



Article

Raindrop-Removal Image Translation Using Target-Mask Network with Attention Module

Hyuk-Ju Kwon  and Sung-Hak Lee 

School of Electronic and Electrical Engineering, Kyungpook National University, 80 Deahakro, Buk-Gu, Daegu 41566, Republic of Korea; olin1223@knu.ac.kr

* Correspondence: shak2@ee.knu.ac.kr; Tel.: +82-53-950-7216

Abstract: Image processing plays a crucial role in improving the performance of models in various fields such as autonomous driving, surveillance cameras, and multimedia. However, capturing ideal images under favorable lighting conditions is not always feasible, particularly in challenging weather conditions such as rain, fog, or snow, which can impede object recognition. This study aims to address this issue by focusing on generating clean images by restoring raindrop-deteriorated images. Our proposed model comprises a raindrop-mask network and a raindrop-removal network. The raindrop-mask network is based on U-Net architecture, which learns the location, shape, and brightness of raindrops. The rain-removal network is a generative adversarial network based on U-Net and comprises two attention modules: the raindrop-mask module and the residual convolution block module. These modules are employed to locate raindrop areas and restore the affected regions. Multiple loss functions are utilized to enhance model performance. The image-quality assessment metrics of proposed method, such as SSIM, PSNR, CEIQ, NIQE, FID, and LPIPS scores, are 0.832, 26.165, 3.351, 2.224, 20.837, and 0.059, respectively. Comparative evaluations against state-of-the-art models demonstrate the superiority of our proposed model based on qualitative and quantitative results.

Keywords: raindrop removal; U-Net; attention mechanism; generative adversarial network

MSC: 68T45



Citation: Kwon, H.-J.; Lee, S.-H. Raindrop-Removal Image Translation Using Target-Mask Network with Attention Module. *Mathematics* **2023**, *11*, 3318. <https://doi.org/10.3390/math11153318>

Academic Editors: Hongang Qi, Yan Liu and Jun Miao

Received: 21 June 2023

Revised: 24 July 2023

Accepted: 26 July 2023

Published: 28 July 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Image processing is used in various fields such as autonomous driving, surveillance cameras, and multimedia. In particular, it is widely used in preprocessing algorithms to improve the performance of models. To obtain good images for recognition and analysis, favorable lighting conditions and a suitable image-capturing environment are crucial. However, selecting an ideal place and time for capturing images is not always feasible. For instance, for images obtained in poor weather conditions such as fog, snow, or rain, it can be challenging for algorithms to differentiate between objects and the background. Research is required to enhance object-recognition performance by restoring images captured in poor weather conditions. The objective of this study is to restore raindrop-deteriorated images and generate clean images.

Image-to-image translation using deep learning can be categorized into models based on convolutional neural networks (CNNs) and those based on generative adversarial networks (GANs) [1]. Image-to-image translation using CNNs involves extracting features from an image and generating a transformed image based on these features. CNNs typically comprise convolutional and pooling layers, which extract low- and high-level features from input images. These features are then used to perform various image transformations using various layers. GANs comprise a generator and a discriminator. The generator is trained to generate images with the desired characteristics, while the discriminator is trained to distinguish between real and generated images. The generator and discriminator compete with each other during training to generate more realistic images.

The representative image-to-image translation models are Pix2Pix [2] and CycleGAN [3]. Pix2Pix is based on GAN, which comprises a generator and discriminator, and combines L1 loss and adversarial loss to enhance image-conversion performance. CycleGAN was developed to overcome the limitations of Pix2Pix, which requires a paired dataset of input and target images for training and has difficulty in obtaining a large dataset. CycleGAN is designed to perform bidirectional image transformation between two different domains. CycleGAN comprises two generators and two discriminators. The first generator transforms an image from one domain into another domain, while the second generator transforms the image back into an image from the original domain. The first discriminator discriminates the difference between the original domain and generated domain image, while the second discriminator discriminates the difference between the reconverted image and original image. Pix2Pix and CycleGAN can be used for raindrop removal in images. In addition, various other models are available for raindrop removal.

