# Hierarchical Semantic-Guided Contextual Structure-Aware Network for Satellite Image Dehazing

Lei Yang , Jianzhong Cao , Hua Wang , Sen Dong , Hailong Ning *

Posted Date: 9 January 2024

doi: 10.20944/preprints202401.0679.v1

Keywords: Satellite Image Dehazing; CNN-Transformer; Hierarchical Semantic Guidance; Cross-Layer Fusion

*Article*

# Hierarchical Semantic-Guided Contextual Structure-Aware Network for Satellite Image Dehazing

**Lei Yang [1,2], Jianzhong Cao [1,2], Hua Wang [1,2], Sen Dong [1], and Hailong Ning [3]***

[1] Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, 710119, China
[2] University of Chinese Academy of Sciences, Beijing, 100049, China
[3] School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, 710121, China
* Correspondence: ninghailong93@gmail

**Abstract:** Haze always shrouds satellite images, obscuring valuable geographic information for military surveillance, natural calamity surveillance and mineral resource exploration. Satellite image dehazing (SID) provides the possibility for better applications of satellite images. Most of the existing dehazing methods are tailored for natural images and are not very effective for satellite images with non-homogeneous haze since the semantic structure information and inconsistent attenuation are not fully considered. To tackle this problem, this study proposes a hierarchical semantic-guided contextual structure-aware network (SCSNet) for satellite image dehazing. Specifically, a hybrid CNN-transformer architecture integrated with a hierarchical semantic guidance (HSG) module is presented to learn semantic structure information by synergetically complementing local representation from non-local features. Furthermore, a cross-layer fusion (CLF) module is specially designed to replace the traditional skip connection during the feature decoding stage, so as to reinforce the attention to the spatial regions and feature channels with more serious attenuation. The results on the SateHaze1k, RS-Haze, and RSID datasets demonstrate that the proposed SCSNet can achieve effective dehazing and outperforms existing state-of-the-art methods.

**Keywords:** satellite image dehazing; CNN-transformer; hierarchical semantic guidance; cross-layer fusion

## 1. Introduction

Satellite images play a pivotal role in numerous fields, from military surveillance [1] and agricultural monitoring [2,3] to natural disaster response [4,5] and resource exploration [6]. However, their value plummets when shrouded in haze. Hazy images suffer from drastically reduced contrast and visibility, crippling their potential for downstream computer vision applications. Therefore, dehazing has become an important topic in satellite images interpretation, promising to enhance the enduring quality of satellite images even under adverse atmospheric conditions.

In recent years, a range of dehazing methods have been developed, roughly divided into prior and deep learning-based methods [7,8]. Most prior-based methods are based on the atmospheric scattering model and require accurate estimation for the transmission and atmospheric light [9]. To guide the estimation process, researchers have devised various prior assumptions about the haze characteristics, such as dark channel prior (DCP) [10] and color attenuation prior [11]. The DCP method [10] is the milestone study in this field, which effectively estimates the transmission map but performs poorly in sky images or white scenes, leading to color distortion. While prior-based methods have proven effective in many scenarios, they can struggle with dense and non-homogeneous haze due to the complexity of estimating multiple haze parameters accurately.

Subsequently, deep learning has revolutionized image dehazing, offering a powerful alternative to prior-based methods. Instead of relying on hand-crafted features, deep neural networks automatically learn the complex relationships within hazy images. Early models like DehazeNet [14] and AOD-Net [15] focused on estimating transmission maps for haze removal. Later approaches like FFA-Net [16]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2024 doi:10.20944/preprints202401.0679.v1



emphasized capturing image details and color fidelity through attention mechanisms. In recent years, the rise of Transformer models in computer vision [17] has sparked interest in their application to dehazing due to the ability to capture non-local interaction information. For instance, DehazeFormer [8] exemplifies this trend with its innovative normalization layer and spatial information aggregation scheme. In the past two years, some dehazing methods tailored for satellite images [18–22] are explored, promoting better applications of satellite images and sharpening our view from space.

In fact, satellite images exhibit more non-homogeneous hazy distributions, more complex terrain information, and more severe object occlusion resulting from excessive fog concentration compared to natural hazy images. Despite the remarkable progress, existing methods are not very effective for SID with the deepening of haze density and haze non-homogeneous. This is due to the significant loss of spectral information and texture details, which are crucial for accurate SID. As shown in Figure 1, all dehazing methods performed well on the first row of images with light homogeneous haze, except for non-learning DCP. However, when processing the dense non-homogeneous hazy images in the second row of Figure 1, both natural image dehazing methods (such as DCP and Restormer) and dehazing methods tailored for satellite images (such as DCRD-Net), exhibited suboptimal performance. Despite tailored for satellite image dehazing, DCRD-Net still exhibits significant color distortion.

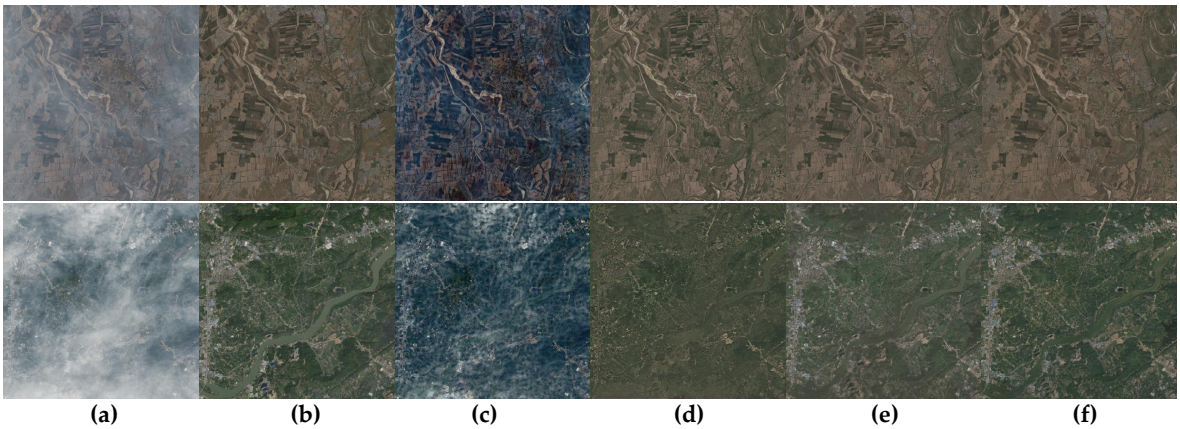


**Figure 1.** Sample results of several representative image dehazing methods. (First row) Light homogeneous haze. (Second row) Dense non-homogeneous haze. (a) Hazy images. (b) Real clean images. (c) DCP [10]. (d) Restormer [12]. (e) DCRD-Net [13]. (f) SCSNet(Ours).

Therefore, this study proposes a hierarchical semantic-guided contextual structure-aware network (SCSNet) for satellite image dehazing (SID) by integrating CNN and Vision Transformer architecture to fully explore the semantic structure information and inconsistent attenuation. The network does not rely on prior knowledge or physical models and is an end-to-end trainable dehazing network capable of directly restoring hazy images into clear ones, even with limited data. As shown in Figure 2, the proposed SCSNet consists of: CNN feature encoder branch, Transformer encoder branch, hierarchical semantic guidance (HSG) module, and image restoration branch. First, the input hazy image is processed through the CNN feature encoder branch for learning hierarchical local structure features and the Transformer encoder branch for extracting hierarchical non-local semantic information. Secondly, the HSG module is devised to learn semantic structure information by synergetically complementing multi-scale local and non-local information. Thirdly, the learned semantic structure information and multi-layer local structure features are combined to restore the hazy-free image. Note that we specially devise a cross-layer fusion (CLF) module to replace the traditional skip connection in the feature decoding stage, which can reinforce the attention to the spatial regions and feature channels with more serious attenuation. The experiments on the SateHaze1k, RS-Haze, and RSID datasets show that our SCSNet outperforming contemporaneous methods.

