electronics

MDPI

Article

Attention Enhanced Serial Unet++ Network for Removing Unevenly Distributed Haze

Wenxuan Zhao, Yaqin Zhao *, Liqi Feng and Jiaxi Tang

College of Mechanical and Electronic Engineering, Nanjing Forestry University, Nanjing 210037, China; kir1160323659@outlook.com (W.Z.); dream6182@163.com (L.F.); tangjiaxi@njfu.edu.cn (J.T.)
* Correspondence: zhaoyaqin@njfu.edu.cn

Abstract: The purpose of image dehazing is the reduction of the image degradation caused by suspended particles for supporting high-level visual tasks. Besides the atmospheric scattering model, convolutional neural network (CNN) has been used for image dehazing. However, the existing image dehazing algorithms are limited in face of unevenly distributed haze and dense haze in real-world scenes. In this paper, we propose a novel end-to-end convolutional neural network called attention enhanced serial Unet++ dehazing network (AESUnet) for single image dehazing. We attempt to build a serial Unet++ structure that adopts a serial strategy of two pruned Unet++ blocks based on residual connection. Compared with the simple Encoder–Decoder structure, the serial Unet++ module can better use the features extracted by encoders and promote contextual information fusion in different resolutions. In addition, we take some improvement measures to the Unet++ module, such as pruning, introducing the convolutional module with ResNet structure, and a residual learning strategy. Thus, the serial Unet++ module can generate more realistic images with less color distortion. Furthermore, following the serial Unet++ blocks, an attention mechanism is introduced to pay different attention to haze regions with different concentrations by learning weights in the spatial domain and channel domain. Experiments are conducted on two representative datasets: the large-scale synthetic dataset RESIDE and the small-scale real-world datasets I-HAZY and O-HAZY. The experimental results show that the proposed dehazing network is not only comparable to state-of-the-art methods for the RESIDE synthetic datasets, but also surpasses them by a very large margin for the I-HAZY and O-HAZY real-world dataset.

Keywords: serial Unet++ module; image dehazing; dep learning; unevenly distributed haze

check for updates

Citation: Zhao, W.; Zhao, Y.; Feng, L.; Tang, J. Attention Enhanced Serial Unet++ Network for Removing Unevenly Distributed Haze. *Electronics* **2021**, *10*, 2868. https://doi.org/10.3390/electronics10222868

Academic Editor: Luis Javier García Villalba

Received: 26 October 2021
Accepted: 16 November 2021
Published: 22 November 2021

1. Introduction

When light spreads in dense suspended particles such as fog, haze, smoke, dust, etc., the image information collected by imaging sensors is seriously degraded due to the scattering of the particles, which causes the loss of a large amount of useful information and greatly limits high-level vision tasks. The purpose of image dehazing is to eliminate the influence of the atmospheric environment on image quality, increase the visibility of images, and provide support for downstream vision tasks such as classification, localization, and self-driving systems. In the past few decades, single image dehazing has been widely used for outdoor video surveillance systems, such as highway traffic, forest, and grassland ecology. As a foundational low-level vision task, single image dehazing has gained more and more attention from the computer vision community and artificial intelligence companies over the world.

Numerous image dehazing methods can be divided into traditional methods and learning-based methods in general. Traditional image dehazing algorithms are mostly based on hypothetical models, among which the atmospheric scattering model introduced in [1,2] is one of the most successful models. The atmospheric scattering model can well explain the formation of haze, therefore it also provides a theoretical basis for traditional dehazing algorithms. Unfortunately, this model inevitably causes errors in estimating the transmission map and global atmospheric light. As a result, the quality of the restored image is not satisfactory. Therefore, much prior knowledge [3–11], varying with atmospheric environment, is utilized to improve the performance of the atmospheric scattering model. Among them, the dark channel prior (DCP) [3] dehazing algorithm is the most successful and famous algorithm. By counting a large number of outdoor hazy images, the author finds that, in the color channels of these images, the brightness value of at least one channel is very small or close to zero. Through such a priori knowledge, the DCP algorithm can locate hazy areas and remove this haze combined with the parameter estimation of the atmospheric scattering model.

In recent years, CNN-based deep learning has achieved excellent results in some high-level visual applications [12–14]. At the same time, it also shows great performance in dealing with some low-level visual tasks such as super resolution [15–18]. To avoid the above disadvantages of the atmospheric scattering model, some end-to-end dehazing networks were proposed to estimate the transmission map or directly predict the hazy-free image [19–21]. Compared with traditional methods, the learning-based image dehazing algorithms have shown more valid, significant and robust visualized improvement. For example, Cai et al. [19] introduced a trainable and end-to-end network (called DehazeNet) that generates a haze-free image using the self-learned transmission map. An all-in-one dehazing network called AODNet was presented by Li et al. [20] to jointly estimate the transmission map and global atmospheric light in one framework. In addition, Zhang et al. [21] presented a densely connected pyramid dehazing network, which is also named DCPDN, to access the transmission map through a pyramid network branch as well as to parallelly estimate atmospheric light via another Unet [22]-based branch.

Although many attempts have been made to improve the dehazing performance of learning-based methods, there still exist some factors that set the limitation for these methods, which causes incomplete dehazing and color distortion in the face of unevenly distributed haze and dense haze in real-world scenes. Accordingly, we proposed a novel end-to-end attention enhanced serial Unet++ Dehazing Network for single image dehazing (called AESUnet). The proposed method can directly generate the dehazed image free of the estimation of middle parameters. We present a serial strategy of two Unet++ modules to fully extract features in different resolutions and promote the information fusion. In order to avoid the loss of shallow features, we established residual connections between two Unet++ modules. Since feature extraction is essential for an end-to-end image restoration task, in this paper, the dehazing model utilizes an enhanced Unet (called Unet++ [23])-based architecture to capture the contextual information between different layers and increase the reception field of each pixel. Although the structure of Unet has been applied to the image dehazing algorithm [24,25], to the best of our knowledge, it is the first time introducing the Unet++ structure for single image dehazing. While retaining the excellent performance of Unet in dealing with the low and deep contextual information at the same time and reducing information loss caused by down-sampling and copy-and-crop strategy through long connections, the Unet++ adds more densely short connections and more skip routes, which improves the efficiency of using features in different resolutions. Moreover, the attention module [26] is introduced into the model to enable the network to learn the uneven distribution of haze.

The contributions of this work are summarized as follows:

- We propose a novel end-to-end attention enhanced serial Unet++ dehazing network. The serial Unet++ module extracts features in different resolutions and effectively fuses them to restore thick hazy images. An attention mechanism is introduced to pay different levels of attention to haze regions with different concentrations;
- We build a serial Unet++ structure that is responsible for fully extracting features of different resolutions and reconstructing them on different scales. The serial Unet++ structure directly transmits the original information of the shallow layers to the subsequent deeper layers, so that the deeper layers focus on residual learning while

reusing shallow contextual information. Thus, the structure can not only avoid the degradation of the model, but also fuse shallow contextual information into deep features, which contributes to generating more realistic images with less color distortion in the faces of dense daze regions;

- To remove the haze and restore the image information as much as possible, we take some improvement measures to the Unet++ module. First, the original Unet++ is pruned to avoid expansion of model parameters. Besides, in the down-sampling operation, we replace the simple convolutional layer with the convolutional module with ResNet in order to prevent the loss of original information in the transmission to the deep network;
- The different pixel values in the spatial domain and different feature channels show different sensitivities to haze regions with different concentrations. We introduce the attention module at the bottom of the decoder to assign different weights to different spaces and channels, which helps to pay different levels of attention to haze regions with different concentrations and further enables the network to learn the uneven haze in images.

