CLIP-HNet: Hybrid Network with Cross-Modal Guidance for Self-Supervised Remote Sensing Dehazing

Shan Wang
Beijing Normal University
Beijing, China
Nanyang Technological University
Singapore

Weisi Lin
Nanyang Technological University
Singapore

Yun Liu
Southwest University
Chongqing, China

Libao Zhang*
Beijing Normal University
Beijing, China
libaozhang@bnu.edu.cn

Abstract

Unsupervised remote sensing dehazing remains a challenging and ill-posed task due to the absence of reliable supervision signals. Existing dehazing methods with unpaired data often oversimplify haze removal as style transfer, limiting generalization in complex scenarios. Moreover, current unimodal frameworks neglect cross-modal cues that could improve contextual reasoning. To address these issues, we propose a novel cross-modal guided self-supervised dehazing framework called CLIP-HNet, which achieves multi-model feature extraction, boundary-focused reconstruction and adaptive sample filtering. Specifically, to capture global-local contextual features, a hybrid feature interaction network is designed, which bridges the feature representations of multi models with global context-aware module (GCAM) and hybrid feature fusion module (HF^2M). Then, based on the hybrid features, a boundary-aware feature reconstruction (BFRec) is proposed to further refine edge details. Furthermore, a CLIP-guided progressive information distillation scheme is presented to dynamically prioritize training samples and distill useful signals, which predicts haze concentration by CLIP and progressively increases sample difficulty during the training stage. Finally, a frequency-domain texture matching (FTM) strategy refines texture and spectral details, enhancing the model's ability to recover fine details. Experiments on synthetic and real RSIs demonstrate that the proposed CLIP-HNet surpasses state-of-the-art approaches, achieving superior visual quality and quantitative performance.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Computer vision tasks**; **Reconstruction**.

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755818

Keywords

Cross-modal guidance, dehazing, remote sensing, self-supervised learning

ACM Reference Format:
Shan Wang, Weisi Lin, Yun Liu, and Libao Zhang. 2025. CLIP-HNet: Hybrid Network with Cross-Modal Guidance for Self-Supervised Remote Sensing Dehazing. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3755818

1 Introduction

With advancements in Earth observation technology, remote sensing now provides high-resolution optical images rich in surface information [19], which are widely used in fields such as environmental monitoring, agricultural production, and geographic information acquisition [1, 10, 30]. However, atmospheric particles like haze, fog, and clouds often degrade these images by absorbing and scattering light, resulting in blurred textures, low contrast, and color shift. This deterioration hampers the effectiveness of remote sensing images (RSIs) in computer vision tasks, making haze removal crucial for restoring image quality and enhancing ground surface visibility.

Early dehazing methods [4, 12, 62] usually use various priors or assumptions to invert the atmospheric scattering model (ASM) [31] to restore clear results on single RSI:

$$I(x) = J(x)t(x) + A(1 - t(x)), \tag{1}$$

where x denotes the pixel coordinates in an image, $I(x)$ is the observed hazy image, $J(x)$ is the clear image without haze, A is the airlight, and $t(x)$ is the trasmission map. The emergence of these prior-based methods [24, 26, 34] has greatly promoted the development of image dehazing technology. However, a significant drawback remains: the performance of these methods is entirely dependent on the accuracy of the proposed priors or assumptions. When these priors fail to adequately represent certain types of hazy images, the dehazing effectiveness significantly diminishes [3, 15].

With the rapid development of deep learning technology, RSI dehazing methods have gradually evolved from traditional physical models to data-driven learning methods. As shown in Fig. 1, based

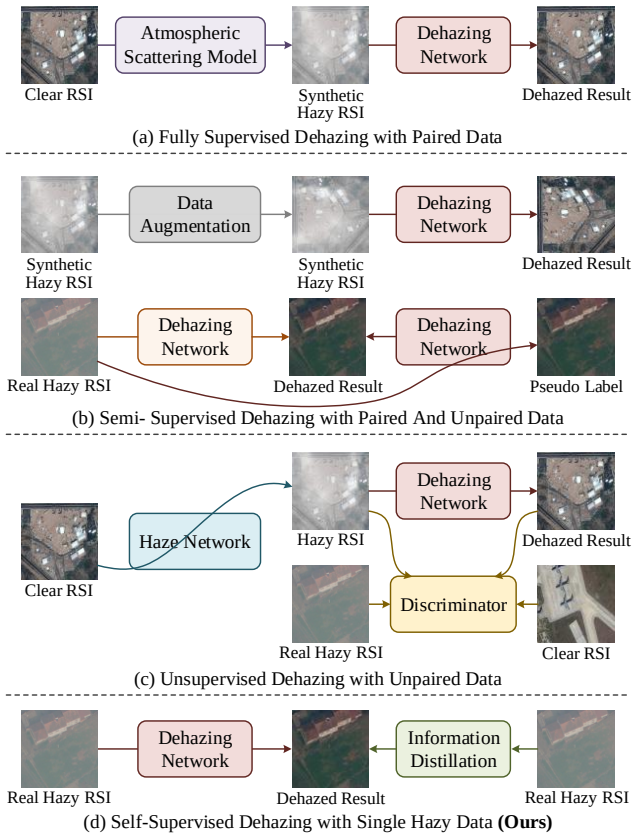


Figure 1: Schematic diagram of different data-driven methods for RSI dehazing.

on the extent to which the model relies on supervisory signals, existing methods can be categorized into three main categories.

Fully Supervised Dehazing with Paired Data. Methods within this paradigm [5, 8, 32, 33, 45, 54] rely on large-scale paired datasets of hazy and clear images, directly learning the mapping from hazy inputs to clear outputs through end-to-end training. These fully supervised learning-based dehazing methods can usually generate very promising results if fed with enough paired data [27, 52]. However, they exhibit poor generalization performance and struggle to handle real-world hazy images.

Semi-Supervised Dehazing with Paired and Unpaired Data. To mitigate limited labeled data, these methods [9, 16, 50] leverage small labeled sets with larger unlabeled data for joint training. Some approaches [2, 41] incorporate physical priors to enhance supervision, while others [6, 17, 39] exploit data distribution characteristics through augmentation and adversarial learning to enforce consistency on unlabeled samples. Despite their effectiveness, label noise and distribution gaps between labeled and unlabeled data can lead to pseudo-labeling errors and degraded performance.

Unsupervised Dehazing with Unpaired Data. Techniques falling within this category [22, 43, 61] achieve dehazing via domain transfer without the need for paired training data by leveraging an adversarial learning framework. By applying a style transfer paradigm between the hazy and clear image domains and enforcing cycle consistency constraints (or using adversarial training), they ensure both the authenticity and semantic consistency of the generated images. Since the aforementioned methods require training the network with unpaired images, thereby converting the dehazing task into a style transformation, the network still struggles to handle real-world hazy images effectively. Moreover, most of these methods, which are designed for natural images, are ill-suited for RSIs, often leading to a loss of rich texture and spectral details during the dehazing process.

In real-world applications, obtaining ground-truth haze-free RSIs is often prohibitively expensive or even unfeasible. Moreover, a notable gap exists between synthetic hazy images and the actual conditions encountered in practice. In light of these challenges, we ask: can we develop a robust dehazing method that relies solely on a single hazy RSI?