In summary, the main contributions of this study are threefolds:

 doi:10.20944/preprints202401.0679.v1



- To better learn semantic structure information in satellite images with non-homogeneous haze, we propose a hybrid CNN-transformer architecture, in which a hierarchical semantic guidance (HSG) module is introduced to synergetically aggregate local structure features and non-local semantic information.
- To fully consider the inconsistent attenuation, we present a cross-layer fusion (CLF) module, which is significantly better than traditional skip connection for integrating cross-layer features and reinforcing the attention to the spatial regions and feature channels with more serious attenuation.
- We establish a hierarchical semantic-guided contextual structure-aware network (SCSNet), which can effectively restore hazy-free images from non-homogeneous hazy satellite ones. Our SCSNet achieves nontrivial performance on three challenging SID datasets.

The remainder of this study is arranged as follows: Section 2 reviews existing dehazing methods. Section 3 elaborates the proposed SCSNet. The experimental results are shown and analyzed in Section 4. Finally, Section 5 gives the conclusion.

## 2. Related Work

Current image dehazing techniques can be broadly classified into two main categories [23]: prior-based methods and deep learning-based methods. This section reviews the advances of these methods during the past decades. Then we also investigate recent dehazing methods tailored for satellite images.

### *2.1. Prior-Based Methods*

Prior-based methods rely on pre-defined priors like dark channel priors (DCP) [10], color attenuation priors [10], and haze-line priors [24,25]. Each approach tackles specific scenarios but faces limitations due to inherent imprecision in prior information. For example, He *et al.* [10] pioneered an efficient dark channel prior method that exploits the low pixel values in non-sky regions of the RGB channels. This enables reliable transmission estimation while potentially introducing color distortion in the sky. Subsequently, Zhu *et al.* [11] introduced a mathematical equation considering brightness and saturation disparities to improve transmission map accuracy. Berman *et al.* [24,25] utilized a non-local dehazing method that analyzes haze lines in degraded images, leveraging the tight clusters of haze-free image colors in the RGB space. Peng *et al.* [26] leveraged regression analysis to model depth-dependent color shifts, then leveraged light differential to estimate medium transmission. Xu *et al.* [27] introduced the concept of "virtual depth" in their iterative dehazing method, quantifying Earth surface coverings to improve transmission estimation and haze removal. Recently, He *et al.* [4] proposed a heterogeneous atmospheric light prior and a side window filter, further enhancing remote sensing image dehazing.

While prior-based methods can enhance image quality, they often struggle with accurate haze parameter estimation in complex environments. In contrast, our SCSNet is data-driven, which overcomes this limitation.

### *2.2. Deep Learning-Based Methods*

Deep learning-based methods use deep neural networks to learn a mapping from hazy images to clear images. Early deep learning-based approaches leverage CNNs to estimate atmospheric light and transmission maps in the atmospheric scattering model [14,15,28]. Cai *et al.* [14] pioneered this approach by designing a CNN to map hazy images to their transmission maps. However, relying on pre-defined priors often leads to inaccuracies in estimating the atmospheric light and transmission map, resulting in dehazed images marred by artifacts, color distortion, and loss of detail. Similarly, Gu *et al.* [28] leveraged both prior knowledge and CNN-based feature extraction for more robust dehazing. However, estimating atmospheric light and transmission maps can be challenging due to the complex

 doi:10.20944/preprints202401.0679.v1



and diverse nature of real-world atmospheric conditions. In addition, inaccurate estimations may lead to artifacts like color distortion or loss in the dehazed image.

To combat image distortions caused by estimation errors of intermediate estimations, end-to-end dehazing methods has emerged as the dominant approach, which directly learns the mapping from hazy to clear images [15,29]. This paradigm increases flexibility and adaptability to complex haze conditions. Specifically, Li *et al.* [15] proposed AOD-Net by reconstructing the atmospheric scattering model directly and achieved superior image clarity. Qin *et al.* [16] presented a feature fusion attention network (FFA-Net), which can restore image details and color fidelity by retaining shallow information. Dong *et al.* [30] incorporated dense skip connections based on the U-Net architecture for better information flow. Recently, Transformer models have made breakthroughs in computer vision [12, 20,31], and many modified Transformer architectures have been proposed for low-level vision tasks. For example, Song *et al.* [8] made some improvements on normalization layer and activation function based on Swin Transformer [32] to adapt image dehazing. Qiu *et al.* [31] employed the Taylor-based approximation to approximate softmax-attention in Tranformers for images dehazing, achieving linear complexity for efficient computation. While existing end-to-end dehazing methods show promise in dehazing, they are not very effective for satellite images with non-homogeneous haze due to the data characteristics such as complex terrain information and severe object occlusion.

In recent years, some dehazing methods tailored for satellite images [18–20] are explored. Jiang *et al.* [33] introduced an empirical haze removal method for visible remote sensing images by applying an additive haze model. Bie *et al.* [18] proposed a Gaussian and physics-guided dehazing network (GPD-Net) to better extract hazy features and guide the model to real-world conditions. Beyond single-stage networks, Li and Chen [19] presented a two-stage dehazing network (FCTF-Net) for haze removal tasks on satellite images by performing coarse dehazing and then refining the results for enhanced performance. Huang and Chen [13] introduced DCRD-Net, a cascaded residual dense network specifically designed for SID. Jiang *et al.* [34] tackle non-uniform haze degradation by combining discrete wavelet transform with a deep learning network, while KFA-Net [6] extends the temporal and spatial scope of dehazing applications. Additionally, Chen *et al.* [35] leveraged multi-scale features to handle varying haze levels, while Li *et al.* [36] employed a multimodel joint estimation and self-correcting framework for improved accuracy and dehazing outcomes. Subsequently, Huang *et al.* [37] developed an adaptive region-based diffusion model for outstanding dehazing performance on both synthetic and real-world images. While these methods are tailored for satellite images, they often neglect crucial factors like semantic structure information and the inherent inconsistencies in attenuation. This oversight leads to subpar dehazing performance.

Compared with existing methods, our SCSNet provides the following unique characteristics: 1) the HSG module bridges the gap between local details and global context, enabling SCSNet to capture the inherent structure of hazy satellite images; 2) replacing traditional skip connections, CLF directly fuses features from different network layers, improving attention towards regions and channels with stronger haze effects.

