



Article

Remote Sensing Image Dehazing via a Local Context-Enriched Transformer

Jing Nie ^{1,*} , Jin Xie ^{2,3}, and Hanqing Sun ⁴ ¹ School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China² School of Big Data and Software Engineering, Chongqing University, Chongqing 400044, China; xiejin@cqu.edu.cn³ Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China⁴ Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China; sunhanqing@ciomp.ac.cn

* Correspondence: jingnie@cqu.edu.cn

Abstract: Remote sensing image dehazing is a well-known remote sensing image processing task focused on restoring clean images from hazy images. The Transformer network, based on the self-attention mechanism, has demonstrated remarkable advantages in various image restoration tasks, due to its capacity to capture long-range dependencies within images. However, it is weak at modeling local context. Conversely, convolutional neural networks (CNNs) are adept at capturing local contextual information. Local contextual information could provide more details, while long-range dependencies capture global structure information. The combination of long-range dependencies and local context modeling is beneficial for remote sensing image dehazing. Therefore, in this paper, we propose a CNN-based adaptive local context enrichment module (ALCEM) to extract contextual information within local regions. Subsequently, we integrate our proposed ALCEM into the multi-head self-attention and feed-forward network of the Transformer, constructing a novel locally enhanced attention (LEA) and a local continuous-enhancement feed-forward network (LCFN). The LEA utilizes the ALCEM to inject local context information that is complementary to the long-range relationship modeled by multi-head self-attention, which is beneficial to removing haze and restoring details. The LCFN extracts multi-scale spatial information and selectively fuses them by the the ALCEM, which supplements more informative information compared with existing regular feed-forward networks with only position-specific information flow. Powered by the LEA and LCFN, a novel Transformer-based dehazing network termed LCEFormer is proposed to restore clear images from hazy remote sensing images, which combines the advantages of CNN and Transformer. Experiments conducted on three distinct datasets, namely DHID, ERICE, and RSID, demonstrate that our proposed LCEFormer achieves the state-of-the-art performance in hazy scenes. Specifically, our LCEFormer outperforms DCIL by 0.78 dB and 0.018 for PSNR and SSIM on the DHID dataset.

Keywords: remote sensing image dehazing; transformer; local context enrichment

Citation: Nie, J.; Xie, J.; Sun, H. Remote Sensing Image Dehazing via a Local Context-Enriched Transformer. *Remote Sens.* **2024**, *16*, 1422. <https://doi.org/10.3390/rs16081422>

Academic Editors: Sidike Paheding and Ashraf Saleem

Received: 25 February 2024

Revised: 5 April 2024

Accepted: 13 April 2024

Published: 17 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Remote sensing images captured by the satellite or unmanned aerial vehicle are degraded by the existing haze or cloud [1–4], which destroys the surface information acquisition and further degrades the downstream tasks including image classification [5–7], object detection [8–10], change detection [11,12], object tracking [13,14], image segmentation [15,16], and so on. Remote image dehazing methods are to recover the clean image from its haze or cloud-polluted variants, which could be applied in applications with environment monitoring, military surveillance, and so on.

Image dehazing methods roughly are divided into prior-based methods [17–20] and deep learning-based methods [21–24] based on whether utilizing deep learning structures.

Prior-based dehazing methods apply various prior constraints to predict the parameters of the atmosphere scattering model [25], and then restore clear images based on the physical model. In recent years, deep learning-based image dehazing approaches have shown significant progress in this area. Previous deep learning-based image dehazing methods [21,22] conduct the dehazing process according to the atmosphere scattering model and performs not well in real hazy scenes because the physical model does not fit the actual hazy scenes. Therefore, end-to-end dehazing methods [23,24] are proposed. The success of these deep learning-based approaches can largely be attributed to their capabilities to generate discriminative features through a series of convolution operations. Since the convolution is a local operation, it does not explicitly capture the global structural information. To address this, a few recent works [26] explore the multi-scale pooling module to integrate the global structural information for image dehazing, which can be effectively utilized to model the long-range dependencies. FFANet [27] utilizes channel attention and BidNet [28] models non-local relationships to introduce global information to improve dehazing performance. Recently, several vision transformers-based approaches [29,30] have pervaded different areas of computer vision by utilizing a self-attention mechanism [31] that captures the long-range dependencies to model the global structural information in an image. The transformer-based dehazing methods [32,33] have achieved great success. However, local contextual information is of great importance in the image dehazing process. Researchers attempt to combine the advantages of CNN and transformer to conduct image dehazing. CloudFormer [34] cascades convolution blocks and transformer blocks to extract shallow and deep features, respectively, to remove the cloud. Dehamer [35] utilizes the features extracted from the Transformer to modulate the CNN features, which is superior to the fusion manner such as addition or concatenation.

We argue that both the local contextual information as well as the global structural information are desired for accurate image dehazing. As discussed above, the convolution operation captures the fine details by focusing on the local contextual information within a local region, whereas self-attention strives to model the long-range dependencies for capturing the global information. A straightforward way to obtain the benefits of convolution and self-attention, in a single image dehazing architecture, is to aggregate the features from convolution and self-attention by an element-wise summation. However, such a straightforward fusion strategy also introduces redundancy and noise during the aggregation. Therefore, we look into an alternative approach to effectively fuse the complementary local and global structural information for remote sensing image dehazing. In this paper, we design a U-shape transformer architecture composed of stacked enhanced transformer blocks called local context-enriched transformer blocks (LCTBs). The proposed LCTB contains a locally enhanced attention (LEA) and a local continuous-enhancement feed-forward network (LCFN), which embeds a local detail enrichment module (LEDM) into the classical attention module and the feed-forward network, respectively. The LEDM extracts multi-scale local contextual information and selectively fuses them. The LEDM is utilized to model local-range attention in the LEA. Besides, the LEDM supplements local contextual information in the feed-forward network.

The contributions of this paper could be summarized as follows:

- We propose a novel transformer-based U-shape remote sensing dehazing network, namely Local Context-Enriched Transformer (LCEFormer). LCEFormer stacks local context-enriched transformer blocks (LCTBs), each comprising a locally enhanced attention (LEA) and a local continuous enhancement feed-forward network (LCFN). Both LEA and LCFN are equipped with an adaptive local context enrichment module (ALCEM) that extracts multi-scale local contextually enriched features and fuses them selectively.
- Different from the common self-attention module, the LEA module employs the ALCEM to extract more informative local context, thus enhancing the discriminative power of the query, key, and value vectors used for computing multi-head attention, which helps in effectively removing haze from the input image, resulting in cleaner re-

sults. In contrast to regular feed-forward networks that only perform position-specific information flow, our LCFN enriches multi-scale local context. This enhancement proves beneficial in refining regions by leveraging neighborhood information inference, resulting in cleaner outputs.

- We validate the effectiveness of the proposed LCEFormer by conducting comprehensive experiments on three remote image dehazing benchmarks: DHID [36], ERICE [37], and RSID [38]. Our LCEFormer outperforms existing image dehazing methods on both benchmarks. Additionally, to demonstrate the scalability of the proposed LCEFormer, experiments on the UCMERGED dataset [39] demonstrate that our LCEFormer achieves the state-of-the-art performance in the remote sensing image super-resolution task.

2. Related Work

In this section, we first review related dehazing methods. Because our proposed method is based on the vision transformer architecture, we introduce recent advancements in vision transformer methods.

2.1. Image Dehazing Methods

With the development of deep learning, dehazing methods based on CNN [22–24,28,40] and transformer [33,41] overwhelm the traditional dehazing methods based on priors [17,18]. GridDehaze [23] directly learns the clear counterpart from the hazy input by an attention-based grid network. FFANet [27] adaptively fuses features according to the channel and pixel attention. MSBDN [24] utilizes a dense feature fusion module based on CNN and a boosting strategy to excavate spatial information to achieve high dehazing performance. PFDN [42] introduces the atmospherical scattering model-based dehazing network and removes the haze in the feature space. CNN has the limitation of capturing long-range dependencies. To overcome the shortcoming of CNN, the dehazing method Uformer [41] is proposed and utilizes a Locally enhanced Window transformer block to extract context information from multi-scale features, which not only reduces computation but also achieves the state-of-the-art performance in various image restoration tasks. DehazeFormer [33] makes several improvements on the elements of SwinTransformer [30] for the dehazing task, which achieves the best performance on both the homogeneous image dehazing dataset and the non-homogeneous remote sensing image dehazing dataset. Dehamer [35] employs the DCP prior [17] into the Transformer position embedding and fuses the CNN features and the Transformer features by a modulation module. Recently, the diffusion module has been widely adopted in image restoration tasks for its generative power. Wang et al. [43] proposed a frequency compensation block to facilitate the diffusion model to restore high-frequency details.

