Proxy and Cross-Stripes Integration Transformer for Remote Sensing Image Dehazing

Xiaozhe Zhang, *Graduate Student Member, IEEE*, Fengying Xie, *Member, IEEE*, Haidong Ding, Shaocheng Yan, and Zhenwei Shi, *Senior Member, IEEE*

***Abstract*— Existing Transformer-based dehazing methods for remote sensing (RS) images, to avoid quadratic computation complexity with respect to the feature map size, either perform self-attention mechanisms within local windows or capture long-range dependencies in the channel dimension rather than spatial. Each of these methods has its drawbacks. To address these limitations, we propose the Proxy and Cross-Stripes Integration Transformer (PCSformer) for RS image dehazing. PCSformer introduces two innovative Transformer blocks, i.e., *sliding cross-stripes Transformer block* and *local proxy-based global Transformer block*. The former allows us to directly model long-range dependencies and capture rich contextual information for large-scale objects in RS images. The latter seeks valuable information for thick haze regions within the whole feature map, generating more consistent and realistic scene details for such regions. Both achieve a large receptive field with cost-effective computational complexity within a single Transformer block. Furthermore, we introduce a shallow deep model with a small receptive field to conduct local refinement, which can mitigate artifacts associated with a large receptive field. Finally, to facilitate the better application of dehazing models to downstream visual tasks, we contribute two large-scale datasets for RS image dehazing. Experiments indicate that the dehazing models trained on our datasets can better assist downstream visual tasks under hazy atmospheric conditions compared to the dehazing models trained on existing datasets. Quantitative and qualitative experiments demonstrate that the proposed PCSformer significantly outperforms existing state-of-the-art techniques on dehazing benchmarks, particularly excelling in the restoration of thick haze scenes. The code and datasets are available at https://github.com/SmileShaun/PCSformer.**

***Index Terms*— Deep learning, dehazing dataset, remote sensing (RS) image dehazing, vision Transformer (ViT).**

Received 15 January 2024; revised 3 April 2024, 16 May 2024, and 5 June 2024; accepted 5 September 2024. Date of publication 11 September 2024; date of current version 26 September 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62475006 and Grant 62125102, and in part by the National Key Research and Development Program of China under Grant 2022ZD0160401. *(Corresponding author: Fengying Xie.)*

Xiaozhe Zhang, Fengying Xie, and Haidong Ding are with the Image Processing Center, School of Astronautics, Beihang University, Beijing 100191, China, and also with the Tianmushan Laboratory, Hangzhou 311115, China (e-mail: xfy_73@buaa.edu.cn).

Shaocheng Yan is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430072, China.

Zhenwei Shi is with the Image Processing Center, School of Astronautics, State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China, and also with Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China.

Digital Object Identifier 10.1109/TGRS.2024.3457868

I. Introduction

REMOTE sensing (RS) images captured in hazy scenarios inevitably suffer from poor visibility and low contrast. These adverse effects severely impede the performance of high-level computer vision tasks (e.g., object detection [1] and semantic segmentation [2]). Consequently, RS image dehazing has garnered significant attention in recent years.

The haze imaging equation, utilized to characterize the degradation process in hazy images, is given by [3], [4], [5]

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

where x represents the coordinates of a pixel's position in the image, I represents the captured hazy image, J represents scene radiance image (i.e., haze-free image), A represents atmospheric light, and t represents medium transmission map. Due to spatially variant t and atmospheric light A, image dehazing is typically an underconstrained problem. To resolve this ambiguity, early image dehazing methods utilize preexisting knowledge and assumptions (i.e., priors) to impose additional constraints among the unknown variables. Examples include dark channel prior (DCP) [6], color-line prior [7], nonlocal prior [8], and elliptical boundary prior [9]. However, prior knowledge does not hold for certain images, leading to suboptimal dehazing performance.

Since AlexNet [10] achieved victory in the ILSVRC-2012, many image dehazing methods based on convolutional neural networks (CNNs) and vision Transformers (ViTs) have been proposed. Initially, CNN-based methods employed deep neural networks to estimate $t(x)$ and used prior assumptions to estimate A. Pioneering methods in this category, such as DehazeNet [11] and the one proposed by Ren et al. [12], achieved improved estimation for transmission maps compared to prior-based methods. Current methods favor end-to-end models to directly learn the haze-free image or the residual of hazy image [13], [14], [15], [16]. With the advent of ViT [17], numerous image dehazing methods based on the Transformer architecture have emerged [18], [19], [20], [21], [22], [23], [24], [25]. To circumvent the quadratic relationship between computation complexity of full self-attention mechanism and feature map size, existing methods either perform self-attention in a local window [18], [19], [20], [21], [22], [23] or capture global long-range dependencies in the channel dimensions rather than spatial [24], [25]. The former sacrifices one fundamental characteristic of ViT, i.e.,

the direct modeling of long-range dependencies. Moreover, its effectiveness may be decreased when a local window is entirely covered by a thick haze region, as valuable information may be lacking within the confined local window. The latter incurs high computational costs when dealing with a large number of channels, and performing self-attention in the channel dimension leads to a lack of model expressive power. In the context of thick and nonhomogeneous RS haze removal, there is an urgent need to design a specialized architecture to address the aforementioned issues.

Our key insight is that under constrained computational overhead, achieving global receptive fields in a single Transformer block is crucial rather than relying on a stack of multiple Transformer blocks. Following this insight not only captures rich contextual information for large-scale objects (such as bridges, football fields, and so on) commonly present in RS hazy images to enhance the model's expressive power but also prevents a decrease in accuracy caused by the absence of valid content when attention windows encompass numerous tokens projected from thick haze regions.

Based on the aforementioned insights, we propose the Proxy and Cross-Stripes Integration Transformer (PCSformer) for RS image dehazing. Drawing inspiration from Cswin Transformer [26], we introduce a sliding cross-stripes Transformer block. This block performs self-attention calculations in both vertical and horizontal stripes concurrently, where each stripe provides abundant contextual information for large-scale objects. In addition, we propose a shifted stripes scheme and efficient batch computation approach, further enhancing the modeling power of PCSformer. Then, leveraging a crucial observation that the attention weight map exhibits cross-scale similarity, we introduce a local proxy-based global Transformer block. Specifically, for each nonoverlapping small patch, we employ the multiexpert system (MOE) to select a proxy token and conduct self-attention with all other proxy tokens in the feature map. By controlling the small patch size, we attain global self-attention in a highly cost-effective manner. The local proxy-based global Transformer block efficiently propagates valuable information from thin haze regions to thick haze regions, avoiding the accuracy degradation issue caused by the complete coverage of the local attention window by thick haze during dehazing. This results in finer and more consistent texture and structural details. Subsequently, we incorporate a shallow deep network with a small receptive field to restore better local texture details that may not be restored by the large receptive field of the global Transformer blocks and eliminate artifacts. The local refinement stage is computationally economical to have a small overhead and sufficiently expressive to provide additional flexibility when handling diverse types of information.

Researchers currently utilize pairs of synthetic hazy images and corresponding ground-truth images to train deep RS image dehazing networks. However, existing open-source RS image dehazing datasets either have a small number of images [22], [27], [28] or lack commonly occurring objects in downstream tasks (such as object detection [1] or semantic segmentation [2]) in the images [18]. The former limitation may lead to overfitting, resulting in poor generalization. The latter limitation results in the inability to significantly improve the performance of downstream computer vision tasks in hazy scenes after using a trained dehazing model to preprocess hazy images compared with not preprocessing hazy images. However, an important role of image dehazing algorithms is to serve as a preprocessing step to assist downstream computer vision tasks in hazy scenarios. To overcome these challenges, we propose synthesizing hazy images using large-scale RS object detection datasets (e.g., DIOR [29]) and RS semantic segmentation datasets (e.g., LoveDA [30]). This approach not only facilitates the acquisition of large-scale datasets but also allows the dehazing model to learn prior knowledge about downstream task scenarios during training. Consequently, it can better preprocess hazy images, thereby improving the performance of downstream tasks in hazy scenes. Our experiments demonstrate the effectiveness of this approach.

The contributions of our work can be summarized as follows.

1) We establish an expressive and general RS image dehazing framework named PCSformer. PCSformer has achieved state-of-the-art results on multiple benchmarks, particularly excelling in thick haze removal, while maintaining a lower parameter count and FLOPs.
2) We propose two innovative Transformer blocks, i.e., sliding cross-stripes Transformer block and local proxy-based global Transformer block. These blocks are capable of capturing rich contextual relationships for large-scale objects commonly observed in RS images and providing better restoration results for thick haze regions.
3) We incorporate a local refinement stage with a small receptive field, which significantly enhances visual quality and facilitates the recovery of rich texture and structural details, mitigating artifacts associated with a large receptive field.
4) We contribute two large-scale nonhomogeneous RS image dehazing benchmarks. After preprocessing hazy images with a dehazing model trained on our datasets rather than existing datasets, there is a more significant improvement in the performance of downstream computer vision tasks.

II. Related Work

A. Prior-Based Methods

As discussed in the Introduction, the prior-based dehazing methods rely on haze imaging equation [i.e., (1)] and impose one or more prior knowledge or assumptions on it to reduce the uncertainty of haze removal [31]. Kopf et al. [32] directly utilized 3-D geographic scene models extracted from Google Earth or Microsoft's Virtual Earth to calculate the scene depth map corresponding to the hazy images. However, accurately pinpointing the locations of many captured hazy images is challenging, leading to a lower generality of the algorithm. Tan [5] observed that hazy images have higher contrast compared to haze-free images. Within the Markov random field framework, image dehazing is achieved by

