electronics

Article

A Novel Approach to Maritime Image Dehazing Based on a Large Kernel Encoder–Decoder Network with Multihead Pyramids

Wei Yang [1], Hongwei Gao [2], Yueqiu Jiang [3,*] and Xin Zhang [4]

[1] Graduate School, Shenyang Ligong University, Shenyang 110159, China
[2] School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110159, China
[3] School of Information Science and Engineering, Shenyang Ligong University, Shenyang 110159, China
[4] School of Automobile and Traffic, Shenyang Ligong University, Shenyang 110159, China
* Correspondence: yueqiujiang@sylu.edu.cn

Abstract: With the continuous increase in human–robot integration, battlefield formation is experiencing a revolutionary change. Unmanned aerial vehicles, unmanned surface vessels, combat robots, and other new intelligent weapons and equipment will play an essential role on future battlefields by performing various tasks, including situational reconnaissance, monitoring, attack, and communication relay. Real-time monitoring of maritime scenes is the basis of battle-situation and threat estimation in naval battlegrounds. However, images of maritime scenes are usually accompanied by haze, clouds, and other disturbances, which blur the images and diminish the validity of their contents. This will have a severe adverse impact on many downstream tasks. A novel large kernel encoder–decoder network with multihead pyramids (LKEDN-MHP) is proposed to address some maritime image dehazing-related issues. The LKEDN-MHP adopts a multihead pyramid approach to form a hybrid representation space comprising reflection, shading, and semanteme. Unlike standard convolutional neural networks (CNNs), the LKEDN-MHP uses many kernels with a 7×7 or larger scale to extract features. To reduce the computational burden, depthwise (DW) convolution combined with re-parameterization is adopted to form a hybrid model stacked by a large number of different receptive fields, further enhancing the hybrid receptive fields. To restore the natural hazy maritime scenes as much as possible, we apply digital twin technology to build a simulation system in virtual space. The final experimental results based on the evaluation metrics of the peak signal-to-noise ratio, structural similarity index measure, Jaccard index, and Dice coefficient show that our LKEDN-MHP significantly enhances dehazing and real-time performance compared with those of state-of-the-art approaches based on vision transformers (ViTs) and generative adversarial networks (GANs).

Keywords: image dehazing; large kernel encoder–decoder network; multihead pyramids; re-parameterization; digital twin

Citation: Yang, W.; Gao, H.; Jiang, Y.; Zhang, X. A Novel Approach to Maritime Image Dehazing Based on a Large Kernel Encoder–Decoder Network with Multihead Pyramids. *Electronics* **2022**, *11*, 3351. https://doi.org/10.3390/electronics11203351

Academic Editor: George A. Papakostas

Received: 20 September 2022
Accepted: 11 October 2022
Published: 17 October 2022

1. Introduction

Maritime situational real-time monitoring enables naval battlefield threat appraisal. The core task of real-time monitoring of offshore ship targets is to obtain real-time high-definition (HD) images of the naval surface, which have promising applications in fishery management, maritime rescue, offshore traffic monitoring, naval battlefield situational awareness, and other fields. However, aerial images possess large amounts of data and are susceptible to various factors, such as complex sea surface conditions, extreme weather, poor lighting conditions, and loading errors from imaging detectors. Some disturbances in natural environments, such as haze, sea clutter, and clouds, will negatively influence decisions based on visual features in a real-time system. In this work, we will limit our discussion to interference in a hazy environment under low and ultra-low altitudes.

1.1. Mathematical Model of Atmospheric Scattering

There are two root causes for haze in sea surface images: (1) The light reflected by the target is absorbed and scattered by suspended particles during transmission, resulting in energy attenuation as well as reductions in image brightness and contrast; and (2) sunlight, skyglow, and other sources of light are scattered by particles to form stray light, causing blurred images and unnatural colors. The atmospheric scattering model commonly used in computer vision and computer graphics is as follows [1]:

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

where $I(x)$ is the observed image intensity at pixel x, and $J(x)$ is the scene radiance. Generally, we can regard haze near ground as homogeneous. The transmission $t(x)$ represents the ability that light interactions with the atmosphere can be expressed as:

$$t(x) = e^{-\beta(\lambda)d}. \tag{2}$$

It describes the part of the light that is not scattered and reaches the camera. $\beta(\lambda)$ denotes the scattering coefficient. According to Rayleigh's law [2] of atmospheric scattering, we know about the relationship between the scattering coefficient β and the wavelength λ:

$$\beta(\lambda) \propto \frac{1}{\lambda^{\gamma}}, \tag{3}$$

where $0 \leq \gamma \leq 4$, and γ depends on particles size suspended in the gas. In clear days, $\gamma = 4$. The molecules' scattering is selective, and the blue wavelengths are scattered more compared to other visible wavelengths. On the other hand, fog and haze scatter all visible wavelengths more or less the same way, $\gamma \approx 0$ [3]. In this paper, we assume the scattering coefficient $\beta(\lambda) = \beta$ to be constant since the camera usually has narrow spectral bands in most cases. d is the distance between the object and the imaging system. We can infer from (2) that light reflected from the object surface is attenuated exponentially with the scene depth d. A is the global atmospheric light and can be regarded as a constant if images are taken under the same weather condition. The first term $J(x)t(x)$ on the right-hand side of (1) is called direct attenuation, which describes the scene radiance and its attenuation in the atmosphere. The latter term $A[1 - t(x)]$ is called airlight [4], which is the addition of a white atmosphere veil.

1.2. Contributions

A large kernel encoder–decoder network with multihead pyramids (LKEDN-MHP) combining convolutional neural networks (CNNs) and vision transformers (ViTs) can powerfully narrow the performance gap between CNNs and ViTs. Notably, CNNs possess translational invariance and a local receptive field, whereas ViTs have a global receptive field and a self-attention mechanism. First, we introduce a guidance map, combined with hazy images, to estimate the thickness of haze directly and generate a realistic haze-free image in the data input stage. Then, the preprocessed images are input into the global aggregation and local convolution paths. Multiple large kernel convolution layers are used to further obtain a larger effective receptive field (ERF), forcing the model to focus on capturing the local region's fine-grained and deep semantic information in the local convolution path. In contrast, in the global aggregation path, global average pooling (GAP) is used for each channel, followed by an upsampling process. The global features are merged with local features generated by the two paths to obtain a wealthy semantic representation space, allowing for superior performance in downstream tasks. Shading and texture perceptions are the essential visual cues for recognizing objects and interpreting scenes. Therefore, the proposed LKEDN-MHP jointly integrates the principles of the haze-free approach from global and local paths for input hazy maritime images. We design a multihead pyramid for predictions of reflectance, shading, and semanteme in the local

path. Since the reflectance component contains the factual color information of the scene, it encourages the reflectance prediction subnetwork to learn more intermediate features, which helps restore proper color information in the degraded images. Furthermore, the intermediate features can serve as complementary features to the semantic subnetwork and thus improve the dehazing results with high color contrast. Similarly, the shading prediction subnetwork can provide wealthy complementary features beneficial to texture enhancement, thus improving the dehazing results with fine details.