Recently, the Contrastive Language-Image Pre-training (CLIP) [35] model has demonstrated remarkable capabilities in cross-modal understanding between text and images. Its architecture, trained on 400 million image-text pairs, enables powerful zero-shot transfer to various visual tasks without task-specific fine-tuning [7, 11, 59]. This dataset-agnostic nature of CLIP [25, 51], combined with the semantic patch relationships captured by multi-head self-attention (MHSA) in its Vision Transformer (ViT) backbone [28], makes it particularly suitable for self-supervised image dehazing. The strong correlation between textual descriptions and corresponding visual elements indicates that CLIP can effectively identify the degree of degradation of hazy images, which is expected to guide self-supervised dehazing models to learn samples more efficiently.

This paper proposes a novel cross-modal guided self-supervised dehazing framework CLIP-HNet specifically designed for RSI. As shown in Fig. 1(d), our method distills inherent self-supervised cues from the input hazy images and incorporates CLIP-guided progressive information distillation to adapt dynamically to the unique characteristics of each scene. This enables the framework to generate high-quality, haze-free outputs without the requirement of paired training data.

The key contributions are summarized as follows:

- To fully extract missing context information and seamlessly integrate global-local features from hazy RSIs, we propose a novel hybrid feature interaction network, which bridges the feature representations of transformer and convolutional network, enabling a more comprehensive feature fusion to improve dehazing results.
- To effectively derive pseudo-supervised signals, we propose a CLIP-guided progressive information distillation scheme that leverages cross-modal guidance to create an adaptive training framework. This framework dynamically selects samples based on multimodal haze density assessment, while progressively escalating learning challenges via a difficulty-adaptive curriculum.
- We introduce frequency-domain texture matching (FTM) to enhance model performance by refining the texture and spectral details. By leveraging the frequency domain, FTM decomposes images into sub-bands to extract and preserve critical high-frequency components, ensuring improved restoration of fine details and overall spectral consistency.

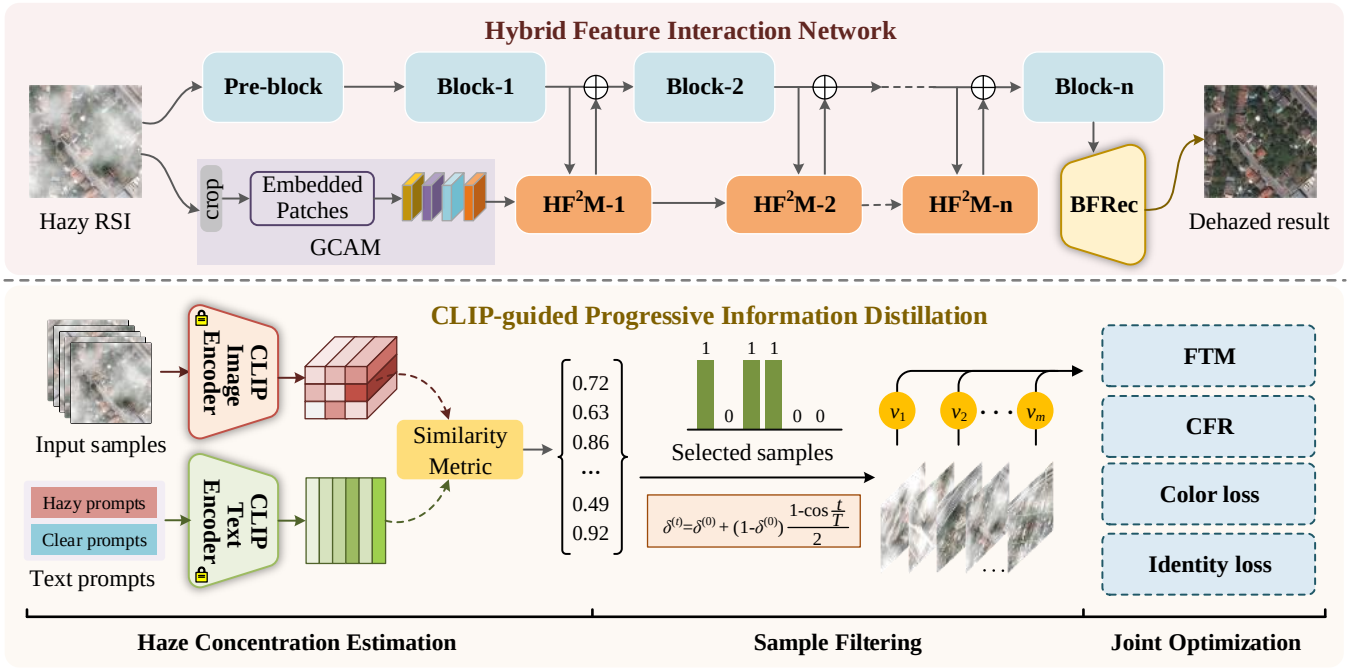


Figure 2: The framework of the proposed model.

2 Related Work

2.1 Fully Supervised RSI Dehazing

With the success of CNNs in computer vision, fully supervised learning has become the mainstream for RSI dehazing. The key challenge lies in jointly modeling the physical laws of ASM and the semantic features of surface scenes. Early work like DehazeNet [5] pioneered end-to-end CNN-based learning. Subsequent methods enhanced feature extraction via multi-scale structures (FFA-Net [33]) and dense connections (DCPDN [53]). Recent advances introduce Transformer and diffusion models: DehazeFormer [38] captures global degradation via window attention, PCSformer [57] uses cross-stripe and proxy Transformers, and RSHazeDiff [46] incorporates Fourier priors to improve texture and color. Despite strong performance, these methods heavily rely on high-quality paired data, which is often scarce in real-world scenarios.

2.2 Unsupervised RSI Dehazing

Recently, some scholars have begun to focus on unsupervised learning-based dehazing techniques [21, 22, 43, 49, 60]. Li et al. [21] proposed a zero-shot dehazing method, which disentangles hazy images into multiple layers eliminating the need for paired training data. Zhao et al. [60] presented RefineDNet, which combines prior-based visibility restoration with learning-based realism enhancement. Li et al. [22] developed an unsupervised single image dehazing network (USID-Net), which leverages disentangled representations without requiring paired training images. Yang et al. [48] proposed a self-supervised enhanced dehazing framework, which re-rendered the haze effect through depth estimation to improve the generalization ability of the model, and introduced contrast-aware loss to optimize the dehazing quality. Wang et al. [43] proposed an unsupervised contrastive learning paradigm for image dehazing using unpaired images to train the model. Zheng et al. [61] presented Dehaze-TGGAN, a Transformer-enhanced CycleGAN framework that removes haze from satellite imagery by incorporating spatial-spectral attention mechanisms.

2.3 CLIP based Image Dehazing

The integration of CLIP models into image dehazing tasks has gained significant attention due to their powerful cross-modal representation capabilities [29]. Wang et al. [40] proposed a language-guided adaptation framework for image dehazing. Zhang et al. [56] proposed DehazeMatic, where Mamba and Transformer are integrated by CLIP for image dehazing. In addition, Ren et al. [36] introduced a triplane-smoothed video dehazing model based on CLIP-enhanced generalization. These approaches typically leverage the rich semantic representations learned by CLIP to provide more informed guidance for dehazing task.

