM	0.0153	0.932	36.27	0.9629
	w/ SDM+FDM(**Ours**)	**0.0138**	**0.880**	**36.88**	**0.9648**
RICE-II	Base	0.0154	1.183	36.53	0.9199
	w/ SDM	0.0154	1.205	36.56	0.9197
	w/ SDM+FDM(**Ours**)	**0.0152**	**1.179**	**36.71**	**0.9201**

directly attributable to our core innovation: the Dual-view Prompting mechanism. Unlike methods that apply a uniform restoration strategy, our Frequency Domain Modulator (FDM) with its learnable spectral prompt allows the network to adaptively generate guidance based on the nature of the cloud obstruction. It learns to apply subtle corrections for thin clouds (as validated on RICE-I) and to guide a more comprehensive reconstruction for thick, structure-occluding clouds (as validated on RICE-II). This adaptive, frequency-aware guidance is the key reason our model consistently delivers superior restoration quality across diverse and challenging real-world cloud conditions.

C. Ablation Studies

To evaluate the effectiveness of the proposed modules, we conducted ablation studies on the RICE-I and RICE-II datasets by incrementally incorporating the Spatial Domain Module (SDM) and the Frequency Domain Module (FDM) into a baseline encoder-decoder framework.

Quantitative results, presented in Table IV, demonstrate the impact of each component. On RICE-I, the SDM-only variant ("w/ SDM") resulted in slight performance degradation, with MAE increasing to 0.0153 and SSIM decreasing to 0.9629. In contrast, the full model ("w/ SDM+FDM (Ours)") achieved substantial improvements across all metrics: MAE of 0.0138 (-4.2%), SAM of 0.880 (-5.7%), PSNR of 36.88 (+1.04%), and SSIM of 0.9648 (+0.08%). A similar trend was observed on RICE-II, where the SDM-only configuration offered negligible benefits and even worsened SAM to 1.205 (+1.86%).

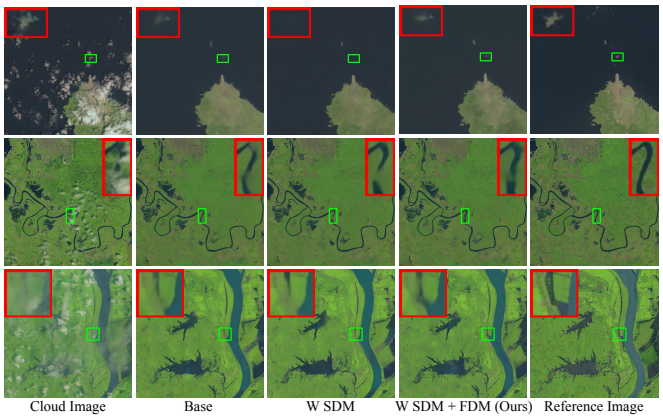


Fig. 9: Ablation study results visualized on the RICE-II dataset. Key local details, highlighted in red boxes, are best viewed when zoomed in.

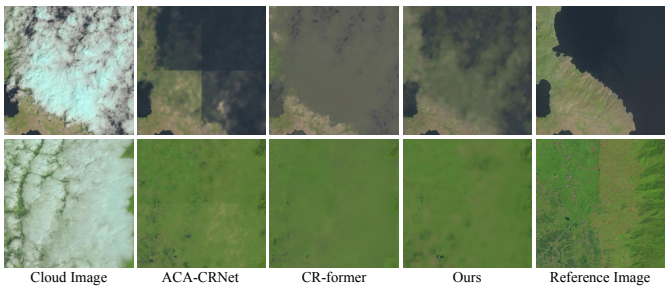


Fig. 10: Failure cases under extreme thick-cloud scenarios.

Incorporating both SDM and FDM yielded the best results, with MAE of 0.0152 (-1.3%), SAM of 1.179 (-0.34%), PSNR of 36.71 (+0.5%), and SSIM of 0.9201 (+0.02%).