The main contributions of this paper include the following:

- We propose a novel CNN architecture for the maritime image dehazing task, namely, the LKEDN-MHP, which combines global and local information. Predictions of reflection, shading, and semanteme are connected to provide rich complementary information for the dehazing task so that the high-quality haze-free images generated by the LKEDN-MHP have natural colors and fine details for a human eye.
- We propose an improved pure CNN paradigm and demonstrate that using a few large convolutional kernels instead of a stack of small kernels can be more powerful. The LKEDN-MHP uses multiple large kernels to obtain a larger ERF, similar to the multihead self-attention (MHSA) method of ViTs. By combining the advantages of CNNs and ViTs, the CNN paradigm is enhanced further to improved dehazing and articulated hazy images.
- We establish a 3D digital twin system to verify the performance of the LKEDN-MHP. We build 3D models of oceanic scenarios, moving ocean-going vessels, inshore ships, and maritime reconnaissance aircraft. Furthermore, the real offshore reconnaissance scene is restored, which includes a haze rendering method to simulate the impact of haze in the real world. A description of research and analysis of modern naval warfare, monitoring, and warning is given. This approach creates considerable numbers of datasets related to naval warfare situations, which will be helpful for ViT technology-related development of guidance and precision strike weapons.

The remainder of the paper is organized as follows: Section 2 briefly reviews related work, focusing on the methods of image dehazing based on CNNs and ViTs. The proposed method, LKEDN-MHP, is described in detail in Section 3. Section 4 reports our experiments and analysis of the results. Finally, Section 5 summarizes our main conclusions and highlights future work.

2. Related Work

Deep learning (DL) algorithms are extensively implemented in multiple fields and have achieved incredible results, such as facial expression recognition and human pose estimation [5,6]. The state-of-the-art dehazing algorithm is wholly based on data-driven learning [7], including the haze-free image generation method based on CNNs and ViTs. ViTs have been appreciated by many researchers due to their powerful attention mechanism and flexibility in time series modeling, but they still cannot achieve supremacy in the field of image processing [8–10].

2.1. Image Dehazing Approaches Based on CNNs

DL approaches can fit training data well with GPU hardware acceleration capability, and the haze-free technique based on CNNs has been ubiquitously studied. The state-of-the-art dehazing network based on DL implemented an end-to-end framework and achieved better performance than traditional methods, including the Enhanced Pix2pix Dehazing Network (EPDN) [11], dense attentive dehazing network (DADN) [12], and multiscale boosted dehazing network MSBDN approaches [13]. These algorithms are trained end-to-end on haze-free images directly. With recent advancements, generative adversarial networks (GANs) for image style transfer, such as Pix2pix and CycleGAN [14,15], have become increasingly popular. Image dehazing can also be considered style transformation: An image is transferred from a hazy domain to a haze-free domain. Engine et al. attempted to combine cycle consistency and perceptual losses in the CycleGAN framework [16].

In addition, approaches derived from image segmentation, such as the feature pyramid network (FPN) approach, have been demonstrated to be influential in image dehazing applications. Image segmentation networks usually use an encoder–decoder to learn the embedded representation of features after input data are mapped to higher dimensions. Chaurasia and Culurciello proposed an efficient semantic segmentation architecture based on a fully convolutional encoder–decoder framework [17]. Their encoder uses a ResNet18 model for feature encoding and avoids spatial information loss by reintroducing residuals from each encoder to the output of its corresponding decoder [18].

2.2. Image Dehazing Approaches Based on ViTs

ViTs have recently become active in visual processing due to their powerful attention mechanism and flexibility in time series modeling, including object detection, image classification, and semantic segmentation [19,20]. The MHSA mechanism plays a vital role in ViTs and has been well documented. MHSA is flexible [21], has powerful performance (less inductive bias) [22], is robust to distortion [23,24], or is able to model long-term dependencies [25,26]. However, some works have questioned the necessity of MHSA and attributed the high performance of ViTs to proper building blocks or sparse dynamic weights [27–30].

To the best of our knowledge, the computational complexity of ViTs increases exponentially with image pixel size, which restrains their application in dehazing tasks. Many versions of ViTs have been proposed to alleviate the problem of high computing costs. For example, the pyramid vision transformer (PVT) applies a transformer to lower-resolution features, significantly reducing computational costs [31]. Another solution proposed in Swin Transformer [32] is locally grouped self-attention, where the input features are divided into a grid of nonoverlapping windows and then the visual transformer works only within each window. LocalViT [33] applies a local mechanism to ViTs by introducing a CNN into the feed-forward network. CvT [34] applies convolutional token embedding to relate local context and a convolution projection layer to provide efficiency benefits. In addition, it has been shown that the fusion of global and local information is significant to upstream visual tasks. However, ViTs do not focus on local textures [8] and cuts images into sequences as input, causing them to lack transitional invariance when used in image processing tasks. Approaches based on ViTs have many limitations in interpretability [7] and are not strong candidates for image dehazing. These limitations prompted us to investigate a more powerful approach for addressing the issue of maritime image dehazing.

3. Proposed Approach

In ViTs, MHSA is commonly designed as a fusion of global features [8,31,35] but with large kernels [9,32,36]so that each output from a single MHSA layer can gather information from a large region. However, large kernels are not popularly used in CNNs (except for the first layer [18]). Instead, a stack of many small spatial convolutions [18,24,37,38] (e.g., 3×3) is typically used to enlarge the ERF in state-of-the-art CNNs. ViTs and CNNs do not have to be applied independently. An aim of this paper is to analyze the respective advantages of the two algorithms and improve standard image dehazing performance. Therefore, based on the above principles, the LKEDN-MHP is proposed in this paper to address the issue of extensive computation and global-local ERFs.

3.1. Architecture Specification

GANs, such as the EPDN and FD-GAN algorithms, are extensively implemented in single-image dehazing tasks. The LKEDN-MHP architecture proposed in this paper is shown in Figure 1.

According to Figure 1, the LKEDN-MHP has five components: (1) Block_1, (2) Block_2, (3) a multihead pyramid, (4) self-attention, and a (5) global aggregation path. The input dataset is a $512 \times 512 \times 3$ RGB image. Noticeably, the guidance map is combined with the

hazy image as the input of the local convolution path to estimate haze thickness and the haze-free image directly.

- Block_1 refers to the beginning layers. Since we target a high-performance backbone of downstream dense-prediction tasks, and commonly, the input data size of the downstream dense-prediction task is significant, we hope to capture more details by implementing several convolutional layers in the initial stage. The number of channels C_1 in Block_1 is 64. After processing the first 3×3 layer with $2 \times$ downsampling, we design a 7×7 DW layer to capture low-level patterns, with a 1×1 conv and another 7×7 DW layer for downsampling, where the process of DW convolution is shown in Figure 2.