## 3. The Proposed Method

As illustrated in Figure 2, the SCSNet is composed of four main parts: CNN encoder branch, Transformer encoder branch, hierarchical semantic guidance (HSG) module, and image restoration branch. In this section, we first elaborate on the encoder branches in Subsection 3.1, including both CNN and Transformer encoder branches. Then, the HSG module is described in detail in Subsection 3.2. In addition, the image restoration branch with coarse-to-fine fusion (CFF) module is introduced in Subsection 3.3. Finally, we will describe the loss function in detail.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2024 doi:10.20944/preprints202401.0679.v1



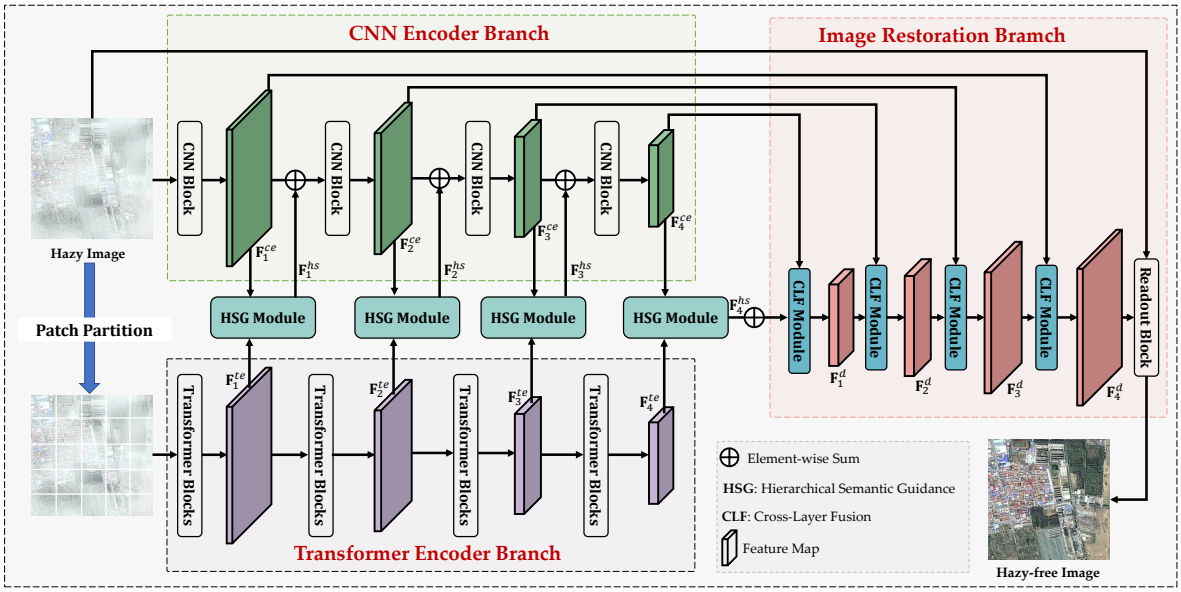


**Figure 2.** The overall framework of the proposed SCSNet. First, the CNN and Transformer encoder branches are used to extract the hierarchical local structure and global non-local semantic features, respectively. Next, the HSG module is devised to learn semantic structure information by synergetically complementing multi-scale local and non-local information. Finally, the semantic structure information is combined with the multi-layer local features for high-fidelity image restoration and obtaining hazy-free image.

### *3.1. Image Encoders*

Motivated by the advantages of CNNs in extracting local structure features and the advantages of Transformers in extracting non-local semantic features, we designed a hybrid CNN-Transformer image encoder. In the following, we will provide a detailed description of the two encoders.

#### 3.1.1. CNN Encoder Branch

In satellite scenes, the distribution of haze across different pixels in an image is non-homogeneous. However, most existing image dehazing methods primarily target homogeneous haze distribution in natural images, resulting in subpar performance when dealing with the non-homogeneous distributed haze in satellite images. Inspired by [38], we design a nested u-structure CNN encoder branch that effectively models the complex distribution of haze in satellite images by introducing residual and attention mechanisms to learn hierarchical local structure features. The nested u-structure CNN encoder branch consists of 4 cascaded u-structure CNN blocks, which are motivated by the strength of U-shape architecture in capturing multi-scale contextual information.

As shown in Figure 3, each u-structure CNN block includes three fundamental stages: 1) The input convolution layer applies basic convolutions to the input feature map $\mathbf{X}_i^{\text{in}} \in \mathbb{R}^{H\times W\times C_{in}}$ for generating a locally intermediate feature $\mathbf{X}_{\text{local}}^{\text{in}} \in \mathbb{R}^{H\times W\times C_{out}}$. 2) A symmetric U-Net-like structure with $D$ layers is designed as the intra-block encoding-decoding stage for progressively extracting and encoding multi-scale contextual feature $\mathcal{U}(\mathbf{X}_i^{\text{local}}) \in \mathbb{R}^{H\times W\times C_{out}}$ from the intermediate features, where $\mathcal{U}$ represents the symmetric U-Net-like structure. Increasing $D$ expands receptive fields, leading to richer representations encompassing both local details and global context. Notably, configuring the parameter $D$ enables effective multi-scale feature extraction from diverse input resolutions through step-wise upsampling, feature fusion, and subsequent convolutions. This mitigates the loss of details often encountered in direct large-scale upsampling. It is worth noting that channel attention and spatial attention are introduced into the fusion process of intra-block encoding and decoding features. The combination of channel attention and spatial attention contributes to guiding the model attending the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2024 doi:10.20944/preprints202401.0679.v1



dense haze area, and reducing the attention to the thin haze area, thereby improving the discriminative performance of feature representation. 3) The local feature $\mathbf{X}^{\text{in}}_{\text{local}}$ and the multi-scale feature $\mathcal{U}(\mathbf{X}^{\text{local}}_i)$ extracted by the U-Net-like structure are fused via summation to obtain the rich local structure feature:

$$\mathbf{F}^{ce}_i = \mathbf{X}^{\text{local}}_i + \mathcal{U}(\mathbf{X}^{\text{local}}_i). \tag{1}$$

This summation serves to preserve detailed information that might be lost through direct upsampling.

With the u-structure CNN block, the local structure feature obtained by the CNN encoder branch not only contains local structure information, but also exhibits spatial adaptability. For efficiency, the CNN encoder branch is achieved by four cascading u-structure CNN blocks, obtaining four hierarchical local structure features $\mathbf{F}^{ce}_i, i \in \{1, \cdots, 4\}$.

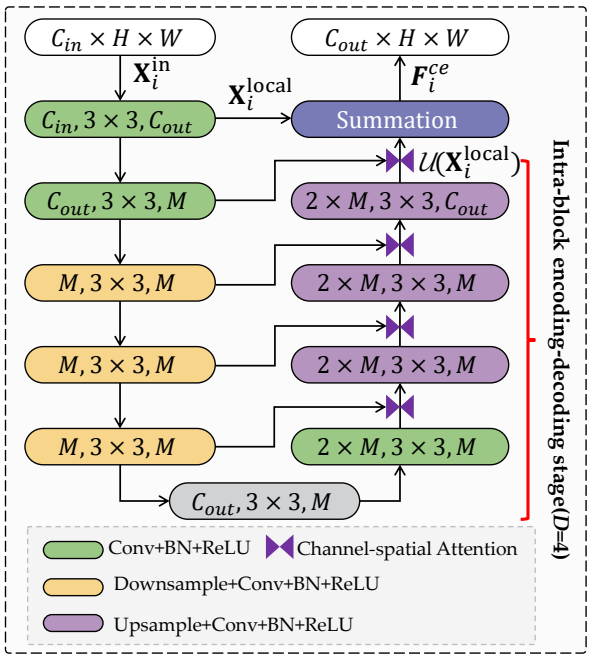


**Figure 3.** The designed u-structure CNN block.

### 3.1.2. Transformer Encoder Branch

Given the advantages of transformer models in capturing global semantic features and the excellent performance of Dehazeformer [8] in image dehazing, we adopt an improved dehazeformer as the basic module of the transformer encoder branch. As shown in Figure 2, the transformer encoder branch consists of four groups of cascaded transformer blocks. Unlike the original DehazeFormer, the devised transformer encoder branch only adopts the first 3 groups of transformer blocks from the original Dehazeformer, and the fourth group is self-designed. This design is intended to keep the output features of the same spatial size as the CNN encoder branch in 3.1.1.