The rest of the paper is organized in the following way. Section 2 describes recent studies related to our work. Section 3 presents the proposed network, including a Unet++-based structure of the Encoder–Decoder and the learnable attention modules. The experiment results are discussed in Section 4, while the ablation studies are drawn in Section 5. The conclusion and acknowledgement are put at the end.

2. Relate Works

The atmospheric scattering model was firstly introduced by McCartney [1,2] and further developed by Narasimhan [27] and Nayar [28]. It is widely used for describing the formation of hazy images and formulated as:

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

where $I(z)$ is the observed hazy image, $J(z)$ is the recovered hazy-free image, $t(z)$ is the medium transmission map, and A is the global atmospheric light. When the atmospheric light A is homogeneous, the transmission map can be expressed as:

$$t(z) = e^{-\beta d(z)}, \tag{2}$$

where β is the scattering coefficient of the atmosphere, and $d(z)$ represents the scene depth. Given a hazy image I, the target hazy-free image J can be calculated from the above two formulas.

From the formula, we can see that, in order to solve the recovered hazy-free image J, we need to calculate three key parameters correctly. However, in practice, we usually cannot obtain these parameters directly. Therefore, many scholars use different prior knowledge.

He et al. [3] discovered DCP (dark channel prior) according to statistical law to reckon the transmission map. However, DCP will be invalid when it comes to the regions with high brightness. Zhu et al. [4] introduced CAP (color attenuation prior) to describe the relationship among brightness, saturation, and the density of haze. Berman et al. [5] proposed a non-local prior that means the color of a haze-free image can form tight, non-local clusters in RGB space, and their varying distances can translate to different transmission coefficients in the presence of haze. Derived from a local linear model, He et al. further put forward a guided filtering method [6] which is cost-efficient in haze removal without the use of a complex atmospheric model. Although many improved atmospheric models [7–11] have achieved a large amount of success, they also showed the problem of insufficient robustness in dealing with more complex real-world scenes. In the meantime, the prior error is still not be completely avoided, which directly causes color distortions in restored images.

Since image dehazing is a highly ill-posed problem, the existing methods often use strong priors or assumptions as additional constraints to restore the transmission map, global atmospheric light, and scene radiance. Due to unavoidable errors caused by the estimation of some middle parameters, the atmospheric scattering model has been gradually replaced by end-to-end models [29–37] to directly generate the dehazed image. For instance, Surez et al. [24] employed a triplet of Generative Adversarial Network (GAN) [38] to remove the haze on each color channel independently. A GAN-based enhanced pix2pix dehazing network (EPDN) [34] was designed to have a multi-resolution generator and a multi-scale discriminator followed by the pyramid pooling enhancer module. Dong et al. [35] also borrowed the structure of GAN for image dehazing. They introduced the frequency domain information into the generator network as a priori knowledge to deal with the problem of color distortion. Inspired by knowledge distillation, Wu et al. [36] designed a two-stream dehaze network, KTDN, to transfer the knowledge learned from abundant haze-free images. Chen et al. [37] adopted a smoothed dilation technique to help to remove the gridding artifacts and leverage a gated sub-network to fuse the features from different levels. The methods mentioned above have significantly improved the performance of dehazed images; however, these generic methods suffered from the problems of complex models, unevenly distributed haze and insufficient dehazing degree after reconstruction.

The Unet model was first proposed for application in biomedical image segmentation [39–41] and soon stretched to a variety of visual tasks [42,43]. On account of its mirrored down-and-up-sampling structure, the Unet structure can pay more attention to the contextual information in one image and restore the features' scale to the size of the original image, which is significant to the end-to-end tasks. Additionally, long connections are also used to fuse the features extracted by the previous down-sampling parts to the later up-sampling parts with the same resolution. Unet++ redesigns the network by adding more skip routes and short connections between different resolutions. Therefore, such an operation can improve the efficiency of feature utilization and avoid the introduction of too many parameters.

3. Method

3.1. Architecture

In this section, we present the AESUnet in detail, including the pipeline and the stabilization of the whole network, the encoder-decoder structure of serial Unet++ with local residual learning, and the attention module.

3.1.1. Pipeline Overview

The pipeline of the network, shown in Figure 1, consists of two Unet++ blocks connected in serial. The input of this network is hazy images. Two serial Unet++ blocks are responsible for fully extracting features of different resolutions and reconstructing them on different scales. When the output features (shallow features) of the first Unet++ block are passed to the second Unet++ block, they are also passed backward to be concatenated with the output features (deep features) of the second block. Through this residual connection, shallow contextual information can be used again; that is, the serial Unet++ structure allows the original information of the shallow layers to be directly transmitted to the subsequent deeper layers, so that the deeper layers can focus on residual learning and avoid the degradation of the model. After obtaining the concatenated features, we take an attention module to pay different attention to haze regions with different concentrations and adopt two convolutional layers to reduce the channels to three. At last, we add the original hazy images to the finally extracted feature channels and obtain the haze-free images.

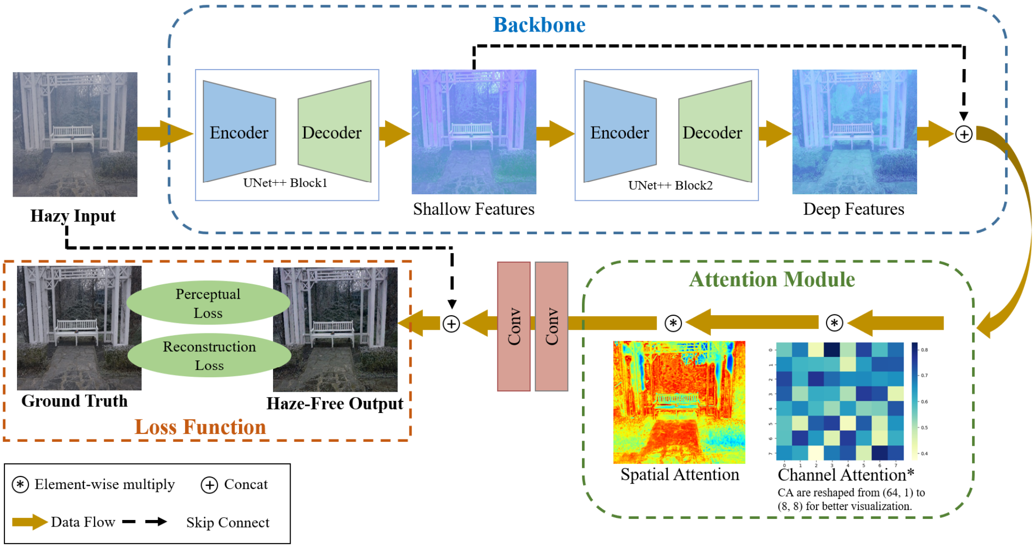

Figure 1. The whole structure of the AESUnet including two Unet++ blocks, an attention module, two convolutional layers, and some skip connections between them. The network is a fully end-to-end structure. Two Unet++ structures are used to extract shallow features and deep features, respectively, and these features are concatenated in channel dimension. Attention module is then employed to have the network learn the distribution of haze. Finally, perceptual loss and reconstruction loss are used to help the training process.