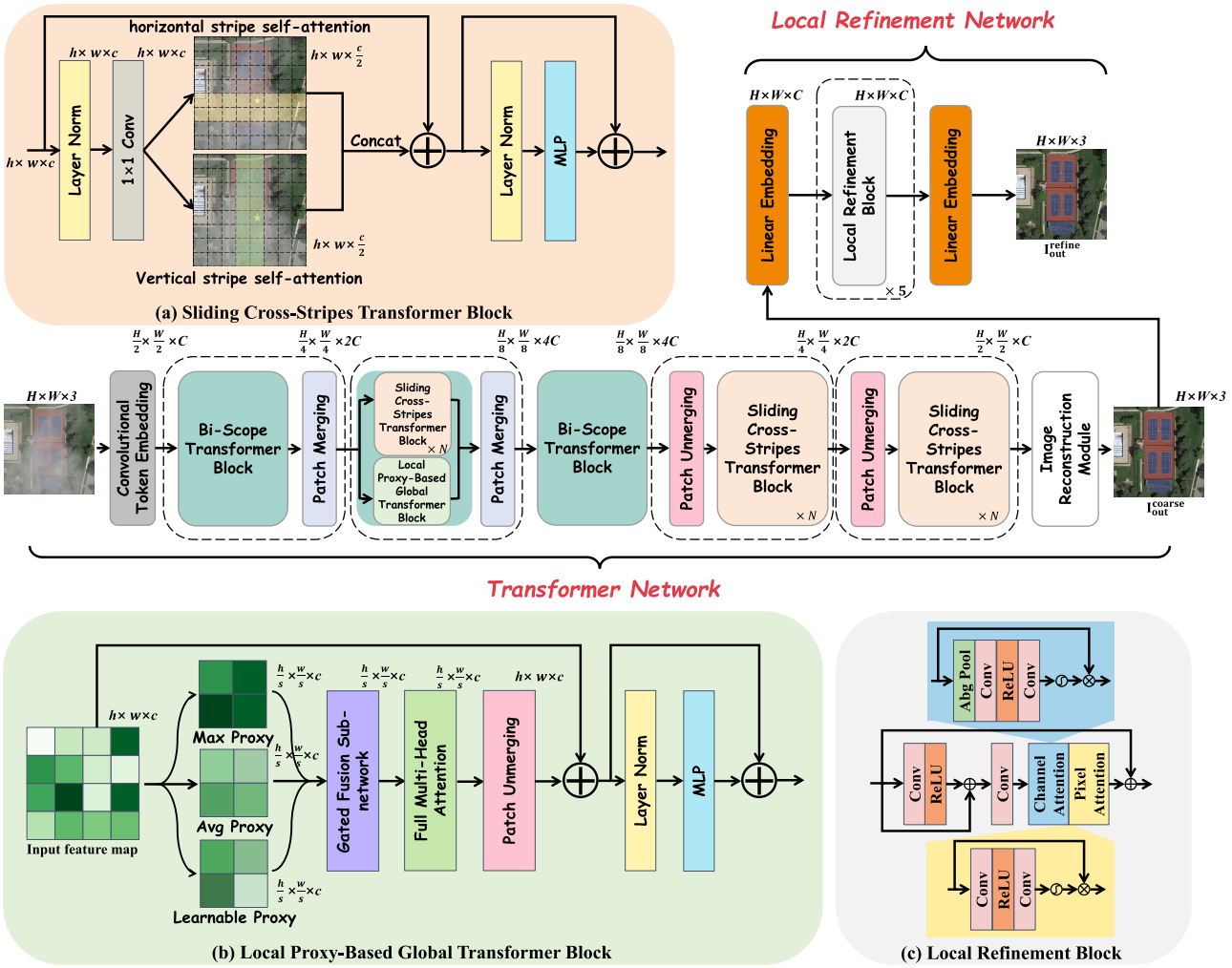

Fig. 1. Architecture of our proposed PCSformer for RS image dehazing. The Transformer network mainly consists of the proposed sliding cross-stripes Transformer block (SCSB) and local proxy-based global Transformer block (LPGB), which produce the coarse dehazed image $\mathbf{I}_{\text{out}}^{\text{coarse}}$. In the local refinement network, local refinement blocks are utilized to generate finer structural and texture details, resulting in the refined dehazed image $\mathbf{I}_{\text{out}}^{\text{refine}}$. (a) Detailed structure of SCSB. (b) Detailed structure of LPGB. (c) Detailed structure of local refinement block. Please note that blocks of the same color represent the same functionality.

maximizing the contrast of local image patches. However, the dehazed images often exhibit significant saturation and may introduce halo artifacts at depth discontinuities. Based on extensive observations of haze-free outdoor images, He et al. [6] proposed a simple yet effective DCP. This means that in most of the nonsky patches, at least one color channel of a certain pixel has very low intensity, even close to zero. At the same time, to eliminate the block artifacts in the estimated transmission map, He et al. [33] proposed the use of a guided filter for fine-tuning the edge regions, significantly improving the visual quality of the dehazed images.

Although prior-based methods can partially remove haze and enhance perceptual quality, they struggle to handle complex scenes or thick haze. More importantly, handcrafted prior assumptions do not always hold.

B. Deep Learning-Based Methods

With the rapid advancement of deep learning, initial CNN-based methods achieved better estimation of the transmission map t and atmospheric light A. For example, [34] employs a U-shaped network for predicting t and A and utilizes a discriminator for joint optimization. Subsequently, deep learning-based methods no longer rely on the estimation of t and A but instead directly predict the haze-free images or the residual of hazy images. For instance, FFA-Net [15] designs a feature attention module containing channel attention and pixel attention modules. These attention mechanisms make FFA-Net more focused on high-frequency information and regions with dense haze.

With the increasing potential demonstrated by ViT in various visual tasks, more Transformer-based dehazing models are being introduced. Dehazeformer [18] is one of the most representative works. It utilizes Swin Transformer [35] as a backbone and introduces several key design modifications, such as normalization layers, activation functions, and spatial information aggregation schemes. To obtain a reasonable estimation of the haze parameters, Trinity-Net [22] feeds the prior information obtained from DCP [6] into the Swin Transformer. In addition, a gradient guidance module is designed for the Swin Transformer blocks to prevent potential blurring of details that may be caused by the Swin

Transformer. AIDTransformer [20] introduces spatially attentive offset extraction in the deformable attentive Transformer block to extract relevant spatial features crucial for effective dehazing. While these Transformer-based methods have achieved impressive performance, they have not fully exploited the advantages of the self-attention mechanism in modeling long-range dependencies. In contrast, our PCSformer captures long-range dependencies directly within a single Transformer block through carefully designed modules.

III. Method

A. Overall Architecture

An overview of the PCSformer architecture is illustrated in Fig. 1. Our goals are threefold: 1) capturing long-range dependencies for large-scale objects in a single Transformer block; 2) providing more reliable and consistent restoration results for thick haze regions; and 3) generating finer details and eliminating potential artifacts caused by a large receptive field. Moreover, achieving the above three goals is computationally friendly. To address the first two goals, we design a *biscope Transformer block*, including sliding cross-stripes Transformer block (SCSB) and local proxy-based global Transformer block (LPGB). To improve computational efficiency and maintain model simplicity, we only utilize addition operations to fuse the two parallel feature maps. Addressing the latter, based on the coarse dehazed image restored by the first Transformer network, we tailor a local refinement stage using a shallow deep network with a small receptive field.

Given a hazy image $\mathbf{I_{in}} \in \mathbb{R}^{H\times W\times 3}$, based on the local smoothness prior of images and considering that early convolution can help the Transformer be more robust and easier to optimize [36], we utilize overlapped depthwise separable convolutional token embedding (7×7 with stride 2) to obtain $\mathbf{X_0} \in \mathbb{R}^{(H/2)\times(W/2)\times C}$. To reduce computational complexity and obtain hierarchical representation, we adopt a U-shaped encoder–decoder structure. Each encoder stage comprises a biscope Transformer block and a patch merging layer. The biscope Transformer block consists of a stack of SCSBs and an LPGB arranged side by side. Subsequently, feature maps are downsampled via the patch merging layer. For example, for input feature maps $\mathbf{X_0} \in \mathbb{R}^{(H/2)\times(W/2)\times C}$, the lth stage of the encoder outputs feature maps $\mathbf{X_l} \in \mathbb{R}^{(H/2^{l+1})\times(W/2^{l+1})\times 2^l C}$.

Moving on, the bottleneck stage exclusively utilizes a biscope Transformer block. For feature reconstruction, the LPGB is deemed unnecessary, resulting in a decoder stage comprising only a stack of SCSBs and a patch unmerging layer. We employ PixelShuffle [37] for upsampling, reducing half of the feature maps channels and doubling the size of the feature maps. At the end of Transformer network, the output feature maps from the last encoder stage are projected back to the original image size ($H \times W \times 3$) via the image reconstruction module, yielding a coarse dehazed image $\mathbf{I_{out}^{coarse}} \in \mathbb{R}^{H\times W\times 3}$. It is noteworthy that we connect each stage of the Transformer network in a densely connected manner [38] (not shown in Fig. 1). This dense connection enhances information flow throughout the network, contributing to the effective restoration of intricate details.

As for the local refinement network, we initially project the coarse dehazed image $\mathbf{I_{out}^{coarse}}$ back into the feature space. Subsequently, a stack of local refinement blocks is employed to enhance the restoration of local structures, intricate texture details, and eliminate potential artifacts. Finally, the refined dehazed image $\mathbf{I_{out}^{refine}} \in \mathbb{R}^{H\times W\times 3}$ is obtained. The same loss function is applied to both $\mathbf{I_{out}^{coarse}}$ and $\mathbf{I_{out}^{refine}}$.

B. Sliding Cross-Stripes Transformer Block

Since large-scale objects (such as bridges and football fields) are frequently present in RS hazy images, abundant contextual information becomes crucial for accurate haze removal. Building upon this understanding, we introduce an SCSB specifically designed for RS image dehazing.

In contrast to existing methods [18], [19], [20], [21], [22], [23] that are confined to local windows, given the input $X \in \mathbb{R}^{H\times W\times C}$, the SCSB performs self-attention in both horizontal stripes ($\text{sh} \times W$) and vertical stripes ($H \times \text{sw}$), where sh and sw represent the height and width of the stripe. To introduce no extra computation cost while enlarging the area for computing self-attention within each Transformer block, we employ them to different attention heads in parallel.

Specifically, we execute the attention operation M (the heads number) times in parallel, with $(M/2)$ devoted to horizontal stripes and the remaining $(M/2)$ dedicated to vertical stripes. For horizontal stripe self-attention, we evenly split X into nonoverlapping $\text{sh} \times W$ horizontal stripes, denoting the ith horizontal stripe feature as $X_i \in \mathbb{R}^{(\text{sh}\times W)\times C}$, where $i = 1, \ldots, (H/\text{sh})$. The output of X_i can be computed as

$$\begin{aligned}\left(Q_i^k, K_i^k, V_i^k\right) &= \left(X_i W_k^Q, X_i W_k^K, X_i W_k^V\right)\\ Y_i^k &= \text{Attention}\left(Q_i^k, K_i^k, V_i^k\right)\\ &= \text{Softmax}\left(\frac{Q_i^k \left(K_i^k\right)^T}{\sqrt{d}}\right) V_i^k\\ Y_i &= \text{Concat}\left(Y_i^1, \ldots, Y_i^{\frac{M}{2}}\right)\end{aligned} \tag{2}$$

where $Y_i^k \in \mathbb{R}^{(\text{sh}\times W)\times D}$ is the attention feature of X_i in the kth head and $D = (C/M)$ is the channel dimension in each head. $W_k^Q, W_k^K, W_k^V \in \mathbb{R}^{C\times(C/M)}$ represent the projection matrices of query, key, and value for the kth head. $Y_i \in \mathbb{R}^{(\text{sh}\times W)\times(C/2)}$ represents the self-attention output for the horizontal stripe X_i. d is a learnable parameter. Performing the attention operation on all X_i ($i = 1, \ldots, (H/\text{sh})$), reshaping, and merging them, we obtain the horizontal stripe attention feature $Y^{\text{horizontal}} \in \mathbb{R}^{(H\times W)\times(C/2)}$ of X.

Similar to the horizontal stripe self-attention, the vertical stripe self-attention partitions X into nonoverlapping $H \times \text{sw}$ vertical stripes and denote the ith vertical stripe feature as $X_i \in \mathbb{R}^{(H\times \text{sw})\times C}$, where $i = 1, \ldots, (W/\text{sw})$. The output of X_i can be computed as

$$\begin{aligned}\left(Q_i^k, K_i^k, V_i^k\right) &= \left(X_i W_k^Q, X_i W_k^K, X_i W_k^V\right)\\ Y_i^k &= \text{Attention}\left(Q_i^k, K_i^k, V_i^k\right)\end{aligned}$$

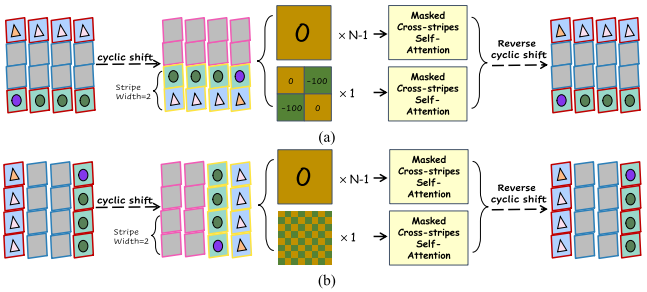

Fig. 2. Illustration of an efficient batch computation approach for (a) horizontal self-attention and (b) vertical self-attention in shifted stripe partitioning. Having the same border color indicates belonging to the same strip when strip partitioning, and having the same geometry indicates that attention should be calculated between them. The same geometry with different colors is used to demonstrate the rules of cyclic shift. N represents the number of stripes.