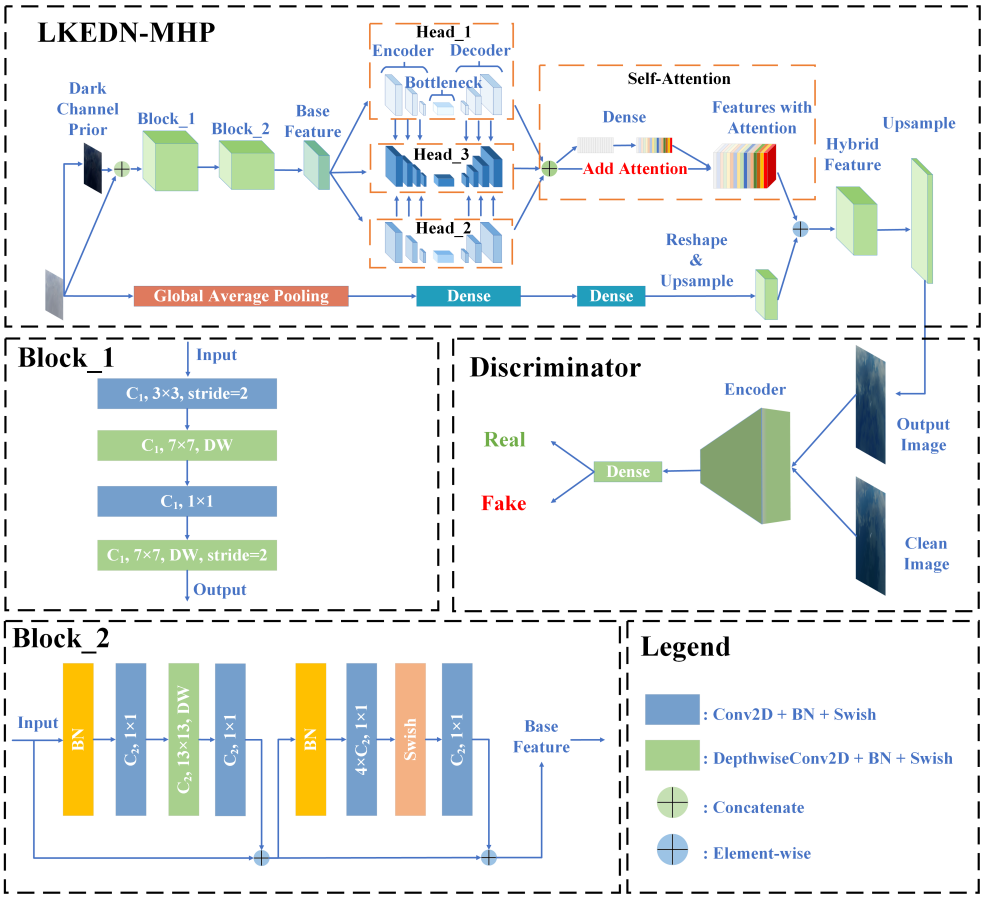


Figure 1. The architecture of the LKEDN-MHP. The guidance map is concatenated with the original input in the local path. The predictions of reflectance, shading, and semanteme are generated by the multihead pyramid. We use self-attention to mark notable information and then merge the local and global features. The generated haze-free image and the GT are the input of the discriminator.

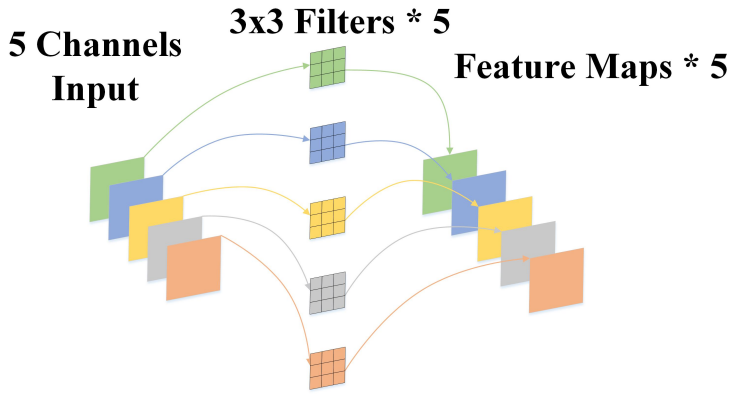


Figure 2. The process of DW convolution. Take the number of input data channels 5 and the kernel size 3×3 as an example. Each channel is convoluted by a filter.

As shown in Figure 2, in contrast to the conventional convolution operation, a DW convolution kernel is responsible for one channel, i.e., a channel is convoluted by only one kernel. It is believed that large-kernel convolutions are computationally expensive because the kernel size quadratically increases the number of parameters and floating point operations (FLOPs). This drawback can be significantly overcome by applying DW convolutions [39].

- Block_2 uses the identity shortcut and DW large kernel convolution. After DW convolution, we desire to use a 1×1 convolution layer (pointwise, PW) as a standard component to increase depth and provide more nonlinearities and information communications across channels. Except for the large convolution layers, which provide a sufficient receptive field and have the ability to aggregate spatial information, the model's representational capacity is closely related to depth. The number of channels C_2 in Block_2 is 128. Inspired by the feed-forward network (FFN) widely used in transformers and MLPs [40], we use a similar CNN-style block consisting of a shortcut, batch normalization (BN) [41], double 1×1 layers, and Swish activation functions. Compared to the classic FFN, which uses layer normalization [38] before the fully connected layers, BN has the advantage that it can be fused into convolution for efficient inference.
- The multihead pyramid contains three heads: reflection prediction, shading prediction, and semantic prediction. The reflection and shading heads provide rich complementary shade and texture information for image dehazing tasks, enabling the network to generate high-quality haze-free images with natural colors and fine details. The architectures of $Head_1$ and $Head_2$ are shown in Figure 3, and the architecture of $Head_3$ is shown in Figure 4.

  According to Figures 3 and 4, $Head_1$ and $Head_2$ contain six stage blocks, and the encoder and decoder each contain three stage blocks. Each stage contains several LK blocks, and the DW convolution in each LK block uses a 5×5 kernel for reparameterization. A ConvFFN block is placed after LK Block, while eight stage modules are included in $Head_3$, and elementwise fusion is performed with the outputs of $Head_1$ and $Head_2$.