The architecture of each transformer block is shown in Figure 4. To empower image dehazing, the transformer block equipped with a powerful combination of techniques: re-scale layer normalization, window-shifted multi-head self-attention with parallel convolution (W-MHSA-PC) as a pivotal component, affine transformation, and multi-layer perceptron (MLP). The details can be found in [8]. For efficiency, each group of transformer blocks consists of 8 cascading transformer blocks. With the transformer branch based on 4 groups of transformer blocks, the non-local semantic features $\mathbf{F}^{te}_i, i \in \{1, \cdots, 4\}$ can be captured to emphasize spatially-varying haze distribution features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2024 doi:10.20944/preprints202401.0679.v1



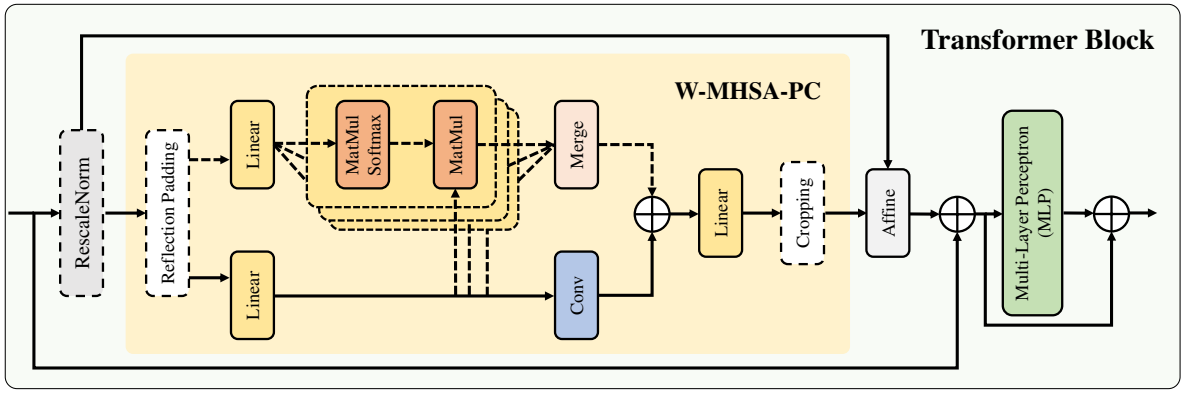


**Figure 4.** The workflow of the transformer block.

### *3.2. Hierarchical Semantic Guidance Module*

In order to guide the CNN encoder branch capture hierarchical local structure features with non-local semantic information, a hierarchical semantic guidance (HSG) module is proposed, which can dynamically combine shared information and faithfully preserve their unique complementary properties. As shown in Figure 5, the HSG module mainly consists of two main steps: coarse-guidance and fine-guidance. The coarse-guidance step generates a global feature descriptor by combining information from local structure features and non-local semantic features. The fine-guidance step uses these descriptors to re-calibrate and aggregate local structure features and non-local semantic features. The implementation is as follows.

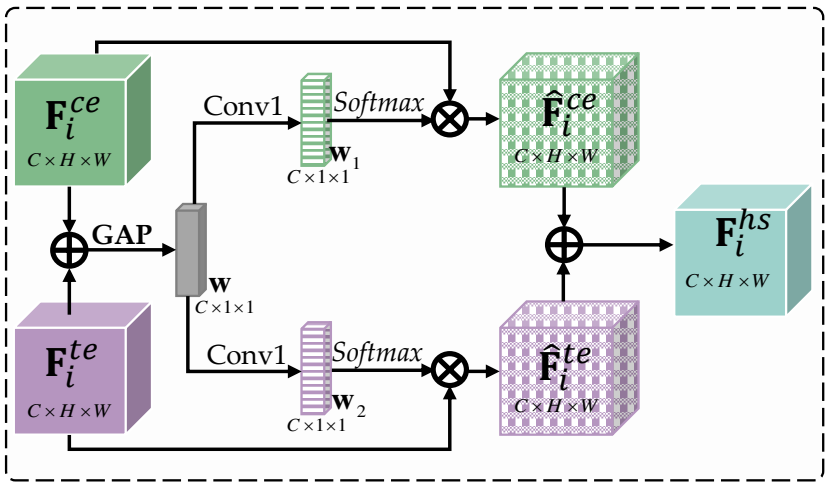


**Figure 5.** The proposed hierarchical semantic guidance (HSG) module.

1) Coarse-guidance: First, the local structure features $\mathbf{F}_i^{ce}$ and non-local semantic features $\mathbf{F}_i^{d}$ are combined using concatenation operation and compressed via channel reduction by:

$$\mathbf{F}_i^{coarse} = \hbar_{1\times 1}(\text{concat}(\mathbf{F}_i^{ce}, \mathbf{F}_i^{te})), \tag{2}$$

where $\mathbf{F}_i^{coarse}$ is the coarse-guidance feature, $\hbar_{1\times 1}$ denotes the $1 \times 1$ convolutional layer with batch normalization for channel reduction, and $\text{concat}(\cdot,\cdot)$ represents the concatenation operation. Then, we use global average pooling on the spatial dimension of the coarse-guidance $\mathbf{F}_i^{coarse}$ to compute the channel-wise descriptor. Finally, we use two parallel 1x1 convolution layers to process the channel-wise descriptor, obtaining $\mathbf{w}_1$ and $\mathbf{w}_2$, which enhance the interaction between local structure features and non-local semantic features.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2024 doi:10.20944/preprints202401.0679.v1



2) Fine-guidance: We apply the softmax function to $\mathbf{w}_1$ and $\mathbf{w}_2$ generating attention activations, which are leveraged to adaptively re-calibrate the local structure features $\mathbf{F}_i^{ce}$ and the non-local semantic features $\mathbf{F}_i^{te}$. Finally, the hierarchical semantic guidance feature obtained after fine-grained is:

$$\mathbf{F}_i^{hs} = \text{softmax}(\mathbf{w}_1) \cdot \mathbf{F}_i^{ce} + \text{softmax}(\mathbf{w}_2) \cdot \mathbf{F}_i^{te}, \tag{3}$$

Through the coarse-to-fine guidance strategy, we can efficiently integrate local structure features and non-local semantic features, and achieve efficient interaction between the two ones, which is beneficial for better preserving structural details and color information of reconstructed hazy-free images.

### *3.3. Image Restoration*

In satellite scenes, terrain information is quite complex. In order to preserve more detailed information and fully leverage the features from non-adjacent levels in the reconstructed hazy-free images, we propose a cross-layer fusion (CLF) module. Different form the simply concatenation, the CLF module introduces the dense contextual feature learning and context residual learning in the fusion stage, which is more effective. The workflow of the proposed CLF module is shown in Figure 6.

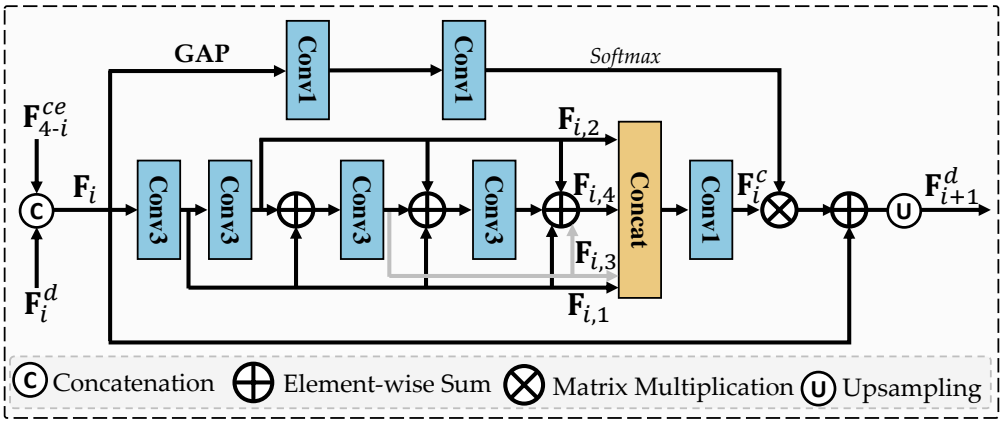


**Figure 6.** The proposed cross-layer fusion (CLF) module.

Firstly, to explore the potential of features from non-adjacent layers, the combined feature $\mathbf{F}_i$ is captured by concatenating the decoder feature $\mathbf{F}_i^d$ and local structure features $\mathbf{F}_i^{ce}$. Note that $\mathbf{F}_i^d = \mathbf{F}_i^{hs} + \mathbf{F}_i^{ce}$, when $i = 0$.