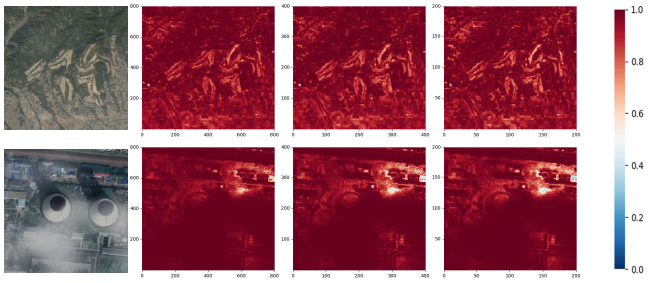

Fig. 3. Illustration of attention maps' cross-scale similarity. (a) Hazy images. (b) Attention map between the center pixel and the original resolution image. (c) Attention map between the center pixel and the image after 2× average pooling. (d) Attention map between the center pixel and the image after 4× average pooling. For better visibility, we normalize the values of the attention map to [0, 1]. Attention maps at different scales exhibit similar structures.

$$= \text{Softmax}\left(\frac{Q_i^k \left(K_i^k\right)^T}{\sqrt{d}}\right) V_i^k$$

$$Y_i = \text{Concat}\left(Y_i^{\frac{M}{2}+1}, \ldots, Y_i^M\right). \tag{3}$$

The vertical stripe attention feature, denoted as $Y^{\text{vertical}} \in \mathbb{R}^{(H\times W)\times(C/2)}$, is derived similarly. Finally, the outputs of these two parallel groups are concatenated along the channel dimension. The process is formulated as

$$\text{SCS-Attention} = \text{Concat}(Y^{\text{horizontal}}, Y^{\text{vertical}})W^P \tag{4}$$

where $W^P \in \mathbb{R}^{C\times C}$ represents the projection matrix for feature fusion and SCS-Attention represents the output of self-attention calculations in SCSB.

The stripe-based self-attention module lacks connections across stripes, which limits its modeling power. To establish connections between stripes while preserving the efficient computation of nonoverlapping stripes, we draw inspiration from [35] and propose an efficient batch computation approach, as illustrated in Fig. 2. Cyclic shift displaces the stripes by $([\text{sh}/2], [\text{sh}/2])$ or $([\text{sw}/2], [\text{sw}/2])$ pixels, respectively, from the regularly partitioned stripes in horizontal stripe self-attention or vertical stripe self-attention. After this shift, the bottom stripe with height sh or the rightmost stripe with width sw in the feature map is composed of two nonadjacent substripes in the feature map before the shift. Because the self-attention computation between these substripes is meaningless, applying a masking mechanism is necessary to limit it.

C. Local Proxy-Based Global Transformer Block

For thick and nonhomogeneous RS hazy images, propagating clear and effective information from thin haze regions to thick haze regions helps to restore more realistic and consistent texture details and structures for thick haze regions. To avoid attention windows encompassing numerous tokens projected from thick haze regions, the intuitive idea is to perform full attention across the entire feature map, but its computational complexity is unacceptable. Hence, we propose the LPGB as a solution to address this challenge.

As shown in Fig. 3, the attention map exhibits cross-scale similarity due to the homogeneous semantics carried by each small patch. This observation suggests that achieving global self-attention does not necessitate performing full attention on the original input feature map. Specifically, given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, we choose a proxy token for each small patch, resulting in a proxy feature map $X_P \in \mathbb{R}^{(H/s)\times(W/s)\times C}$, where s represents the small patch size. Subsequently, we conduct full attention on X_P.

Taking inspiration from the classical Mixture-of-Experts (MoEs) [39], we adopt a multiscale approach to select proxy tokens for small patches and utilize a simple gating network for aggregation, as illustrated in Fig. 1(b). The average proxy can extract background information, but it may dilute significant feature responses; Max proxy excels in texture and edge detection, but it tends to be oversensitive. Leveraging learnable proxy is an adaptive feature representation modulation approach, which provides greater flexibility and generalization. Then, we employ a gated fusion subnetwork, denoted by G, to determine the contributions of each proxy token selection strategy. G is computationally inexpensive yet sufficiently expressive to make informative decisions, which can be expressed as

$$\begin{aligned}(\sigma_1, \sigma_2, \sigma_3) &= G(\text{Avg}(X), \text{Max}(X), \text{Conv}(X)) \\ X_P &= \sigma_1 * \text{Avg}(X) + \sigma_2 * \text{Max}(X) + \sigma_3 * \text{Conv}(X)\end{aligned} \tag{5}$$

where $\text{Avg}(X)$, $\text{Max}(X)$, and $\text{Conv}(X)$ represent preliminary proxy feature maps obtained using average pooling, max pooling, and learnable convolutional layers, respectively. In order to reduce computational load, G is composed of depthwise separable convolution layers [40]. Finally, we execute full self-attention [17] on the proxy feature map X_P

$$\begin{aligned}(Q, K, V) &= (X_P W_Q, X_P W_K, X_P W_V) \\ \text{LPG-Attention} &= \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V\end{aligned} \tag{6}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C\times C}$ represent the projection matrices of query, key, and value. d is a learnable parameter. LPG-Attention represents the output of self-attention calculations in LPGB.

The computational complexity of a proxy-based global multihead self-attention is

$$\Omega = 4HWC^2 + 2\left(\frac{HW}{s^2}\right)^2 C \tag{7}$$

where H and W are the height and width of the original feature map, s is the patch size, and C is the channel dimension. Therefore, the proposed LPGB has roughly the same computational complexity as the window-based Transformer block. In other words, we generate more consistent and realistic dehazed results for thick haze regions without incurring high computational complexity.

D. Local Refinement Network

For local refinement, we adopt a shallow deep network. Considering that FFA-Net [15] has achieved impressive results in image dehazing, we only use a stack of five basic blocks proposed in FFA-Net [15] to construct our local refinement network. We would like to emphasize that a network with small receptive fields is an important supplement to image dehazing, so we do not care about the specific structural design of the local refinement network. Therefore, other modules can be used here to replace the modules in FFA-Net [15].

We begin by projecting the coarse dehazed image $\mathbf{I}_{\mathbf{out}}^{\mathbf{coarse}} \in \mathbb{R}^{H\times W\times 3}$ back into the feature space, yielding $F_{\text{in}} \in \mathbb{R}^{H\times W\times C}$. This F_{in} is then fed into a stack of five FFA basic blocks, and the architecture of each FFA basic block is illustrated in Fig. 1(c). Finally, a simple convolutional layer is utilized for reconstruction, yielding the refined dehazed image $\mathbf{I}_{\mathbf{out}}^{\mathbf{refine}} \in \mathbb{R}^{H\times W\times 3}$. The local refinement network introduces only additional 215k parameters, but ablation experiments demonstrate its significant capacity to enhance the texture and structural details of the dehazing results while eliminating artifacts.

E. Loss Function

The loss function of PCSformer includes three terms, i.e., $L1$ loss $\mathcal{L}_1$, robust loss $\mathcal{L}_R$, and perceptual loss $\mathcal{L}_P$.

Robust loss is proposed in [41], which can described as

$$\mathcal{L}_R\left(J_d, \hat{J}, \alpha, c\right) = \frac{|\alpha-2|}{\alpha}\left(\left(\frac{\left((J_d-\hat{J})/c\right)^2}{|\alpha-2|}+1\right)^{\alpha/2}-1\right) \tag{8}$$

where $\alpha \in \mathbb{R}$ is a shape parameter that controls the robustness of the loss and $c > 0$ is a scale parameter that controls the size of the loss's quadratic bowl near $J_d - \hat{J} = 0$. In the experiment, α and $c > 0$ are learnable parameters. J_d is the dehazed image, and $\hat{J}$ is the corresponding ground-truth image.

Perceptual loss is first proposed in [42] to keep image contents for style transfer and is now widely used for image dehazing [13], [43], [44] to minimize the perceptual difference between the reconstructed image and the ground-truth image. The perceptual loss computes the $L1$ loss in the feature level

$$\mathcal{L}_P\left(J_d, \hat{J}\right) = \sum_{p=0}^{P-1} \frac{\left|\Psi_p^{J_d} - \Psi_p^{\hat{J}}\right|}{N_{\Psi_p^{\hat{J}}}} \tag{9}$$

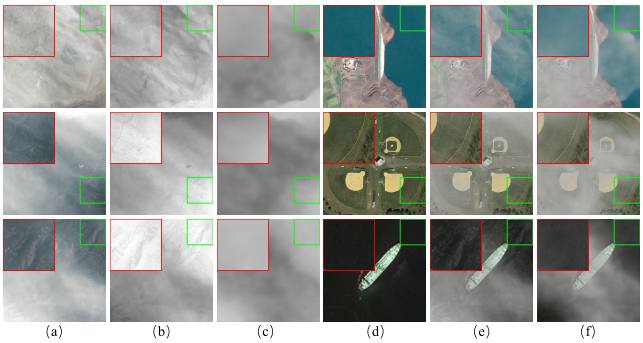

Fig. 4. Illustration of hazy image synthesis based on postprocessed transmission maps. (a) Collected RS hazy image. (b) Estimated transmission map by FSID [47]. (c) Refined transmission map obtained by postprocessing (b). (d) Collected RS haze-free image. (e) Hazy image synthesized using (b). (f) Hazy image synthesized using (c).

TABLE I
SUMMARY OF HAZY-DIOR AND HAZY-LOVEDA DATASETS

Dataset	Training Set	Validation Set	Test Set		
			Thin Haze	Moderate Haze	Thick Haze
Hazy-DIOR	56,310	6,258	2,607	2,607	2,607
Hazy-LoveDA	12,330	1,419	577	577	577

where $\Psi_p^{\hat{J}}$ denotes the activation from the pth selected layer of the pretrained network given the input $\hat{J}$ and $N_{\Psi_p^{\hat{J}}}$ is the number of elements in the pth layer. We use layer relu_{2_2}, relu_{3_3}, and relu_{4_3} from the VGG [45] network pretrained on ImageNet [10].

The overall loss function is defined as

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_1 + \lambda_2\mathcal{L}_R + \lambda_3\mathcal{L}_P \tag{10}$$

where λ_1, λ_2, and λ_3 are empirically set to 1, 0.4, and 0.5 to balance the scales of multiple losses, respectively.

The currently available large-scale datasets lack frequently occurring objects in downstream visual tasks, limiting the assistance of dehazing models for downstream visual tasks in foggy weather scenarios.

IV. EXPERIMENTS

A. Data Generation

A large-scale nonhomogeneous RS image dehazing dataset is crucial for training a deep network. However, the currently available open-source large-scale dataset [18] lacks frequently occurring objects in downstream visual tasks, limiting the assistance of dehazing models for downstream visual tasks in haze weather scenarios. To address this issue, we propose collecting RS haze-free images from the DIOR (RS object detection dataset) [29] and the LoveDA (RS semantic segmentation dataset) [30]. Inspired by the method of synthesizing hazy images in [46], we synthesize the RS hazy images based on (1).