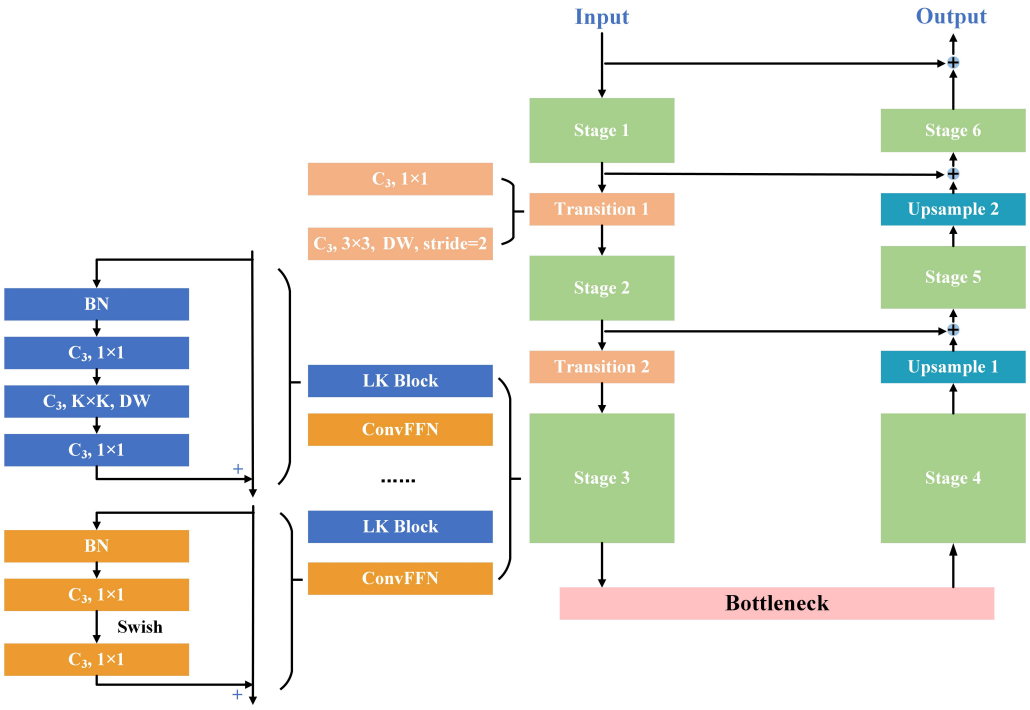


Figure 3. The architecture of $Head_1$ and $Head_2$. Each stage contains several LK blocks, and the DW convolution in each LK block uses a 5 × 5 kernel for re-parameterization.

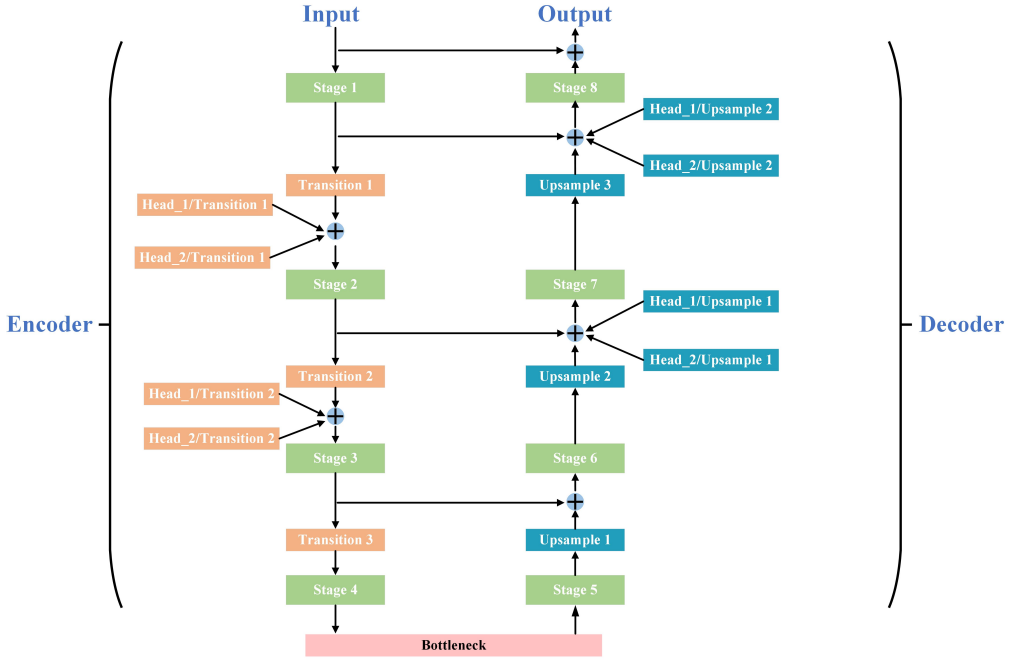


Figure 4. The architecture of $Head_3$. An elementwise fusion is performed with the outputs of $Head_1$ and $Head_2$.

- In the self-attention block, the attention score is generated by the GAP layer and the full connection of the two layers, and then the attention score is added to the fused MHP output to force the model to focus on critical information. To effectively integrate complementary features, we introduce the self-attention block. The self-attention block significantly improves the effectiveness of feature fusion by adaptively boosting the weights of the appropriate complementary channel while eliminating

unrelated channels. The self-attention block highlights essential information, cuts redundant information, and optimizes feature fusion performance.

- The global aggregation path provides the global feature to the local convolution path, and the local-global hybrid feature can be obtained by elementwise aggregation of features that contain the attention score and features obtained through GAP, a full connection layer, and upsampling. The global aggregation path improves efficiency while being able to aggregate global information. The hybrid output of the local and global paths is converted to the original size by 3×3 convolution and upsampling to obtain the generated HD haze-free image.

The discriminator takes a generated haze-free image or GT as its input, and then the output is obtained by the encoder and fully connected layers. The multihead pyramid encoder is employed to encourage the discriminator to have the same capacity to extract and analyze sophisticated features such as generators, thus causing the two networks to compete to improve their performance.

3.2. Loss Function

Since adversarial loss has achieved significant success in the image restoration field, for the sake of both pixel quality and human perception, we adopted a combination of adversarial loss, MSE loss, and perceptual loss [14,42,43] in the proposed scheme. Adversarial loss is defined as [14]:

$$L_{adv} = \frac{1}{B}\sum_{i=1}^{B}\{1 - D[G(z_i)]\}, \tag{4}$$

where B is the number of samples in the mini-batch, $D(\cdot)$ is the output of the discriminator, $G(\cdot)$ is the generated image, and z is the input hazy image. The MSE loss function is shown as [42]:

$$L_{MSE} = \frac{1}{K}\sum_{j=1}^{K}\left[G_j(z) - R_j\right], \tag{5}$$

where K is the number of pixels in the generated image and R is the GT. The perceptual loss used to measure perceptual similarity in feature space is shown as [43]:

$$L_{per} = \frac{1}{P}\sum_{n=1}^{P}\{\varphi_n[G(z)] - \varphi_n(R)\}^2, \tag{6}$$

where P is the number of elements in the feature map φ in layer conv3-3 of the VGG 16 model. By combining all the related loss functions, the integral loss function optimizing the generator can be formulated as [10,42,43]:

$$L_{total} = \alpha_1 L_{adv} + \alpha_2 L_{MSE} + \alpha_3 L_{per}, \tag{7}$$

The critical loss function of the discriminator is shown as [44]:

$$L_D = \alpha_4 \frac{1}{B}\sum_{m=1}^{B}\{D(R_m) - D[G(z_m)]\}, \tag{8}$$

For our present scheme, the weights of integral loss and critical loss functions are set as $\alpha_1 = 500$, $\alpha_2 = 500$, $\alpha_3 = 500$, $\alpha_4 = 1$.

4. Experimental Study

4.1. Datasets and Experimental Setup

With the constant influence of novel technologies such as 3D simulation incorporated into research, the limited perception of previous vision approaches has been overcome, and a new world of 3D interaction has arisen. Realistic 3D virtual dynamic displays and immersive experiences have gradually become more common. We perform 3D simulation

with hazy and haze-free environments to display the target captured under the offshore situation, as shown in Figure 5.

A total of 18,540 images with different angles, diverse target scales, and various lighting situations were collected, each with a size of 1920 × 1080, including 9648 hazy images and 8892 GT images. The training set, test set, and validation set are divided into 7000 pairs of hazy and hazy-free images, 756 hazy images and 1892 pairs of hazy and hazy-free images, respectively. To improve the performance of the proposed LKEDN-MHP, data augmentation is adopted. First, each hazy image is randomly clipped into five patches with the same width-to-height ratio as the origin. Then, the horizontal flip method is applied to double the number of training samples and create a new geometric texture in the training dataset. To verify the universality of the LKEDN-MHP, we add a public outdoor-haze dataset (O-Haze) [45].