The dehazing methods proposed to recover natural scene images could be utilized to remove the haze in remote sensing images. However, the remote sensing images have different visual angles and different scene depths. Zhang et al. [36] proposed a dynamic collaborative inference learning (DCIL) framework to remove dense haze that existed in remote sensing images, which efficiently restores the texture details, spectral characteristics, and small-scale objects. Trinity-Net [38] employs Swin Transformer to estimate the parameters of the physical model and introduces the gradient maps to enhance the detail information for the Transformer features, which obtains great performance in the remote sensing image dehazing task. AMGAN-CR [44] embeds the attention into the generative adversarial networks to remove thin clouds. McGAN [45] fuses the information of RGB and multi-spectral images to improve the cloud removal performance. Rice dataset [37] is collected by Lin et al., and is a remote sensing cloud removal dataset. Tao et al. [46] proposed to use a self-paced learning mechanism to train the cloud removal network across easier to harder difficulty levels. CloudFormer [34] combines the advantages of CNN and transformer to remove the cloud, which cascades convolution blocks and transformer blocks to extract shallow and deep features, respectively.

2.2. Vision Transformer

Transformer [31] is originally proposed to be applied in the natural language processing task due to its advantage of capturing long-range dependencies. Alexey et al. [29] proposed Vision Transformer to process image patches in sequence, based on which a series of Transformers [30,47,48] are developed for image processing tasks including recognition, object detection, and image segmentation. Although vision Transformers have achieved significant success in image dehazing, image deraining, and so on, Transformer-based image restoration methods need high computational costs when the input images are of high resolution. Besides, Transformer can model long-range dependencies while lacking local detail information. To enhance the transformer-based dehazing method with the patch-level feature, patch-level attention is proposed in [49]. To combine the advantages of CNN and Transformer, Dehamer [35] utilizes Transformer features to learn the modulation matrices to modulate CNN features. In contrast, we propose an adaptive local context enrichment module (ALCEM) that extracts multi-scale local contextually-enriched features and fuses them selectively. Syed et al. [50] proposed an efficient image restoration Transformer termed Restormer that efficiently computes self-attention in channel dimension in a multi-Dconv head transposed attention (MDTA) module. In contrast, we design an adaptive local context enrichment module (ALCEM) to extract multi-scale local context. The ALCEM strengthens both the attention module, referred to as the locally enhanced attention (LEA) module, and the feed-forward network, known as the local continuous-enhancement feed-forward network (LCFN). The proposed method contributes to improved dehazing performance.

3. Method

In this section, we first introduce the overall pipeline of the proposed LCEFormer for remote sensing image dehazing. Then, we describe the basic block of the LCEFormer, i.e., local context-enriched transformer block (LCTB) with a locally enhanced attention module and local continuous-enhancement feed-forward network. Finally, we present the loss function for the entire framework.

3.1. Overall Pipeline

Figure 1 shows the overall architecture of the proposed U-shape transformer-based dehazing framework, denoted as LCEFormer. Firstly, a hazy remote sensing image $I \in \mathbb{R}^{H \times W \times 3}$ undergoes a 3×3 convolution operation to generate intermediate features with a channel number of C . These intermediate features then go through four-level encoders, each consisting of a specified number of stacked LCTBs N_i , where $i \in 0, 1, 2, 3$, to extract multi-scale features. Between two-level encoders, a 3×3 convolution operation followed by a pixel-unshuffle operation is utilized to down-sample the features to reduce their size to half of the original feature size. Subsequently, decoders comprising N_i stacked LCTBs ($i \in 2, 1, 0$). Between two-level decoders, a 3×3 convolution operation followed by a pixel-shuffle operation is utilized to recover high-resolution features. The addition of features with the same feature size in both the encoder and its corresponding decoder is conducted. Finally, the features outputted by the last LCTB undergo a 3×3 convolution to restore the haze-free image.

3.2. Local Context-Enriched Transformer Block

Rich local context information can help restore image details, while global structural information can provide clues for removing haze from the entire image. In this paper, we propose a local context-enriched transformer block (LCTB) that can simultaneously extract rich local context information and global structural information, ensuring clear image details while removing haze.

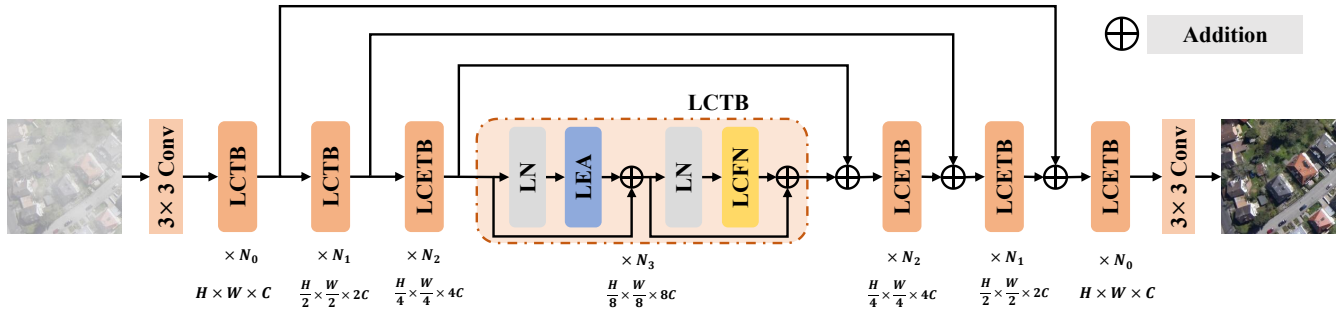


Figure 1. The overall architecture of the proposed LCEFormer. The LCEFormer is a U-shape architecture consisting of stacked of local context-enriched transformer blocks (LCTBs). The key components of LCTB are a locally enhanced attention (LEA) and a local continuous-enhancement feed-forward network (LCFN).

To enhance the capability of extracting local detail information within the Transformer block, we introduce an adaptive local context enrichment module (ALCEM) designed to extract local contextual information. This module is seamlessly integrated into two key components of the Transformer block: the multi-head attention module and the feed-forward network. The modified modules are referred to as locally enhanced attention (LEA) and Local Continuous-Enhancement Feed-Forward Network (LCFN).

Next, we will first provide a detailed description of our ALCEM, followed by introductions to LEA and LCFN.

Adaptive Local Context Enrichment Module: To incorporate enriched and broader local context information into the self-attention module and feed-forward network of Transformer, we propose the adaptive local context enrichment module (ALCEM). This module is designed to capture intricate details and local contextual information effectively. This proves to be advantageous in the restoration of local details by utilizing information from the neighborhood information. As shown in Figure 2, the input feature and the output feature of the ALCEM are denoted as x and y , respectively. The computational process of the ALCEM is formulated as follows:

$$\begin{aligned}
 \{x_1, x_2, x_3, x_4\} &= \mathbf{S}(x), \\
 y_1 &= \sigma_l(\mathbf{Conv}_{3 \times 3}(x_1)), \\
 y_2 &= \sigma_l(\mathbf{Conv}_{3 \times 3}(x_2) + y_1), \\
 y_3 &= \sigma_l(\mathbf{Conv}_{3 \times 3}(x_3) + y_3), \\
 y_4 &= \sigma_l(\mathbf{Conv}_{3 \times 3}(x_4) + y_3), \\
 y &= \sum_{i=1}^4 w_i \odot y_i,
 \end{aligned} \tag{1}$$

where $\mathbf{S}(\cdot)$ denotes a splitting operation that partitions the input feature maps into four features maps along the channel axis, σ_l represents the Leaky ReLU activation function, and $\mathbf{Conv}_{3 \times 3}$ represents a 2D convolution operation with a kernel size of 3×3 . $\odot$ denotes the element-wise multiplication, and $\mathbf{Conv}_{1 \times 1}$ represents a 2D convolution operation with a kernel size of 1×1 .