To synthesize nonhomogeneous haze, we collect 20 000 nonhomogeneous RS hazy images and estimate their transmission maps using FSID [47]. However, estimated transmission maps using prior-based method may contain undesirable and data-dependent scene texture details, which lead to synthetic hazy images having artifacts, as shown in Fig. 4. We perform smoothing and refinement operations on coarse

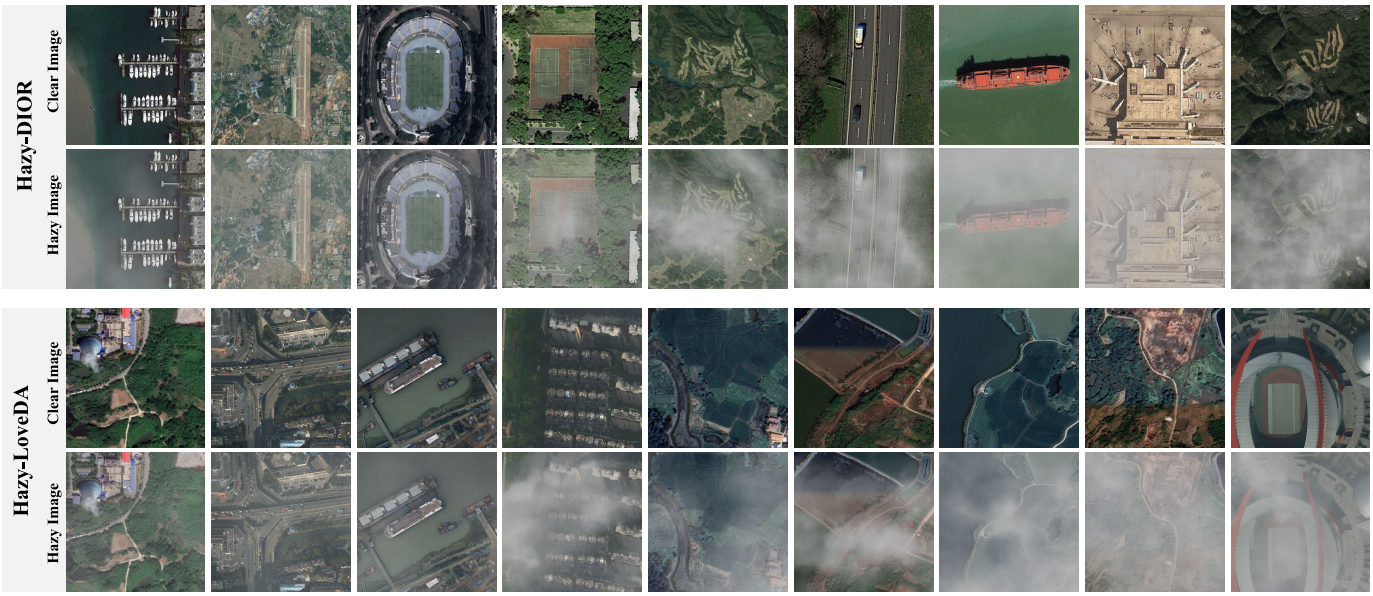

Fig. 5. Samples from the Hazy-Dior and Hazy-LoveDA datasets.

transmission maps acquired from FSID [47] to obtain the refined transmission maps. From Fig. 4, it can be observed that the hazy images synthesized using the refined transmission map are more realistic and visually appealing.

We collect 23 463 haze-free images from the DIOR [29] and 5987 haze-free images from the LoveDA [30] to create the Hazy-DIOR dataset and the Hazy-LoveDA dataset, respectively. For each haze-free image, we randomly select one from the 20 000 refined transmission maps and multiply the transmission map t in (1) by a coefficient to synthesize thin, moderate, and thick haze images. Hazy-DIOR and Hazy-LoveDA are divided into the training set, the validation set, and the test set according to the ratio of 8:2:1. The summary of Hazy-DIOR and Hazy-LoveDA is shown in Table I. Some example images of the constructed datasets are presented in Fig. 5.

B. Implementation Details

1) Parameter Settings: The proposed PCSformer is implemented with the PyTorch framework on an Intel Gold 6252 CPU and an NVIDIA A100 GPU. We use the Adam [51] optimizer with default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.99$) and cosine annealing strategy [52] to train the PCSformer.

For data augmentation, we randomly crop the input into a size of 256 × 256 and undergo horizontal flipping with a certain probability. The batch size and epoch were set to 10 and 100, respectively.

2) Benchmarks and Metrics: To validate the effectiveness of our proposed PCSformer, we train and test the model on the proposed Hazy-DIOR and Hazy-LoveDA datasets, respectively. For a more convincing comparison, we further extend evaluation to other real-world datasets, such as WHUS2-CR [53] and RICE-I [28]. The WHUS2-CR dataset is obtained from the Sentinel-2A satellite, and the period of acquisition of image pairs was the revisit time of the satellite (ten days) to a minimum of the difference between cloud and clean images. It contains a total of 24 450 images, with resolutions of 10, 20, and 60 m. We performed random cropping, generating 5000 image patches with dimensions of 256 × 256 pixels from the original high-resolution image pairs. In our experimentation, 4000 pairs were allocated for training, while the remaining 1000 pairs were reserved for testing. The RICE-I [28] dataset contains 500 image pairs from Google Earth, and each pair has cloudy and cloud-free images with sizes of 512 × 512. From RICE-I [28], 400 pairs were randomly allocated for training, while the remaining 100 pairs were reserved for testing purposes.

To evaluate the effectiveness of different image dehazing algorithms objectively, we use four metrics: peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) [54], learned perceptual image patch similarity (LPIPS) [55], and CIDE2000 [56]. In general, a higher PSNR or SSIM while a lower LPIPS or CIDE2000 indicates more authentic restored results and higher quality details.

C. Comparison With the State-of-the-Art

We compare our PCSformer with many state-of-the-art methods, including DCP [6], FFA-Net [15], UHD [48], Dehamer [19], Uformer [49], Restormer [25], UMW-Transformer [23], FocalNet [50], Trinity-Net [22], and DehazeFormer [18]. To verify scalability, we provide two architectures of PCSformer including PCSformer-B (the basic model) and PCSformer-S (a smaller variant).

1) Quantitative Comparison: We quantitatively compare the performance of PCSformer and challenging baselines. Table II displays the results on the Hazy-DIOR and Hazy-LoveDA datasets, while Tables III and IV present the results on the WHUS2-CR and RICE-I datasets, respectively. Our method achieves nearly SOTA performance in all metrics across all benchmarks. Overall, DehazeFormer [18],

TABLE II

AVERAGE PSNR (dB), SSIM, LPIPS ($\times 10^2$), AND CIDE2000 ON HAZY-DIOR AND HAZY-LOVEDA DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Methods	Publication	Haze Density	Hazy-DIOR				Hazy-LoveDA				Overhead	
			PSNR↑	SSIM↑	LPIPS↓	CIDE2000↓	PSNR↑	SSIM↑	LPIPS↓	CIDE2000↓	#Param	MACs
DCP [6]	TPAMI-10	thin	17.717	0.857	5.659	12.259	17.872	0.842	2.991	10.596	-	-
FFA-Net [15]	AAAI-20	thin	31.596	0.952	2.111	3.188	31.250	0.985	0.627	2.516	4.64M	288.86G
UHD [48]	CVPR-21	thin	31.131	0.951	2.108	3.231	32.339	0.986	0.537	2.328	34.55M	103.83G
Dehamer [19]	CVPR-22	thin	30.503	0.931	2.307	3.375	29.540	0.948	1.043	3.707	132.45M	59.25G
Uformer [49]	CVPR-22	thin	32.177	0.952	2.155	3.039	33.678	0.989	0.366	1.958	5.29M	10.76G
Restormer [25]	CVPR-22	thin	32.101	0.952	2.516	3.085	34.229	0.989	0.331	1.829	26.14M	140.99G
UMWTransformer [23]	ECCV-22	thin	31.822	**0.953**	**2.033**	3.097	33.505	0.988	0.404	1.995	17.51M	95.11G
FocalNet [50]	ICCV-23	thin	31.529	0.948	3.098	3.179	33.693	0.988	0.410	2.093	3.74M	30.67G
Trinity-Net [22]	TGRS-23	thin	30.573	0.952	2.527	3.323	33.042	0.989	0.514	2.492	20.14M	204.74G
DehazeFormer-M [18]	TIP-23	thin	32.253	0.952	2.470	3.048	33.632	0.989	0.374	1.878	4.63M	48.64G
PCSformer-S	-	thin	32.045	0.951	2.489	3.082	34.405	0.990	0.348	1.816	1.54M	12.03G
PCSformer-B	-	thin	**32.291**	**0.953**	2.106	**3.016**	**34.601**	**0.991**	**0.305**	**1.725**	3.73M	27.66G
DCP [6]	TPAMI-10	moderate	18.670	0.830	10.924	10.935	20.218	0.877	7.468	7.641	-	-
FFA-Net [15]	AAAI-20	moderate	26.923	0.893	6.720	4.292	22.168	0.906	8.708	5.990	4.64M	288.86G
UHD [48]	CVPR-21	moderate	25.491	0.886	7.744	4.856	25.083	0.932	6.101	4.367	34.55M	103.83G
Dehamer [19]	CVPR-22	moderate	26.392	0.873	7.064	4.569	25.450	0.900	6.312	4.769	132.45M	59.25G
Uformer [49]	CVPR-22	moderate	27.475	**0.895**	6.327	4.122	26.489	0.940	5.455	3.617	5.29M	10.76G
Restormer [25]	CVPR-22	moderate	27.508	0.894	6.722	4.129	26.610	0.941	5.287	3.535	26.14M	140.99G
UMWTransformer [23]	ECCV-22	moderate	26.941	0.893	6.247	4.263	24.629	0.928	6.224	4.428	17.51M	95.11G
FocalNet [50]	ICCV-23	moderate	26.114	0.881	10.295	4.565	26.177	0.934	6.022	3.996	3.74M	30.67G
Trinity-Net [22]	TGRS-23	moderate	23.981	0.870	11.401	5.750	25.054	0.921	7.964	4.685	20,14M	204.74G
DehazeFormer-M [18]	TIP-23	moderate	27.563	**0.895**	6.592	4.066	26.870	0.939	5.390	3.343	4.63M	48.64G
PCSformer-S	-	moderate	27.756	0.894	6.639	4.045	28.041	0.949	4.619	2.975	1.54M	12.03G
PCSformer-B	-	moderate	**27.835**	**0.895**	**6.246**	**4.011**	**28.151**	**0.952**	**4.240**	**2.953**	3.73M	27.66G
DCP [6]	TPAMI-10	thick	19.489	0.827	11.171	9.899	21.913	0.911	7.250	5.897	-	-
FFA-Net [15]	AAAI-20	thick	26.106	0.843	9.473	4.728	24.658	0.913	8.146	5.522	4.64M	288.86G
UHD [48]	CVPR-21	thick	24.707	0.841	10.550	5.496	25.265	0.926	6.398	4.502	34.55M	103.83G
Dehamer [19]	CVPR-22	thick	25.863	0.831	10.049	4.975	25.551	0.897	6.384	4.858	132.45M	59.25G
Uformer [49]	CVPR-22	thick	26.371	0.844	9.463	4.668	26.914	0.936	5.249	3.909	5.29M	10.76G
Restormer [25]	CVPR-22	thick	26.462	0.844	9.623	4.637	26.578	0.935	5.357	3.899	26.14M	140.99G
UMWTransformer [23]	ECCV-22	thick	26.114	0.844	9.129	4.693	25.913	0.930	5.446	4.232	17.51M	95.11G
FocalNet [50]	ICCV-23	thick	25.572	0.836	14.619	5.017	26.246	0.922	6.823	4.468	3.74M	30.67G
Trinity-Net [22]	TGRS-23	thick	23.973	0.826	15.493	6.519	25.573	0.913	9.056	4.923	20,14M	204.74G
DehazeFormer-M [18]	TIP-23	thick	26.471	0.844	9.441	4.601	26.890	0.933	5.589	3.599	4.63M	48.64G
PCSformer-S	-	thick	26.576	0.843	10.035	4.521	28.260	0.946	4.788	**3.072**	1.54M	12.03G
PCSformer-B	-	thick	**26.736**	**0.845**	**9.040**	**4.504**	**28.292**	**0.948**	**4.391**	3.136	3.73M	27.66G