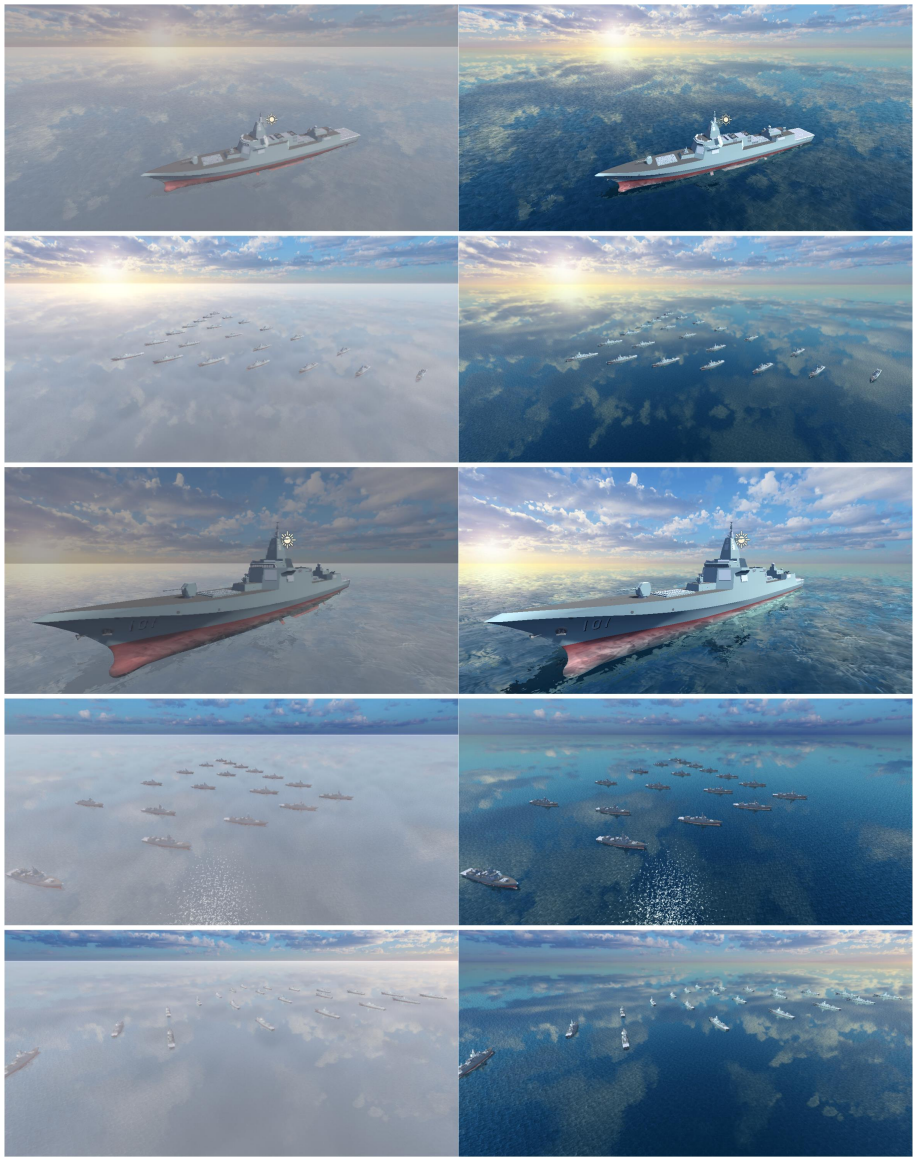


Figure 5. Samples of datasets. Left: Hazy input. Right: Ground-truth. The datasets contain multi targets and a variety of situations to be as realistic as possible.

4.2. Experimental Settings

The experiments were performed on a GeForce GTX TITAN X graphics card with a network input size of 512 × 512 × 3 and 1000 epochs. In the first 800 epochs, the learning rate is fixed at 10^{-4}, and in the final 200 epochs, we linearly decay the learning rate to

zero. The peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) were used to measure the dehazing performance of the LKEDN-MHP. The PSNR, however, is chosen as our primary performance measure because we consider the pixel quality of the restored image that can benefit other computer vision tasks such as object detection. The network is trained with a batch size of one sample, which empirically improves the validation results in the image restoration task [46]. In the MHP, each stage has three architectural hyperparameters: the number of LK blocks N, the channel dimension C_3, and the kernel size K. The model with the above parameters is called LKEDN-MHP-B (with B for Base), and the more comprehensive model is called LKEDN-MHP-L (with L for Large). Table 1 shows the setting of architectural hyperparameters for each stage in the experiment.

Table 1. The setting of architectural hyperparameters.

Model Name	Parameter Name			
	$Head_{num}$	N	C_3	K
LKEDN-MHP-B	Head_ 1 & Head_ 2	[2,2,8,8,2,2]	[128,256,512, 512,256,128]	[13,13,13, 13,13,13]
	Head_ 3	[2,2,4,8,8,4,2,2]	[128,256,512,1024, 1024,512,256,128]	[31,29,27,25, 21,19,17,13]
LKEDN-MHP-L	Head_ 1 & Head_ 2	[2,2,16,16,2,2]	[256,512,1024, 1024, 512,256]	[13,13,13, 13,13,13]
	Head_ 3	[2,2,8,16,16,8,2,2]	[256,512,1024,2048, 2048,1024,512,256]	[31,29,27,25, 21,19,17,13]

4.3. Results on the Self-Built and O-Haze Datasets

The flow of information over different layers of the LKEDN-MHP is shown in Figure 6.

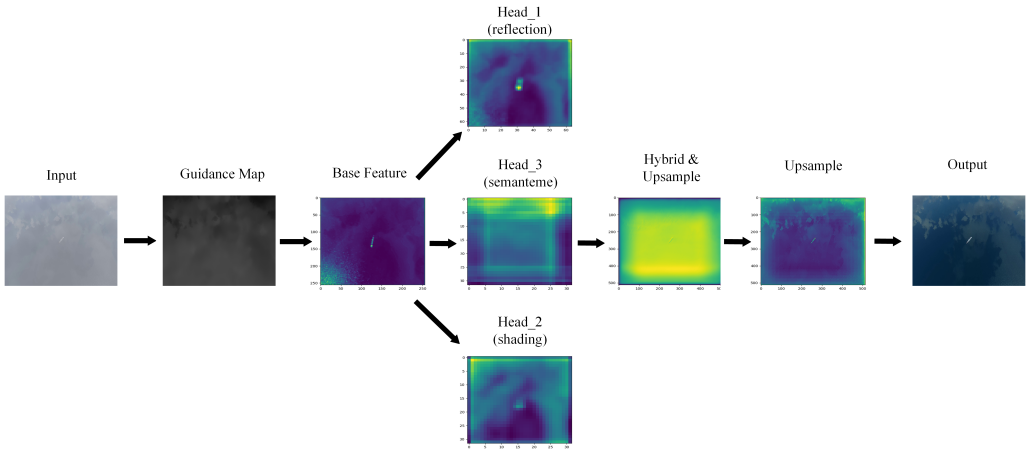
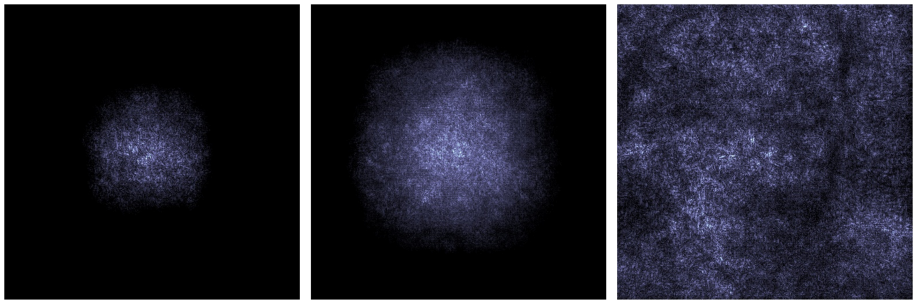


Figure 6. The flow of information over different layers of the LKEDN-MHP.