3 Methodology

Our goals are twofold: (1) enhancing multi-model feature extraction and boundary-focused reconstruction for effective haze removal, and (2) developing a dynamic sample filtering mechanism that leverages contrastive cues for efficient unsupervised optimization. Fig. 2 shows our model structure. For the first goal, we design a hybrid feature interaction network (HFIN) that fuses multi-model representations via a global context-aware module (GCAM) and a hybrid feature fusion module (HF^{2}M). Additionally, a boundary-aware feature reconstruction (BFRec) module refines edge details to boost restoration accuracy. For the second goal, we propose a CLIP-guided progressive information distillation scheme, adaptively prioritizing training samples by estimating haze concentration and increasing

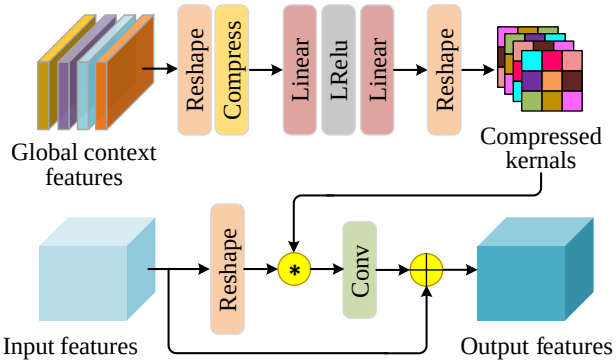


Figure 3: Hybrid Feature Fusion Module (HF^{2}M).

sample difficulty during training. This strategy distills more reliable supervision signals in an unsupervised setting.

3.1 Hybrid Feature Interaction Network

As illustrated in Fig. 2, the proposed HFIN is built upon an encoder-decoder architecture, designed to extract and integrate both local and global features. The encoder consists of three main components: the backbone, GCAM, and HF^{2}Ms. These components work collaboratively to produce high-quality, multi-scale features.

The backbone is responsible for extracting local features. Initially, the input image is processed by a pre-block, which generates shallow feature maps through a combination of a convolution layer and a max-pooling layer. These feature maps are then passed through n encoder layers (typically $n = 4$) adapted from ResNet [13], enabling the extraction of hierarchical local features.

Global Context-Aware Module. The GCAM is designed to capture global context information by collecting D-dimensional context features from multiple layers. The input image is divided into non-overlapping patches of size 32×32 during patch embedding. These patches are flattened and projected into D-dimensional tokens, which are further enhanced with position embeddings. The tokens are then processed through four attention blocks, leveraging the long-range dependency modeling capabilities of transformers [37, 38, 42]. Finally, the feature tokens from all attention blocks are concatenated to serve as inputs for feature interaction. In each encoder layer, the HF^{2}M integrates these global context features with local features.

Hybrid Feature Fusion Module. The HF^{2}M fuses global and local features by dynamically generating filter weights from global context features. As shown in Fig. 3, the i-th block of the HF^{2}Ms first compresses the global context features $\mathcal{G}$ into a smaller vector $\bar{\mathcal{G}}$ using a multi-layer perceptron (MLP). These compressed features are then transformed via two linear layers and reshaped into compressed kernals Θ, which are used to convolve the local feature maps. This convolution operation produces hybrid feature maps that combine global and local information. A 1×1 convolutional layer is subsequently applied to adjust the channel dimensions of the hybrid feature maps, which are then added back to the original local feature maps via a residual connection. This process ensures efficient feature integration and propagation through the network.

Boundary-aware Feature Reconstruction. In general, the haze-free image can be obatined by the reversing ASM:

$$J(x) = \frac{I(x) - A}{t(x)} + A. \tag{2}$$

However, directly estimating $t(x)$ via a neural network can lead to numerical instability when $t(x)$ approaches zero, causing exploding gradients. Enforcing valid ranges for $t(x)$ also becomes non-trivial. Without proper regularization or priors, the network may learn physically inconsistent solutions, compromising both interpretability and generalization.

To address above issues, we propose a novel BFRec as the decoder of HFIN that estimates an "inverse transmission map" $T(x)$ instead. By avoiding direct prediction of $t(x)$ in the denominator, $T(x)$ can reduce numerical instability and simplifies the training process, while also adhering more closely to physical constraints.

In contrast to $t(x)$, which often benefits from being relatively blurred [12], the inverse transmission map $T(x)$ requires sharper texture details to more accurately reconstruct the haze-free images. Therefore, we extracte gradient features from RSI as boundary information to further enhance the generated $T(x)$:

$$T_e(x) = T(x) + T(x) * B(x), \tag{3}$$

where $B(x)$ is boundary information, which is calculated by Sobel operator. Then, the enhanced $T_e(x)$ is substituted into $\hat{J}(x) = A - (A - I(x)) * T(x)$ to obtain the dehazed results. In this method, A is chosen as the average of the first 0.1% brightest pixels in the hazy RSI, leveraging the bright-spot scattering phenomenon commonly observed under uneven haze conditions. Physically, this scattering elevates intensity in certain regions, creating bright spots that provide a suitable approximation of the global atmospheric light.

3.2 CLIP-guided Progressive Information Distillation

While conventional training strategies treat all samples equally regardless of difficulty (e.g., thin-haze vs. dense-haze samples), this uniform approach proves inefficient. Challenging samples with dense haze introduce excessive noise into the training process, hindering optimal learning, especially in early stages.

To address the aforementioned issues, we introduce a CLIP-guided progressive information distillation approach, which integrates haze concentration estimation and a sample filtering strategy into a unified optimization process (see Fig. 2). This method selectively updates training samples based on their haze concentration, gradually increasing sample difficulty over time. By prioritizing easier samples initially and progressively incorporating more challenging ones, the model achieves a smoother learning curve and enhanced robustness.

Haze Concentration Estimation. As illustrated in Fig. 4, CLIP can effectively distinguish varying levels of haze concentration using carefully crafted text prompt sets. This ability is attributed to CLIP's extensive pre-training on diverse image-text pairs, enabling it to capture subtle atmospheric characteristics. By utilizing contrastive prompt sets, CLIP can infer the presence and intensity of haze in a given image with impressive sensitivity.

Algorithm 1 CLIP-guided Sample Filtering

Require: Inputs $I = \{I^{(m)}\}_{m=1}^{n}$; total training epochs $T_{\max}$.
Ensure: Updated model parameters θ.

Initialization:
  Given initial threshold $\delta^{(0)}$ and scaling factor T.
for each epoch $t = 1$ to $T_{\max}$ **do**
  if $t <= T$ **then**
    Update the control factor $\delta^{(t)}$ by (5).
    for each hazy image $I^{(m)}$ **do**
      Compute the haze concentration $P(haze \mid I^{(m)})$ by (4).
      if $P(haze \mid I^{(m)}) \leqslant \delta^{(t)}$ **then**
        Set binary selection vector $v_m \leftarrow 1$.
      else
        Set $v_m \leftarrow 0$.
      end if
    end for
  else if $t > T$ **then**
    for each hazy image $I^{(m)}$ **do**
      Set $v_m \leftarrow 1$.
    end for
  end if
  Compute joint loss $\mathbb{E}(\mathbf{w}, \mathbf{v})$ by (7).
  Update model parameters θ using backpropagation.
end for

To fully harness this capability, we propose a prompt-based framework tailored for haze concentration estimation. Specifically, we construct two complementary categories of textual prompts: hazy prompts (T_h) and clear prompts (T_c). The hazy prompts are designed to describe aerial scenes with haze, using expressions such as "non-uniform fog" and "bright scattering" while the clear prompts capture haze-free conditions with phrases like "clear scene" and "no haze".