3.1.2. Encoder-Decoder of the Serial Unet++ Structure

In order to remove the haze and restore the image as much as possible, the feature extractor must make full use of the information in one image. Inspired by several previous dehazing networks that utilized the encoder–decoder structure as the feature extractor and achieved great performance, we build a serial Unet++ block with an encoder–decoder structure. Especially, we use a variant of the original Unet model called Unet++, which adds more short connections and skip routes to promote the information for contacting and fusing. As shown in Figure 2, different from the original Unet++, we do some pruning to the model. Specifically, since the input patches were resized to 256 × 256 pixels, we cut the deepest layer of the Unet++ and just keep three layers to down-sample the resolution to 1/8 scale. Therefore, the expression of the j-th feature at the i-th layer is formulated as:

$$x^{i,j} = \mathscr{G}\left(x^{i,0} \oplus x^{i,1} \oplus \ldots \oplus x^{i,j-1} \oplus up\left(x^{i+1,j-1}\right)\right), \tag{3}$$

where $\mathscr{G}(\cdot)$ means convolution layer, $up(\cdot)$ means Upsampling operation, and $\oplus$ represents the concat operation.

Moreover, in the Down-sampling operation, a convolutional module with ResNet [44] structure is used to replace the simple convolutional layer. As shown in Figure 3a, the Down-sampling operation contains three convolutional layers, and immediately after every convolutional layer are batch normalization (BN) and ReLU layers. To keep the gradient from dispersion, a residual learning strategy is introduced. The input features transferred from the upper encoder are pooled to half size and fed to the first two convolutional layers. Further information extracted by two series of convolutional, batch normalization (BN), and ReLU layers are then added to the input and sent to the next convolutional layer together. The structure of the Up-sampling operation is similar to that of the Down-sampling operation, as shown in Figure 3b, except that the pool operation is replaced by interpolation to restore the feature size to the original resolution. Through assigning different weight to different spaces and channels, the attention module at the bottom of the decoder helps to learn the uneven distribution of haze.

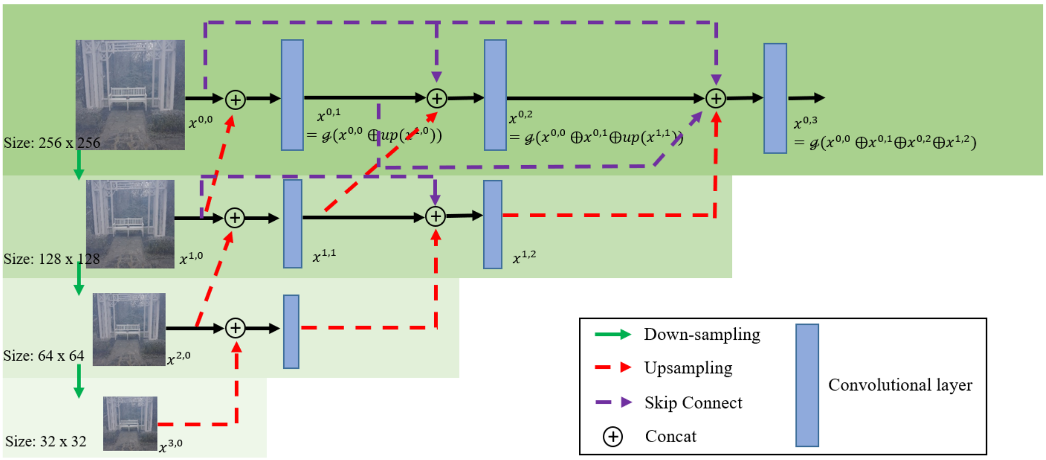

Figure 2. The encoder–decoder architecture of the serial Unet++ block. Compared to the original Unet++ model, we reduced the layers of the network from four to three for the consideration of less parameters and resolution of the input. The three-layer down-sampling structure reduces the image resolution to one eighth of the original size. Through the Unet++ structure, the features of different resolutions can be fully fused, which gives the model the ability to capture deep features and retain shallow features.

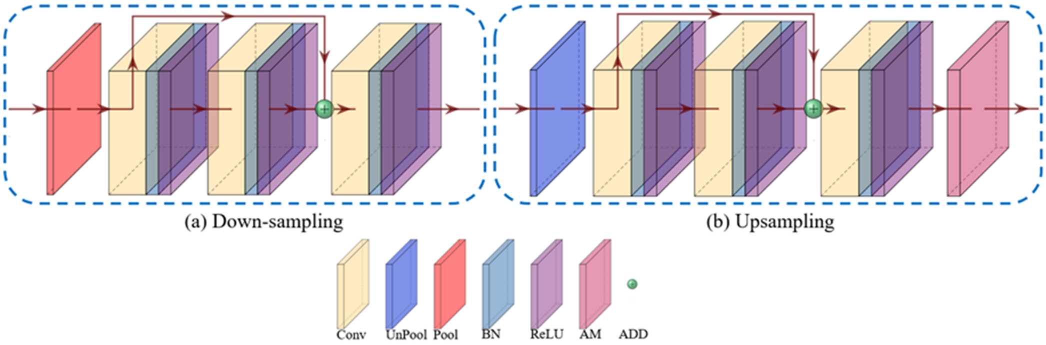

Figure 3. The detailed structure of the Down-sampling operation and Up-sampling operation in serial Unet++ module. Compared to original network, we replaced the convolutional layer with residual convolutional layer. After the Down-sampling operation or the Up-sampling operation, the size of the features is reduced to half of the original size or doubled accordingly. An attention module is added at the bottom of the Up-sampling operation to facilitate learning the distribution of haze in different spaces or channels.

3.1.3. Attention Mechanism

In most cases, the distribution of haze is uneven, especially for dense haze. This makes it difficult to apply CNN-based dehazing network to the real scene. At the same time, different feature channels also have different sensitivities to haze regions with different concentrations. Therefore, assigning different weights to corresponding channels also has an effect on the dehazing performance. Many works [45–47] have applied the attention mechanism to the Unet structure and achieved good results in different visual tasks. Inspired by [48,49], we introduce the attention mechanism into our network so that it can focus more on the dense haze areas when the distribution of haze is uneven in one image. As shown in Figure 4, in the process of keeping input features passed back, channel attention and spatial attention are multiplied in turn to obtain refined features as the output of the feature module.

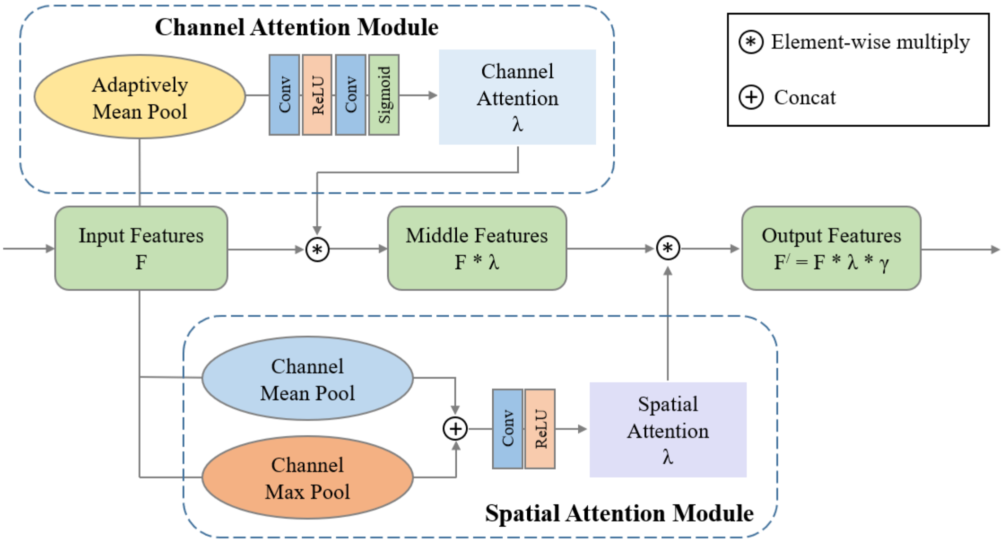

Figure 4. The structure of attention module. In the process of forwarding the input features, the channel attention and spatial attention are added in turn. Therefore, the network has the ability to give bigger weight to the important information and has more flexibility to deal with the unevenly distributed haze in image.