Qualitative results on the RICE-II dataset, shown in Figure 9, further support the complementary roles of SDM and FDM. The SDM-only output (column 3) introduces visible artifacts: in row 1, the island enclosed in the green box is entirely smoothed out; in row 2, the reconstructed river exhibits fragmentation and gaps. These visual flaws are consistent with the quantitative trends, suggesting that SDM's contextual operations may suppress fine-grained structures in complex scenes. In contrast, the full model (column 4)

restores structural integrity more effectively: the river in row 2 is spatially continuous, and the tributary in row 3 displays sharper boundaries and a more coherent topology than both the baseline and SDM-only models. These improvements are attributed to FDM's frequency-aware design, which integrates spectral prompt learning to retain high-frequency edges and adaptive feature calibration to balance global context with local detail.

Figure 8 presents the amplitude spectra of error maps as SDM and FDM are added to the base model. These maps use a color scale from black (low error) to red (high error), with darker areas indicating better performance. Our model ("w/ SDM+FDM (Ours)") has the darkest error maps (lowest error) and a distinct amplitude spectrum versus models without these modules. This unique spectral pattern suggests optimizing spatial and frequency components reduces cross-band errors, correlating with enhanced restoration quality.

Functionally, SDM enhances spatial representation, while FDM introduces complementary frequency-domain information to preserve structural sharpness. Their combined use consistently improves both accuracy (MAE, SAM) and perceptual quality (PSNR, SSIM), especially under complex conditions. The ablation results confirm the effectiveness of the dual-domain design and demonstrate that the synergy between SDM and FDM contributes significantly to the fidelity and robustness of cloud removal.

D. Limitation Analysis

Although DVPNet achieves strong cloud removal performance, it remains limited under extreme thick-cloud conditions where optical data offers no discernible cues (Figure 10). In such cases, all compared methods—including ACA-CRNet, CR-former, and ours—fail to reconstruct fully obscured terrain, as in Row 1 where hidden sea surfaces are incorrectly restored. This reflects an inherent inability to perform terrain-type conversion without contextual information. Moreover, in vegetation regions (Row 2), these methods preserve macro-scale patterns but lose fine, density-dependent textures. Such limitations underscore persistent structural and textural reconstruction challenges under optically opaque clouds, pointing to the potential of integrating elevation data or physics-guided hybrid frameworks to improve feature recovery.

V. Conclusion

In this paper, we addressed the prevalent oversight of frequency-domain information in remote sensing cloud removal by proposing a novel Dual-view Prompting Network (DVPNet). The cornerstone of our network is the innovative Dual-view Prompting Block (DVPB), which employs an ingenious cascaded design. It first refines key features via a spatial gating mechanism, then transforms these features into the Fourier domain for deep calibration using a learnable spectral prompt. This architecture enables the network to adaptively identify and enhance cloud-relevant components, which are then seamlessly fused with spatial features through an efficient windowed attention mechanism. Our method establishes a new state-of-the-art (SOTA) on three public benchmarks: RICE-I, RICE-II, and T-CLOUD. DVPNet demonstrates a significant improvement in preserving structural continuity and textural fidelity in complex scenes involving fine-grained features like rivers and islands, effectively mitigating the spectral distortions and artifacts common in prior methods. In conclusion, this work not only robustly demonstrates the immense potential of fusing spatial and frequency domains for cloud removal but also pioneers a promising research avenue for tackling other atmospheric degradation problems, such as haze and shadows.

References

[1] M. D. King, S. Platnick, W. P. Menzel, S. A. Ackerman, and P. A. Hubanks, "Spatial and temporal distribution of clouds observed by modis onboard the terra and aqua satellites," *IEEE transactions on geoscience and remote sensing*, vol. 51, no. 7, pp. 3826–3852, 2013.

[2] S. Ji, P. Dai, M. Lu, and Y. Zhang, "Simultaneous cloud detection and removal from bitemporal remote sensing images using cascade convolutional neural networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 732–748, 2020.