In addition, w_i can be expressed as follows:

$$\begin{aligned}
 y_c &= \mathbf{Concat}(y_1, y_2, y_3, y_4), \\
 l_i &= \mathbf{Reshape}(\mathbf{Conv1D}_3(\mathbf{GAP}(y_c))), \\
 w_i &= \frac{e^{l_i}}{\sum_{i=1}^4 e^{l_i}},
 \end{aligned} \tag{2}$$

where σ_r denotes the ReLU activation function, and $\mathbf{GAP}$ represents global average pooling operation. $\mathbf{Conv1D}_3$ is a one-dimension convolution with a kernel size of 3.

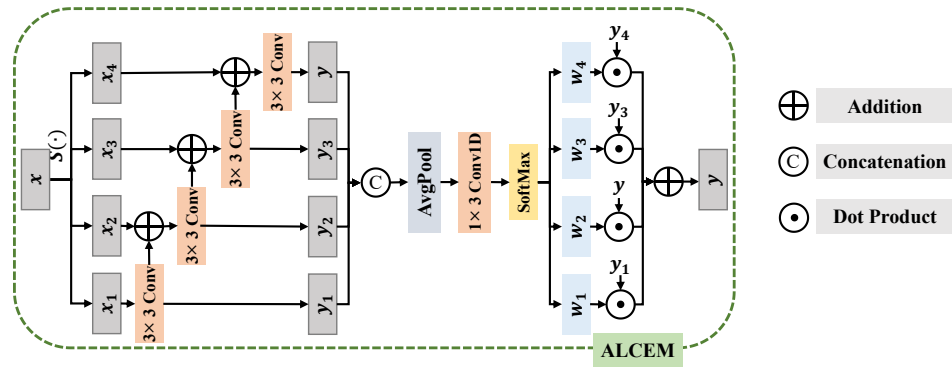


Figure 2. The structures of the proposed adaptive local context enrichment module (ALCEM).

Locally Enhanced Attention Module: The regular multi-head self-attention module is designed to capture long-range dependencies. We provide an overview of the computation process of the self-attention module. The first step in self-attention involves computing the query, key, and value vectors from the input features by 1×1 convolutions. Subsequently, attention weights are calculated using the query and key vectors. These weights are then used to compute a weighted sum of the values to obtain the output. It can be observed that the quality of output features depends on the robustness and discriminative capabilities of the query, key, and value vectors. However, existing methods typically utilize 1×1 convolution operations or 3×3 depth-wise convolution operations, which often lack enriched and broader local information. To address this issue, we propose the Locally Enhanced Attention (LEA) module, which integrates our ALCEM into the self-attention module to enhance the local contextual information of the query, key, and value vectors. This enhancement proves beneficial for recovering local details by inferring from neighborhood information. The structure of the proposed LEA module is shown in Figure 3. Next, we will detail the computational process of our LEA module. The LEA module first employs our proposed ALCEM, as described in the above section, to enrich the local context. Subsequently, it generates query (Q), key (K), and value (V) vectors. To reduce computational complexity, following the approach outlined in [50], we compute the channel-aware attention weights. Finally, similar to the regular attention module, we compute the output by employing the attention weights to perform matrix multiplication of the value vectors.

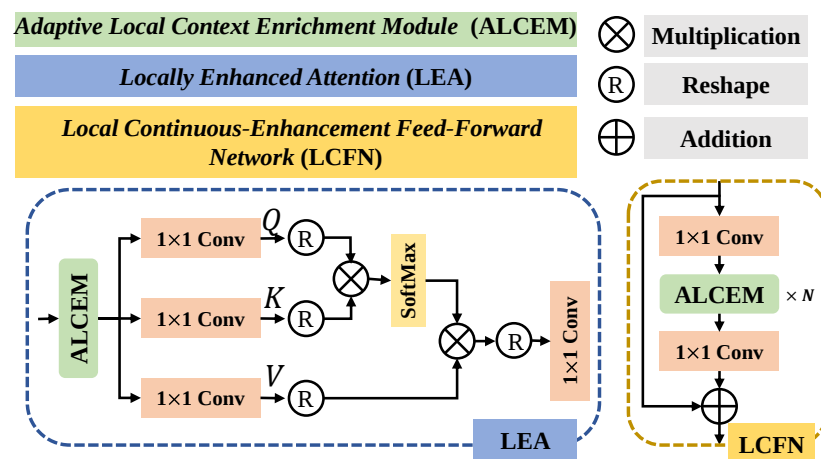


Figure 3. The structures of the proposed locally enhanced attention (LEA) module, and Local Continuous-Enhancement Feed-Forward Network (LCFN).

Given an input tensor $x \in \mathbb{R}^{H \times W \times C}$, The specific process could be formulated as

$$\begin{aligned}
 x_l &= \mathbf{ALCEM}(x), \\
 Q &= \mathbf{Conv}_{1 \times 1}(x_l), \\
 K &= \mathbf{Conv}_{1 \times 1}(x_l), \\
 V &= \mathbf{Conv}_{1 \times 1}(x_l), \\
 A &= \mathbf{Softmax}(\mathbf{R}(K) \otimes \mathbf{R}(Q)), \\
 u &= \mathbf{Conv}_{1 \times 1}(\mathbf{R}(\mathbf{R}(V) \otimes A)),
 \end{aligned} \tag{3}$$

where $\otimes$ represents the matrix multiplication, $\mathbf{Conv}_{1 \times 1}$ denotes the convolution layer with a kernel size of 1×1 , $\mathbf{Softmax}$ denotes the softmax operation, and $\mathbf{R}(\cdot)$ denotes the reshape operation. $x_l \in \mathbb{R}^{H \times W \times C}$ represents the feature vector with enriched local context. Three parallel 1×1 convolution layers are then utilized to compute the query $Q \in \mathbb{R}^{H \times W \times C}$, key $K \in \mathbb{R}^{H \times W \times C}$, and value $V \in \mathbb{R}^{H \times W \times C}$ vectors. The query and key vectors are firstly reshaped and then utilized to compute the attention weights $A = \mathbf{Softmax}(\mathbf{R}(K) \otimes \mathbf{R}(Q)) \in \mathbb{R}^{C \times C}$, where the sizes of reshaped query and key are $HW \times C$ and $C \times HW$, respectively. These attention weights are subsequently employed to perform matrix multiplication to the reshaped value $\mathbf{R}(V) \in \mathbb{R}^{HW \times C}$. $u \in \mathbb{R}^{H \times W \times C}$ is the output features containing enriched local context and long-range information.

Local Continuous-Enhancement Feed-Forward Network: Existing feed-forward networks [31] only learn complex interactions between different features within each position of an image, which ignores local context information that is important for image dehazing. Therefore, a local continuous-enhancement feed-forward network (LCFN) is designed as shown in Figure 3. Our LCFN consists of two 1×1 convolution layers and an adaptive local context enrichment module (ALCEM), the 1×1 convolution layer before ALCEM is utilized to expand the feature channels (usually by factor $\gamma = 1$ in LEA and $\gamma = 4$ in LCFN) and the 1×1 convolution layer after ALCEM to reduce channels back to the original input dimension. The detailed computation can be described as

$$z = \mathbf{Conv}_{1 \times 1}(\mathbf{ALCEM}(\mathbf{Conv}_{1 \times 1}(x))) + x \tag{4}$$

3.3. Loss Function

We train the proposed LCEFormer in an end-to-end way with the Charbonnier loss [51] and the SSIM loss. The loss function is formulated as

$$\mathcal{L} = \alpha \mathcal{L}_{ssim} + \beta \mathcal{L}_{char}, \tag{5}$$

where $\mathcal{L}_{ssim}$ and $\mathcal{L}_{char}$ demote the SSIM loss and the Charbonnier loss, respectively. α and β are the balance factors, which are both set to 1 for all experiments.

The SSIM loss $\mathcal{L}_{ssim}$ can be computed as

$$\mathcal{L}_{ssim} = -\mathbf{SSIM}(\hat{J}, J) \tag{6}$$

where $\hat{J}$ represents the predicted haze-free images predicted by our LCEFormer, J denotes the ground-truth clear image, and $\mathbf{SSIM}$ represents compute the Structural Similarity Index (SSIM) value using Equation (9).