Given an input image I, we first extract global visual features using CLIP's image encoder E_i, and compute their similarity with text embeddings produced by the text encoder E_t. This process yields a probabilistic estimate of haze presence:

$$P(haze|I) = \frac{\sum_{h \in T_h} S(E_i(I), E_t(h))}{\sum_{k \in \{T_h, T_c\}} S(E_i(I), E_t(k))}, \tag{4}$$

where $S(\cdot, \cdot)$ denotes the cosine similarity between image and text embeddings. This formulation allows CLIP to serve as a semantic haze detector, laying the foundation for reliable and interpretable haze concentration estimation.

CLIP-guided Sample Filtering. Inspired by self-paced learning [18], we introduce a control factor δ to dynamically adjust the selection of training samples. The value of δ evolves during training and is defined as:

$$\delta^{(t)} = \delta^{(0)} + (1 - \delta^{(0)}) \frac{1 - \cos \frac{t}{T}}{2}, \tag{5}$$

where $\delta^{(0)}$ is the initial threshold, t is the current epoch, and T serves as a scaling factor to modulate the cosine term. At each epoch t, the value of δ determines which samples are included in the training process. The sample inclusion criterion is expressed as:

$$v_m = \begin{cases} 1, & \text{if } P(haze|I^{(m)}) \leqslant \delta. \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where $P(haze|I^{(m)})$ is the probabilistic estimate of the m-th sample $I^{(m)}$, calculated by CLIP. Smaller $P(haze|I^{(m)})$ values indicate lower haze concentration, making such samples easier for the model to process. Therefore, the process of progressive information distillation can be expressed as:

$$\min_{\mathbf{w},\mathbf{v}} \mathbb{E}(\mathbf{w}, \mathbf{v}) = \sum_{m=1}^{n} v_m \mathcal{L}_{Joint}(f(I^m, \theta_f)) - \delta \sum_{m=1}^{n} v_m, \tag{7}$$

where I^m is the m-th training sample, and $\mathbf{v} = [v_1, v_2, \ldots, v_m]$ is a binary selection vector determined by (6). We show the pseudo-code implementation of CLIP-guided sample filtering in Algorithm 1.

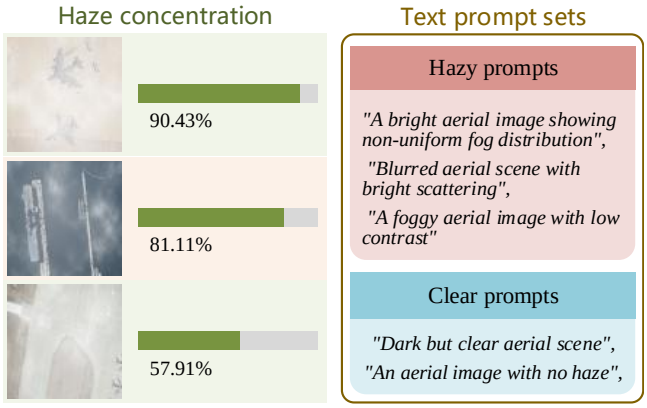


Figure 4: Sample haze concentration estimation by CLIP and corresponding text prompt sets.

3.3 Joint Optimization Strategy

Frequency-domain Texture Matching. RSIs inherently preserve rich texture and spectral details, even in the presence of haze. These textures and details are crucial for maintaining the visual and structural integrity of dehazed results. However, these features are often degraded during the dehazing process. To address this, we propose a frequency-domain texture matching (FTM) strategy that leverages frequency-domain analysis to enhance texture and spectral refinement in a self-supervised manner.

Our approach utilizes the discrete wavelet transform (DWT) to decompose an image $I \in \mathcal{R}^{W \times H \times C}$ into four sub-bands: one low-frequency approximation sub-band and three high-frequency detail sub-bands, which can be expressed as:

$$I_{LL}, I_{LH}, I_{HL}, I_{HH} = \Psi(I), \tag{8}$$

where $\Psi(\cdot)$ represents the DWT operation, I_{LL} is the low-frequency sub-band containing global structure and smooth intensity variations, and I_{LH}, I_{HL}, I_{HH} are high-frequency sub-bands capturing finer details and edges along horizontal, vertical, and diagonal

orientations, respectively. These high-frequency components are particularly valuable for extracting texture and edge features.

For hazy images, the high-frequency texture features are defined as $T^h = \{I_{LH}, I_{HL}, I_{HH}\}$, while for the corresponding dehazed results, the texture features are $T^d = \{J_{LH}, J_{HL}, J_{HH}\}$. To ensure that the dehazing model retains the fine textures and edges inherent in the original hazy images, we introduce a FTM loss defined as:

$$\mathcal{L}_{ftm} = \frac{1}{3}\sum_{i=1}^{3}\left\|T_i^h - T_i^d\right\|_1, \tag{9}$$

where T_i^h and T_i^d are the high-frequency sub-bands of the hazy and dehazed images, respectively.

Contrastive Feature Representation. In the context of unsupervised contrastive representation for image dehazing, the anchor is defined as the recovered result produced by the proposed dehazing network, the positive as the recovered result from an existing dehazing model, and the negative as the hazy input. The goal of this contrastive representation is to minimize the L_1 distance between the embeddings of the anchor and the positive while maximizing their distance from the negative. This can be formulated as:

$$\mathcal{L}_{cfr} = \sum_{i=1}^{n}\omega_i \cdot \frac{\left\|\Phi_i(g(I,\theta_g)) - \Phi_i(f(I,\theta_f))\right\|_1}{\left\|\Phi_i(I) - \Phi_i(f(I,\theta_f))\right\|_1 + \epsilon}, \tag{10}$$

where $f(\cdot,\theta_f)$ denotes the proposed dehazing network with parameters θ_f, and $g(\cdot,\theta_g)$ denotes an existing dehazing model with parameters θ_g. ϵ is a small positive constant added to prevent division by zero. For ease of implementation, we use the results generated by [12] as $g(\cdot,\theta_g)$. Additionally, $\Phi_i(\cdot)$ represents the i-th latent feature map extracted from VGG-19, and ω_i is the weight coefficient for the i-th feature map, empirically set it to 1/16, 1/8, 1/4, 1/2, and 1.

In addition to the two aforementioned loss functions, we also incorporate a color loss $\mathcal{L}_{color}$ and an identity loss $\mathcal{L}_{iden}$ to preserve the color fidelity and structural consistency of the dehazed output. Specifically, $\mathcal{L}_{color}$ measures the discrepancy between the U and V channels of the dehazed image and the original image in the YUV color space, ensuring color consistency. Meanwhile, $\mathcal{L}_{iden}$ is computed by comparing the dark channel of the dehazed result with that of the reference image generated by the method in [12], thereby maintaining structural integrity. The overall optimazation $\mathcal{L}_{Joint}$ is the sum of all the above loss functions.