The performance of the LKEDN-MHP proposed in this paper is compared with that of other state-of-the-art algorithms on the self-built dataset and O-Haze dataset [45]. To evaluate the superiority of the proposed method, state-of-the-art algorithms based on CNN and ViT were made a qualitative comparison. We believe that the high performance of LKEDN-MHP is mainly because of the large ERFs we build via large kernels, as compared in Figure 7.



(a) EDN-GTM (b) DehazeFormer (c) LKEDN-MHP

Figure 7. The ERFs of EDN-GTM, DehazeFormer, and LKEDN-MHP, respectively. A more widely distributed bright area indicates a larger ERF. We measure the effective receptive field of different layers as the absolute value of the gradient of the center location of the feature map with respect to the input. Results are averaged across all channels in each map for 32 randomly-selected images.

In many classical computer vision tasks, the Dice score and Jaccard index are used to evaluate the segmentation performance [47]. The results compared with the state-of-the-art methods are shown in Figure 8 and Table 2.

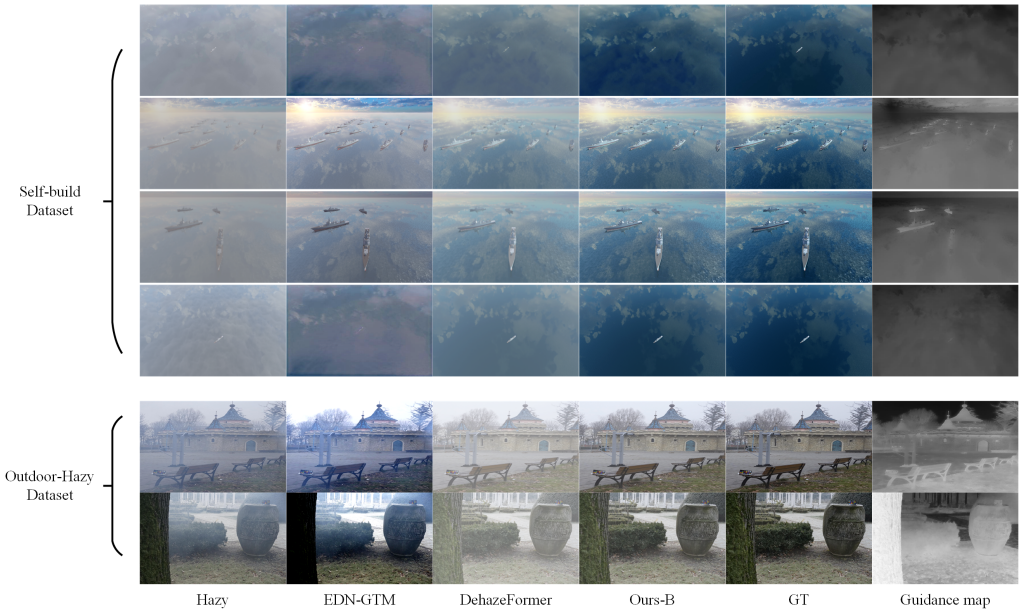


Figure 8. The results obtained with the state-of-the-art algorithms on the self-built dataset and O-Haze dataset. The top four rows show the dehazing results on the self-built dataset, and the bottom two rows show the dehazing result on the O-Haze public dataset.

Table 2. Quantitative defogging results of the self-built data set and O-Haze data set.

Model	Self-Built Datasets		O-Haze [45]		Jaccard Index	Dice Coefficient	Params	MACs
	PSNR	SSIM	PSNR	SSIM				
DCP [48]	20.15	0.7247	22.98	0.7948	0.244	0.362	—	—
PCFAN [49]	22.59	0.7695	25.46	0.8498	0.407	0.538	36.475 M	68.71 G
Grid-DehazeNet [50]	25.41	0.7966	28.32	0.8845	0.571	0.712	0.958 M	21.49 G
FFA-Net [51]	29.72	0.8148	31.13	0.9392	0.577	0.717	4.456 M	287.8 G
EDN-GTM [52]	30.68	0.8331	33.07	0.9537	0.598	0.656	17.362 M	83.48 G
Dehaze-Former [1]	31.23	0.8654	34.95	0.9840	0.780	0.870	2.514 M	25.79 G
LKEDN-MHP-B	35.63	0.9146	38.77	0.9881	0.897	0.944	0.635 M	12.84 G
LKEDN-MHP-L	38.57	0.9270	39.16	0.9896	0.913	0.962	0.985 M	21.74 G

Figure 8 shows that our scheme achieves the best dehazing results on both the self-built dataset and the O-Haze dataset. As shown in Table 2, in the experiment performed on the self-built maritime hazy image dataset obtained by the simulation environment based on the digital twin, the PSNR of LKEDN-MHP-B reached 35.63 dB. The PSNR of LKEDN-MHP-L in our scheme is further improved by 2.94 dB, and the SSIM reaches the optimal value of 0.9896, indicating the best performance in the maritime dehazing task. To thoroughly verify the reliability and universality of the scheme, the proposed approach was tested on the public dataset O-Haze, and it performed best on this dataset, with the large model PSNR reaching 39.16 dB.

Compared with those of the state-of-the-art algorithms, the PSNR and SSIM of our large model reached 39.16 dB and 0.9896 on the O-Haze dataset, respectively. In addition, the large model under the PSNR and SSIM yielded values of 35.63 dB and 0.9146 on the self-built dataset, respectively, which were better than those of the state-of-the-art algorithms. Compared with the public dataset, the self-built maritime hazy image dataset tested in this paper was more challenging. It provides rich data for object detection and other downstream tasks under extreme weather conditions, which is of great significance to the study of maritime situational reconnaissance, monitoring and detection tasks. The experimental results show that the proposed LKEDN-MHP algorithm yields promising results on the self-built and natural outdoor hazy datasets.

4.4. Ablation Study

We conducted some ablation studies to prove the effectiveness of our method and the rationality of selecting parameters on the dataset of self-built.

- The number of LK blocks N. In Table 3, we compare performances with different number of LK blocks N. The rest of the hyperparameters refer to the baseline.

Table 3. Ablation experiments for the number of LK blocks N on the self-built dataset for maritime image dehazing task.