[3] J. Liu, B. Pan, and Z. Shi, "Cascaded memory network for optical remote sensing imagery cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

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

[5] Y. Wu, Y. Deng, S. Zhou, Y. Liu, W. Huang, and J. Wang, "Cr-former: Single image cloud removal with focused taylor attention," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[6] A. Meraner, P. Ebel, X. X. Zhu, and M. Schmitt, "Cloud removal in sentinel-2 imagery using a deep residual neural network and sar-optical data fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 333–346, 2020.

[7] W. Li, Y. Li, D. Chen, and J. C.-W. Chan, "Thin cloud removal with residual symmetrical concatenation network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 153, pp. 137–150, 2019.

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

[9] Q. Yan, A. Niu, C. Wang, W. Dong, M. Woźniak, and Y. Zhang, "Kgsr: A kernel guided network for real-world blind super-resolution," *Pattern Recognition*, vol. 147, p. 110095, 2024.

[10] Z. Gong, P. Zhong, Y. Yu, W. Hu, and S. Li, "A cnn with multiscale convolution and diversified metric for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3599–3618, 2019.

[11] K. Enomoto, K. Sakurada, W. Wang, H. Fukui, M. Matsuoka, R. Nakamura, and N. Kawaguchi, "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*, 2017, pp. 48–56.

[12] C. Grohnfeldt, M. Schmitt, and X. Zhu, "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*. IEEE, 2018, pp. 1726–1729.

[13] P. Singh and N. Komodakis, "Cloud-gan: Cloud removal for sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2018, pp. 1772–1775.

[14] J. Zheng, X.-Y. Liu, and X. Wang, "Single image cloud removal using u-net and generative adversarial networks," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 8, pp. 6371–6385, 2020.

[15] X. Wen, Z. Pan, Y. Hu, and J. Liu, "Generative adversarial learning in yuv color space for thin cloud removal on satellite imagery," *Remote Sensing*, vol. 13, no. 6, p. 1079, 2021.

[16] Z. He, P. Wang, Y. Zou, B. Huang, D. Zhu, H. F. Lee, and H. Leung, "Dadigan: A dual attention blocks-based disentangled iterative generative adversarial network for cloud and shadow removal on sar and optical images," *Information Fusion*, p. 103487, 2025.

[17] A. Niu, T. X. Pham, K. Zhang, J. Sun, Y. Zhu, Q. Yan, I. S. Kweon, and Y. Zhang, "Acdmsr: Accelerated conditional diffusion models for single image super-resolution," *IEEE Transactions on Broadcasting*, vol. 70, no. 2, pp. 492–504, 2024.

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

[19] W. Huang, Y. Deng, Y. Wu, and J. Wang, "Attentive contextual attention for cloud removal," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[20] M. Wang, Y. Song, P. Wei, X. Xian, Y. Shi, and L. Lin, "Idf-cr: Iterative diffusion process for divide-and-conquer cloud removal in remote-sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[21] J. Wu, Z. Zhang, Y. Xia, X. Li, Z. Xia, A. Chang, T. Yu, S. Kim, R. A. Rossi, R. Zhang *et al.*, "Visual prompting in multimodal large language models: A survey," *arXiv preprint arXiv:2409.15310*, 2024.

[22] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European conference on computer vision*. Springer, 2022, pp. 709–727.

[23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4015–4026.

[24] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[25] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[26] Y.-Y. Tsai, C. Mao, and J. Yang, "Convolutional visual prompt for robust visual perception," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27 897–27 921, 2023.

[27] C. Xu, Y. Zhu, H. Shen, B. Chen, Y. Liao, X. Chen, and L. Wang, "Progressive visual prompt learning with contrastive feature re-formation," *International Journal of Computer Vision*, vol. 133, no. 2, pp. 511–526, 2025.

[28] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 113–19 122.

[29] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, "Biomedcoop: Learning to prompt for biomedical vision-language models," *arXiv preprint arXiv:2411.15232*, 2024.