4 Experiments and Results

4.1 Experimental Settings

Datasets. We evaluate our proposed framework on multiple publicly available datasets, including SateHaze1k [14] and RSID [8]. SateHaze1k contains 1,200 pairs of nonuniformly hazy RSIs captured by the GF-2 satellite, categorized into three haze concentration levels: thin, medium, and thick. From this dataset, we allocate 1,078 image pairs for training and 120 pairs for testing. Notably, CLIP-HNet requires only the hazy images during training, without utilizing the corresponding clear images. RSID provides 900 synthetic hazy images for training and 100 for testing. To further validate our model's restoration quality and generalization capability, we incorporate two additional real-world hazy RSI datasets: MODIS [54] and UAV.

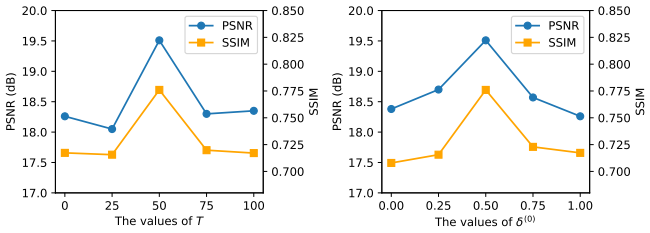


Figure 5: The relationship between hyperparameters (T, $\delta^{(0)}$) and the dehazing quality.

Implementation Details. We implement CLIP-HNet using PyTorch and train on an NVIDIA RTX 4090 GPU. The decoder of CLIP-HNet is mainly based on RDN [58]. The optimizer of CLIP-HNet is Adam optimizer [20] with a batch size of 16. The initial learning rate is set to 2×10^{-4} and decreases by a factor of 0.5 whenever validation performance deteriorates for three consecutive epochs. To enhance model robustness and generalization capability, we apply data augmentation during training by randomly rotating input images at angles of 0°, 90°, 180°, 270°.

4.2 Exploration of Model Hyperparameters

Setting T in CLIP-guided sample filtering. The parameter T in (5) significantly influences the sample filtering rate's progression. As a temporal scaling factor, T determines how rapidly the filtering threshold $\delta^{(t)}$ transitions from its initial value $\delta^{(0)}$ toward its maximum. When T is larger, the cosine term $\cos\frac{t}{T}$ decreases more gradually as t increases, resulting in a slower, more measured change in $\delta^{(t)}$. Conversely, a smaller T accelerates this transition, causing $\delta^{(t)}$ to reach higher values more quickly. As shown in Fig. 5, the experimental results clearly support selecting $T = 50$ as the optimal value. This choice achieves a significantly higher dehazing quality compared to other settings. It provides the optimal balance between early training stability and progressive refinement, allowing the model to establish foundational learning before gradually incorporating more challenging hazy instances while avoiding both premature difficulty introduction and excessively delayed filtering.

Setting $\delta^{(0)}$ in CLIP-guided sample filtering. The parameter $\delta^{(0)}$ in (5) serves as the initial threshold value for sample filtering, fundamentally determining the starting point of the adaptive selection process. This parameter directly impacts both the initial filtering strictness and the overall trajectory of threshold evolution throughout training. As shown in Fig. 5, our experimental results support choosing $\delta^{(0)} = 0.5$ as optimal. This value achieves the highest dehazing quality, outperforming other configurations by at least 1.16 dB. Setting $\delta^{(0)} = 0.5$ provides a balanced approach that filters approximately half the samples initially, allowing the model to focus on moderately challenging examples from the beginning while maintaining sufficient diversity in the training set. This balanced initial filtering avoids both the overly permissive inclusion of all samples ($\delta^{(0)} = 0$) and the excessively strict filtering ($\delta^{(0)} = 1$) that could limit learning from valuable training signals.

4.3 Comparison with the State-of-the-arts

The comparative methods are categorized into two groups: supervised and unsupervised approaches. The supervised algorithms

Table 1: Quantitative comparison on two paired nonuniform hazy RSI datasets.The best performance is shown in red and the second best is shown in blue.

Method	Type	SateHaze1k			RSID		
		PSNR (dB) ↑	SSIM ↑	FSIM ↑	PSNR (dB) ↑	SSIM ↑	FSIM ↑
FFA-Net$_{(AAAI2020)}$ [33]	Supervised	23.12	0.9151	0.9685	20.85	0.9058	0.9871
FCTF-Net$_{(GRSL2021)}$ [23]	Supervised	22.43	0.9051	0.9593	20.56	0.8984	0.9812
DCIL$_{(TGRS2022)}$ [54]	Supervised	24.85	0.9124	0.9782	20.46	0.7271	0.9749
Trinity-Net$_{(TGRS2023)}$ [8]	Supervised	23.12	0.8904	0.9641	21.52	0.8774	0.9795
RSHazeNet$_{(ICASSP2024)}$ [45]	Supervised	23.94	0.8963	0.9710	20.39	0.8898	0.9821
ZID$_{(TIP2020)}$ [21]	Unsupervised	12.32	0.5281	0.8997	11.46	0.3259	0.9274
USID-Net$_{(TMM2022)}$ [22]	Unsupervised	17.36	0.6489	0.8976	16.58	0.6335	0.9535
UCLD$_{(TIP2024)}$ [43]	Unsupervised	16.43	0.7256	0.9563	17.50	0.7269	0.9586
D4+$_{(IJCV2024)}$ [48]	Unsupervised	13.15	0.6048	0.8940	15.90	0.8259	0.9608
UR2P$_{(Arvix2025)}$ [47]	Unsupervised	11.87	0.4640	0.8601	14.59	0.6582	0.9000
CLIP-HNet (Ours)	Unsupervised	19.51	0.7760	0.9682	17.63	0.8367	0.9665