Qian et al. [4] proposed an attentive generative adversarial network (ATTGAN) that utilizes adversarial training to address the issues of raindrops overlapping with the background and the resulting loss of background information. ATTGAN employs an LSTM model to identify raindrop regions in an image. The generated raindrop area is utilized as a visual attention mechanism in the generator to focus on the raindrops and their surrounding structures. In the discriminator, the attention map is incorporated into the network to assess the local consistency of the restored image.

Alletto et al. [5] proposed a spatiotemporal architecture for removing raindrops in images and utilized computer graphics to insert realistic virtual raindrop images to supplement the limited raindrop data; moreover, they proposed a competitive single-image deraining baseline to gather information about the raindrop region. Optical flow and image-synthesis techniques were then applied to enhance the performance of the raindrop-removal model.

Quan et al. [6] proposed a joint shape-channel attention mechanism based on CNN for raindrop removal. This attention mechanism leverages the physical characteristics of raindrops to enhance model performance.

Shao et al. [7] proposed a selective skip connection GAN (SSCGAN) for restoring raindrop-degraded images. SSCGAN utilizes gated recurrent units to capture raindrop information. A selective connection model is employed to extract a raindrop binary mask. Self-attention blocks are then utilized to focus on the global structure of raindrops. The generator focuses on the global structure of raindrops and enhances the performance of the raindrop binary mask.

Anwar et al. [8] proposed a single-stage blind real-image-restoration network (R²Net) for restoring real degraded photographs. R²Net incorporates a residual structure to alleviate the flow of low-frequency information and employs feature attention to exploit channel dependencies.

Yang [9] increased the number of layers in the visual attentive-recurrent network of ATTGAN to prevent gradient sparsity; such modification allows the network to generate raindrop removal images more reliably.

Xia et al. [10] proposed a hierarchical supervision network to enhance the balance between raindrop removal and image inpainting; this model combines dense network blocks with U-Net architecture and inserts a dense block into the skip connection of U-Net. By applying a loss function to each layer, the model achieves improved performance without a substantial increase in the number parameters.

Chen et al. [11] proposed a GAN model based on different learning to remove raindrops. The generative network learns the difference between images with raindrops and clean ones, leveraging the simpler distribution of raindrop scenes. The final raindrop-free image is obtained by subtracting this learned difference from the original raindrop image.

Xu et al. [12] proposed raindrop removal from transmission lines based on unmanned-aerial-vehicle inspection. They employed an attention-recurrent network to generate the

raindrop attention map. Additionally, a generation countermeasure network based on GAN was used to remove raindrops from the images.

In this study, we propose a model to effectively remove raindrops from images and restore the areas affected by the raindrops. The proposed model comprises a raindrop-mask network and raindrop-removal network. The raindrop-mask network is a CNN model based on U-Net [13] architecture, which learns the location, shape, and brightness of raindrops. U-Net has also been widely adopted in image translation, super-resolution imaging, and image enhancement [14–19]. The generated raindrop mask serves as the attention mechanism for the raindrop-removal network [4,20]. The raindrop-removal network is a GAN model that combines U-Net and an attention mechanism; it utilizes two attention modules to identify the raindrop area and restore the degraded regions. The first attention module is the raindrop mask, which is used as an input along with the raindrop image. The second attention module is the residual convolution block module (RCAM) [21], a self-attention mechanism that is inserted before the max-pooling layers of U-Net. Furthermore, various loss functions are employed to enhance model such as the adversarial loss [1], perceptual loss [22,23], structural-similarity-index-measure (SSIM) loss, and multiscale-mean-square-error (MSE) loss [4]. We compared the performance of the proposed model with the state-of-the-art models [2,4,8,24], utilizing not only quantitative but also qualitative comparisons through six image quality index metrics [25–29]. The contributions of our model are as follows:

- The proposed model utilizes two networks to separate the raindrop-mask network and raindrop-removal network.
- The raindrop-mask network serves as an attention module to accurately represent the location, size, and brightness of raindrops. The raindrop-mask network is based on U-Net and learns the raindrop-mask area in the raindrop image by training on the difference between the raindrop image and clean image.
- The raindrop-removal network is based on GAN, and the attention mechanisms are applied to the input and the internal layers of the generator. The input attention of the generator is the raindrop mask, while the internal attention is the residual convolution block attention module (RCBAM). These two modules contribute to enhancing the performance of the raindrop-removal network.