[30] J. Ma, T. Cheng, G. Wang, Q. Zhang, X. Wang, and L. Zhang, "Prores: Exploring degradation-aware visual prompt for universal image restoration," *arXiv preprint arXiv:2306.13653*, 2023.

[31] Z. Li, Y. Lei, C. Ma, J. Zhang, and H. Shan, "Prompt-in-prompt learning for universal image restoration," *arXiv preprint arXiv:2312.05038*, 2023.

[32] R. Wu, T. Yang, L. Sun, Z. Zhang, S. Li, and L. Zhang, "Seesr: Towards semantics-aware real-world image super-resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 25 456–25 467.

[33] F. Yu, J. Gu, Z. Li, J. Hu, X. Kong, X. Wang, J. He, Y. Qiao, and C. Dong, "Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 25 669–25 680.

[34] G. Wu, J. Jiang, K. Jiang, X. Liu, and L. Nie, "Learning dynamic prompts for all-in-one image restoration," *IEEE Transactions on Image Processing*, 2025.

[35] A. Jiang, H. Chen, Z. Chen, J. Ye, and M. Wang, "Multi-dimensional visual prompt enhanced image restoration via mamba-transformer aggregation," *arXiv preprint arXiv:2412.15845*, 2024.

[36] K. Zhang, L. Li, L. Jiao, X. Liu, W. Ma, F. Liu, and S. Yang, "Csct: Channel-spatial coherent transformer for remote sensing image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2025.

[37] Z. Qin, P. Zhang, F. Wu, and X. Li, "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 783–792.

[38] Y. Rao, W. Zhao, Z. Zhu, J. Zhou, and J. Lu, "Gfnet: Global filter networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 960–10 973, 2023.

[39] M. Cai, H. Zhang, H. Huang, Q. Geng, Y. Li, and G. Huang, "Frequency domain image translation: More photo-realistic, better identity-preserving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 930–13 940.

[40] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 4504–4513.

[41] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 5052–5060.

[42] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with fourier convolutions," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2149–2159.

[43] Y. Jeong, D. Kim, Y. Ro, and J. Choi, "Frepgan: robust deepfake detection using frequency-level perturbations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 1, 2022, pp. 1060–1068.

[44] X. Mao, Y. Liu, F. Liu, Q. Li, W. Shen, and Y. Wang, "Intriguing findings of frequency selection for image deblurring," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 1905–1913.

[45] X. Mao, J. Wang, X. Xie, Q. Li, and Y. Wang, "Loformer: Local frequency transformer for image deblurring," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 10 382–10 391.

[46] H. Yu, J. Huang, F. Zhao, J. Gu, C. C. Loy, D. Meng, C. Li *et al.*, "Deep fourier up-sampling," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 995–23 008, 2022.

[47] J. Guo and H. Chao, "Building dual-domain representations for compression artifacts reduction," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*. Springer, 2016, pp. 628–644.

[48] W. Xie, D. Song, C. Xu, C. Xu, H. Zhang, and Y. Wang, "Learning frequency-aware dynamic network for efficient super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4308–4317.

[49] Y. Xiao, Q. Yuan, K. Jiang, Y. Chen, Q. Zhang, and C.-W. Lin, "Frequency-assisted mamba for remote sensing image super-resolution," *IEEE Transactions on Multimedia*, 2024.

[50] B. Jiang, X. Li, H. Chong, Y. Wu, Y. Li, J. Jia, S. Wang, J. Wang, and X. Chen, "A deep-learning reconstruction method for remote sensing images with large thick cloud cover," *International Journal of Applied Earth Observation and Geoinformation*, vol. 115, p. 103079, 2022.

[51] H. Shen, H. Li, Y. Qian, L. Zhang, and Q. Yuan, "An effective thin cloud removal procedure for visible remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 96, pp. 224–235, 2014.

[52] J. Li, Z. Wu, Z. Hu, J. Zhang, M. Li, L. Mo, and M. Molinier, "Thin cloud removal in optical remote sensing images based on generative adversarial networks and physical model of cloud distortion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 166, pp. 373–389, 2020.