In the channel attention module (see Figure 4), we first adopted an adaptive mean pooling operation to obtain the raw weight of each channel. Through the adaptive mean pooling operation, for the feature map of size H × W × C, we extract a feature matrix of size 1 × 1 × C, where each value is a weight of all the pixel values in the corresponding feature map. Then, the raw weights are sent to a learning module consisting of one convolutional layer, a ReLU activation function unit followed by the other convolutional layer and Sigmoid activation function. Finally, the learned feature weights are channel-wise multiplied into the input features so that different channels have different degrees of attention to the haze.

After the channel attention module, a spatial attention module (see Figure 4) is employed to measure the degree of attention to different locations of the feature map. We first perform the max-pooling and mean-pooling operations along the channel axis on the feature map fused with channel attention. In this way, two spatial attention maps of H × W × 1 are obtained from the original feature map of H × W × C. Immediately after concatenating them, a convolutional layer and Sigmoid activation function are utilized to learn the distribution of haze in the whole image. At last, the spatial attention map is pixel-wise multiplied into the input features. In summary, the attention feature is computed as:

$$\begin{gathered} F' = F * \lambda * \gamma, \\ \lambda = \delta(Conv(\sigma(Conv(AMP(F))))), \\ \gamma = \sigma(Conv(CAT[Max(F),\ Mean(F)])), \end{gathered} \tag{4}$$

where F is the input features of the attention module, and F' is the output features with channel attention and spatial attention. $\delta(\cdot)$ is the Sigmoid activation function and $\sigma(\cdot)$ is the ReLU activation function. $AMP(\cdot)$ represents the adaptively mean pooling operation and $CAT(\cdot)$ means the concatenation in channel dimension. *Max* and *Mean* mean max-pooling and mean-pooling operations, respectively.

3.2. Loss Function

We use reconstruction loss L_r and perceptual loss L_p as the composition of the integrated loss function and formulate them as:

$$L_{total} = \alpha L_r + \beta L_p, \tag{5}$$

Reconstruction loss measures the mean absolute error (MAE), which is also called L1 Loss, between ground truth and corresponding image, and is formulated as:

$$L_r = \frac{1}{n} \sum_i L1(G(I_i) - J_i), \tag{6}$$

where I_i is the input hazy image, $G(\cdot)$ means the network operating on the input, and J_i is the corresponding ground truth.

Perceptual loss was proposed in [50] to measure the perceptual similarity in features space and calculate the mean square error, also called L2 Loss. The $vgg(\cdot)$ means pretrained VGG16 [51] network. It is defined as:

$$L_p = \frac{1}{n} \sum_i L2(vgg(G(I_i)) - vgg(J_i)), \tag{7}$$

Finally, we use a weight combination of the two loss functions. In this work, the parameters α, β are set to 1, 1 correspondingly.

4. Experiments

In this section, we will introduce the datasets used in training and testing our network. At the same time, the detail parameters of the training process are given. Finally, we compare the results of the network with several representative methods in the same objective metrics.

4.1. Datasets and Metrics

Similar to the existing learn-based dehazing methods, we utilized two of the most commonly used dehazing datasets, RESIDE datasets [52] and I-HAZY and O-HAZY image dehazing datasets [53,54], for training our model.

The RESIDE dataset is a large-scale benchmark consisting of both synthetic and real-world hazy images. It is divided into five subsets, each serving different training or evaluation purposes. In our experiment, we used Indoor Train Set (ITS) and Outdoor Train Set (OTS) as training datasets, and Synthetic Objective Testing Set (SOTS) for evaluating. In ITS, there are 10,000 different hazy indoor images and 10 corresponding synthesized hazed images to each. In OTS, there are 8970 different hazy outdoor images and 35 corresponding synthesized hazed images to each. Therefore, there are, totally, 100,000 images in ITS and 313,950 in OTS. In SOTS, there are 500 hazed images and their corresponding ground truth images are used for calculating the metrics partly.

Compared with the RESIDE dataset, I-HAZY and O-HAZY datasets are real-world datasets. I-HAZY and O-HAZY datasets are proposed to address the limitation that is currently considered, both for assessment and training of learning-based dehazing techniques which all exclusively rely on synthetic hazy images. They are composed of pairs of real hazy and corresponding haze-free images. The real hazy images are all generated by professional haze machines and captured under the same illumination parameters with the corresponding haze-free images, therefore they are closer to the actual application. I-HAZY datasets have 30 images, among which 25 are for training and 5 are for evaluating. O-HAZY dataset have 45 images, among which 40 are for training and the rest are for evaluating.

To objectively evaluate the performance of the proposed method, we adopt two metrics widely used in image dehazing task: the Peak Signal to Noise Ratio (PSNR) and the Structural Similarity index (SSIM).

PSNR is the most common and widely used image objective evaluation index, and it is based on the error between corresponding pixels, which is an error-sensitive based image quality evaluation matric. PSNR can be formulated as:

$$PSNR = 10log_{10}(\frac{(2^n - 1)^2}{MSE}), \tag{8}$$

where n represents the bit width of the pixel. MSE stands for the mean absolute error and it can be formulated as:

$$MSE = \frac{1}{hw} \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} \| I_{i,j} - J_{i,j} \|, \tag{9}$$

where h, w mean the height and weight of the image, and $I_{i,j}$, $J_{i,j}$ mean the pixel value of the input hazy image and the corresponding output haze-free image at position $(i,\ j)$.

SSIM is also a full-reference image quality evaluation index, which measures image similarity from three aspects: brightness, contrast, and structure. SSIM can be formulated as:

$$\text{SSIM} = l(x,y) * c(x,y) * s(x,y), \tag{10}$$

where l, c, and s stand for brightness, contrast, and structure, respectively.

To evaluate and compare the proposed model with previous methods from a more comprehensive perspective, except for the above two most commonly used reference subjective evaluation metrics, we also selected two additional evaluation metrics: Natural Image Quality Evaluator (NIQE), a non-reference image quality index, and Natural Image Quality Evaluator (LPIPS) [55], a subjective evaluation index.

The design idea of NIQE is to construct a series of features to measure image quality and use these features to fit a multivariate Gaussian model. These features are extracted from some simple and highly regular natural landscapes. The smaller the value of NIQE, the more the characteristics of the image conform to the natural image with high rules, which means that its quality is better. LPIPS uses the similarity measurement of high-dimension image structure to replace the distance measurement that cannot be formed in practice, which means the difference of pixel values is not always consistent with people's subjective perception. In practical use, LPIPS uses the deep network pre-trained on ImageNet datasets to extract the deep features of images and reference images. The lower the LPIPS value, the higher the feature similarity between the generated image and the corresponding reference image, and the more similar the subjective perception.