TABLE III

AVERAGE PSNR (dB), SSIM, LPIPS ($\times 10^2$), AND CIDE2000 ON THE WHUS2-CR DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Methods	Publication	PSNR↑	SSIM↑	LPIPS↓	CIDE2000↓
DCP [6]	TPAMI-10	12.970	0.753	19.867	13.738
FFA-Net [15]	AAAI-20	30.700	0.939	6.083	4.453
UHD [48]	CVPR-21	30.121	0.937	6.252	4.865
Dehamer [19]	CVPR-22	30.132	0.937	**5.589**	4.809
Uformer [49]	CVPR-22	30.053	0.938	6.163	4.736
Restormer [25]	CVPR-22	30.227	0.937	6.180	4.676
UMWTransformer [23]	ECCV-22	30.119	0.938	5.757	4.790
FocalNet [50]	ICCV-23	29.746	0.938	6.370	4.868
Trinity-Net [22]	TGRS-23	30.232	0.938	5.796	4.764
DehazeFormer-M [18]	TIP-23	30.550	0.939	6.141	4.671
PCSformer-S	-	30.603	0.939	5.665	**4.011**
PCSformer-B	-	**30.840**	**0.940**	5.663	4.227

TABLE IV

AVERAGE PSNR (dB), SSIM, LPIPS ($\times 10^2$), AND CIDE2000 ON THE RICE-I DATASET. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Methods	Publication	PSNR↑	SSIM↑	LPIPS↓	CIDE2000↓
DCP [6]	TPAMI-10	17.431	0.908	9.128	11.049
FFA-Net [15]	AAAI-20	36.649	0.978	2.542	1.797
UHD [48]	CVPR-21	26.284	0.938	5.512	5.281
Dehamer [19]	CVPR-22	36.201	0.975	2.140	1.912
Uformer [49]	CVPR-22	37.048	0.980	2.359	1.851
Restormer [25]	CVPR-22	37.185	0.978	2.336	1.768
UMWTransformer [23]	ECCV-22	36.935	0.979	2.248	1.796
FocalNet [50]	ICCV-23	36.118	0.978	2.404	1.949
Trinity-Net [22]	TGRS-23	32.049	0.962	3.310	3.002
DehazeFormer-M [18]	TIP-23	37.291	0.980	2.024	1.760
PCSformer-S	-	37.384	0.980	2.035	**1.745**
PCSformer-B	-	**37.482**	**0.981**	**1.951**	1.758

Uformer [49], and Restormer [25] are the best-performing methods among the baselines. In thin haze scenes of the Hazy-DIOR and Hazy-LoveDA datasets, our PCSformer-B only shows a slight advantage over them. However, PCSformer-B achieves a significant lead in moderate haze scenes and thick haze scenes. As stated in the introduction, Dehazeformer [18] and Uformer [49], which perform self-attention within local windows, face limitations in capturing rich contextual information for large-scale objects commonly found in RS images. Moreover, their effectiveness may be compromised when a local window is entirely covered by a thick haze region, as valuable information may be lacking within the confined local window. Restormer [25] applies self-attention across channels rather than the spatial dimension, which results in a lower expressive capacity for the model. This limitation makes it unsuitable for handlingchallenging restoration tasks. Thanks to the designs of SCSB and LPGB, PCSformer achieves a large receptive field with lower

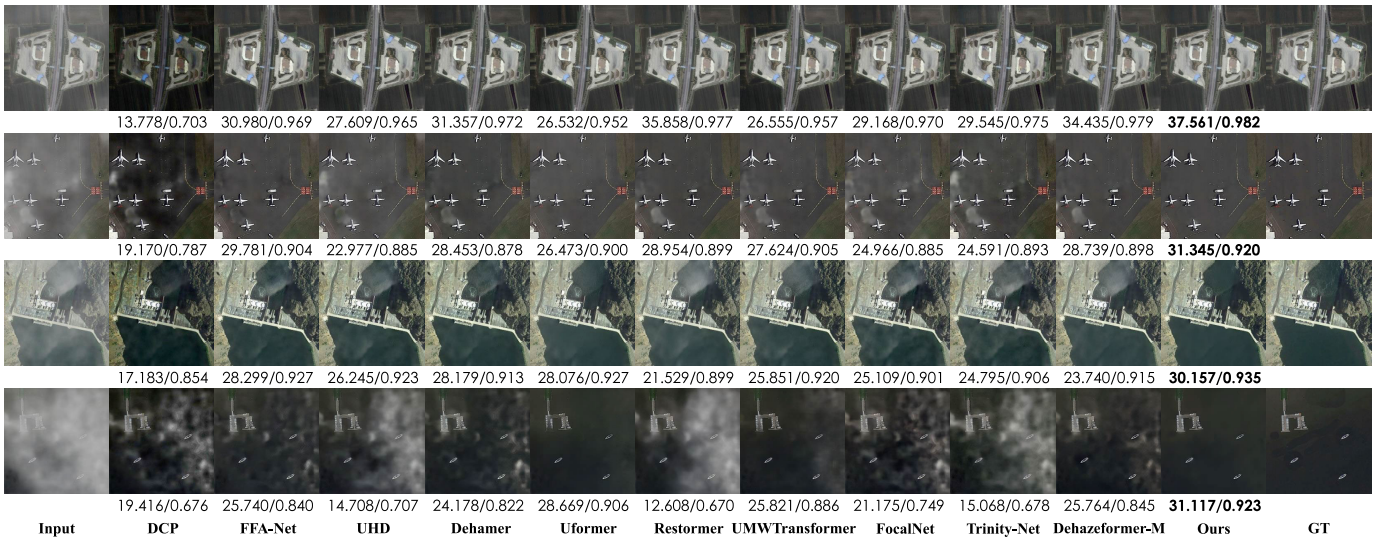

Fig. 6. Visual comparisons on the Hazy-DIOR dataset. The values below the images represent the PSNR and SSIM metrics.

Fig. 7. Visual comparisons on the Hazy-LoveDA dataset. The values below the images represent the PSNR and SSIM metrics.

computational complexity within a single Transformer block. In this process, it significantly enhances the ability to remove thick haze without sacrificing performance in scenes with thin haze. In summary, we establish an expressive and general RS image dehazing framework.

2) Qualitative Comparison: For a more intuitive comparison, we report the visual results of all approaches in Figs. 6, 7, and 8. Our method can handle challenging thick and nonhomogeneous haze, producing fewer artifacts and better haze removal compared to the baseline. For example, in the third and fourth rows of Fig. 6, only our method achieves nearly complete removal of haze in the water surface area. This once again confirms the superiority of our PCSformer in thick haze removal.

3) Performance and Efficiency Tradeoffs: Fig. 9 reflects the performance and efficiency tradeoffs of several SOTA methods, with the bar graph representing inference time and the line graph representing PSNR. The inference time is calculated on 256×256 images. Distinctly, only the proposed PCSformer and DehazeFormer [18] have achieved excellent performance and efficiency tradeoffs across all datasets (i.e., the line graph being notably higher than the bar graph). The inference time of PCSformer is only slightly higher than DehazeFormer [18] yet still achieves 29 frames per second (FPS), while the dehazing performance has significantly improved. These results demonstrate the efficiency and practicality of PCSformer.

D. Ablation Study

We conduct comprehensive ablation studies on the proposed dataset to verify the effectiveness of core components.

1) Training Dataset: To assess whether there is a more significant improvement in the performance of downstream computer vision tasks in hazy scenes after preprocessing hazy images with a dehazing model trained on our proposed datasets rather than existing datasets, we conduct a comprehensive ablation study on object detection and semantic segmentation

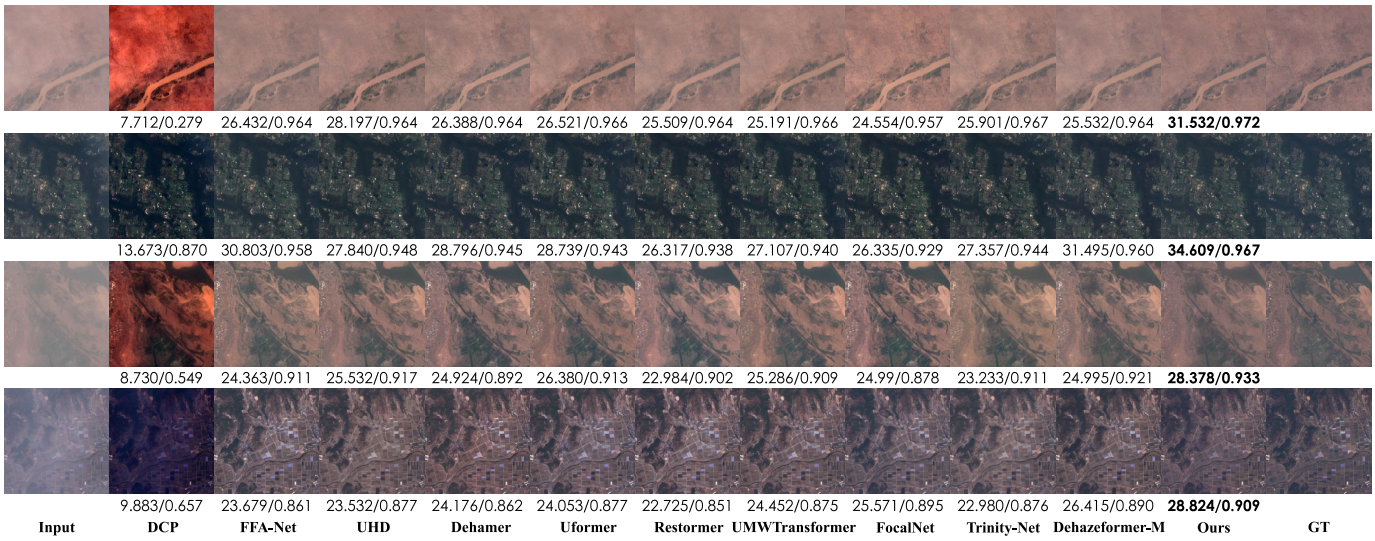

Fig. 8. Visual comparisons on the WHUS2-CR dataset. The values below the images represent the PSNR and SSIM metrics.

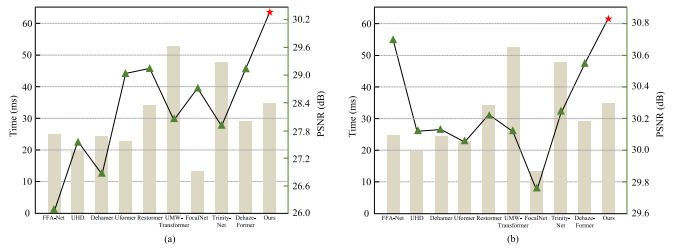

Fig. 9. Comparison of performance and efficiency tradeoffs on (a) Hazy-LoveDA and (b) WHUS2-CR [53] datasets. The PSNR on the Hazy-LoveDA dataset is the average over thin, moderate, and thick haze.