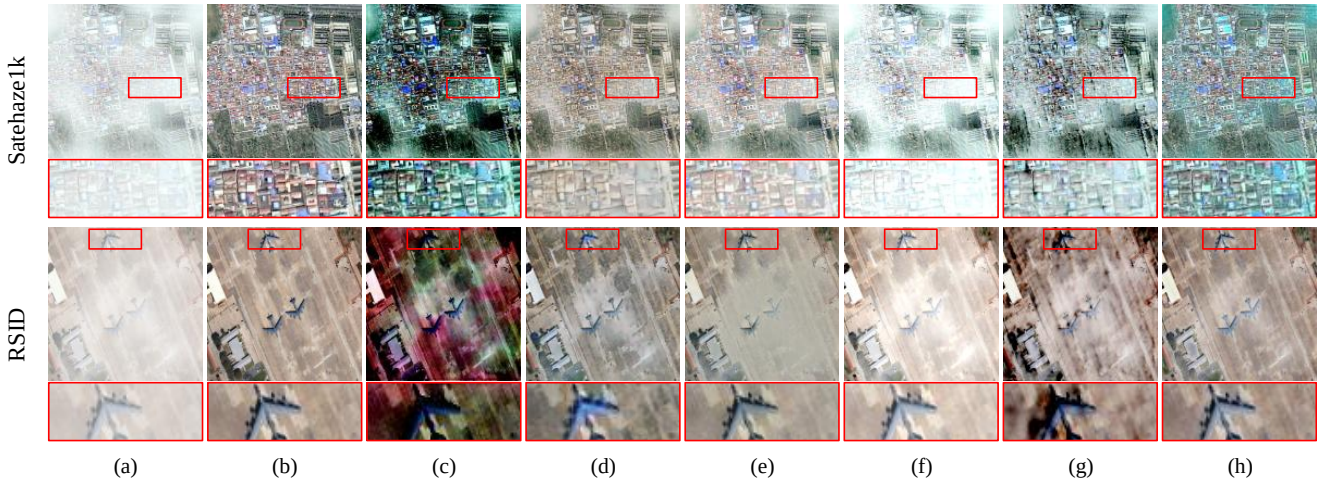


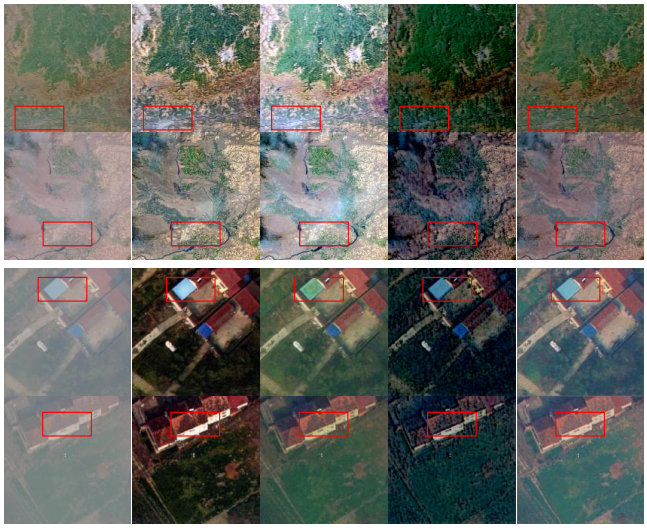
Figure 6: Visual comparisons of our CLIP-HNet and other unsupervised dehazing methods on restoring RSIs from SateHaze1k [14] and RSID [8]. (a) Hazy RSIs. (b) Ground Truths. (c) ZID [21]. (d) USID-Net [22]. (e) UCLD [43]. (f) D4+ [48]. (g) UR2P [47]. (h) CLIP-HNet (Ours).

include FFA-Net [33], FCTF-Net [23], DCIL [54], Trinity-Net [8], and RSHazeNet [45], while the unsupervised algorithms comprise ZID [21], USID-Net [22], UCLD [43], D4+ [48] and UR2P [47].

Evaluation on Synthetic Datasets. We evaluate performance using established quantitative metrics including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [44], and Feature Similarity Index (FSIM) [55]. Table 1 provides a comprehensive quantitative comparison across two paired nonuniform hazy RSI datasets. Supervised approaches (e.g., DCIL [54] and Trinity-Net [8]) consistently show strong performance but depend on paired training data. In contrast, unsupervised methods face greater challenges but offer broader applicability; among them, UCLD [43] and D4+ [48] perform relatively well yet still struggle in more complex scenes. Our proposed CLIP-HNet surpasses all unsupervised counterparts, achieving the highest PSNR and superior SSIM/FSIM, due to its cross-modal guidance. Notably, its performance narrows the gap with supervised methods or even exceeds them, indicating robust haze removal without explicit supervision.

Visual comparisons (Fig. 6) confirm that CLIP-HNet restores sharper textures and more natural colors. For instance, other approaches often show residual haze or missing details in the red boxes, while CLIP-HNet preserves building rooftops and airplane contours with greater fidelity and fewer artifacts, highlighting the advantages of its cross-modal design.

Evaluation on Real-world RSIs. From the visual results in Fig. 7, CLIP-HNet removes haze more effectively in challenging regions, such as mountainous terrain or densely vegetated areas. Compared with UCLD and UR2P, for instance, CLIP-HNet retains more realistic colors and sharper textures, mitigating color shifts or over-smoothing artifacts. Additionally, in urban environments, edges of buildings and other structures appear crisper, reflecting CLIP-HNet's capacity to preserve subtle boundaries while effectively suppressing haze-induced distortions.



(a) (b) (c) (d) (e)

Figure 7: Visual comparison on real-world hazy RSIs. (a) Real-world hazy RSIs. (b) USID-Net [22]. (c) UCLD [43]. (d) UR2P [47]. (e) CLIP-HNet (Ours).

Table 2: The ablation experiments of modules on SateHaze1k dataset.

GCAM	HF^{2}Ms	BFRec	CLIP-PID	PSNR (dB) ↑	SSIM ↑	FSIM ↑
✗	✗	✓	✓	17.43	0.7076	0.9492
✓	✗	✓	✓	18.54	0.7155	0.9579
✓	✓	✗	✓	18.62	0.7335	0.9614
✓	✓	✓	✗	18.26	0.7173	0.9569
✓	✓	✓	✓	**19.51**	**0.7760**	**0.9682**

4.4 Ablation Studies

We performed ablations on different components of the model and loss functions on the SateHaze1k dataset.

Investigation of GCAM. As shown in the first two rows of Table 2, adding the GCAM leads to a notable improvement in PSNR from 17.43 dB to 18.54 dB, as well as consistent gains in SSIM and FSIM. This indicates that capturing long-range dependencies via transformer-based attention blocks significantly enhances dehazing performance. Furthermore, the global context features from GCAM provide complementary information that refines local patch representations in subsequent modules.

Investigation of HF^{2}Ms. The HFIMs are replaced with a pixel unshuffle operation and a convolution layer in ablation model. Comparing the second row (w/o HF^{2}Ms) with the final row (full model) reveals that incorporating the HF^{2}Ms raises the PSNR from 18.54 dB to 19.51 dB. This improvement underscores the value of dynamically blending local and global features, where filter weights are adaptively generated from global context. As a result, HF^{2}Ms effectively consolidate multi-scale information, yielding clearer structural details in the dehazed output.

Table 3: Ablation studies of the contribution of loss functions.

Loss function	PSNR (dB) ↑	SSIM ↑	FSIM ↑
w/o $\mathcal{L}_{ftm}$	17.49	0.7071	0.9522
w/o $\mathcal{L}_{cfr}$	18.45	0.7167	0.9571
w/o $\mathcal{L}_{color}$	18.32	0.7145	0.9577
Full Loss	**19.51**	**0.7760**	**0.9682**

Investigation of BFRec. In this ablation model, the boundary informtion is removed from BFRec. By contrasting the third row and the full model, one observes a jump in PSNR from 18.62 dB to 19.51 dB and accompanying increases in SSIM/FSIM. This confirms that the proposed BFRec with boundary informtion further refines output quality by preserving edge information. Consequently, BFRec not only improves visual sharpness but also ensures consistent physical interpretation in haze removal.