4.2. Implement Details

We implement our framework in Pytorch 1.7.1 and train our model in a computer equipped with a RTX 2080Ti GPU and an Intel i9-9900K CPU. We utilize ADAM [56] as an optimizer where β_1 and β_2 are set to 0.9 and 0.999. The default learning rate is set to 0.0001. To better adjust the learning rate, we adopt CosineAnnealingLR [57] as a scheduler.

Every image is randomly rotated by 0°, 90°, 180°, or 270° and flipped horizontally with 50% probability to improve the robustness of the model and prevent overfitting. The batch size is set to 2 and the thread of the CPU is set to 16. Other hyperparameters are different in training different datasets. When training on RESIDE dataset, we randomly take a pair of images from the dataset and train our network for 1,000,000 iterations. All the patches transferred to the network are resized to 256 × 256pixels. In I-HAZY and O-HAZY datasets, all the images are resized to 512 × 512, and the size of patches transferred to the network is also 256 × 256. Due to the small number of samples in the datasets, we train for 125,000 iterations on each dataset. Code will be made available at https://github.com/kirqwer6666/Image-dehazing-pytorch.

4.3. Experiments Results

We compare the proposed method with several representative and state-of-the-art methods: DCP [3], AODNet [20], DCPDN [21], FD-GAN [37], and GCANet [30].

4.3.1. Experiment on Synthetic RESIDE Datasets

The experimental results of our method AESUnet and the other comparative methods on the RESIDE dataset are shown in Figure 5 and Table 1. As shown in Table 1, AESUnet can achieve state-of-the-art performance in all four metrics. The performance of AESUnet on the indoor RESIDE dataset is comparable with GCANet. Moreover, when it comes to the outdoor dataset, AESUnet can reach significant improvements over the other comparative methods.

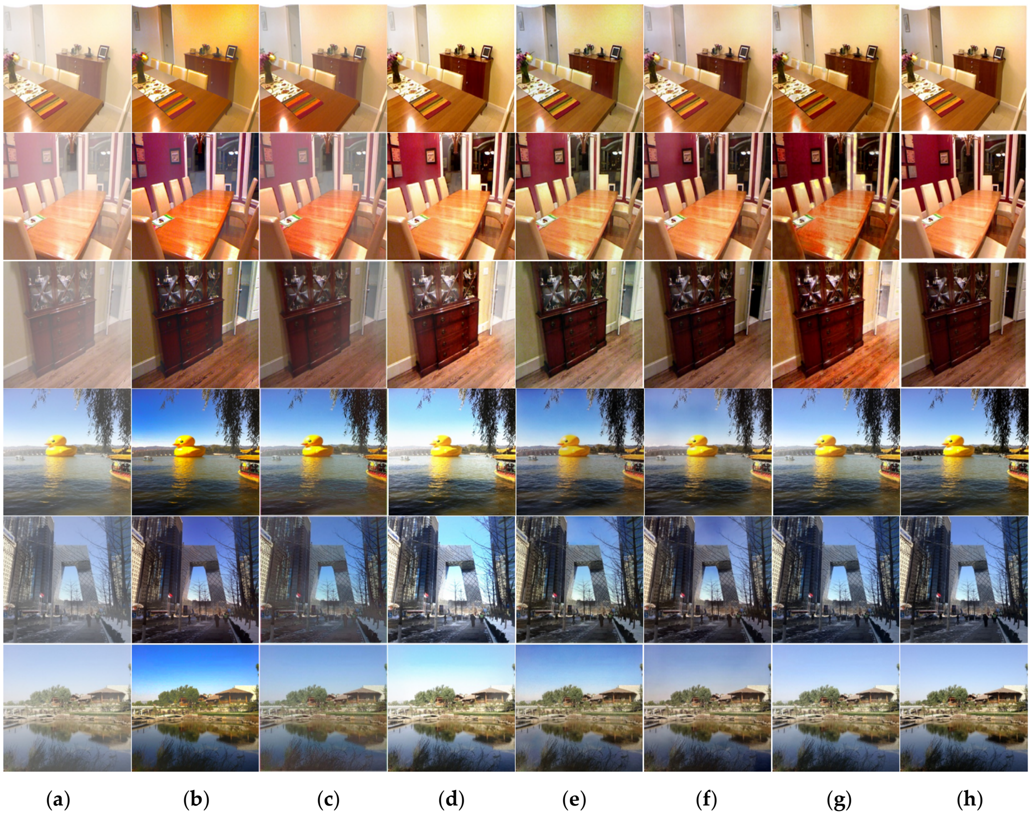

Figure 5. Comparisons on SOTS dataset. (**a**) Hazy, (**b**) DCP [3], (**c**) AODNet [20], (**d**) DCPDN [21], (**e**) FDGAN [37], (**f**) GCANet [30], (**g**) ours, and (**h**) GT.

Table 1. Metrics comparisons of the dehazing results on SOTS dataset. In this table, "↑" and "↓" respectively mean that the larger the metric, the better and the smaller the metric, the better. The best results are shown in bold.

Method	**Indoor**				**Outdoor**			
	PSNR (dB) ↑	**SSIM** ↑	**LPIPS** ↓	**NIQE** ↓	**PSNR (dB)** ↑	**SSIM** ↑	**LPIPS** ↓	**NIQE** ↓
DCP	16.62	0.8179	0.268	\	19.13	0.8148	0.257	\
AODNet	19.06	0.8504	0.228	13.5746	20.29	0.8765	0.243	12.6212
DCPDN	19.98	0.8565	0.243	13.9711	20.67	0.9098	0.239	13.2775
FDGAN	23.15	0.9207	0.203	13.7742	23.43	0.9285	0.212	14.0714
GCANet	**30.23**	**0.9800**	**0.176**	13.8669	28.13	0.9450	**0.184**	12.4203
Ours	29.60	0.9549	0.200	**11.6144**	**30.75**	**0.9639**	0.191	**10.9311**

Specifically, as seen from Figure 5, the DCP method can achieve relatively soft visual performance, but in the face of areas with high brightness such as sky (the image in the first row) and wall (the image in the sixth row), it causes serious color distortion compared with ground truth. The convolutional module used in AODNet is at the same resolution, therefore its ability to characterize features is weak. AODNet's dehazing performance is not thorough enough, which makes the images still present a hazy sense. The feature extraction module used in DCPDN comprehensively considers different resolutions, and multi-scale fusion is used in feature reconstruction. However, due to the error accumulation of parameter estimation in the atmospheric scattering model, DCPDN will bring obvious color distortion. This error is more obvious when the haze becomes thicker (see Figures 5d and 6d). Although DCPDN has achieved good results in some images, there is still much color distortion that cannot be ignored, and a large amount of haze remains in some areas with high-density haze, such as the lower right area of the image in the third row. FDGAN and GCANet use more advanced feature extraction modules. The former uses a GAN network with a deep encoder–decoder structure, while the latter uses a gated fusion module to optimize the features extracted by encoder–decoder structure. FDGAN and GCANet perform well in the indoor dataset; however, the effect of dehazing is not ideal in the outdoor dataset, especially in the areas with obvious gradient changes, such as the junction of objects and the sky.