$Head_{num}$	N	PSNR	Time (ms)	Params (M)
Head_ 1 & Head_ 2 Head_ 3	[2,2,2,2,2,2] [2,2,2,2,2,2,2,2]	26.79	97	0.378
Head_ 1 & Head_ 2 Head_ 3	[2,2,4,4,2,2] [2,2,4,4,4,4,2,2]	32.67	113	0.451
Head_ 1 & Head_ 2 Head_ 3	[2,2,8,8,2,2] [2,2,4,8,8,4,2,2]	35.63	148	0.635
Head_ 1 & Head_ 2 Head_ 3	[2,2,16,16,2,2] [2,2,8,16,16,8,2,2]	38.57	232	0.985
Head_ 1 & Head_ 2 Head_ 3	[2,2,32,32,2,2] [2,2,16,32,32,16,2,2]	35.51	421	1.587

The large conv kernel size K and 5×5 re-parameterization in LK block. In Table 4, we compare performances with different large conv kernel size K and 5×5 re-parameterizationin LK block. The rest of the hyperparameters refer to the baseline. We introduced a 5×5 kernel for re-parameterization in each LK block. In this way, we make the very large kernel capable of capturing small-scale patterns, hence improving the performance of the model. Table 4 shows that directly increasing the kernel size from 13 to 31 reduces the accuracy, while re-parameterization addresses the issue.

Table 4. Ablation experiments for the large conv kernel size K and 5×5 re-parameterization in LK blocks.

K	5×5 Re-Param	PSNR
[13,13,13,13,13,13,13,13]		28.79
[13,13,13,13,13,13,13,13]	✓	32.74
[31,29,27,25,21,19,17,13]		31.89
[31,29,27,25,21,19,17,13]	✓	38.57

- The channel number C_3. In Table 5; we compare performances with different number of channels C_3 in each $Stage$. The rest of the hyperparameters refer to the baseline.

Table 5. Ablation experiments for the channels C_3 in each *Stage* on the self-built dataset for the maritime image dehazing task.

$Head_{num}$	C_3	PSNR	Time (ms)	Params (M)
Head_ 1 & Head_ 2	[128,128,128, 128,128,128]	25.44	71	0.306
Head_ 3	[128,256,256,256, 256,256,256,128]			
Head_ 1 & Head_ 2	[128,256,256, 256,256,128]	31.38	105	0.424
Head_ 3	[128,256,512,512, 512,512,256,128]			
Head_ 1 & Head_ 2	[128,256,512, 512,256,128]	35.63	148	0.635
Head_ 3	[128,256,512,1024, 1024,512,256,128]			
Head_ 1 & Head_ 2	[256,512,1024, 1024,512,256]	38.57	232	0.985
Head_ 3	[256,512,1024,2048, 2048,1024,512,256]			
Head_ 1 & Head_ 2	[512,512,2048, 2048,512,512]	34.85	476	1.734
Head_ 3	[512,1024,2048,4096, 4096,2048,1024,512]			

5. Conclusions

This paper proposes a novel large kernel encoder–decoder network with multihead pyramids (LKEDN-MHP) for maritime image dehazing tasks. The proposed LKEDN-MHP scheme utilizes a transmission map extracted by using a guidance map as additional input to the network to achieve improved performance. The architecture is inspired by U-NET and includes several improvements to the network, including a multihead spatial pyramid cell and Swish activation to achieve optimal dehazing performance. To restore the actual hazy offshore situation, 3D simulation was used for ocean scene model construction, haze rendering, and offshore target animation. High-definition images with different angles, illumination conditions, locations, and other parameters were collected by our simulation of the natural scene. To improve performance, data augmentation methods such as random clipping and flipping were used during preprocessing. The real public outdoor haze dataset (O-Haze [49]) was supplemented to verify model performance. Experiments on the datasets showed that the proposed LKEDN-MHP algorithm was superior to the state-of-the-art algorithms in terms of the PSNR and SSIM indicators.

Future Work. The database and the improved algorithm are two indispensable parts of the maritime image dehazing task. Our future work will focus on the following two aspects.

- Prospective studies based on digital twin technology will be of great significance in effectively promoting the transformation and reconstruction of research techniques

under the new technology in future war situations. We will focus on refining the rendering result of the target under hazy to generate a more realistic world.
- Our work has proven that a large kernel could help to obtain a large ERF. We will focus on improving the architecture to take full advantage of effective information on a maritime image dehazing task.

Author Contributions: Conceptualization, W.Y. and Y.J.; methodology, W.Y. and H.G.; software, W.Y. and X.Z.; validation, W.Y., X.Z., and Y.J.; writing—original draft preparation, W.Y. and Y.J.; writing—review and editing, W.Y., H.G., and X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was jointly supported by the following projects: Program for Liaoning Innovative Talents in University under Grant No. LR2019058, Scientific Research Fund of Liaoning Provincial Education Department under Grant No. 2021-KF-12-05, Program for Liaoning Innovative Talents in University under Grant No. LG202107, Liaoning Revitalization Talents Program under Grant No. XLYC1902095, and Shenyang Science and Technology Bureau under Grant No. RC200386.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: A publicly available dataset was provided by O-Haze [49] in this paper. The data can be found from the github of the above references: http://www.vision.ee.ethz.ch/ntire18/o-haze/(accessed on 18 June 2018).

Acknowledgments: The authors would like to acknowledge support from the following projects: Program for Liaoning Innovative Talents in University under Grant No. LR2019058, Scientific Research Fund of Liaoning Provincial Education Department under Grant No. 2021-KF-12-05, Program for Liaoning Innovative Talents in University under Grant No. LG202107, Liaoning Revitalization Talents Program under Grant No. XLYC1902095, and Shenyang Science and Technology Bureau under Grant No. RC200386.

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

References

1. Song, Y.; He, Z.; Qian, H.; Du, X. Vision Transformers for Single Image Dehazing. *arXiv* **2022**, arXiv:2204.03883.
2. Jeans, M.A.; James, H. On the partition of energy between matter and Æther. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1905**, *10*, 91–98. [CrossRef]
3. Tan, R.T. Visibility in bad weather from a single image. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
4. Tarel, J.P.; Hautière, N. Fast visibility restoration from a single color or gray level image. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Kyoto, Japan, 29 September–2 October 2009; pp. 2201–2208.
5. Yang, W.; Gao, H.W.; Jiang, Y.Q.; Yu, J.H.; Sun, J.; Liu, J.G.; Ju, Z.J. A cascaded feature pyramid network with non-backward propagation for facial expression recognition. *IEEE Sens. J.* **2020**, *10*, 11382–11392. [CrossRef]
6. Yu, J.H.; Gao, H.W.; Yang, W.; Jiang, Y.Q.; Chin, W.H.; Kubota, N.; Ju, Z.J. A discriminative deep model with feature fusion and temporal attention for human action recognition. *IEEE Access* **2020**, *8*, 43243–43255. [CrossRef]
7. Li, P.Y.; Tian, J.D.; Tang, Y.D.; Wang, G.L.; Wu, C.D. Deep retinex network for single image dehazing. *IEEE Trans. Image Process.* **2020**, *30*, 1100–1115. [CrossRef] [PubMed]
8. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.H.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 3–7 May 2021; pp. 1–22.
9. Vaswani, A.; Ramachandran, P.; Srinivas, A.; Parmar, N.; Hechtman, B.; Shlens, J. Scaling local self-attention for parameter efficient visual backbones. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 12894–12904.
10. Chu, X.X.; Tian, Z.; Wang, Y.Q.; Zhang, B.; Ren, H.B.; Wei, X.L.; Xia, H.X.; Shen, C.H. Twins: Revisiting the design of spatial attention in vision transformers. In Proceedings of the 35th Conference on Neural Information Processing Systems, Sydney, Australia, 6–14 December 2021; pp. 9355–9366.
11. Ren, W.Q.; Pan, J.S.; Zhang, H.; Cao, X.C.; Yang, M. Single Image Dehazing via Multi-scale Convolutional Neural Networks with Holistic Edges. *Int. J. Comput. Vis.* **2020**, *128*, 240–259. [CrossRef]
12. Shao, Y.J.; Li, L.; Ren, W.Q.; Gao, C.X.; Shang, N. Domain adaptation for image dehazing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2808–2817.