The Charbonnier loss $\mathcal{L}_{char}$ is computed as

$$\mathcal{L}_{char} = \sqrt{\|\hat{J} - J\|^2 + \varepsilon^2}, \tag{7}$$

where the constant ε is set to 10^{-3} .

4. Experiments

4.1. Datasets and Evaluation Metrics

We perform experiments on three remote image dehazing datasets: DHID [36], ERICE [37], and RSID [38] datasets. The detailed distributions of all datasets are presented in Table 1. Figure 4 shows the example images of all datasets.

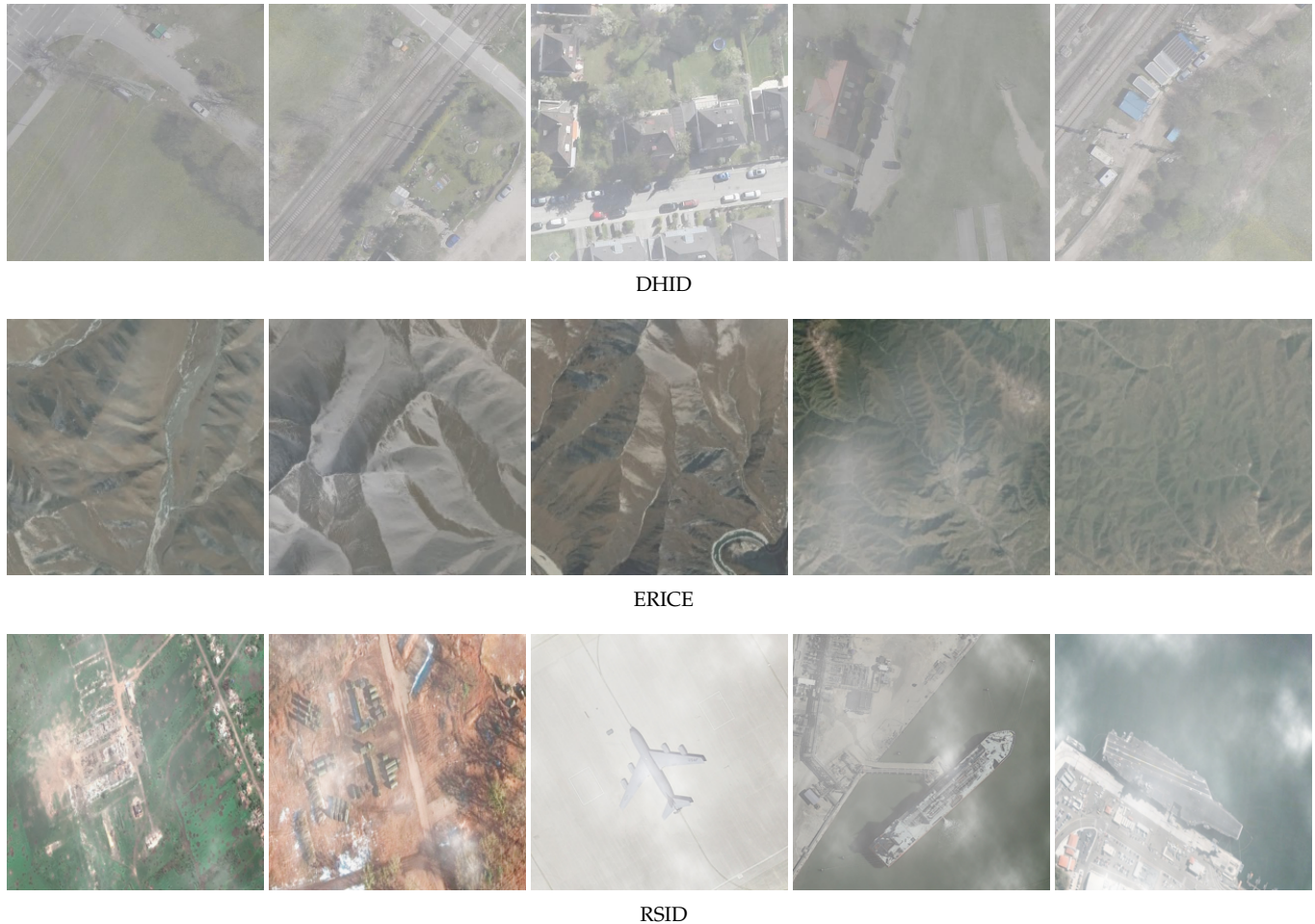


Figure 4. Example images from the remote sensing image dehazing datasets DHID, ERICE, and RSID.

Table 1. Detail of the training and test datasets. # represents the number of images.

Datasets	# of Train	# of Test	Resolution
DHID [36]	14,490	500	512×512
ERICE [37]	1600	400	256×256
RSID [38]	1760	100	$256 \times 256, 512 \times 512$

DHID dataset: The DHID dataset is a dense hazy remote sensing dataset with 14,990 images, in which the training set has 14,490 images and the test set has 500 images. The resolution of the images is 512×512 .

ERICE dataset: The ERICE dataset is a hazy remote sensing dataset derived from the RICE dataset [37] by cropping its images. It comprises 1600 images in the training set and 400 images in the test set.

RSID dataset: The RSID dataset is a real-world remote sensing image dehazing benchmark. To ensure a fair comparison with other methods, we follow the dataset processing method outlined in the paper introducing the dataset [38]. Specifically, both

900 images from the RSID dataset and 860 images from the SateHaze1k [52] dataset are utilized for training. Furthermore, 100 images from the RSID dataset are used for testing.

Evaluation Metrics: In all experiments, we report performance by utilizing the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Next, we describe the computational process of PSNR and SSIM.

The Peak Signal-to-Noise Ratio (PSNR) can measure the difference between two images. Computing the PSNR values between predicted images with ground-truth images serves as a common metric for evaluating the performance of image dehazing methods. A higher PSNR value indicates superior image dehazing performance. The PSNR can be formulated as follows:

$$\text{PSNR}(\hat{J}, J) = 10 \cdot \log_{10} \left(\frac{1}{\frac{1}{HW} \sum_h^H \sum_w^W \hat{J}(h, w) - J(h, w)} \right), \quad (8)$$

where H and W are the dimensions of the images, $\hat{J}(h, w)$ and $J(h, w)$ represent the pixel values at position (h, w) in predicted images $\hat{J}$ and the ground-truth clear images J , respectively. It can be noted that the values of $\hat{J}$ and J range from 0 to 1.

The Structural Similarity Index (SSIM) can measure the similarity between two images. Computing the SSIM between predicted images and ground-truth images is a widely used evaluation metric in image dehazing tasks. A higher SSIM value denotes superior image dehazing performance. The formulation of the Structural Similarity Index (SSIM) is as follows:

$$\text{SSIM}(\hat{J}, J) = l(\hat{J}, J)c(\hat{J}, J)s(\hat{J}, J), \quad (9)$$

where $l(\hat{J}, J) = \frac{(2\mu_{\hat{J}}\mu_J + c_1)}{(\mu_{\hat{J}}^2 + \mu_J^2 + c_1)}$ represents the luminance similarity, $c(\hat{J}, J) = \frac{(2\sigma_{\hat{J}}\sigma_J + c_2)}{(\sigma_{\hat{J}}^2 + \sigma_J^2 + c_2)}$ is the contrast similarity, $s(\hat{J}, J) = \frac{(\sigma_{\hat{J}J} + c_3)}{(\sigma_{\hat{J}}\sigma_J + c_3)}$ represents the structure similarity, $\mu_{\hat{J}}$ and μ_J denote the average values of the predicted image $\hat{J}$ and the ground-truth image J , respectively. $\sigma_{\hat{J}}$ and σ_J denote the standard deviation of the values in the predicted image $\hat{J}$ and the ground-truth image J , respectively. $\sigma_{\hat{J}J}$ is the covariance between predicted image $\hat{J}$ and the ground-truth image J . The constants c_1 , c_2 , and c_3 are utilized to avoid division by zero and to scale the SSIM value to the range from 0 to 1. In general, we set c_1 , c_2 and c_3 to 0.01^2 , 0.03^2 , and $\frac{c_2}{2}$, respectively. Similar to computing PSNR, the values of $\hat{J}$ and J are normalized to range from 0 to 1.