Secondly, to obtain better dense context features, the dense connection operation is applied to integrate the feature maps from the previous layers, which increases the variation of the input to subsequent layers and thus enhances the representation ability. Because the concatenation operation in the dense block increases the channels of the obtained features, we adopt an element-wise summation operation instead of the concatenation operation to reduce its channel number. This strategy has fewer parameters and faster computation speed. In order to capture more information, we combine all the output feature maps. Then, the $1 \times 1$ convolution is utilized to achieve dense context feature fusion as follows:

$$\mathbf{F}_i^c = \text{conv1}(\text{concat}(\mathbf{F}_{i,1}, \mathbf{F}_{i,2}, \mathbf{F}_{i,3}, \mathbf{F}_{i,4})), \tag{4}$$

where $\mathbf{F}_i^c$ represents the dense context fusion feature, and $\mathbf{F}_{i,3}(\cdot)$ denotes the the $1 \times 1$ convolution operation.

Thirdly, to preserve previous information, we also conduct the context residual learning. Based on the input feature $\mathbf{F}_i$, the channel weight descriptor is computed with the residual learning mechanism guided by channel attention. Then, the leaned channel weight descriptor is applied to re-weight the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2024 doi:10.20944/preprints202401.0679.v1



dense context fusion feature $\mathbf{F}_i^c$, and then the next decoder feature $\mathbf{F}_i^e$ is obtained by context residual learning. The context residual learning can be represented as:

$$\mathbf{F}_i^d = \text{Upsample}(\mathbf{F}_i + \text{conv1}_{\times 2}(\text{GAP}(\mathbf{F}_i)) \cdot \mathbf{F}_i^c), \tag{5}$$

where $\text{GAP}(\cdot)$ denotes the global average pooling operation, Upsample stands for the operation of increasing the resolution of feature by a factor of 2, and $\text{conv1}_{\times 2}$ represents 2 stacked $1 \times 1$ convolutional layers. The context residual learning ensures that the model adapts to skip less important information such as thin fog or low frequency areas and focuses more on dense fog areas. After four cascaded CLF module, the final decoder feature $\mathbf{F}_4$ is obtained for restoring haze-free images. To this end, a readout network is designed for post processing and restore haze-free images. Specifically, the operation of the readout network can be expressed as:

$$\mathbf{I}^{\text{pred}} = (\tanh(\text{conv3}_{\times 2}(\mathbf{D}^e)) + 1)/2, \tag{6}$$

where $\text{conv3}_{\times 2}$ means 2 stacked $3 \times 3$ convolutional layers.

### *3.4. Loss Function*

We adopt two different loss functions to train the proposed network, including smooth L1 loss and perceptual loss. The smooth L1 loss can ensure that the reconstructed hazy-free image is close to the clean image and has been proven to be superior to L2 loss in many image restoration tasks, and more robust than L1 loss. Let $\mathbf{I}^{\text{pred}}$ indicate the reconstructed hazy-free image and $\mathbf{I}^{\text{gt}}$ represent the clean image. The smooth L1 loss can be written as:

$$\mathcal{L}_{l1} = \frac{1}{N} \sum_{x=1}^{N} \sum_{j=1}^{3} \text{smooth}_{L1}(\mathbf{I}_j^{\text{pred}}(x) - \mathbf{I}_j^{\text{gt}}(x)), \tag{7}$$

in which,

$$\text{smooth}_{L1}(y) = \begin{cases} 0.5y^2, & if\ |y| < 1 \\ |y| - 0.5, & otherwise \end{cases} \tag{8}$$

where $j$ indexes the $j$-th color channel for images and $N$ is the total number of pixels.

The perceptual loss can provide additional supervision in the high-level feature space, which is used to quantify the visual difference between the reconstructed hazy-free image and the clean image. The perceptual loss function is described as follows:

$$\mathcal{L}_{perc} = \sum_{j=1}^{3} \frac{1}{W_j H_j C_j} \| f_j(\mathbf{I}^{\text{gt}}) - f_j(\mathbf{I}^{\text{pred}}) \|_2^2, \tag{9}$$

where $f_j(\mathbf{I}^{\text{gt}})$ and $f_j(\mathbf{I}^{\text{pred}})$ mean the feature maps extracted by pre-trained VGG16 model associated with the clean image and the reconstructed hazy-free image, respectively.

Finally, the total loss function is defined as:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{l1} + \lambda \cdot \mathcal{L}_{perc} \tag{10}$$

where $\lambda$ is the coefficient that balances the two losses. The optimization of the entire model is achieved by minimizing the total loss.

 doi:10.20944/preprints202401.0679.v1



## 4. Experiment and Results

In this section, we first describe the experimental settings, including the dataset, training details, and evaluation metrics. Then, we compare the proposed SCSNet with other state-of-the-art dehazing methods quantitatively and qualitatively. Finally, ablation studies is conducted to further demonstrate the superiority of the proposed SCSNet and the effectiveness of different modules in the proposed SCSNet.

### *4.1. Datasets*

We evaluated the proposed SCSNet using the SateHaze1k [39], RS-Haze [8], and RSID [1] datasets.

*1) SateHaze1k:* The SateHaze1k dataset [39] consists of 1200 pairs of hazy images, corresponding clear images, and SAR images. The dataset has three levels of haze, covering thin, medium, and thick haze, each with 400 pairs, which is beneficial for evaluating the robustness of the proposed method. Following the previous work [39], we divided the training, validation, and testing data ratio into 8:1:1 for each level of haze. In addition, in order to better evaluate the dehazing effect in real situations, we mixed the data of the three different haze levels together.

*2) RS-Haze:* The RS-Haze dataset [8] is a challenging and representative large-scale image dehazing dataset consisting of 51,300 paired images, of which 51,030 are for training and 270 are for testing. The dataset covers a variety of scenes and haze intensities, including urban, forest, beach, and mountain scenes.

*3) RSID:* The RSID dataset [1] offers a collection of 1000 image pairs, each consisting of a hazy and haze-free counterpart. Following the previous work [40], we randomly selected 900 of these pairs. The remaining 100 pairs were reserved as a distinct test set to assess the model's ability to generalize to unknown data.

### *4.2. Implementation Details and Evaluation Metrics*

During the training process, in order to avoid overfitting, we augmented the training set with random rotation at 90, 180, and 270 degrees, horizontal flips, and vertical flips. The Adam optimizer is utilized to optimize the model. The initial learning rate is set to 0.0001, and we use the cosine annealing strategy to adjust the learning rate until convergence. The batch size is set to 64, and the training epoch is set to 30. Considering the high cost of adjusting the weighting coefficients, we adopted a multi-task learning strategy with equal variances and uncertainties [41] to learn the optimal weighting coefficients. The proposed SCSNet is implemented using the PyTorch library on an RTX 3090 GPU with 24GB of memory.

We quantitatively evaluated the dehazing performance using commonly used metrics, including peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM).