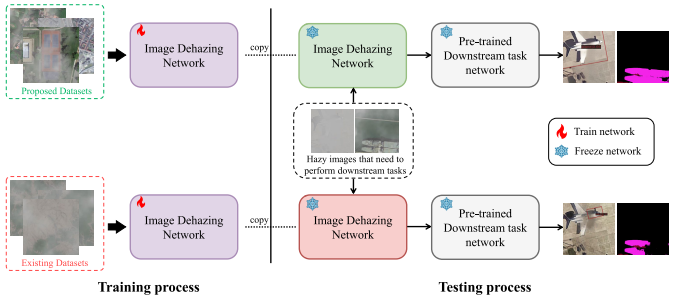

Fig. 10. Pipeline of ablation experiment that demonstrates the effectiveness of proposed RS image dehazing datasets. We trained the same image dehazing network using the proposed datasets (i.e., Hazy-DIOR or Hazy-LoveDA) and an existing dataset (e.g., RSHaze [18]), respectively. These two networks are then used to preprocess images for downstream visual tasks in hazy scenes. Finally, the preprocessed images (dehazed images) were fed into a pretrained downstream task network. We compare the results on downstream tasks, and it can be seen that the dehazing network trained on the proposed datasets can better help downstream computer vision tasks in hazy scenes.

tasks. The pipeline for this ablation experiment is shown in Fig. 10.

In this study, we adopt RSHaze [18] as the baseline dataset, which lacks commonly occurring objects in downstream tasks. This helps us validate that reasonably selecting clear images (i.e., selecting clear images that contain objects frequently appearing in downstream tasks) is crucial to support downstream computer vision tasks in hazy scenes. Some example images of RSHaze [18] are presented in Fig. 11. To ensure a fair comparison, we resynthesize hazy images on the clear images from RSHaze [18] using the method

Fig. 11. Examples of clear images from the RSHaze [18] dataset.

in Section IV-A, denoted as RSHaze-New. This is consistent with how Hazy-DIOR and Hazy-LoveDA were constructed. Considering that RSHaze [18] contains 5700 clear images from different scenes, and each clear image is synthesized into three different concentrations of hazy images, RSHaze-New comprises 17 100 training image pairs. Due to the unequal number of image pairs in RSHaze-New compared to Hazy-DIOR or Hazy-LoveDA, for a fair comparison, we randomly select image pairs from the dataset with a larger number of pairs to match the number of pairs in the dataset with a smaller number. Subsequently, we train the PCSformer on RSHaze-New and proposed datasets for 100 epochs.

Next, we need to synthesize the test sets for downstream computer vision tasks in hazy scenes. We use the test set of DOTAv1.0 [57] (object detection), the validation set of iSAID [58] (semantic segmentation), and the validation set of ISPRS Potsdam (semantic segmentation) as clear images and use the method in Section IV-A to synthesize thick-level hazy images. The obtained test sets are denoted as Hazy-DOTA, Hazy-iSAID, and Hazy-Potsdam. To validate the proposed Hazy-DIOR dataset, we conduct an ablation study on object detection and semantic segmentation tasks using the Hazy-DOTA and Hazy-iSAID datasets, respectively. To validate the proposed Hazy-LoveDA dataset, we conduct an ablation study on the semantic segmentation task using the Hazy-Potsdam dataset.

We use Oriented RepPoints [59] and HRNet [60] as object detection and semantic segmentation models, respectively. We utilize public codebase MMDetection [61] and

TABLE V

ABLATION EXPERIMENT ON THE PROPOSED DATASET HAZY-DIOR USING THE HAZY-DOTA DATASET, INVESTIGATING THE IMPACT OF TRAINING DATASET ON THE PERFORMANCE OF RS OBJECT DETECTION IN HAZY SCENES. THE BEST RESULTS AMONG HAZY, RSHAZE-NEW, AND HAZY-DIOR SETTINGS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Setting	mAP↑	AP per category ↑														
		PL	BD	BR	GTF	SV	LV	SH	TC	BC	ST	SBF	RA	HA	SP	HC
Clear	75.18	88.27	83.73	53.28	67.96	80.56	77.33	87.10	90.84	82.62	85.65	59.41	69.19	69.54	73.07	59.12
Hazy	37.97	68.72	26.62	22.06	22.57	39.72	14.53	63.10	81.56	58.58	50.45	18.41	31.63	26.65	14.68	30.26
RSHaze-New	58.78	80.69	53.42	34.17	43.06	61.62	47.09	83.39	**90.81**	68.69	70.67	**41.92**	50.63	59.66	55.79	40.02
Hazy-DIOR	**68.12**	**86.19**	**64.61**	**38.15**	**49.94**	**79.21**	**76.48**	**86.63**	90.72	**74.04**	**82.76**	41.36	**55.85**	**65.81**	**73.08**	**56.92**

PL: plane. BD: baseball diamond. BR: bridge. GTF: ground track field. SV: small vehicle. LV: large vehicle. SH: ship. TC: tennis court. BC: baseball court. ST: storage tank. SBF: soccer ball field. RA: roundabout. HA: harbor. SP: swimming pool. HC: helicopter.

TABLE VI

ABLATION EXPERIMENT ON THE PROPOSED DATASET HAZY-DIOR USING THE HAZY-IDAID DATASET, INVESTIGATING THE IMPACT OF TRAINING DATASET ON THE PERFORMANCE OF RS SEMANTIC SEGMENTATION IN HAZY SCENES. THE BEST RESULTS AMONG HAZY, RSHAZE-NEW, AND HAZY-DIOR SETTINGS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Setting	mIoU↑	IoU per category ↑															
		LV	SV	Plane	HC	Ship	ST	Bridge	RA	Harbor	BD	TC	GTF	SBF	SP	BC	Background
Clear	67.80	65.03	52.01	85.18	44.22	72.96	72.91	45.25	71.63	59.36	79.57	88.21	58.74	78.57	46.98	65.04	99.15
Hazy	51.69	39.15	32.95	79.85	32.70	56.44	60.50	23.94	41.56	44.11	60.90	81.95	45.32	59.57	13.11	56.29	98.66
RSHaze-New	62.12	59.15	46.87	83.17	36.94	67.63	66.39	33.70	61.73	54.57	73.18	**87.08**	**53.12**	**72.30**	37.96	61.19	98.99
Hazy-DIOR	**64.53**	**64.48**	**50.70**	**84.44**	**39.48**	**70.06**	**71.55**	**36.04**	**64.99**	**57.21**	**74.72**	86.65	52.58	71.41	**43.73**	**65.31**	**99.06**

Categories in iSAID dataset: Large Vehicle (LV), Small Vehicle (SV), Plane, Helicopter (HC), Ship, Storage Tank (ST), Bridge, Roundabout (RA), Harbor, Baseball Diamond (BD), Tennis Court (TC), Ground Track Field (GTF), Soccerball Field (SBF), Swimming Pool (SP), Basketball Court (BC) and Background.

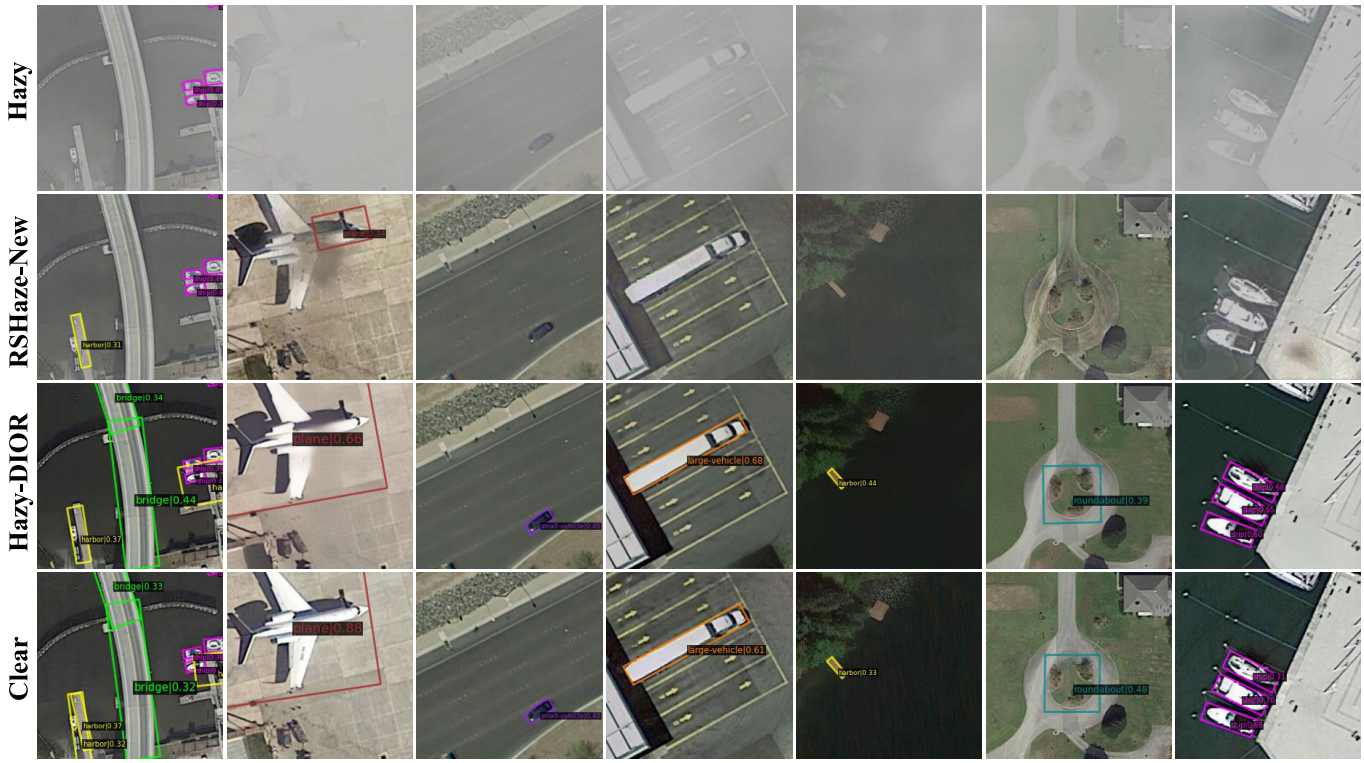

Fig. 12. Qualitative comparison of the impact of training dataset for dehazing models using the Hazy-DOTA dataset. **Hazy** denotes direct object detection on hazy images. **RSHaze-New** denotes object detection on dehazed images generated by PCSformer trained on the RSHaze-New dataset. **Hazy-DIOR** denotes object detection on dehazed images generated by PCSformer trained on the proposed Hazy-DIOR dataset. **Clear** denotes object detection on clear images.

MMSegmentation [62] in our implementation, respectively. We use mean Average Precision (mAP) and mean Intersection over Union (mIoU) as evaluation metrics for object detection and semantic segmentation, which are the most commonly used metrics.