4.2. Training Details

Our proposed method is implemented using MMagic [53] (<https://github.com/open-mmlab/mmagic/>, accessed on 24 February 2024), an open-source image and video editing toolbox based on PyTorch [54].

For both datasets, our proposed methodology entails end-to-end training conducted on two NVIDIA RTX 3090 GPUs, with a mini-batch size of 16 images per GPU. To enrich the diversity of the training images, we utilize various data augmentation methods, specifically employing random rotations of the images within intervals of 45° . The training process is facilitated by the AdamW optimizer [55], where weight decay is set to 10^{-3} , and the exponential decay rates for the first and second moments are both set to 0.9. The initial learning rate is set to 1×10^{-3} , and is subsequently decayed to 1×10^{-7} employing the cosine annealing strategy [56]. This process can be formulated as

$$l_t = l_{\min} + \frac{1}{2}(l_{\text{init}} - l_{\min}) \left(1 + \cos \left(\frac{t}{T} \pi \right) \right) \quad (10)$$

where t is the current iteration number, T is the total number of iterations, l_t represents the learning rate at iteration t , $l_{\min}$ and l_{init} are the minimum and initial learning rates, respectively.

For the 4-level encoder stage of our LCEFormer, We set the number of LCTBs as follows: $N_0 = 1, N_1 = 2, N_2 = 2, N_3 = 2$. Attention heads in LEA are $[1, 2, 4, 8]$, and the number of channels are $[24, 48, 96, 192]$. The decoder stage is symmetric to the encoder stage.

Next, we describe the experimental settings specific to the DHID, ERICE, and RSID.

DHID: The model is trained for 50,000 iterations. The DHID dataset comprises images with a resolution of 512×512 pixels; therefore, the input images are randomly cropped from the original images to a resolution of 256×256 pixels during training.

ERICE: The model is trained for 30,000 iterations. Image resizing or random cropping is not utilized during training.

RSID: The model is trained for 30,000 iterations. The training dataset comprises 2 distinct resolutions: 512×512 and 256×256 pixels. For images with a resolution of 256×256 , neither image resizing nor random cropping is employed during the training process. For images with a resolution of 512×512 , input images are randomly cropped from the original images to a resolution of 256×256 pixels during training.

4.3. Experimental Results

4.3.1. Results on the DHID Dataset

Our LCEFormer is compared to the recent state-of-the-art methods, namely Y-Net [57], FCTF-Net [58], AFDN [59], Dehamer [35], DehazeFormer [33], and DCIL [36]. Table 2 presents the results. In the case of Y-Net, FCTF-Net, AFDN, and DCIL, the results are taken from [36]. Furthermore, for a fair comparison, we evaluate the performance of Dehamer and DehazeFormer by utilizing the publicly available code provided by their respective authors. Our proposed LCEFormer significantly outperforms other methods in terms of PSNR and SSIM, demonstrating the effectiveness and superiority of our LCEFormer. In detail, Y-Net, FCTF-Net, AFDN, and DCIL are CNN-based methods, among these methods, DCIL achieves the highest PSNR and SSIM with 28.18 dB in PSNR and 0.892 in SSIM. Compared with DCIL, our method increases the PSNR score by 0.78 dB and SSIM score by 0.018, demonstrating that our method achieves better results compared to the CNN-based methods. DehazeFormer is a popular Transformer-based image dehazing method, achieving 26.29 dB in PSNR and 0.889 in SSIM. Compared with DehazeFormer, our LCEFormer obtains better dehazing performance with 28.96 dB in PSNR and 0.910 in SSIM, demonstrating superior performance than Transformer-based methods. Dehamer, combining CNN and Transformer, achieves 26.19 dB in PSNR and 0.886 in SSIM. Our LCEFormer outperforms Dehamer by 2.77 dB in PSNR and 0.024 in SSIM, showcasing that our LCEFormer combines CNN and Transformer more effectively than Dehamer.

Table 2. The experimental results on the DHID dataset. Bold numbers represent the best performance, while underlined numbers indicate the second best.

Type	Method	PSNR	SSIM
CNN-based	Y-Net [57]	18.31	0.783
	FCTF-Net [58]	18.77	0.794
	AFDN [59]	20.03	0.803
	DCIL [36]	<u>28.18</u>	<u>0.892</u>
Transformer	DehazeFormer [33]	26.29	0.889
CNN + Transformer	Dehamer [35]	26.19	0.886
	LCEFormer	28.96	0.910

Comparison with other multi-head attention modules: To verify the effectiveness of the proposed LEA, we conduct experiments to compare our proposed LEA with other multi-head attention modules including spatial-reduction attention modules (SRA) [47] and multi-dconv head transposed attention modules (MDTA) [50]. The results are reported in Table 3. For fair comparisons, except for the multi-head attention modules, we employ the same experimental settings. It can be noted that our proposed LCFN is selected as the

feed-forward network for all experiments in Table 3. It can be observed that compared with SRA, our proposed L2RA obtains an improvement of 0.57 dB in the PSNR value and 0.003 in the SSIM value. In addition, our L2RA outperforms MDTA by 0.21 dB in terms of PSNR. The remarkable improvement demonstrates the effectiveness of our proposed LEA. Here, we analyze the reasons for the superior performance of our LEA as follows: SRA reduces the computational complexity of the traditional multi-head attention module by reducing the spatial dimension of the key and value. However, the spatial reduction results in a loss of detailed information, thereby weakening its ability to extract local contextual information. MDTA employs self-attention across channel dimensions instead of the spatial dimension. Additionally, it utilizes depth-wise convolutions to encode information from neighboring pixel positions in the spatial dimension. Nevertheless, a depth-wise convolution layer can only extract features within a fixed, small range, and its receptive field is insufficient for capturing enriched contextual information. In contrast to SRA and MDTA, our L2RA employs the adaptive local context enrichment module (ALCEM) to extract enriched local contextual information during computing multi-head attention, which is useful for learning local image structure for effective image dehazing.

Computational Complexity Analysis: The Floating Point Operations per Second (FLOPS) of the model utilizing our proposed LEA is 25.08 G, whereas the model employing SRA [47] incurs 42.31 G FLOPS. Notably, our model enhances PSNR from 28.39 dB to 28.96 dB while reducing computational costs. Additionally, the FLOPS of the model employing MDTA [50] stands at 24.43 G. Despite our proposed method having slightly higher computational complexity than MDTA, the model utilizing LEA outperforms the one employing MDTA with a significant gain of 0.21 dB in PSNR.

Table 3. Comparison (in PSNR and SSIM) with different multi-head attention methods on DHID *test* set. Bold numbers represent the best performance.

Methods	PSNR	SSIM
SRA [47]	28.39	0.907
MDTA [50]	28.75	0.910
LEA	28.96	0.910

Comparison with other feed-forward networks: To verify the effectiveness of the proposed LCFN, we conduct experiments to compare our proposed LCFN with other feed-forward networks including the regular feed-forward network (FN) [31], locally-enhanced feed-forward network (LeFF) [32], and gated-dconv feed-forward network (GDFN) [50]. The results are reported in Table 4. For fair comparisons, except for the feed-forward networks, we employ the same experimental settings. It can be noted that our proposed LEA is selected as the multi-head attention module for all experiments in Table 4. From Table 4, it can be observed that compared with the existing feed-forward network, our LCFN obtains higher PSNR and SSIM scores, demonstrating the superiority and effectiveness of our approach. In contrast to the regular feed-forward network (FN) [31], our LCFN enriches multi-scale local context with the ALCEM. As presented in Table 4, our LCFN outperforms FN by 2.1 dB in terms of PSNR and 0.012 in terms of SSIM. LeFF [32], integrating a depth-wise convolutional layer into the regular feed-forward network to enhance single-scale local context, obtains 28.01 dB and 0.905 in terms of PSNR and SSIM, respectively. Comparatively, our LCFN obtains an improvement of 0.95 dB in PSNR and 0.005 in SSIM over LeFF. This experiment demonstrates the effectiveness of multi-scale local context.