### *4.3. Comparison with State-of-the-Arts*

In this subsection, we quantitatively and qualitatively compare the proposed SCSNet with state-of-the-art methods on the SateHaze1k, RS-Haze, and RSID datasets. These methods include: 1) methods for natural image dehazing, including DCP [10], AOD-Net [15], Light-DehazeNet [42], FFA-Net [16], Restormer [12], and DehazeFormer [8]; 2) tailored methods for satellite image dehazing, such as DCRD-Net [13] and FCTF-Net [19]. All compared methods are trained using the public code provided by the corresponding author.

**Quantitative analysis:** Table 1 shows the quantitative comparison results of different methods on the SateHaze1k dataset. According to the results in the table, the DCP method performs poorly in dehazing, with both low PSNR and SSIM. This may be because DCP is a prior-based method that is difficult to learn discriminative semantic features from the image. The dehazing performance of AOD-Net and Light-DehazeNet is also mediocre, which may be due to the fact that these two methods are among the early and lightweight neural networks used for dehazing and are limited by

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2024 doi:10.20944/preprints202401.0679.v1



the depth of the network. Other comparison methods consciously designed their network structures, so they achieved impressive results with higher PSNR and SSIM. It is worth noting that DCRD-Net and FCTF-Net are models specifically tailored for satellite image dehazing and also show good performance. However, compared to them, our proposed SCSNet achieves best results in all three levels of haze. It not only outperforms tailored methods for satellite image dehazing, but also achieves higher metrics than natural image dehazing methods. This indicates that the proposed SCSNet has high efficiency and generalization in removing haze with different degrees. In addition, our proposed SCSNet also shows optimal performance in mixed haze data, indicating better practical application. Table 2 shows the quantitative comparison results of different methods on RS-Haze and RSID datasets. Based on the result, we can find our proposed SCSNet delivers substantial PSNR improvements of 2.5724 and 2.6962 on RS-Haze and RSID datasets, respectively. This highlighting the exceptional generalization capabilities of our proposed SCSNet.

**Table 1.** The Qualitative Comparison on the SateHaze1k Dataset.

| Methods | Thin fog | | Moderate fog | | Thick fog | | Mixed fog | |
|---|---|---|---|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM | PSNR | SSIM |
| DCP [10] | 19.1183 | 0.8518 | 19.8384 | 0.8812 | 16.793 | 0.7701 | 18.5833 | 0.8344 |
| AOD-Net [15] | 19.0548 | 0.7777 | 19.4211 | 0.7015 | 16.4672 | 0.7123 | 17.4859 | 0.6332 |
| Light-DehazeNet [42] | 18.4868 | 0.8658 | 18.3918 | 0.8825 | 16.7662 | 0.7697 | 17.8132 | 0.8352 |
| FFA-Net [16] | 20.141 | 0.8582 | 22.5586 | 0.9132 | 19.1255 | 0.7976 | 21.2873 | 0.8663 |
| Restormer [12] | 20.9829 | 0.8686 | 23.1574 | 0.9036 | 19.6984 | 0.7739 | 20.7892 | 0.8379 |
| DehazeFormer [8] | 21.9274 | 0.8843 | 24.4407 | 0.9268 | 20.2133 | 0.8049 | 22.0066 | 0.8659 |
| DCRD-Net [13] | 20.8473 | 0.8767 | 23.3119 | 0.9225 | 19.725 | 0.8121 | 21.7468 | 0.8812 |
| FCTF-Net [19] | 18.3262 | 0.8369 | 20.9057 | 0.8553 | 17.2551 | 0.6922 | 19.5883 | 0.8434 |
| **SCSNet** | **26.1460** | **0.9415** | **28.3501** | **0.9566** | **24.6542** | **0.9015** | **25.1759** | **0.9223** |

**Table 2.** The Qualitative Comparison on the RS-Haze and RSID Datasets.

| Methods | RS-Haze | | RSID | |
|---|---|---|---|---|
| | PSNR | SSIM | PSNR | SSIM |
| DCP [10] | 18.1003 | 0.6704 | 17.3256 | 0.7927 |
| AOD-Net [15] | 23.9677 | 0.7207 | 18.7037 | 0.7424 |
| Light-DehazeNet [42] | 25.5965 | 0.8209 | 17.9279 | 0.8414 |
| FFA-Net [16] | 29.1932 | 0.8846 | 21.2876 | 0.9042 |
| Restormer [12] | 25.6700 | 0.7563 | 11.7240 | 0.5971 |
| DehazeFormer [8] | 29.3419 | 0.8730 | 22.6859 | 0.9118 |
| DCRD-Net [13] | 29.6780 | 0.8878 | 22.1643 | 0.8926 |
| FCTF-Net [19] | 29.6240 | 0.8958 | 20.2556 | 0.8397 |
| **SCSNet** | **32.2504** | **0.9271** | **25.3821** | **0.9585** |

**Qualitative analysis:** Figure 7 shows qualitative comparison results of different methods on the SateHaze1k dataset. The red boxes in the figure highlight the regions where the dehazing results of various methods differ significantly. Distinctly, the proposed SCSNet generate better results with more structural details and color information, which indicates our method is superior to all comparison methods and produces significant visibility improvement. In addition, our proposed SCSNet performs well on images with thin haze, moderate haze, and thick haze, further demonstrating its better generalization in removing haze with different degrees.

Figure 8 shows qualitative comparison results of different methods on the RS-Haze and RSID datasets. As illustrated, our proposed SCSNet achieves relatively satisfactory dehazing results, which is more closely resemble the groundtruth. It is noteworthy that our proposed SCSNet significantly outperforms other comparative methods in terms of heavy haze removal. As shown in the red-boxed areas of the second row images in Figure 7, other comparison methods, except ours, fail to effectively recover the real objects covered by fog. While DCRD-Net [13] and FCTF-Net [19] produces reasonably

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 January 2024 doi:10.20944/preprints202401.0679.v1



effective dehazing, it falls short of fully eliminating non-homogeneous haze, as shown in the second row images. Other comparison methods either struggle to remove dense and non-homogeneous haze or introduce unwanted color shifts and structural details loss. This further demonstrates the superiority of our method in detail and color restoration, as we better consider semantic structure information and inconsistent attenuation.