13. Dong, H.; Pan, J.S.; Xiang, L.; Hu, Z.; Zhang, X.Y.; Wang, F.; Yang, M. 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.
14. Isola, P.; Zhu, J.Y.; Zhou, T.H.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
15. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223-2232.
16. Engin, D.; Genc, A.; Ekenel, H.K. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 825–833.
17. Chaurasia, A.; Culurciello, E. Linknet: Exploiting encoder representations for efficient semantic segmentation. In Proceedings of the IEEE Visual Communications and Image Processing, St. Petersburg, FL, USA, 10–13 December 2017; pp. 1–4.
18. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
19. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko , S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
20. Zheng, S.X.; Lu, J.C.; Zhao, H.S.; Zhu, X.T.; Luo, Z.K.; Wang, Y.B; Fu, Y.W.; Feng, J.F.; Xiang, T.; Torr, P.H.S.; Zhang, L. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6881–6890.
21. Jia, C.; Yang, Y.F.; Xia, Y.; Chen, Y.T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.H.; Li, Z.; Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In Proceedings of the 38th International Conference on Machine Learning, Vienna, Austria, 18–24 July 2021; pp. 4904–4916.
22. Cordonnier, J.B.; Loukas, A.; Jaggi, M. On the relationship between self-attention and convolutional layers. In Proceedings of the Eighth International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020; pp. 1–18.
23. Paul, S.; Chen, P.Y. Vision transformers are robust learners. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; pp. 2071–2081.
24. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.M.; Dollar, P. Designing network design spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10428–10436.
25. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.Y.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks. In Proceedings of the 35th Conference on Neural Information Processing Systems, Sydney, Australia, 6–14 December 2021; pp. 12116–12128.
26. Vaswani, A.; Vaswani, N.; Parmar, N.; Parmar, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31th Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 1–11.
27. Zhu, X.Z.; Cheng, D.Z.; Zhang, Z.; Lin, S.; Dai, J.F. An empirical study of spatial attention mechanisms in deep networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2017; pp. 6688–6697.
28. Dong, Y.H.; Cordonnier, J.B.; Loukas, A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In Proceedings of the 38th International Conference on Machine Learning, Vienna, Austria, 18–24 July 2021; pp. 2793–2803.
29. Han, Q.; Fan, Z.J.; Dai, Q.; Sun, Lei; Cheng, M.M.; Liu, J.Y.; Wang, J.D. On the connection between local attention and dynamic depth-wise convolution. In Proceedings of the Eighth International Conference on Learning Representations, Virtual, 25–29 April 2022; pp. 1–25.
30. Zhao, Y.; Wang, G.; Tang, C.; Luo, C.; Zeng, W.; Zha, Z.-J. A battle of network structures: An empirical study of CNN, Transformer, and MLP. *arXiv* **2021**, arXiv:2108.13002.
31. Wang, W.H.; Xie, E.Z.; Li, X.; Fan, D.P.; 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/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 568–578.
32. Liu, Z.; Lin, Y.T.; Cao, Y.; Hu, H.; Wei, Y.X.; Zhang, Z.; Lin, S.; Guo, B.N. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
33. Li, Y.; Zhang, K.; Cao, J.; Timofte, R.; van Gool, L. LocalViT: Bringing locality to vision transformers. *arXiv* **2021**, arXiv:2104.05707.
34. Wu, H.P.; Xiao, B.; Codella, N.; Liu, M.C.; Dai, X.Y.; Yuan, L.; Zhang, L. Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 22–31.
35. Srinivas, A.; Lin, T.Y.; Parmar, N.; Shlens, J.; Abbeel, P.; Vaswani, A. Bottleneck transformers for visual recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16519–16529.
36. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. In Proceedings of the 31th Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 1–13.

37. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
38. Huang, G.; Liu, Z.; Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
39. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
40. Ding, X.H.; Chen, H.H.; Zhang, X.Y.; Han, J.G.; Ding, G.G. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 578–587.
41. Ba, J.L.; Kiros, J.R.; Hinton, G.E. Layer normalization. *arXiv* **2016**, arXiv:1607.06450.
42. Ephraim, Y.; Malah, D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. Acoust. Speech Signal Process.* **1985**, *33*, 443–445. [CrossRef]
43. Johnson, J.; Alahi, A.; Li, F.F. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 694–711.
44. Lucas, A.; Lopez-Tapia, S.; Molina, R.; Katsaggelos, A.K. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Trans. Image Process.* **2019**, *28*, 3312–3327. [CrossRef] [PubMed]
45. Ancuti, C.O.; Ancuti, C.; Timofte, R.; Vleeschouwer, C.D. O-haze: A dehazing benchmark with real hazy and haze-free outdoor images. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 754–762.
46. Kupyn, O.; Budzan, V.; Mykhailych, M.; Mishkin, D.; Matas, J. DeblurGAN: Blind motion deblurring using conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8183–8192.
47. Eelbode, T.; Bertels, J.; Berman, M.; Vandermeulen, D.; Maes, F. Optimization for medical image segmentation: theory and practice when evaluating with Dice score or Jaccard index. *IEEE Trans. Med. Imaging* **2020**, *39*, 3679–3690. [CrossRef] [PubMed]
48. He, K.M.; Sun, J.; Tang, X.O. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2341–2353. [CrossRef] [PubMed]
49. Zhang, X.Q.; Wang, T.; Wang, J.X.; Tang, G.Y.; Li, Z. Pyramid channel-based feature attention network for image dehazing. *Comput. Vis. Image Underst.* **2020**, *197*, 103003. [CrossRef]
50. Liu, X.H.; Ma, Y.R.; Shi, Z.H.; Chen, J. Griddehazenet: Attention-based multi-scale network for image dehazing. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 7314–7323.
51. Qin, X.; Wang, Z.L.; Bai, Y.C.; Xie, X.D.; Jia, H.Z. 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; pp. 11908–11915.
52. Tran, L.A.; Moon, S.; Park, D.C. A novel encoder–decoder network with guided transmission map for single image dehazing. *Procedia Comput. Sci.* **2022**, *204*, 682–689. [CrossRef]