Table 4. Comparison (in PSNR and SSIM) with different feed-forward networks on DHID *test* set. Bold numbers represent the best performance.

Methods	PSNR	SSIM
FN [31]	26.86	0.898
LeFF [32]	28.01	0.905
GDFN [50]	28.17	0.906
LCFN	28.96	0.910

Additionally, GDFN [50], which integrates a gated mechanism and depth-wise convolutions into the regular feed-forward network to enrich local information, achieves PSNR and SSIM scores of 28.17 dB and 0.906, respectively. Compared with GDFN, our LCFN achieves higher dehazing performance with PSNR and SSIM scores of 28.96 and 0.910, respectively. Both LeFF and GDFN rely on 3×3 depth-wise convolution operations to extract local information, leading that the extracted features are from a small fixed region. In contrast, our LCFN can extract local features across a wider and more varied range while selectively aggregating local information from different receptive fields. Meanwhile, the enriched information proves to be effective in preserving the desired local continuity in image dehazing tasks. Therefore, our LCFN is theoretically superior to LeFF and GDFN, and the experimental results in Table 4 further validate this assertion.

4.3.2. Results on the ERICE Dataset

To validate the effectiveness and generality of our proposed LCEFormer method, we conduct a comprehensive experimental comparison with several state-of-the-art techniques: GridDehaze [23], Uformer [32], Dehamer [35], and DehazeFormer [33] on the ERICE dataset. The evaluation metrics are the same as those used in the experiments on the DHID dataset, namely Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). The results are presented in Table 5. To guarantee a comprehensive and reliable performance comparison, we train all of the existing methods in Table 5 using the official code and experimental settings provided by the authors. Additionally, all methods employ the same training datasets. It can be observed that our LCEFormer achieves the highest PSNR and SSIM values, demonstrating the superiority of our approach. Specifically, compared with the CNN-based method GridDehaze, our LCEFormer outperforms it by 2.2 dB in terms of PSNR and 0.011 in terms of SSIM. Among Transformer-based methods, Uformer achieves 34.14 dB and 0.954 in PSNR and SSIM, respectively, while Dehazeformer achieves 36.49 dB and 0.958 in PSNR and SSIM, respectively. In contrast, our LCEFormer achieves better dehazing performance with 37.23 dB and 0.965 in PSNR and SSIM respectively. Furthermore, compared with Dehamer, which combines CNN and Transformer, our LCEFormer increases PSNR from 33.43 dB to 37.23 dB and SSIM from 0.953 to 0.965, demonstrating that our approach effectively combines the advantages of both the CNN and Transformer.

Table 5. The state-of-the-art comparison on the ERICE dataset. Bold numbers represent the best performance, while underlined numbers indicate the second best.

Type	Method	PSNR	SSIM
CNN	GridDehaze [23]	35.03	0.954
Transformer	Uformer [41]	34.14	0.954
	DehazeFormer [33]	<u>36.49</u>	<u>0.958</u>
CNN + Transformer	Dehamer [35]	33.43	0.953
	LCEFormer	37.23	0.965

4.3.3. Results on the RSID Dataset

To further validate the effectiveness and generality of our proposed LCEFormer method, we conduct a comprehensive experimental comparison with several state-of-the-

art techniques: FCTF-Net [58], FFANet [27], UHD [60], Dehamer [35], and Trinity-Net [38] in Table 6. The evaluation metrics are the same as those used in the experiments on the DHID dataset and ERICE dataset, namely Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). The results of all other methods are taken from [38]. It can be observed that our LCEFormer achieves the highest PSNR and SSIM values by achieving 27.55 dB in terms of PSNR and 0.960 in terms of SSIM. This outcome underscores the superiority of our approach. In comparison to Trinity-Net, which integrates both CNNs and SwinTransformer while introducing structural priors to generate rich details, our LCEFormer exhibits superior performance, surpassing it by 0.026 in terms of SSIM. This result demonstrates the efficacy of our LCEFormer in effectively introducing rich details.

Table 6. The state-of-the-art comparison on the RSID dataset. Bold numbers represent the best performance, while underlined numbers indicate the second best.

Type	Method	PSNR	SSIM
CNN	FCTF-Net [58]	19.31	0.856
	FFANet [27]	24.05	0.899
	UHD [60]	26.66	0.923
CNN + Transformer	Dehamer [35]	23.75	0.899
	Trinity-Net [38]	<u>27.24</u>	<u>0.934</u>
	LCEFormer	27.55	0.960

4.3.4. Experimental Results on the Remote Sensing Image Super-Resolution

Finally, we also evaluate our approach for the remote sensing image super-resolution task. We report the results on UCMERGED [39], following the same protocol as in [61], with the upsampling scale ratio set to 4. Table 7 shows the comparison of our approach with several state-of-the-art methods: Bicubic, SC [62], SRCNN [63], FSRCNN [64], LGCNet [65], DCM [66], DGANet-ISE [67], HSENet [61] on UCMERGED dataset. Our approach outperforms other state-of-the-art methods, in terms of both PSNR and SSIM.

Table 7. State-of-the-art comparison on the UCMERGED dataset for x4 upsampling. Bold numbers represent the best performance, while underlined numbers indicate the second best.

Method	PSNR	SSIM
Bicubic	25.65	0.673
SC [62]	25.51	0.715
SRCNN [63]	26.78	0.722
FSRCNN [64]	26.93	0.727
LGCNet [65]	27.02	0.733
DCM [66]	27.22	0.753
DGANet-ISE [67]	27.31	<u>0.767</u>
HSENet [61]	<u>27.73</u>	0.762
LCEFormer	27.80	0.774

4.3.5. Qualitative Comparison

Figure 5 shows the qualitative results on the DHID dataset including DCIL [36], DehazeFormer [33], and our proposed LCEFormer. It could be found that there is some haze left in the results of DCIL and DehazeFormer from the roof of the first two examples and the tree of the third example. In contrast, our LCEFormer removes haze and restores more details, as highlighted by the red boxes in Figure 5. This observation underscores the capability of our proposed models to acquire and leverage more detailed information effectively. Additionally, from the rest examples, our LCEFormer restores more natural colors compared with other methods.



Figure 5. Dehazing results on the DHID dataset.

Figure 6 shows the visual comparison on the ERICE dataset. We compare our method with Dehamer [35] and DehazeFormer [33]. From Figure 6, it could be found that our

model produces much clearer and more natural results. Dehamer removes most haze but some details are indistinct. The results dehazed by DehazeFormer have color distortion, especially for the first and last examples in Figure 6. The regions highlighted by the red boxes underscore the advantage of our LCEFormer.