Investigation of CLIP-PID. As shown in Table 2, the proposed model with CLIP-PID achieves clear performance boost demonstrating the effectiveness of progressively incorporating challenging samples based on haze concentration. By focusing on simpler data early on and gradually introducing heavily hazed samples, the model avoids being overwhelmed by high-noise inputs. Consequently, CLIP-PID not only stabilizes training but also leads to more robust generalization.

Investigation of Loss Functions. The $\mathcal{L}_{ftm}$ preserves critical high-frequency textures and edges, mitigating over-smoothing. Meanwhile, $\mathcal{L}_{cfr}$ enhances network robustness by aligning with reference images while distancing from hazy inputs. In addition, the $\mathcal{L}_{color}$ maintains color fidelity through YUV space comparisons. Ablation studies in Table 3 show removing any loss causes significant drops in PSNR, SSIM, and FSIM metrics. Combining these losses produces high-quality dehazed outputs by balancing texture preservation, color accuracy and structural consistency.

5 Conclusion

In this paper, we introduced a novel unsupervised dehazing framework for RSI, termed CLIP-HNet. By incorporating cross-modal guidance from CLIP, our approach effectively captures both global and local representations through the GCAM and the HF^{2}Ms. The proposed BFRec further refines edge details, enhancing structural fidelity in the final dehazed images. Additionally, the CLIP-guided progressive information distillation scheme dynamically filters training samples based on haze concentration, fostering a smoother learning trajectory that steadily accommodates increasingly complex hazy scenarios. Lastly, the FTM strategy preserves subtle texture and spectral details, mitigating common oversimplifications of haze removal as mere style transfer. Extensive experiments on both synthetic and real-world RSI datasets confirm the superior performance of CLIP-HNet, showcasing its robust generalization and improved visual quality over existing methods.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62271060, the Beijing Natural Science Foundation under Grant 4222046, and the Ministry of Education of Singapore Tier3 Grant MOET32022-0006.

References

[1] Hasan M Abdullah, Nusrat T Mohana, Bhoktear M Khan, Syed M Ahmed, Maruf Hossain, KH Shakibul Islam, Mahadi H Redoy, Jannatul Ferdush, MAHB Bhuiyan, Motaher M Hossain, et al. 2023. Present and future scopes and challenges of plant pest and disease (P&D) monitoring: Remote sensing, image processing, and artificial intelligence perspectives. *Remote Sensing Applications: Society and Environment* 32 (2023), 100996.

[2] Shunmin An, Linling Wang, and Le Wang. 2024. Semi-supervised dehazing network using multiple scattering model and fuzzy image prior. *Applied Intelligence* 54, 7 (2024), 5794–5812.

[3] Abeer Ayoub, Walid El-Shafai, Fathi E Abd El-Samie, Ehab KI Hamad, and El-Sayed M EL-Rabaie. 2024. Review of dehazing techniques: Challenges and future trends. *Multimedia Tools and Applications* (2024), 1–29.

[4] Dana Berman, Shai Avidan, et al. 2016. Non-local image dehazing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1674–1682.

[5] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. 2016. Dehazenet: An end-to-end system for single image haze removal. *IEEE transactions on image processing* 25, 11 (2016), 5187–5198.

[6] Wei-Ting Chen, I-Hsiang Chen, Chih-Yuan Yeh, Hao-Hsiang Yang, Hua-En Chang, Jian-Jiun Ding, and Sy-Yen Kuo. 2022. Rvsl: Robust vehicle similarity learning in real hazy scenes based on semi-supervised learning. In *European Conference on Computer Vision*. Springer, 427–443.

[7] Jun Cheng, Dong Liang, and Shan Tan. 2024. Transfer CLIP for generalizable image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25974–25984.

[8] Kaichen Chi, Yuan Yuan, and Qi Wang. 2023. Trinity-Net: Gradient-guided Swin transformer-based remote sensing image dehazing and beyond. *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), 1–14.

[9] Xiaofeng Cong, Jie Gui, Jing Zhang, Junming Hou, and Hao Shen. 2024. A semi-supervised nighttime dehazing baseline with spatial-frequency aware and realistic brightness constraint. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2631–2640.

[10] Marcelo de Abreu. 2020. Acquiring And Extraction Of Information Collected By Unmanned Aerial Vehicles And Omnidirectional Cameras And Their Applications Through Management Software. In *2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS)*. IEEE, 279–283.

[11] Kanchana Vaishnavi Gandikota and Paramanand Chandramouli. 2024. Text-guided explorable image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25900–25911.

[12] Kaiming He, Jian Sun, and Xiaoou Tang. 2010. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence* 33, 12 (2010), 2341–2353.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[14] Binghui Huang, Li Zhi, Chao Yang, Fuchun Sun, and Yixu Song. 2020. Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

[15] Jehoiada Jackson, Kwame Obour Agyekum, Chiagoziem Ukwuoma, Rutherford Patamia, Zhiguang Qin, et al. 2024. Hazy to hazy free: A comprehensive survey of multi-image, single-image, and CNN-based algorithms for dehazing. *Computer Science Review* 54 (2024), 100669.

[16] Tongyao Jia, Jiafeng Li, and Li Zhuo. 2023. Graph Disentangled Representation Based Semi-supervised Single Image Dehazing Network. In *International Conference on Intelligent Computing*. Springer, 652–663.

[17] Tongyao Jia, Jiafeng Li, Li Zhuo, and Tianjian Yu. 2023. Semi-supervised single-image dehazing network via disentangled meta-knowledge. *IEEE Transactions on Multimedia* 26 (2023), 2634–2647.

[18] Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander Hauptmann. 2015. Self-paced curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[19] Wandong Jiang, Yuli Sun, Lin Lei, Gangyao Kuang, and Kefeng Ji. 2024. Change detection of multisource remote sensing images: a review. *International Journal of Digital Earth* 17, 1 (2024), 2398051.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[21] Boyun Li, Yuanbiao Gou, Jerry Zitao Liu, Hongyuan Zhu, Joey Tianyi Zhou, and Xi Peng. 2020. Zero-shot image dehazing. *IEEE Transactions on Image Processing* 29 (2020), 8457–8466.

[22] Jiafeng Li, Yaopeng Li, Li Zhuo, Lingyan Kuang, and Tianjian Yu. 2022. USID-Net: Unsupervised single image dehazing network via disentangled representations. *IEEE transactions on multimedia* 25 (2022), 3587–3601.

[23] Yufeng Li and Xiang Chen. 2021. A coarse-to-fine two-stage attentive network for haze removal of remote sensing images. *IEEE Geoscience and Remote Sensing Letters* 18, 10 (2021), 1751–1755.

[24] Shan Liang, Tao Gao, Ting Chen, and Peng Cheng. 2024. A remote sensing image dehazing method based on heterogeneous priors. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[25] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei He. 2023. Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15305–15314.

[26] Pengyang Ling, Huaian Chen, Xiao Tan, Yi Jin, and Enhong Chen. 2023. Single image dehazing using saturation line prior. *IEEE Transactions on Image Processing* 32 (2023), 3238–3253.

[27] Yi Liu, Jiachen Li, Yanchun Ma, Qing Xie, and Yongjian Liu. 2024. HcaNet: Haze-concentration-aware Network for Real-scene Dehazing with Codebook Priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 9136–9144.