The corresponding quantitative results are presented in Tables V–VII, respectively. Specifically, the setting "Clear" indicates performing downstream tasks directly on clear images, representing the highest performance achievable by the selected pretrained downstream task network. The

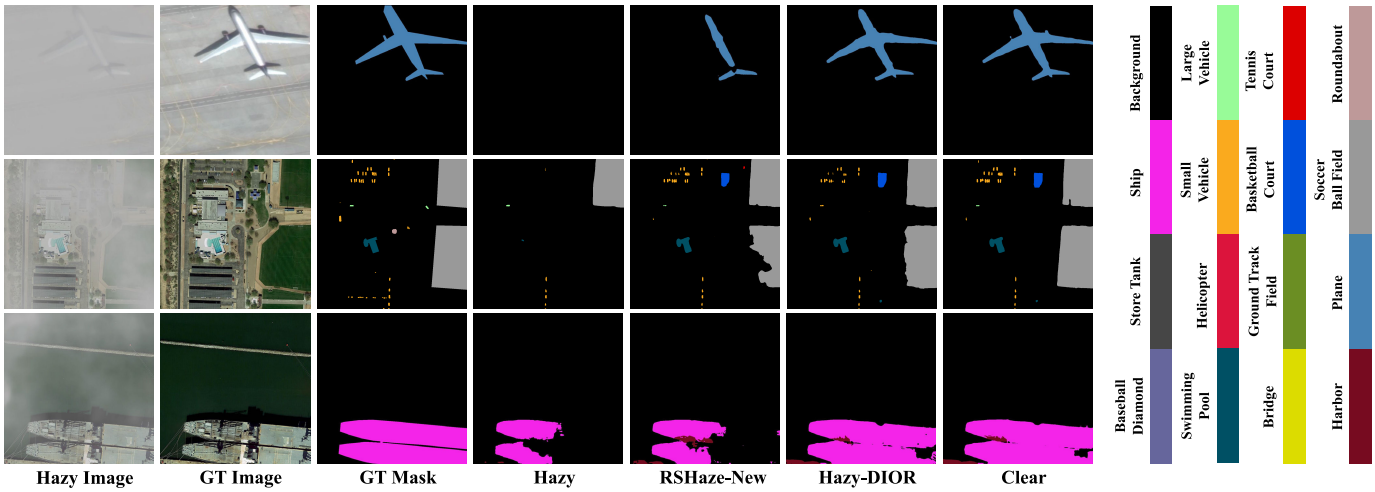

Fig. 13. Qualitative comparison of the impact of training dataset for dehazing models using the Hazy-iSAID dataset. **Hazy** denotes direct semantic segmentation on hazy images. **RSHaze-New** denotes semantic segmentation on dehazed images generated by PCSformer trained on the RSHaze-New dataset. **Hazy-DIOR** denotes semantic segmentation on dehazed images generated by PCSformer trained on the proposed Hazy-DIOR dataset. **Clear** denotes semantic segmentation on clear images.

TABLE VII

ABLATION EXPERIMENT ON THE PROPOSED DATASET HAZY-LOVEDA USING THE HAZY-POTSDAM DATASET, INVESTIGATING THE IMPACT OF TRAINING DATASET ON THE PERFORMANCE OF RS SEMANTIC SEGMENTATION IN HAZY SCENES. THE BEST RESULTS AMONG HAZY, RSHAZE-NEW, AND HAZY-LOVEDA SETTINGS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Setting	mIoU↑	IoU per category ↑					
		Impervious_Surface	Building	Low_Vegetation	Tree	Car	Clutter
Clear	78.39	87.22	93.75	76.86	79.54	92.65	40.32
Hazy	47.38	50.03	81.14	15.71	53.67	68.46	15.27
RSHaze-New	69.95	73.34	91.46	58.30	74.08	87.81	34.70
Hazy-LoveDA	**73.66**	**80.99**	**91.77**	**67.93**	**74.42**	**90.06**	**36.78**

TABLE VIII

ABLATION STUDY ON WINDOW PARTITIONING STRATEGIES. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Setting	Thin		Moderate		Thick		#Param	MACs
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
PCSformer	**32.291**	**0.953**	**27.835**	**0.895**	**26.736**	**0.845**	3.73M	27.66G
→ Swin Transformer Block	32.178	0.952	27.082	0.893	26.163	0.841	4.85M	28.02G

setting "Hazy" indicates performing downstream tasks directly on hazy images without using a dehazing model for image preprocessing. The setting "RSHaze-New" indicates performing downstream tasks on the dehazed images produced by the dehazing model trained on the RSHaze-New dataset. Similarly, the setting "Hazy-DIOR" or "Hazy-DIOR" indicates performing downstream tasks on the dehazed images produced by the dehazing model trained on the proposed Hazy-DIOR or Hazy-LoveDA dataset.

The visualization results of object detection and semantic segmentation are shown in Figs. 12–14, respectively. It is evident that the dehazing model trained on the proposed Hazy-DIOR and Hazy-LoveDA datasets can significantly enhance the performance of downstream tasks in hazy scenes. Most importantly, we provide a new insight for selecting clear images to create large-scale dehazing datasets.

2) Window Partitioning Strategy: We investigate the impact of window partitioning within the Transformer block on the proposed Hazy-DIOR dataset, and the results are displayed in Table VIII. We replace the SCSB with the Swin Transformer block [35] as the baseline. Notably, despite PCSformer having 1 million fewer parameters, it outperforms in performance across all three haze densities. This implies that SCSB can replace the commonly used Swin Transformer block in the image dehazing backbone as a fundamental building block. Moreover, SCSB significantly enhances the model's expressive power, achieving better dehazing performance with fewer computational expenses.

TABLE IX

ABLATION STUDY ON LOCAL PROXY-BASED GLOBAL TRANSFORMER BLOCK. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Setting	Moderate		Thick		#Param	MACs
	PSNR	SSIM	PSNR	SSIM		
PCSformer	**28.151**	**0.952**	**28.292**	**0.948**	3.73M	27.66G
− LPBG	27.873	0.951	27.843	0.945	4.51M	28.08G

3) Local Proxy-Based Global Transformer Block: This ablation study was conducted on the proposed Hazy-LoveDA dataset, and the results showcasing the effectiveness of the LPGB are presented in Table IX. To establish a baseline, we exclude the LPGB and adjust the model hyperparameters to make their parameter counts approximately equal for a fair comparison. To validate that LPGB can generate superior restoration results in regions with thick haze, we report the metrics on the Hazy-LoveDA dataset for moderate and thick levels of haze. This implies that LPGB has the potential to serve as a supplementary block for dehazing models, bringing finer and more consistent restoration results to challenging dense haze regions.

4) Proxy Selection Strategy: We investigate the impact of three proxy selection strategies and the gating network on the proposed Hazy-DIOR dataset, and the quantitative comparison results are presented in Table X. The fusion of a multiscale selection strategy and gating network enhances the robustness of proxy token selection, significantly improving dehazing performance with only minimal additional computational cost.

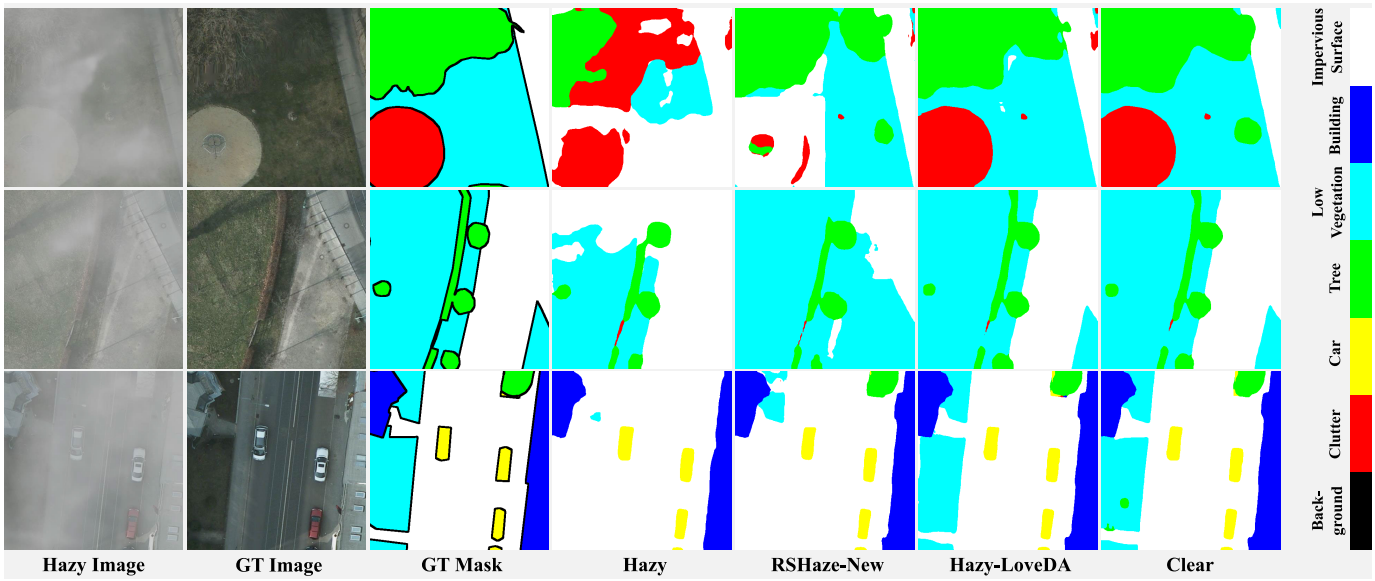

Fig. 14. Qualitative comparison of the impact of training dataset for dehazing models using the Hazy-Potsdam dataset. **Hazy** denotes direct semantic segmentation on hazy images. **RSHaze-New** denotes semantic segmentation on dehazed images generated by PCSformer trained on the RSHaze-New dataset. **Hazy-LoveDA** denotes semantic segmentation on dehazed images generated by PCSformer trained on the proposed Hazy-LoveDA dataset. **Clear** denotes semantic segmentation on clear images.

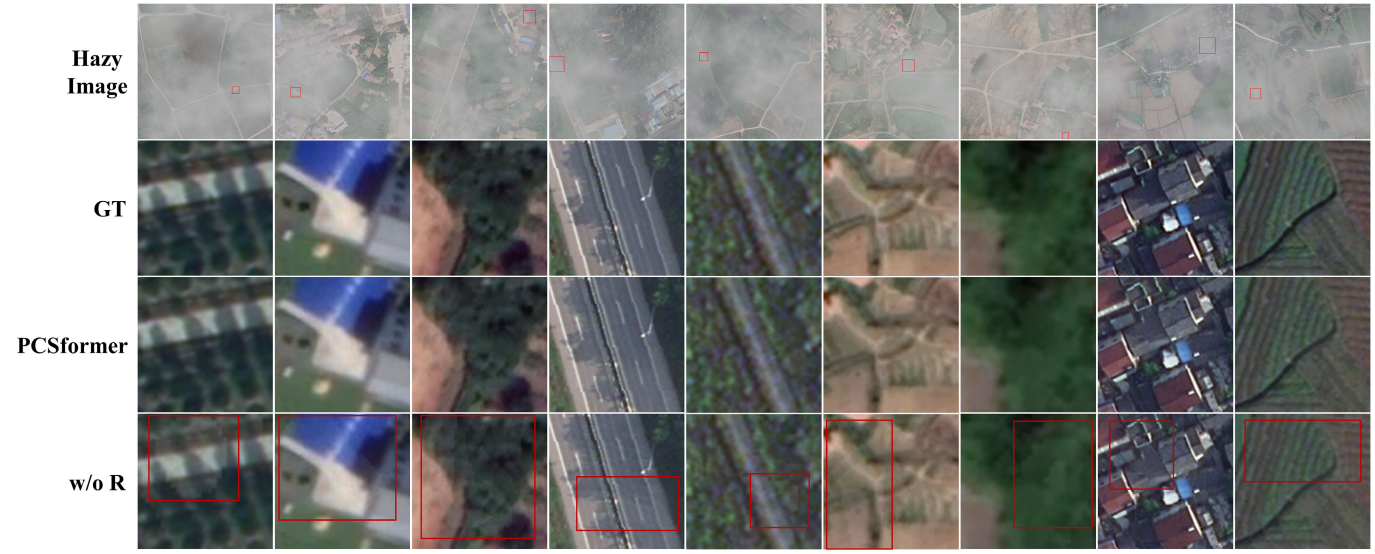

Fig. 15. Ablation study on the local refinement network, where w/o R represents without the local refinement network. The local refinement network is able to capture finer local textures and structural details while reducing artifacts. Zoom in for the best view and artifact areas marked in red box.