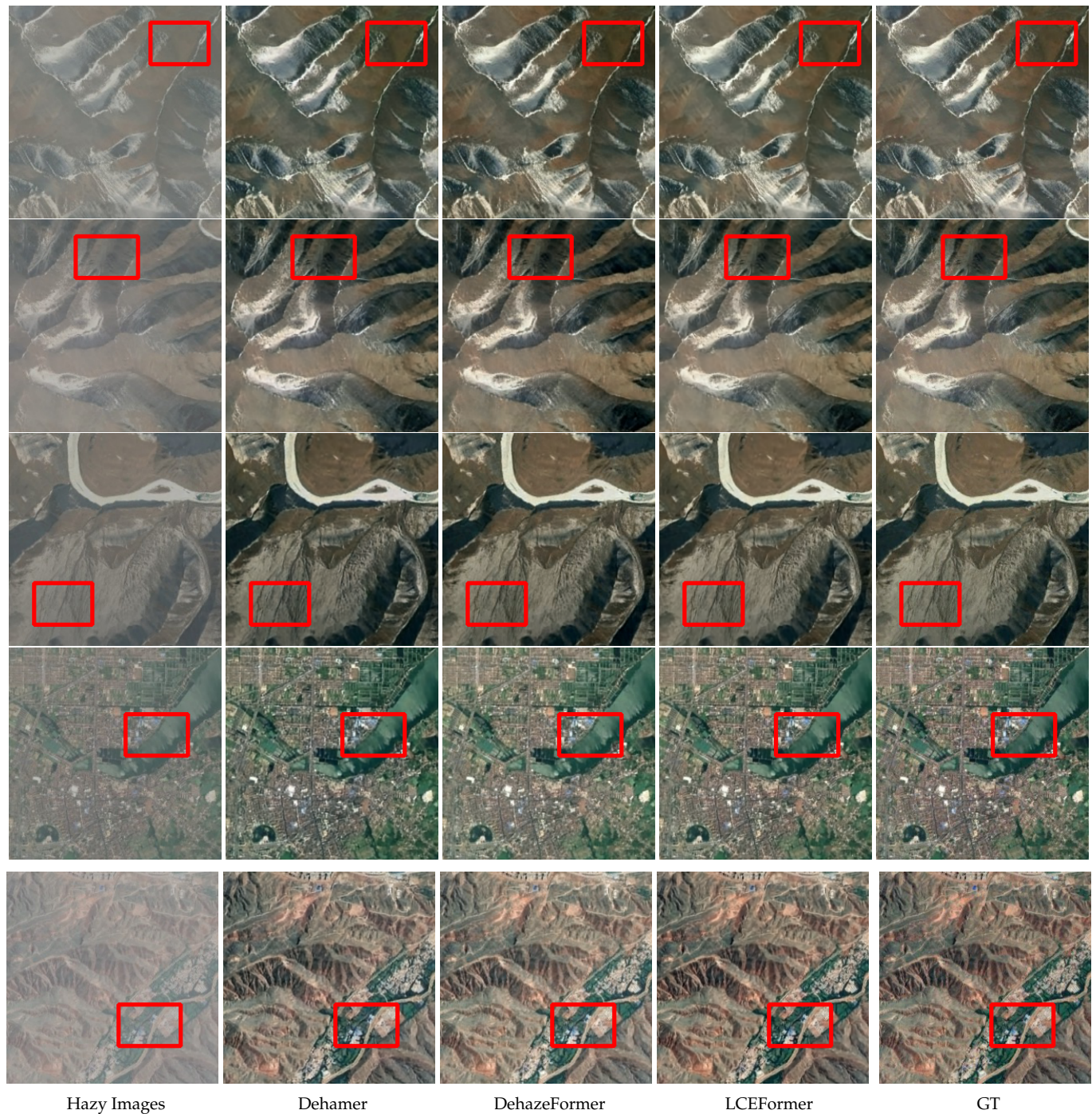


Figure 6. Dehazing results on the ERICE dataset.

The qualitative comparisons on the DHID and ERICE datasets in Figures 5 and 6 verify the superiority of our LCEFormer in the remote sensing image dehazing task.

5. Conclusions

In this paper, we have designed a local context-enriched transformer (LCEFormer) to remove the fog in remote sensing images. The proposed LCEFormer stacks local context-enriched transformer blocks to construct a U-shape dehazing framework. To enhance the transformer blocks with the local contextual information, we have proposed a CNN-based adaptive local context enrichment module (ALCEM) to extract multi-scale features and fuse them in a gated way. The proposed ALCEM is utilized to supplement long-range information with local context and construct a locally enhanced attention (LEA). Moreover, a local continuous-enhancement feed-forward network (LCFN) is devised to introduce more local context information flow. Extensive experiments conducted on the DHID, ERICE, and RSID datasets demonstrate the effectiveness of the proposed LEA and LCFN. Quantitative and qualitative analyses show that our LCEFormer significantly surpasses the state-of-the-art remote sensing image dehazing methods.

Author Contributions: Methodology, J.N. and J.X.; writing—original draft preparation, J.N.; writing—review and editing, J.X. and H.S.; funding acquisition, J.N. and J.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (grant No. 2022ZD0160400), the National Natural Science Foundation of China (grant Nos. 62206031, and 62301092), the China Postdoctoral Science Foundation (grant Nos. 2021M700613, 2022M720581, and 2023T160762), and the Fundamental Research Funds for the Central Universities (grant No. 2023CDJXY-036).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The RICE dataset is available online at https://github.com/BUPTLDy/RICE_DATASET. The DHID dataset is available online at <https://github.com/Shan-rs/DCI-Net>. The RSID dataset is available online at <https://github.com/chi-kaichen/Trinity-Net>.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wei, J.; Cao, Y.; Yang, K.; Chen, L.; Wu, Y. Self-Supervised Remote Sensing Image Dehazing Network Based on Zero-Shot Learning. *Remote Sens.* **2023**, *15*, 2732. [\[CrossRef\]](#)
2. Yu, J.; Liang, D.; Hang, B.; Gao, H. Aerial image dehazing using reinforcement learning. *Remote Sens.* **2022**, *14*, 5998. [\[CrossRef\]](#)
3. Jia, J.; Pan, M.; Li, Y.; Yin, Y.; Chen, S.; Qu, H.; Chen, X.; Jiang, B. GLTF-Net: Deep-Learning Network for Thick Cloud Removal of Remote Sensing Images via Global-Local Temporality and Features. *Remote Sens.* **2023**, *15*, 5145. [\[CrossRef\]](#)
4. Saleem, A.; Paheding, S.; Rawashdeh, N.; Awad, A.; Kaur, N. A Non-Reference Evaluation of Underwater Image Enhancement Methods Using a New Underwater Image Dataset. *IEEE Access* **2023**, *11*, 10412–10428. [\[CrossRef\]](#)
5. Paheding, S.; Reyes, A.A.; Kasaragod, A.; Oommen, T. GAF-NAU: Gramian Angular Field Encoded Neighborhood Attention U-Net for Pixel-Wise Hyperspectral Image Classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, New Orleans, LA, USA, 18–24 June 2022; pp. 409–417.
6. Bazi, Y.; Bashmal, L.; Rahhal, M.M.A.; Dayil, R.A.; Ajlan, N.A. Vision transformers for remote sensing image classification. *Remote Sens.* **2021**, *13*, 516. [\[CrossRef\]](#)
7. Zheng, X.; Sun, H.; Lu, X.; Xie, W. Rotation-invariant attention network for hyperspectral image classification. *IEEE Trans. Image Process.* **2022**, *31*, 4251–4265. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Liu, Y.; Jiang, W. OII: An Orientation Information Integrating Network for Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 731. [\[CrossRef\]](#)
9. Xu, C.; Zheng, X.; Lu, X. Multi-level alignment network for cross-domain ship detection. *Remote Sens.* **2022**, *14*, 2389. [\[CrossRef\]](#)
10. Xie, J.; Nie, J.; Ding, B.; Yu, M.; Cao, J. Cross-Modal Local Calibration and Global Context Modeling Network for RGB-Infrared Remote-Sensing Object Detection. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 8933–8942. [\[CrossRef\]](#)
11. Asokan, A.; Anitha, J. Change detection techniques for remote sensing applications: A survey. *Earth Sci. Inform.* **2019**, *12*, 143–160. [\[CrossRef\]](#)
12. Zheng, X.; Chen, X.; Lu, X.; Sun, B. Unsupervised change detection by cross-resolution difference learning. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–16. [\[CrossRef\]](#)
13. Ma, J.; Liu, D.; Qin, S.; Jia, G.; Zhang, J.; Xu, Z. An Asymmetric Feature Enhancement Network for Multiple Object Tracking of Unmanned Aerial Vehicle. *Remote Sens.* **2023**, *16*, 70. [\[CrossRef\]](#)