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*. 10012–10022.

[29] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. 2023. Controlling Vision-Language Models for Universal Image Restoration. *arXiv preprint arXiv:2310.01018* (2023).

[30] Yuchi Ma, Shuo Chen, Stefano Ermon, and David B Lobell. 2024. Transfer learning in environmental remote sensing. *Remote Sensing of Environment* 301 (2024), 113924.

[31] Srinivasa G Narasimhan and Shree K Nayar. 2000. Chromatic framework for vision in bad weather. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, Vol. 1. IEEE, 598–605.

[32] Jiangtao Nie, Wei Wei, Lei Zhang, Jianlong Yuan, Zhibin Wang, and Hao Li. 2022. Contrastive haze-aware learning for dynamic remote sensing image dehazing. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–11.

[33] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. 2020. FFA-Net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 11908–11915.

[34] Zifeng Qiu, Tianyu Gong, Zichao Liang, Taoyi Chen, Runmin Cong, Huihui Bai, and Yao Zhao. 2024. Perception-oriented UAV image dehazing based on super-pixel scene prior. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[36] Jingjing Ren, Haoyu Chen, Tian Ye, Hongtao Wu, and Lei Zhu. 2025. Triplane-Smoothed Video Dehazing with CLIP-Enhanced Generalization. *International Journal of Computer Vision* 133 (2025), 475–488.

[37] Tianyu Song, Shumin Fan, Pengpeng Li, Jiyu Jin, Guiyue Jin, and Lei Fan. 2023. Learning an effective transformer for remote sensing satellite image dehazing. *IEEE Geoscience and Remote Sensing Letters* (2023).

[38] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. 2023. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing* 32 (2023), 1927–1941.

[39] Ming Tong, Xuefeng Yan, Yongzhen Wang, and Mingqiang Wei. 2024. Semi-uformer: Semi-supervised uncertainty-aware transformer for image dehazing. In *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[40] Ruiyi Wang, Wenhao Li, Xiaohong Liu, Chunyi Li, Zicheng Zhang, Xiongkuo Min, and Guangtao Zhai. 2025. HazeCLIP: Towards Language Guided Real-World Image Dehazing. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5. doi:10.1109/ICASSP49660.2025.10889509

[41] Xudong Wang, Xi'ai Chen, Weihong Ren, Zhi Han, Huijie Fan, Yandong Tang, and Lianqing Liu. 2024. Compensation atmospheric scattering model and two-branch network for single image dehazing. *IEEE Transactions on Emerging Topics in Computational Intelligence* (2024).

[42] Yongzhen Wang, Jiamei Xiong, Xuefeng Yan, and Mingqiang Wei. 2023. Uscformer: Unified transformer with semantically contrastive learning for image dehazing. *IEEE Transactions on Intelligent Transportation Systems* 24, 10 (2023), 11321–11333.

[43] Yongzhen Wang, Xuefeng Yan, Fu Lee Wang, Haoran Xie, Wenhan Yang, Xiao-Ping Zhang, Jing Qin, and Mingqiang Wei. 2024. Ucl-dehaze: Towards real-world image dehazing via unsupervised contrastive learning. *IEEE Transactions on Image Processing* (2024).

[44] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[45] Yuanbo Wen, Tao Gao, Ziqi Li, Jing Zhang, and Ting Chen. 2024. Encoder-minimal and Decoder-minimal Framework for Remote Sensing Image Dehazing. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 36–40.

[46] Jiamei Xiong, Xuefeng Yan, Yongzhen Wang, Wei Zhao, Xiao-Ping Zhang, and Mingqiang Wei. 2024. RSHazeDiff: A unified Fourier-aware diffusion model for

remote sensing image dehazing. *IEEE Transactions on Intelligent Transportation Systems* (2024).

[47] Minglong Xue, Shuaibin Fan, Shivakumara Palaiahnakote, and Mingliang Zhou. 2025. UR2P-Dehaze: Learning a Simple Image Dehaze Enhancer via Unpaired Rich Physical Prior. *arXiv preprint arXiv:2501.06818* (2025).

[48] Yang Yang, Chaoyue Wang, Xiaojie Guo, and Dacheng Tao. 2024. Robust unpaired image dehazing via density and depth decomposition. *International Journal of Computer Vision* 132, 5 (2024), 1557–1577.

[49] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. 2022. Self-augmented unpaired image dehazing via density and depth decomposition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2037–2046.

[50] Weichao Yi, Liquan Dong, Ming Liu, Mei Hui, Lingqin Kong, and Yuejin Zhao. 2023. Semi-supervised progressive dehazing network using unlabeled contrastive guidance. *Neurocomputing* 551 (2023), 126494.

[51] Qihang Yu, Xiaohui Shen, and Liang-Chieh Chen. 2024. Towards Open-Ended Visual Recognition with Large Language Models. In *European Conference on Computer Vision*. Springer, 359–376.

[52] Shenghai Yuan, Jijia Chen, Jiaqi Li, Wenchao Jiang, and Song Guo. 2023. Lhnet: A low-cost hybrid network for single image dehazing. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7706–7717.

[53] He Zhang and Vishal M Patel. 2018. Densely connected pyramid dehazing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3194–3203.

[54] Libao Zhang and Shan Wang. 2022. Dense haze removal based on dynamic collaborative inference learning for remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing* 60 (2022), 1–16.

[55] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* 20, 8 (2011), 2378–2386.

[56] Xiaozhe Zhang, Fengying Xie, Haidong Ding, Linpeng Pan, and Zhenwei Shi. 2024. Towards Optimal Aggregation of Varying Range Dependencies in Haze Removal. *arXiv e-prints*, Article arXiv:2408.12317 (Aug. 2024), arXiv:2408.12317 pages. arXiv:2408.12317 doi:10.48550/arXiv.2408.12317

[57] Xiaozhe Zhang, Fengying Xie, Haidong Ding, Shaocheng Yan, and Zhenwei Shi. 2024. Proxy and Cross-Stripes Integration Transformer for Remote Sensing Image Dehazing. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[58] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. 2018. Residual dense network for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2472–2481.

[59] Zhe Zhang, Jianjun Lei, Bo Peng, Jie Zhu, Liying Xu, and Qingming Huang. 2025. Advancing Real-World Stereoscopic Image Super-Resolution via Vision-Language Model. *IEEE Transactions on Image Processing* (2025).

[60] Shiyu Zhao, Lin Zhang, Ying Shen, and Yicong Zhou. 2021. RefineDNet: A weakly supervised refinement framework for single image dehazing. *IEEE Transactions on Image Processing* 30 (2021), 3391–3404.

[61] Yitong Zheng, Jia Su, Shun Zhang, Mingliang Tao, and Ling Wang. 2024. Dehaze-tggan: transformer-guide generative adversarial networks with spatial-spectrum attention for unpaired remote sensing dehazing. *IEEE Transactions on Geoscience and Remote Sensing* (2024).

[62] Qingsong Zhu, Jiaming Mai, and Ling Shao. 2015. A fast single image haze removal algorithm using color attenuation prior. *IEEE transactions on image processing* 24, 11 (2015), 3522–3533.