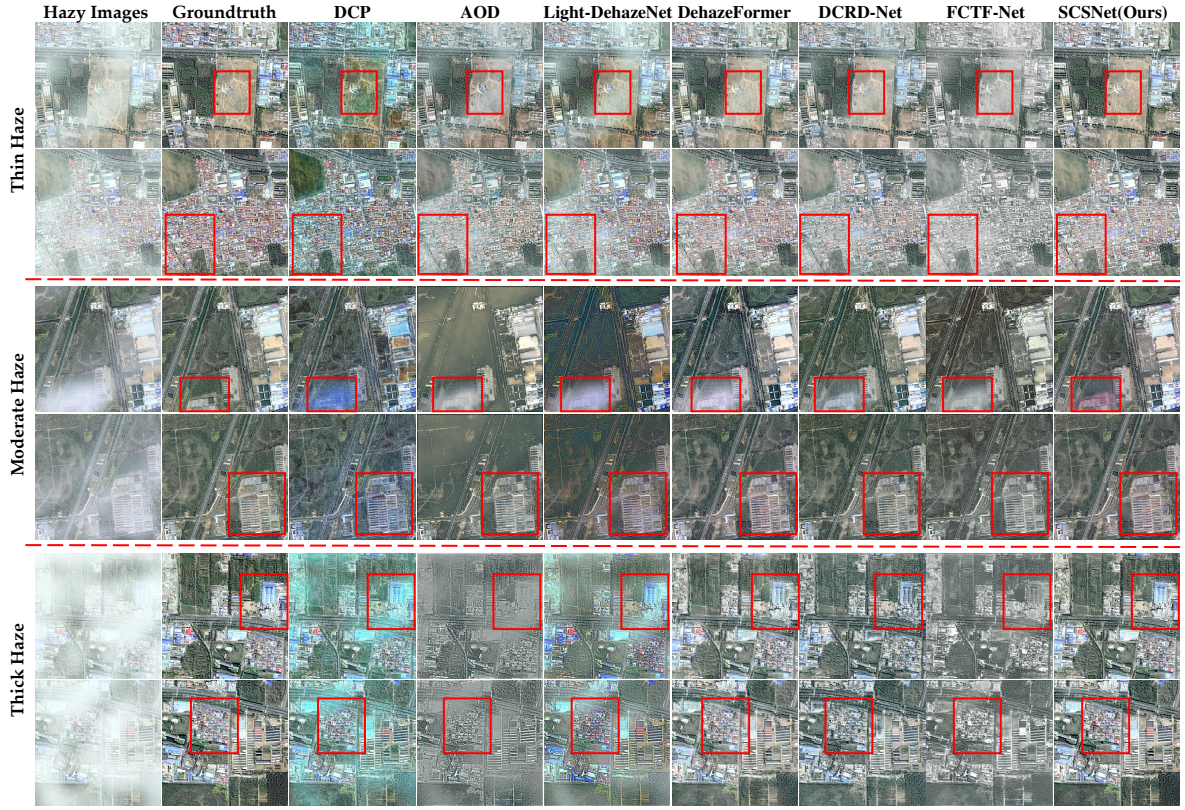


**Figure 7.** Qualitative results of different methods on the SateHaze1k dataset.

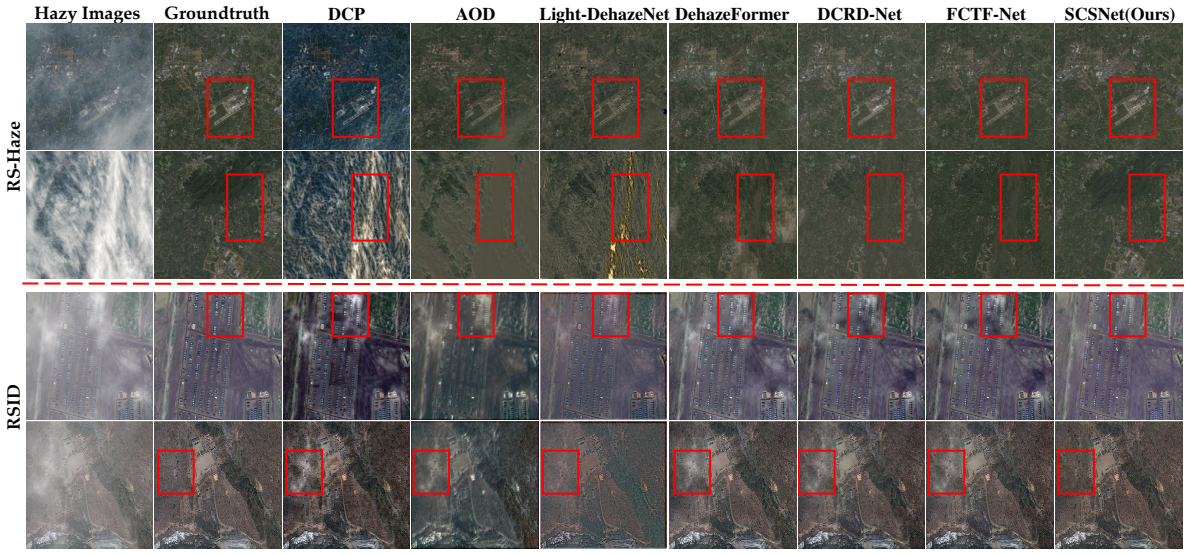


**Figure 8.** Qualitative results of different methods on RS-Haze and RSID datasets.

### *4.4. Ablation Analysis*

In order to further demonstrate the effectiveness of the proposed SCSNet, we conducted ablation studies by considering the different modules. We primarily focused on the following factors: 1) the

 doi:10.20944/preprints202401.0679.v1



designed u-structure CNN block, 2) the hierarchical semantic guidance (HSG) module, and 3) the cross-layer fusion (CLF) module. The corresponding ablation models include: 1) w/ plain ConvE: using plain convolutional encoding module to replace the u-structure CNN block; 2) w/o HSG: using the summation operation to replace the HSG module for semantic guidance; 3) w/ concat: using the concatenation operation to replace the CLF module for feature fusion.

We conducted ablation experiments on the mixed data of the SateHaze1k dataset, and the results are shown in Table 3. We can see that all modules can improve the dehazing performance and the best performance achieves 25.1759 dB by using the full model, which demonstrates the effectiveness of each proposed module. Specifically, when using plain convolutional encoding module to replace the u-structure CNN block, the model's PSNR and SSIM decrease by 3.1543 and 0.0269, respectively. This indicates the superiority of the designed u-structure CNN block in extracting local structural and spatially adaptive features. When using the summation operation to replace the HSG module, the significant drop in model performance indicates the effectiveness of the designed HSG module in integrating local structural features and non-local semantic features. When using the concatenation operation to replace the CLF module, the decrease in model performance indicates the effectiveness of the designed CLF module in preserving more detailed information and fully leveraging the features from non-adjacent levels.

**Table 3.** The Experimental Results for Ablation Study.

| Ablated Modules | Baselines | PSNR | SSIM |
|---|---|---|---|
| Feature Encoding | w/ plain ConvE | 22.0216 | 0.8954 |
| HSG | w summation | 21.6719 | 0.8927 |
| CLF | w/ add | 23.5234 | 0.9026 |
| **Full model (SCSNet)** | | **25.1759** | **0.9223** |

## 5. Conclusions

In this work, we propose a hierarchical semantic-guided contextual structure-aware network (SCSNet) for effective satellite image dehazing. The key insight of this work is designing hybrid CNN-transformer architecture with hierarchical semantic guidance (HSG) module for synergetically complementing local representation from non-local features, and devising cross-layer fusion (CLF) module for reinforcing the attention to the spatial regions and feature channels with more serious attenuation. The experimental results on the SateHaze1k, RS-Haze, and RSID datasets demonstrate that the designed hybrid CNN-transformer architecture, HSG module, and CLF module, are effective. In addition, the proposed SCSNet achieves better performance than other state-of-the-art methods.

**Author Contributions:** Data curation, L.Y.; Conceptualization, L.Y. and H.W.; Methodology, L.Y., J.C. and S.D.; Supervision, J.C. and H.N.; Writing—original draft, L.Y.; writing—review and editing, L.Y., J.C. and H.N.; formal analysis, H.W. and S.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported in part by National Natural Science Foundation of China under Grant 62201452, in part by Scientific Research Program Funded by Shaanxi Provincial Education Department under Grant 22JK0568, and in part by Natural Science Basic Research Program of Shaanxi Province, China under Grant 2022JQ-592.

**Data Availability Statement:** No new data were created or analyzed in this study. Data sharing is not applicable to this article.