14. Zheng, X.; Cui, H.; Lu, X. Multiple Source Domain Adaptation for Multiple Object Tracking in Satellite Video. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4702914. [[CrossRef](#)]
15. Qi, L.; Zuo, D.; Wang, Y.; Tao, Y.; Tang, R.; Shi, J.; Gong, J.; Li, B. Convolutional Neural Network-Based Method for Agriculture Plot Segmentation in Remote Sensing Images. *Remote Sens.* **2024**, *16*, 346. [[CrossRef](#)]
16. Paheding, S.; Reyes, A.A.; Rajaneesh, A.; Sajinkumar, K.; Oommen, T. MarsLS-Net: Martian Landslides Segmentation Network and Benchmark Dataset. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 4–8 January 2024; pp. 8236–8245.
17. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2341–2353.
18. Berman, D.; Avidan, S. Non-local image dehazing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1674–1682.
19. Tan, R.T. Visibility in bad weather from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
20. Fattal, R. Dehazing using color-lines. *ACM Trans. Graph.* **2014**, *34*, 13–31. [[CrossRef](#)]
21. Cai, B.; Xu, X.; Jia, K.; Qing, C.; Tao, D. Dehazenet: An end-to-end system for single image haze removal. *IEEE Trans. Image Process.* **2016**, *25*, 5187–5198. [[CrossRef](#)] [[PubMed](#)]
22. Pang, Y.; Xie, J.; Li, X. Visual Haze Removal by a Unified Generative Adversarial Network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 3211–3221. [[CrossRef](#)]
23. Liu, X.; Ma, Y.; Shi, Z.; Chen, J. GridDehazeNet: Attention-Based Multi-Scale Network for Image Dehazing. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7313–7322.
24. Hang, D.; Jinshan, P.; Zhe, H.; Xiang, L.; Fei, W.; Ming-Hsuan, Y. Multi-Scale Boosted Dehazing Network with Dense Feature Fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2157–2167.
25. McCartney, E.J. *Optics of the Atmosphere: Scattering by Molecules and Particles*; John Wiley and Sons, Inc.: New York, NY, USA, 1976.
26. Zhang, H.; Patel, V.M. Densely Connected Pyramid Dehazing Network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018.
27. Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; Jia, H. FFA-Net: Feature Fusion Attention Network for Single Image Dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; AAAI Press: Washington, DC, USA, 2020; pp. 11908–11915.
28. Pang, Y.; Nie, J.; Xie, J.; Han, J.; Li, X. BidNet: Binocular image dehazing without explicit disparity estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 5931–5940.
29. 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 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
30. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *arXiv* **2021**, arXiv:2103.14030.
31. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. *arXiv* **2017**, arXiv:1706.03762.
32. Wang, Z.; Cun, X.; Bao, J.; Liu, J. Uformer: A General U-Shaped Transformer for Image Restoration. *arXiv* **2021**, arXiv:2106.03106.
33. Song, Y.; He, Z.; Qian, H.; Du, X. Vision transformers for single image dehazing. *IEEE Trans. Image Process.* **2023**, *32*, 1927–1941. [[CrossRef](#)] [[PubMed](#)]
34. Wu, P.; Pan, Z.; Tang, H.; Hu, Y. Cloudformer: A Cloud-Removal Network Combining Self-Attention Mechanism and Convolution. *Remote Sens.* **2022**, *14*, 6132. [[CrossRef](#)]
35. Guo, C.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; Li, C. Image Dehazing Transformer with Transmission-Aware 3D Position Embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 5802–5810.
36. Zhang, L.; Wang, S. Dense Haze Removal Based on Dynamic Collaborative Inference Learning for Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5631016. [[CrossRef](#)]
37. Lin, D.; Xu, G.; Wang, X.; Wang, Y.; Sun, X.; Fu, K. A remote sensing image dataset for cloud removal. *arXiv* **2019**, arXiv:1901.00600.
38. Chi, K.; Yuan, Y.; Wang, Q. Trinity-Net: Gradient-Guided Swin Transformer-Based Remote Sensing Image Dehazing and Beyond. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 4702914. [[CrossRef](#)]
39. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
40. Li, B.; Peng, X.; Wang, Z.; Xu, J.; Feng, D. Aod-net: All-in-one dehazing network. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4770–4778.
41. Wang, Z.; Cun, X.; Bao, J.; Zhou, W.; Liu, J.; Li, H. Uformer: A general u-shaped transformer for image restoration. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17683–17693.

42. Dong, J.; Pan, J. Physics-based feature dehazing networks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 188–204.
43. Wang, J.; Wu, S.; Xu, K.; Yuan, Z. Frequency Compensated Diffusion Model for Real-scene Dehazing. *arXiv* **2023**, arXiv:2308.10510.
44. Xu, M.; Deng, F.; Jia, S.; Jia, X.; Plaza, A.J. Attention mechanism-based generative adversarial networks for cloud removal in Landsat images. *Remote. Sens. Environ.* **2022**, *271*, 112902. [\[CrossRef\]](#)
45. Enomoto, K.; Sakurada, K.; Wang, W.; Fukui, H.; Matsuoka, M.; Nakamura, R.; Kawaguchi, N. Filmy cloud removal on satellite imagery with multispectral conditional generative adversarial nets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 48–56.
46. Tao, C.; Fu, S.; Qi, J.; Li, H. Thick Cloud Removal in Optical Remote Sensing Images Using a Texture Complexity Guided Self-Paced Learning Method. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5619612. [\[CrossRef\]](#)
47. Wang, W.; Xie, E.; Li, X.; Fan, D.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the IEEE International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 548–558.
48. Zhang, Q.; Yang, Y.B. ResT: An Efficient Transformer for Visual Recognition. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 7–10 December 2021.
49. Liu, J.; Yuan, H.; Yuan, Z.; Liu, L.; Lu, B.; Yu, M. Visual transformer with stable prior and patch-level attention for single image dehazing. *Neurocomputing* **2023**, *551*, 126535. [\[CrossRef\]](#)
50. 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 IEEE Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.
51. Charbonnier, P.; Blanc-Feraud, L.; Aubert, G.; Barlaud, M. Two deterministic half-quadratic regularization algorithms for computed imaging. In Proceedings of the 1st International Conference on Image Processing, Austin, TX, USA, 13–16 November 1994; Volume 2, pp. 168–172.
52. Huang, B.; Li, Z.; 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 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 1795–1802.
53. MMLab Contributors. MMLab: OpenMMLab Multimodal Advanced, Generative, and Intelligent Creation Toolbox. 2023. Available online: <https://github.com/open-mmlab/mmagic> (accessed on 24 February 2024).
54. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In Proceedings of the NIPS workshop, Long Beach, CA, USA, 4–9 December 2017.
55. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv* **2017**, arXiv:1711.05101.
56. Loshchilov, I.; Hutter, F. SGDR: Stochastic gradient descent with warm restarts. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
57. Yang, H.H.; Yang, C.H.H.; James Tsai, Y.C. Y-Net: Multi-Scale Feature Aggregation Network With Wavelet Structure Similarity Loss Function For Single Image Dehazing. In Proceedings of the ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual, 4–9 May 2020; pp. 2628–2632.
58. Li, Y.; Chen, X. A Coarse-to-Fine Two-Stage Attentive Network for Haze Removal of Remote Sensing Images. *IEEE Geosci. Remote. Sens. Lett.* **2021**, *18*, 1751–1755. [\[CrossRef\]](#)
59. Wang, S.; Wu, H.; Zhang, L. Afdn: Attention-Based Feedback Dehazing Network For Uav Remote Sensing Image Haze Removal. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 3822–3826.
60. Zheng, Z.; Ren, W.; Cao, X.; Hu, X.; Wang, T.; Song, F.; Jia, X. Ultra-high-definition image dehazing via multi-guided bilateral learning. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 16180–16189.
61. Lei, S.; Shi, Z. Hybrid-scale self-similarity exploitation for remote sensing image super-resolution. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5401410. [\[CrossRef\]](#)
62. Yang, J.; Wright, J.; Huang, T.S.; Ma, Y. Image super-resolution via sparse representation. *IEEE Trans. Image Process.* **2010**, *19*, 2861–2873. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Dong, C.; Loy, C.C.; He, K.; Tang, X. Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 295–307. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Dong, C.; Loy, C.C.; Tang, X. Accelerating the super-resolution convolutional neural network. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 391–407.
65. Lei, S.; Shi, Z.; Zou, Z. Super-resolution for remote sensing images via local-global combined network. *IEEE Geosci. Remote. Sens. Lett.* **2017**, *14*, 1243–1247. [\[CrossRef\]](#)

66. Haut, J.M.; Paoletti, M.E.; Fernández-Beltran, R.; Plaza, J.; Plaza, A.; Li, J. Remote sensing single-image superresolution based on a deep compendium model. *IEEE Geosci. Remote. Sens. Lett.* **2019**, *16*, 1432–1436. [[CrossRef](#)]
67. Qin, M.; Mavromatis, S.; Hu, L.; Zhang, F.; Liu, R.; Sequeira, J.; Du, Z. Remote sensing single-image resolution improvement using a deep gradient-aware network with image-specific enhancement. *Remote Sens.* **2020**, *12*, 758. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.