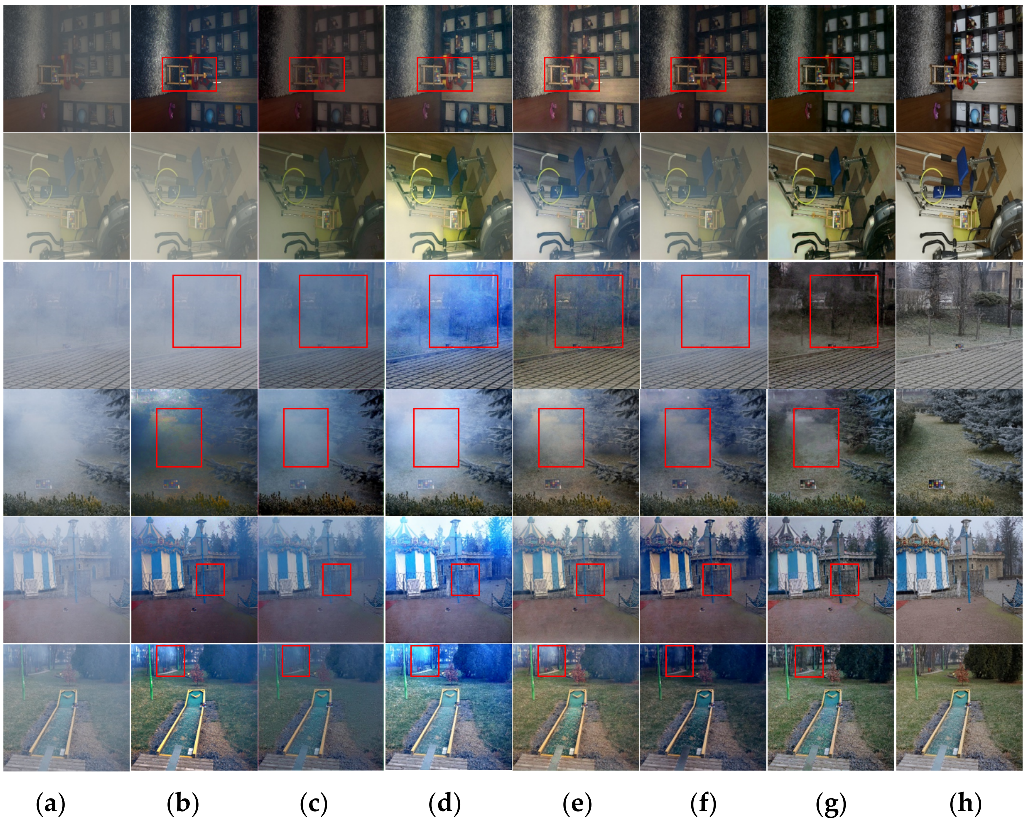

Figure 6. Comparisons on I-HAZY and O-HAZY dataset. (**a**) Hazy, (**b**) DCP [3], (**c**) AODNet [20], (**d**) DCPDN [21], (**e**) FDGAN [37], (**f**) GCANet [30], (**g**) ours, and (h) GT.

In comparation, the dehazed images generated by AESUnet are not only more visually faithful and closer to the ground truth, but the color is changed more smoothly even in the areas with dense haze.

4.3.2. Experiment on Real-World I-HAZY and O-HAZY Datasets

Compared to the RESIDE datasets, the advantage of our method is more obvious on more challenging I-HAZY and O-HAZY datasets. As shown in Table 2, we reach the best performance and surpass the second place by a very large margin, 4.425 dB in PSNR and 0.028 in SSIM as the average. As for LPIPS and NIQE, we also achieved the best results.

Table 2. Metrics comparisons of the dehazing results on I-HAZY and O-HAZY dataset. In this table, "↑" and "↓" respectively mean that the larger the metric, the better and the smaller the metric, the better. The best results are shown in bold.

Method	**I-HAZY**				**O-HAZY**			
	PSNR (dB) ↑	**SSIM** ↑	**LPIPS** ↓	**NIQE** ↓	**PSNR (dB)** ↑	**SSIM** ↑	**LPIPS** ↓	**NIQE** ↓
DCP	14.43	0.752	0.333	\	16.78	0.653	0.411	\
AODNet	13.98	0.732	0.374	10.7116	15.03	0.539	0.445	12.6648
DCPDN	16.21	0.755	0.274	15.3483	15.16	0.673	0.377	18.3160
FDGAN	17.82	**0.757**	0.224	13.9222	18.38	0.682	0.289	14.2454
GCANet	14.95	0.719	0.207	11.3894	16.28	0.645	0.259	12.2723
Ours	**22.08**	0.728	**0.197**	**10.8302**	**22.97**	**0.767**	**0.206**	**12.0232**

As seen from Figure 6, some previous methods, such as DCP, AODNet, and DCPDN, totally fail in this dehazed task of the real-world dataset. Due to the lack of attention module to learn the distribution of uneven haze, FDGAN and GCANet have a certain effect but are accompanied by serious degradation in dealing with the unevenly distributed haze. Stacking more parameters in the feature extraction module (14.07 M for FDGAN and 9.61 M for GCANet) does not bring qualitative change to deal with uneven haze. As marked with red boxes in Figure 6, because of the dense haze attached on the surfaces of some objects, the outlines and texture details cannot be clearly seen on the object surfaces of the FDGAN and GCANet results in row 1 and row 5. In addition, it is worth mentioning that FDGAN and GCANet are not enough to completely restore the original color of the images covered by heavily dense haze. Finally, as shown in row 2 and row 3 in Figure 6, in the face of dense and uneven haze, FDGAN and GCANet almost completely fail. The lack of ability to capture deeper features makes them unable to recover the image information better. Compared with these methods, our model can not only adaptively remove haze in both low-density and high-density areas to the greatest extent, but can also restore more outlines and texture details with less color distortion.

5. Ablation Study

In order to analyze the effectiveness of each module in the proposed network, we conduct the ablation study by consideration of two main factors: (1) Unet + + structure compared with ordinary Unet structure; (2) strategy of module series connection; (3) attention mechanism. Therefore, we design three different models in the ablation study:

- AES-simple-Unet: Serial Unet-based block with attention module;
- AEUnet: Single Unet++ block with attention module;
- SUnet: Serial Unet++ block without attention module.

In order to avoid the positive effect caused by parameter stacking, we adjust the convolutional layers of the three models in the ablation study so that their flops and parameters are almost the same. When calculating the flops and parameters, the size of the input is set to 1 × 3 × 256 × 256. We trained the models in the RESIDE outdoor dataset and tested it in the SOTS outdoor dataset. Other hyperparameter settings are also consistent.

As shown in Table 3, the three factors can bring a significant improvement to the network. This promotion mainly comes from the mechanism of these factors rather than the stacking of parameters. In particular, the introduction of the attention module can bring more obvious performance improvement compared with the Unet++ structure. The performance improvement of serial strategy is greater than the first two. On the one hand, this is due to the positive impact of parameters; on the other hand, it is brought by feature fusion between shallow and deep layers. The results are also reflected in Figure 7. Due to the lack of short connections and more skip routes in the Unet++ structure, although more convolutional layers are added to extract features, the AES-simple-Unet also performs poorly in some regions compared with AESUnet. In the red box of Figure 7b, the color of the sky area around the sun is clearly divided into three layers, while, in Figure 7e, the color changes more naturally and smoothly. By comparison, as marked with the red box in

Figure 7, the image generated by AESUnet is closer to the ground truth. The performance of AEUnet is the worst among the four models in the ablation study. It can be clearly seen from the Figure 7c that the color distortion in the sky area is very serious. This is because a single UNet++ block cannot capture deeper color information well to restore the original color. As for SUnet, because of the lack of attention module, the haze in the high-density areas is left on the image (marked with green box in Figure 7d), which seriously degrades the visual performance, and AESUnet can better take effect (marked with green box in Figure 7e).