[53] Y. Guo, W. He, Y. Xia, and H. Zhang, "Blind single-image-based thin cloud removal using a cloud perception integrated fast fourier convolutional network," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 206, pp. 63–86, 2023.

[54] Y. Zi, H. Ding, F. Xie, Z. Jiang, and X. Song, "Wavelet integrated convolutional neural network for thin cloud removal in remote sensing images," *Remote Sensing*, vol. 15, no. 3, p. 781, 2023.

[55] B. Jiang, H. Chong, Z. Tan, H. An, H. Yin, S. Chen, Y. Yin, and X. Chen, "Fdt-net: Deep-learning network for thin-cloud removal in remote sensing image using frequency domain training strategy," *IEEE Geoscience and Remote Sensing Letters*, 2024.

[56] H. Pfister, "Discrete-time signal processing," *Lecture Note, pfister. ee. duke. edu/courses/ece485/dtsp. pdf*, 2017.

[57] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[59] D. Lin, G. Xu, X. Wang, Y. Wang, X. Sun, and K. Fu, "A remote sensing image dataset for cloud removal," *arXiv preprint arXiv:1901.00600*, 2019.

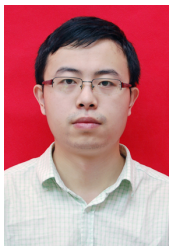
[60] T. Chai and R. R. Draxler, "Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature," *Geoscientific model development*, vol. 7, no. 3, pp. 1247–1250, 2014.

[61] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *2012 Fourth International Workshop on Quality of Multimedia Experience*. IEEE, 2012, pp. 37–38.

[62] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[63] F. A. Kruse, A. Lefkoff, J. Boardman, K. Heidebrecht, A. Shapiro, P. Barloon, and A. Goetz, "The spectral image processing system (sips)—interactive visualization and analysis of imaging spectrometer data," *Remote sensing of environment*, vol. 44, no. 2-3, pp. 145–163, 1993.
[64] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
[65] F. Xu, Y. Shi, P. Ebel, L. Yu, G.-S. Xia, W. Yang, and X. X. Zhu, "Glf-cr: Sar-enhanced cloud removal with global–local fusion," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 192, pp. 268–278, 2022.
[66] P. Ebel, V. S. F. Garnot, M. Schmitt, J. D. Wegner, and X. X. Zhu, "Uncrtaints: Uncertainty quantification for cloud removal in optical satellite time series," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2085–2095.

Ye Deng received his Ph.D. from Xi'an Jiaotong University, Xi'an, China, in 2023. He is currently an Assistant Professor at Southwestern University of Finance and Economics. His research interests include image inpainting, image restoration, and machine learning.

Wenli Huang received her Ph.D. from Xi'an Jiaotong University, Xi'an, China, in 2023. She is currently a lecturer at Ningbo University of Technology. Her research interests include deep learning, image processing, and computer vision, focusing on network structure optimization, image inpainting, image restoration, etc.



Zixin Tang received the M.S. degree from China Academy of Ordnance Science, Beijing, China, in 2016 and the Ph.D. degree in computer science from Sichuan University, Chengdu, China, in 2023. He is currently a Lecturer with the School of Computing and Artificial Intelligence, Southwest University of Finance and Economics, Chengdu, China. His research interests include image restoration, computa tional imaging, and machine learning.

Jiang Duan is a professor in Southwestern University of Finance and Economics, China. He received the BSc degree in mechanical engineering from Southwest Jiaotong University, Chengdu, China, in 2001, the MSc degree from the University of Derby, Derby, U.K., in 2002, and the PhD degree from the University of Nottingham, Nottingham, U.K., in 2006. Professor Duan is a recognized national expert, the winner of the first Sichuan outstanding talent award (the highest talent award of Sichuan province), the winner of the Sichuan youth science and technology award and the standing committee member of Sichuan association for science and technology. Professor Duan has published more than 30 academic papers, won more than 10 international and national patents, and his researches have been funded by about 20 national and provincial funds.