**Acknowledgments:** We would also like to express our gratitude to the anonymous reviewers and the editors for their valuable advice and assistance.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. Chi, K.; Yuan, Y.; Wang, Q. Trinity-Net: Gradient-Guided Swin Transformer-based Remote Sensing Image Dehazing and Beyond. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*.
2. Sun, H.; Luo, Z.; Ren, D.; Hu, W.; Du, B.; Yang, W.; Wan, J.; Zhang, L. Partial Siamese with Multiscale Bi-codec Networks for Remote Sensing Image Haze Removal. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*.
3. Zheng, X.; Sun, H.; Lu, X.; Xie, W. Rotation-invariant attention network for hyperspectral image classification. *IEEE Transactions on Image Processing* **2022**, *31*, 4251–4265.
4. He, Y.; Li, C.; Bai, T. Remote Sensing Image Haze Removal Based on Superpixel. *Remote Sensing* **2023**, *15*, 4680.
5. Zheng, X.; Chen, X.; Lu, X.; Sun, B. Unsupervised change detection by cross-resolution difference learning. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–16.
6. Jiang, B.; Wang, J.; Wu, Y.; Wang, S.; Zhang, J.; Chen, X.; Li, Y.; Li, X.; Wang, L. A Dehazing Method for Remote Sensing Image Under Nonuniform Hazy Weather Based on Deep Learning Network. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–17.
7. Dong, W.; Wang, C.; Sun, H.; Teng, Y.; Liu, H.; Zhang, Y.; Zhang, K.; Li, X.; Xu, X. End-to-End Detail-Enhanced Dehazing Network for Remote Sensing Images. *Remote Sensing* **2024**, *16*, 225.
8. Song, Y.; He, Z.; Qian, H.; Du, X. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing* **2023**, *32*, 1927–1941.
9. Peng, Y.T.; Lu, Z.; Cheng, F.C.; Zheng, Y.; Huang, S.C. Image haze removal using airlight white correction, local light filter, and aerial perspective prior. *IEEE Transactions on Circuits and Systems for Video Technology* **2020**, *30*, 1385–1395.
10. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2010**, *33*, 2341–2353.
11. Zhu, Q.; Mai, J.; Shao, L. A fast single image haze removal algorithm using color attenuation prior. *IEEE transactions on image processing* **2015**, *24*, 3522–3533.
12. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H. Restormer: Efficient transformer for high-resolution image restoration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5728–5739.
13. Huang, Y.; Chen, X. Single Remote Sensing Image Dehazing Using a Dual-Step Cascaded Residual Dense Network. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, pp. 3852–3856.
14. Cai, B.; Xu, X.; Jia, K.; Qing, C.; Tao, D. Dehazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing* **2016**, *25*, 5187–5198.
15. Li, B.; Peng, X.; Wang, Z.; Xu, J.; Feng, D. Aod-net: All-in-one dehazing network. In Proceedings of the Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4770–4778.
16. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature fusion attention network for single image dehazing. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2020, Vol. 34, pp. 11908–11915.
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations (ICLR), 2021.
18. Bie, Y.; Yang, S.; Huang, Y. Single Remote Sensing Image Dehazing using Gaussian and Physics-Guided Process. *IEEE Geoscience and Remote Sensing Letters* **2022**, *19*, 1–5.
19. Li, Y.; Chen, X. A coarse-to-fine two-stage attentive network for haze removal of remote sensing images. *IEEE Geoscience and Remote Sensing Letters* **2020**, *18*, 1751–1755.
20. Kulkarni, A.; Murala, S. Aerial Image Dehazing With Attentive Deformable Transformers. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 6305–6314.
21. Ning, J.; Zhou, Y.; Liao, X.; Duo, B. Single Remote Sensing Image Dehazing Using Robust Light-Dark Prior. *Remote Sensing* **2023**, *15*, 938.

22. Wei, J.; Wu, Y.; Chen, L.; Yang, K.; Lian, R. Zero-shot remote sensing image dehazing based on a re-degradation haze imaging model. *Remote Sensing* **2022**, *14*, 5737.
23. Frants, V.; Agaian, S.; Panetta, K. QCNN-H: Single-image dehazing using quaternion neural networks. *IEEE Transactions on Cybernetics* **2023**, *53*, 5448–5458.
24. Berman, D.; Avidan, S.; et al. Non-local image dehazing. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1674–1682.
25. Berman, D.; Treibitz, T.; Avidan, S. Air-light estimation using haze-lines. In Proceedings of the 2017 IEEE International Conference on Computational Photography (ICCP). IEEE, 2017, pp. 1–9.
26. Peng, Y.T.; Cao, K.; Cosman, P.C. Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing* **2018**, *27*, 2856–2868.
27. Xu, L.; Zhao, D.; Yan, Y.; Kwong, S.; Chen, J.; Duan, L.Y. IDeRs: Iterative dehazing method for single remote sensing image. *Information Sciences* **2019**, *489*, 50–62.
28. Gu, Z.; Zhan, Z.; Yuan, Q.; Yan, L. Single remote sensing image dehazing using a prior-based dense attentive network. *Remote Sensing* **2019**, *11*, 3008.
29. Sun, H.; Li, B.; Dan, Z.; Hu, W.; Du, B.; Yang, W.; Wan, J. Multi-level Feature Interaction and Efficient Non-Local Information Enhanced Channel Attention for image dehazing. *Neural Networks* **2023**, *163*, 10–27.
30. Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; Yang, M.H. Multi-scale boosted dehazing network with dense feature fusion. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2157–2167.
31. Qiu, Y.; Zhang, K.; Wang, C.; Luo, W.; Li, H.; Jin, Z. MB-TaylorFormer: Multi-branch efficient transformer expanded by Taylor formula for image dehazing. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12802–12813.
32. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
33. Jiang, H.; Lu, N.; Yao, L.; Zhang, X. Single image dehazing for visible remote sensing based on tagged haze thickness maps. *Remote Sensing Letters* **2018**, *9*, 627–635.
34. Jiang, B.; Chen, G.; Wang, J.; Ma, H.; Wang, L.; Wang, Y.; Chen, X. Deep dehazing network for remote sensing image with non-uniform haze. *Remote Sensing* **2021**, *13*, 4443.
35. Chen, X.; Li, Y.; Dai, L.; Kong, C. Hybrid high-resolution learning for single remote sensing satellite image Dehazing. *IEEE Geoscience and Remote Sensing Letters* **2022**, *19*, 1–5.
36. Li, S.; Zhou, Y.; Xiang, W. M2SCN: Multi-Model Self-Correcting Network for Satellite Remote Sensing Single-Image Dehazing. *IEEE Geoscience and Remote Sensing Letters* **2023**, *20*, 1–5.
37. Huang, Y.; Xiong, S. Remote sensing image dehazing using adaptive region-based diffusion models. *IEEE Geoscience and Remote Sensing Letters* **2023**.
38. Qin, X.; Zhang, Z.; Huang, C.; Dehghan, M.; Zaiane, O.R.; Jagersand, M. U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern recognition* **2020**, *106*, 107404.
39. Huang, B.; Zhi, L.; Yang, C.; Sun, F.; Song, Y. Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks. In Proceedings of the Proceedings of the IEEE/CVF winter Conference on Applications of Computer Vision, 2020, pp. 1806–1813.
40. Wen, Y.; Gao, T.; Zhang, J.; Li, Z.; Chen, T. Encoder-free Multi-axis Physics-aware Fusion Network for Remote Sensing Image Dehazing. *IEEE Transactions on Geoscience and Remote Sensing* **2023**, *61*, 1–15.
41. Kendall, A.; Gal, Y.; Cipolla, R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7482–7491.
42. Ullah, H.; Muhammad, K.; Irfan, M.; Anwar, S.; Sajjad, M.; Imran, A.S.; de Albuquerque, V.H.C. Light-DehazeNet: a novel lightweight CNN architecture for single image dehazing. *IEEE Transactions on Image Processing* **2021**, *30*, 8968–8982.