Table 3. Ablation study results.

	Flops (G)	Params (M)	PSNR (dB)	SSIM
AES-simple-Unet	173.417	14.10	27.39	0.9586
AEUnet	263.218	15.79	23.26	0.8873
SUnet	234.478	14.05	26.20	0.9117
AESUnet	234.216	14.05	30.74	0.9639

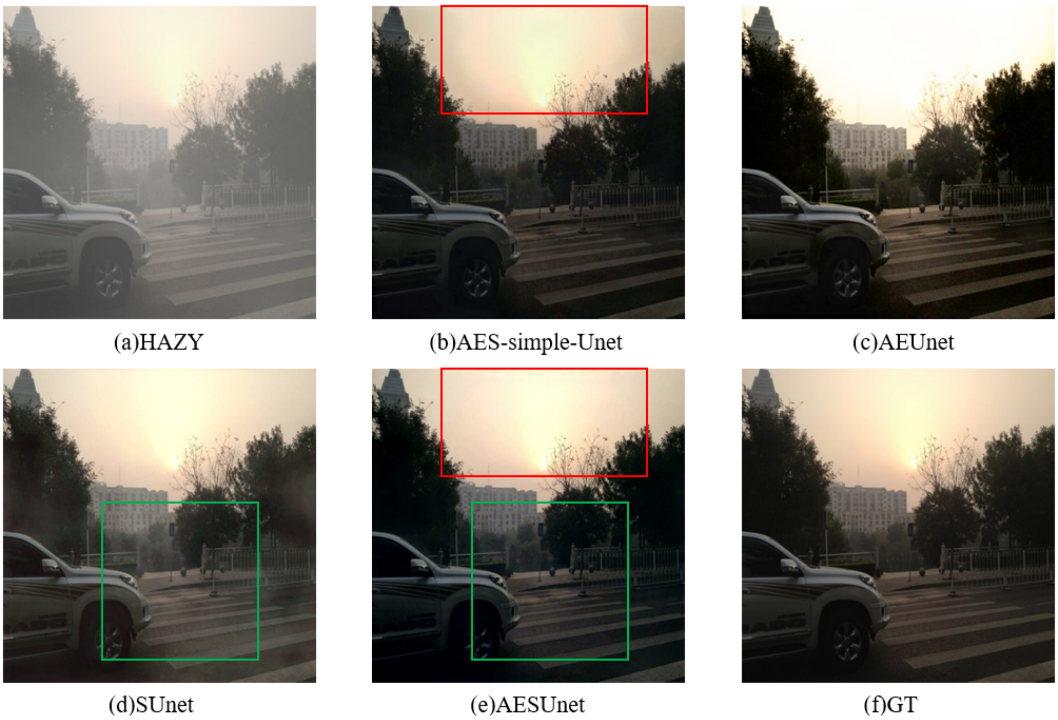

Figure 7. Comparisons of different models in ablation study.

6. Conclusions and Future Work

In this paper, we propose a fully end-to-end Convolution Neural Network called Attention Enhanced Serial Unet++ Dehazing Network (AESUnet) for single image dehazing. To fully make use of the features extracted, we employ serial structure of two Unet++ blocks to replace the simple Encoder–Decoder structures. Moreover, the attention module is introduced to help the network learn the distribution of the uneven haze. Compared with the existing dehazing methods, AESUnet can better remove the dense haze in images with less color distortion. Experiments on both synthetic and real-world datasets show that our method can achieve state-of-the-art performance and generate more visually pleasing results in the image dehazing task.

Although the proposed model can achieve quite pleasing dehazing performance on real-world datasets, it is still worth discussing in the following aspects. Firstly, because the proposed model adopts the end-to-end Unet++ architecture, it will be very computationally expensive, which is not conducive to its industrial deployment. Replacing the heavy-weight Unet structure with a more light-weight structure such as MobileNet [58] will have broader application prospects [59]. Secondly, the running speed of the proposed model is not fast enough to meet the needs of real-time operation. In our experiment, the model can only process images in the RESIDE dataset at a speed of 14 FPS. Therefore, the model still needs to be improved to meet the needs of video dehazing. Considering the existing video image processing methods [60–62], feature fusion methods such as key frame, nearest neighbor frame, or time attention can be used to speed up the processing speed of video frames.

Author Contributions: Conceptualization, W.Z.; Data curation, W.Z.; Formal analysis, W.Z., Y.Z. and L.F.; Funding acquisition, W.Z. and Y.Z.; Investigation, W.Z. and Y.Z.; Methodology, W.Z.; Project administration, W.Z., Y.Z. and L.F.; Resources, W.Z. and J.T.; Software, W.Z.; Supervision, L.F. and J.T.; Validation, J.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by National Natural Science Fund, grant number No. 31200496.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Mccartney, E.J. Scattering phenomena. (book reviews: Optics of the atmosphere. scattering by molecules and particles). *Science* **1977**, *196*, 1084–1085.
2. Cartney, E.J. *Optics of the Atmosphere: Scattering by Molecules and Particles*; John Wiley and Sons, Inc.: New York, NY, USA, 1976; 421p.
3. He, K.; Sun, J.; Tang, X. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *33*, 2341–2353. [PubMed]
4. Zhu, Q.; Mai, J.; Shao, L. Single image dehazing using color attenuation prior. In *BMVC*; Citeseer: University Park, PA, USA, 2014.
5. Berman, D.; Avidan, S. Non-local image dehazing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 27–30 June 2016; pp. 1674–1682.
6. He, K.; Sun, J.; Tang, X. *Guided Image Filtering[C]//European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 1–14.
7. Fattal, R. Dehazing using color-lines. *ACM Trans. Graph.* **2014**, *34*, 13. [CrossRef]
8. Jiang, Y.; Sun, C.; Zhao, Y.; Yang, L. Image dehazing using adaptive bi-channel priorson superpixels. *Comput. Vis. Image Underst.* **2017**, *165*, 17–32. [CrossRef]
9. Ju, M.; Gu, Z.; Zhang, D. Single image haze removal based on the improved atmospheric scattering model. *Neurocomputing* **2017**, *260*, 180–191. [CrossRef]
10. Meng, G.; Wang, Y.; Duan, J.; Xiang, S.; Pan, C. Efficient image dehazing with boundary constraint and contextual regularization. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 617–624.
11. Riaz, S.; Anwar, M.W.; Riaz, I.; Kim, H.-W.; Nam, Y.; Khan, M.A. Multiscale Image Dehazing and Restoration: An Application for Visual Surveillance. *Comput. Mater. Contin.* **2021**, *70*, 1–17. [CrossRef]
12. Jin, X.; Che, J.; Chen, Y. Weed Identification Using Deep Learning and Image Processing in Vegetable Plantation. *IEEE Access* **2021**, *9*, 10940–10950. [CrossRef]
13. Khan, M.A.; Akram, T.; Zhang, Y.-D.; Sharif, M. Attributes based skin lesion detection and recognition: A mask RCNN and transfer learning-based deep learning framework. *Pattern Recognit. Lett.* **2021**, *143*, 58–66. [CrossRef]
14. Gao, J.; Chen, Y.; Wei, Y.; Li, J. Detection of Specific Building in Remote Sensing Images Using a Novel YOLO-S-CIOU Model. Case: Gas Station Identification. *Sensors* **2021**, *21*, 1375. [CrossRef] [PubMed]
15. Dong, C.; Loy, C.C.; He, K.; Tang, X. *Learning a Deep Convolutional Network for Image Super-Resolution[C]//European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014; pp. 184–199.
16. Xie, C.; Liu, Y.; Zeng, W.; Lu, X. An improved method for single image super-resolution based on deep learning. *Signal Image Video Process.* **2019**, *13*, 557–565. [CrossRef]
17. Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Loy, C.C.; Qiao, Y.; Tang, X. Esrgan: Enhanced super-resolution generative adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
18. Christian, L.; Lucas, T.; Ferenc, H.; Jose, C.; Andrew, C.; Alejandro, A.; Andrew, A.; Alykhan, T.; Johannes, T.; Zehan, W.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
19. 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]
20. 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.