TABLE X

ABLATION STUDY ON PROXY SELECTION STRATEGY. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Setting				Thin		Moderate		Thick		#Param	MACs
Avg	Max	Conv	Gate	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM		
✔	✔	✔	✔	**32.291**	0.953	**27.835**	**0.895**	**26.736**	**0.845**	3.73M	27.66G
✔	✔	✔	✗	32.110	0.948	27.709	0.889	26.579	0.839	3.71M	27.64G
✔	✗	✗	✗	32.035	0.951	27.512	0.892	26.546	0.843	3.60M	27.56G
✗	✔	✗	✗	32.250	**0.956**	27.309	0.891	26.139	0.834	3.60M	27.56G
✗	✗	✔	✗	32.242	**0.956**	27.693	0.893	26.494	0.838	3.71M	27.64G

TABLE XI

ABLATION STUDY ON LOSS FUNCTION. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND BEST RESULTS ARE UNDERLINED

Setting			Thin		Moderate		Thick	
Robust Loss	Perceptual Loss	L_1 Loss	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
✔	✔	✔	**34.601**	**0.991**	**28.151**	**0.952**	**28.292**	**0.948**
✗	✔	✔	34.146	0.988	27.757	0.950	28.086	0.946
✗	✗	✔	33.850	0.985	28.042	0.949	28.182	0.943

5) Local Refinement Network: We qualitatively investigate the impact of the local refinement network on the proposed Hazy-DIOR and Hazy-LoveDA datasets. Visual comparisons are illustrated in Fig. 15, where w/o R represents without the local refinement network. We observe that the incorporation of the local refinement network enables the capture of finer local textures and structural details. We have brought a new perspective to image dehazing, showing that while large receptive fields are better suited for handling dense haze regions, complementary small receptive fields are effective in eliminating artifacts.

6) Loss Function: Furthermore, we investigate the influence of the objective function on the network's final recovery

performance on the proposed Hazy-LoveDA dataset. The corresponding experimental results are displayed in Table XI. Perceptual loss [42] may not handle moderate and thick haze scenes well, so we use robust loss [41] as a supplementary measure. We are the first to introduce the robust loss [41] into the field of image dehazing and demonstrate its effectiveness.

V. Conclusion

In this article, we propose an expressive and general RS image dehazing framework called PCSformer, which achieves state-of-the-art results on multiple evaluation metrics across multiple benchmarks with a lower number of parameters and Flops. PCSformer introduces two innovative Transformer blocks, namely, SCSB and LPGB. The former can serve as the fundamental building block for the image dehazing backbone, not only significantly enhancing the model's feature representation capability but also effectively addressing the common limitations of Transformer-based dehazing models. The latter is a supplementary block that can be used in any network, capable of providing finer restoration results for dense haze regions without significantly increasing the model complexity. The local refinement network effectively eliminates artifacts in the dehazing results, emphasizing the importance of small receptive fields and bringing a new perspective to image dehazing. Finally, we provide two large-scale datasets for RS image dehazing. Through numerous experiments, we demonstrate that after preprocessing hazy images with a dehazing model trained on our datasets, rather than existing datasets, there is a more significant improvement in the performance of downstream computer vision tasks in hazy scenes. This provides insights into the creation of RS image dehazing datasets.

References

[1] W. Fang, G. Zhang, Y. Zheng, and Y. Chen, "Multi-task learning for UAV aerial object detection in foggy weather condition," *Remote Sens.*, vol. 15, no. 18, p. 4617, Sep. 2023.

[2] W. Shi, W. Qin, and A. Chen, "Towards robust semantic segmentation of land covers in foggy conditions," *Remote Sens.*, vol. 14, no. 18, p. 4551, Sep. 2022.

[3] S. G. Narasimhan and S. K. Nayar, "Vision and the atmosphere," *Int. J. Comput. Vis.*, vol. 48, no. 3, pp. 233–254, 2002.

[4] R. Fattal, "Single image dehazing," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–9, 2008.

[5] R. T. Tan, "Visibility in bad weather from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[6] K. He, J. Sun, and X. Tang, "Single image haze removal using dark channel prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2341–2353, Dec. 2011.

[7] R. Fattal, "Dehazing using color-lines," *ACM Trans. Graph.*, vol. 34, no. 1, pp. 1–14, Dec. 2014.

[8] D. Berman, T. Treibitz, and S. Avidan, "Non-local image dehazing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1674–1682.

[9] Q. Guo, H.-M. Hu, and B. Li, "Haze and thin cloud removal using elliptical boundary prior for remote sensing image," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 9124–9137, Nov. 2019.

[10] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[11] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "DehazeNet: An end-to-end system for single image haze removal," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5187–5198, Nov. 2016.

[12] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single image dehazing via multi-scale convolutional neural networks," in *Proc. 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 154–169.

[13] X. Liu, Y. Ma, Z. Shi, and J. Chen, "GridDehazeNet: Attention-based multi-scale network for image dehazing," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7314–7323.

[14] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2157–2167.

[15] X. Qin, Z. Wang, Y. Bai, X. Xie, and H. Xie, "FFA-Net: Feature fusion attention network for single image dehazing," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2020, vol. 34, no. 7, pp. 11908–11915.

[16] D. Chen et al., "Gated context aggregation network for image dehazing and deraining," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jun. 2019, pp. 1375–1383.

[17] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[18] Y. Song, Z. He, H. Qian, and X. Du, "Vision transformers for single image dehazing," *IEEE Trans. Image Process.*, vol. 32, pp. 1927–1941, 2023.

[19] C. Guo, Q. Yan, S. Anwar, R. Cong, W. Ren, and C. Li, "Image dehazing transformer with transmission-aware 3D position embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5812–5820.

[20] A. Kulkarni and S. Murala, "Aerial image dehazing with attentive deformable transformers," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6305–6314.

[21] Y. Wang, J. Xiong, X. Yan, and M. Wei, "USCFormer: Unified transformer with semantically contrastive learning for image dehazing," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 10, pp. 11321–11333, Oct. 2023.

[22] K. Chi, Y. Yuan, and Q. Wang, "Trinity-net: Gradient-guided Swin transformer-based remote sensing image dehazing and beyond," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4702914.

[23] A. Kulkarni, S. S. Phutke, and S. Murala, "Unified transformer network for multi-weather image restoration," in *Proc. Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, 2022, pp. 344–360.

[24] T. Song, S. Fan, P. Li, J. Jin, G. Jin, and L. Fan, "Learning an effective transformer for remote sensing satellite image dehazing," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[25] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5728–5739.

[26] X. Dong et al., "CSWin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12124–12134.

[27] B. Huang, Z. Li, C. Yang, F. Sun, and Y. Song, "Single satellite optical imagery dehazing using SAR image prior based on conditional generative adversarial networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1806–1813.

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

[29] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[30] J. Wang, Z. Zheng, A. Ma, X. Lu, and Y. Zhong, "LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation," 2021, *arXiv:2110.08733*.

[31] J. Liu, S. Wang, X. Wang, M. Ju, and D. Zhang, "A review of remote sensing image dehazing," *Sensors*, vol. 21, no. 11, p. 3926, Jun. 2021.

[32] J. Kopf et al., "Deep photo: Model-based photograph enhancement and viewing," *ACM Trans. Graph.*, vol. 27, no. 5, pp. 1–10, 2008.

[33] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 6, pp. 1397–1409, Jun. 2013.

[34] H. Zhang and V. M. Patel, "Densely connected pyramid dehazing network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3194–3203.

[35] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10012–10022.

[36] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400.

[37] W. Shi et al., "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[38] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4700–4708.
[39] S. Masoudnia and R. Ebrahimpour, "Mixture of experts: A literature survey," *Artif. Intell. Rev.*, vol. 42, pp. 275–293, Aug. 2014.
[40] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[41] J. T. Barron, "A general and adaptive robust loss function," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4331–4339.
[42] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. 14th Eur. Conf.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 694–711.
[43] L. Li et al., "Semi-supervised image dehazing," *IEEE Trans. Image Process.*, vol. 29, pp. 2766–2779, 2020.
[44] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 902–911.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[46] Z. Gu, Z. Zhan, Q. Yuan, and L. Yan, "Single remote sensing image dehazing using a prior-based dense attentive network," *Remote Sens.*, vol. 11, no. 24, p. 3008, Dec. 2019.
[47] S. E. Kim, T. H. Park, and I. K. Eom, "Fast single image dehazing using saturation based transmission map estimation," *IEEE Trans. Image Process.*, vol. 29, pp. 1985–1998, 2020.
[48] Z. Zheng et al., "Ultra-high-definition image dehazing via multi-guided bilateral learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16180–16189.
[49] Z. Wang, X. Cun, J. Bao, W. Zhou, J. Liu, and H. Li, "Uformer: A general U-shaped transformer for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17683–17693.
[50] Y. Cui, W. Ren, X. Cao, and A. Knoll, "Focal network for image restoration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 13001–13011.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[52] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[53] J. Li, Z. Wu, Z. Hu, Z. Li, Y. Wang, and M. Molinier, "Deep learning based thin cloud removal fusing vegetation red edge and short wave infrared spectral information for Sentinel-2A imagery," *Remote Sens.*, vol. 13, no. 1, p. 157, Jan. 2021.
[54] W. Zhou, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[55] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[56] G. Sharma, W. Wu, and E. N. Dalal, "The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations," *Color Res. Appl., Endorsed Inter-Soc. Color Council, Colour Group (Great Britain), Can. Soc. Color, Color Sci. Assoc. Jpn., Dutch Soc. Study Color, Swedish Colour Centre Found., Colour Soc. Aust., Centre Français de la Couleur*, vol. 30, no. 1, pp. 21–30, Feb. 2005.
[57] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3974–3983.
[58] S. W. Zamir et al., "iSAID: A large-scale dataset for instance segmentation in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2019, pp. 28–37.
[59] W. Li, Y. Chen, K. Hu, and J. Zhu, "Oriented reppoints for aerial object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1829–1838.
[60] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5693–5703.
[61] K. Chen et al., "MMDetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
[62] M-Contributors. (2020). *MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark*. [Online]. Available: https://github.com/open-mmlab/mmsegmentation

Xiaozhe Zhang (Graduate Student Member, IEEE) received the B.E. degree in electronic and information engineering from Southwest Jiaotong University, Chengdu, China, in 2023. He is currently pursuing the master's degree with the Image Processing Center, School of Astronautics, Beihang University, Beijing, China.

His research interests include deep learning and computer vision.

Fengying Xie (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Beihang University, Beijing, China, in 2009.

From 2010 to 2011, she was a Visiting Scholar with the Laboratory for Image and Video Engineering, The University of Texas at Austin, Austin, TX, USA. She is currently a Professor with the Image Processing Center, School of Astronautics, Beihang University. Her research interests include biomedical image processing, remote sensing image understanding and applications, image quality assessment, and object recognition.

Haidong Ding received the B.E. degree from the Department of Image Processing Center, School of Astronautics, Beihang University, Beijing, China, in 2021, where he is currently pursuing the M.S. degree.

His research interests include image processing and deep learning.

Shaocheng Yan received the B.E. degree in electronic and information engineering from Southwest Jiaotong University, Chengdu, China, in 2023. He is currently pursuing the master's degree with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.

His research interests include image processing and point cloud registration.

Zhenwei Shi (Senior Member, IEEE) is currently a Professor and the Dean of the Image Processing Center, School of Astronautics, Beihang University, Beijing, China. He has authored or co-authored over 200 scientific articles in refereed journals and proceedings, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and the IEEE International Conference on Computer Vision (ICCV). His current research interests include remote sensing image processing and analysis, computer vision, pattern recognition, and machine learning.

Prof. Shi serves as an Editor for IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, *Pattern Recognition*, and *ISPRS Journal of Photogrammetry and Remote Sensing, Infrared Physics and Technology*. His personal website is http://levir.buaa.edu.cn/.