21. Zhang, H.; Patel, V.M.; Patel, V.M.; 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, 18–23 June 2018; pp. 3194–3203.
22. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015.
23. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; pp. 3–11.
24. Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; Yang, M.H. Multi-scale boosted dehazing network with dense feature fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
25. Chen, D.; He, M.; Fan, Q.; Liao, J.; Zhang, L.; Hou, D.; Yuan, L.; Hua, G. Gated context aggregation network for image dehazing and deraining. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, 7–11 January 2019.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
27. Narasimhan, S.G.; Nayar, S.K. Chromatic framework for vision in bad weather. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, SC, USA, 15 June 2000; CVPR 2000 (Cat. No.PR00662). Volume 1, pp. 598–605.
28. Narasimhan, S.G.; Nayar, S.K. Vision and the atmosphere. *Int. J. Comput. Vis.* **2002**, *48*, 233–254. [CrossRef]
29. Ren, W.; Liu, S.; Zhang, H.; Pan, J.; Cao, X.; Yang, M.H. Single image dehazing via multi-scale convolutional neural networks. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016.
30. Ren, W.; Ma, L.; Zhang, J.; Pan, J.; Cao, X.; Liu, W.; Yang, M.H. Gated fusion network for single image dehazing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
31. Zhu, H.; Peng, X.; Chandrasekhar, V.; Li, L.; Lim, J.-H. DehazeGAN: When Image Dehazing Meets Differential Programming. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Stockholm, Sweden, 13–19 July 2018; pp. 1234–1240.
32. Pang, Y.; Nie, J.; Xie, J.; Han, J.; Li, X. BidNet: Binocular image dehazing without explicit disparity estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
33. Suarez, P.L.; Sappa, A.D.; Vintimilla, B.X.; Hammoud, R.I. Deep learning based single image dehazing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018.
34. Qu, Y.; Chen, Y.; Huang, J.; Xie, Y. Enhanced pix2pix dehazing network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
35. Dong, Y.; Liu, Y.; Zhang, H.; Chen, S.; Qiao, Y. FD-GAN: Generative adversarial networks with fusion-discriminator for single image dehazing. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 10729–10736.
36. Wu, H.; Liu, J.; Xie, Y.; Qu, Y.; Ma, L. Knowledge transfer dehazing network for nonhomogeneous dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 478–479.
37. Shao, Y.; Li, L.; Ren, W.; Gao, C.; Sang, N. Domain adaptation for image dehazing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
38. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, W.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. Available online: https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf (accessed on 26 October 2021).
39. Huang, H.; Lin, L.; Tong, R.; Hu, H.; Zhang, Q.; Iwamoto, Y.; Han, X.; Chen, Y.-W.; Wu, J. Unet 3+: A full-scale connected Unet for medical image segmentation. In Proceedings of the ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020.
40. Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.W.; Heng, P.A. H-DenseUnet: Hybrid densely connected Unet for liver and tumor segmentation from CT volumes. *IEEE Trans. Med Imaging* **2018**, *37*, 2663–2674. [CrossRef]
41. Yan, W.; Wang, Y.; Gu, S.; Huang, L.; Yan, F.; Xia, L.; Tao, Q. The domain shift problem of medical image segmentation and vendor-adaptation by Unet-GAN. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019.
42. Guo, S.; Yan, Z.; Zhang, K.; Zuo, W.; Zhang, L. Toward convolutional blind denoising of real photographs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
43. Jiang, K.; Wang, Z.; Yi, P.; Chen, C.; Huang, B.; Luo, M.; Ma, Y.; Jiang, J. Multi-scale progressive fusion network for single image deraining. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Juan, PR, USA, 27–30 June 2016.
45. Li, C.; Tan, Y.; Chen, W.; Luo, X.; Gao, Y.; Jia, X.; Wang, Z. Attention Unet++: A Nested Attention-Aware U-Net for Liver CT Image Segmentation. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020.

46. Khanh, T.L.B.; Dao, D.-P.; Ho, N.-H.; Yang, H.-J.; Baek, E.-T.; Lee, G.; Kim, S.-H.; Yoo, S.B. Enhancing u-net with spatial-channel attention gate for abnormal tissue segmentation in medical imaging. *Appl. Sci.* **2020**, *10*, 5729. [CrossRef]
47. Oktay, O.; Schlemper, J.; Le Folgoc, L.; Lee, M.; Heinrich, M.; Misawa, K.; Mori, K.; McDonagh, S.; Hammerla, N.Y.; Kainz, B.; et al. Attention u-net: Learning where to look for the pancreas. *arXiv* **2018**, arXiv:1804.03999.
48. 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; Volume 34, pp. 11908–11915.
49. Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; Fu, Y. Image super-resolution using very deep residual channel attention networks. In Proceedings of the European Conference on Computer Vision (ECCV) Munich, Germany, 8–14 September 2018; pp. 286–301.
50. Glorot, X.; Bordes, A.; Bengio, Y. Deep sparse rectifier neural networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Ft. Lauderdale, FL, USA, 11–13 April 2011; pp. 315–323.
51. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
52. Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; Wang, Z. Reside: A benchmark for single image dehazing. *arXiv* **2017**, arXiv:1712.04143, 1.
53. Ancuti, C.; Ancuti, C.O.; Timofte, R.; De Vleeschouwer, C. I-HAZE: A dehazing benchmark with real hazy and haze-free indoor images. In Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems, Poitiers, France, 24–27 September 2018; pp. 620–631.
54. Ancuti, C.O.; Ancuti, C.; Timofte, R.; De Vleeschouwer, C. O-haze: A dehazing benchmark with real hazy and haze-free outdoor imag-es. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–22 June 2018; pp. 754–762.
55. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
56. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
57. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
58. 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.
59. Zhou, Y.; Jing, W.; Wang, J.; Chen, G.; Scherer, R.; Damaševičius, R. MSAR-DefogNet: Lightweight cloud removal network for high resolution remote sensing images based on multi scale convolution. *IET Image Process.* **2021**, 1–10. [CrossRef]
60. Bai, Z.; Li, Y.; Chen, X.; Yi, T.; Wei, W.; Wozniak, M.; Damasevicius, R. Real-time video stitching for mine surveillance using a hybrid image registration method. *Electronics* **2020**, *9*, 1336. [CrossRef]
61. Zhu, X.; Wang, Y.; Dai, J.; Yuan, L.; Wei, Y. Flow-guided feature aggregation for video object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
62. Deng, J.; Pan, Y.; Yao, T.; Zhou, W.; Li, H.; Mei, T. Relation distillation networks for video object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.