2. Related Works

2.1. U-Net

U-Net [13] is a deep-learning model that is widely used in image segmentation, particularly for segmenting various cellular objects in medical images. U-Net has also been widely adopted in image translation, super-resolution imaging, and image enhancement [14–19]. Moreover, it comprises an encoder, a decoder, and a skip connection. The encoder comprises convolutional layers and a max-pooling layer; the max-pooling layer is applied after the convolution layers. This design facilitates hierarchical extraction of features from an image, converting the high-level abstract features of an image into low-dimensional feature maps. The decoder uses the feature map obtained from the encoder and generates an output that matches the size of the original image. The decoder consists of a convolutional layer and an upconvolutional layer that increases the size of the feature map. The skip connection is used to concatenate the feature map between the encoder and decoder. Furthermore, the skip connection effectively restores information that may be lost in the hierarchical structure when transferring the feature map from the encoder to the decoder. Figure 1 illustrates the structure of U-Net.

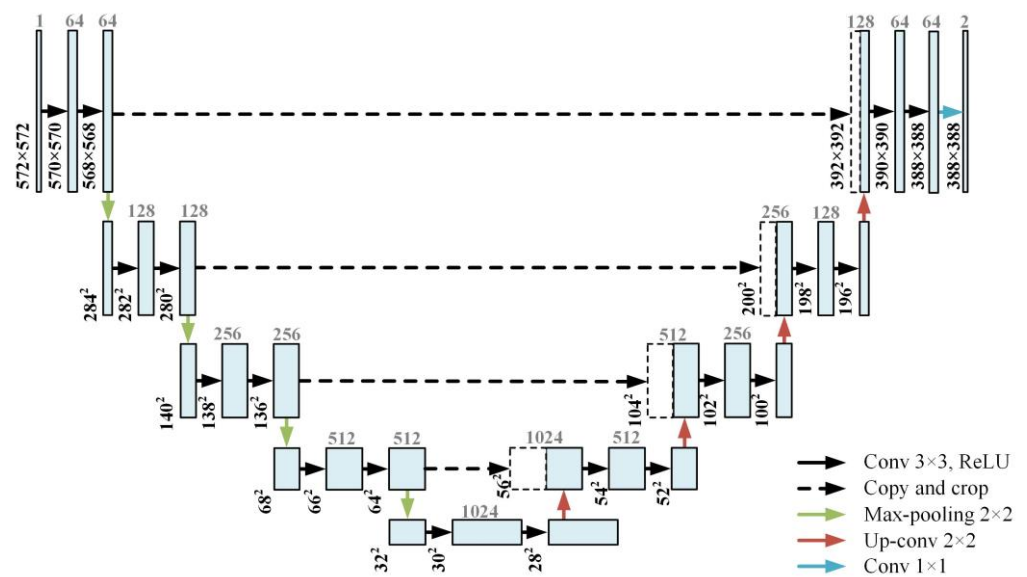


Figure 1. Architecture of U-Net.

2.2. Convolution Block Attention Module

Attention is a method of using the information of each position separately by assigning different weights to each position in the input image. This improves the localization and expression of the region of interest in the image. The performance of a CNN model can be improved by incorporating attention mechanisms. The convolution block attention module (CBAM) [21] is an attention module applicable to feed-forward CNN. CBAM comprises channel- and spatial-attention modules. The channel-attention module is used to find areas with important meaning in the image, while the spatial-attention module is used to find the position of a meaningful region in the image. The formulas presented below represent the channel- and spatial-attention modules of CBAM. Figure 2 illustrates the diagrams of CBAM and RCBAM. RCBAM represents the integration CBAM with the residual block using a skip connection. RCBAM is a CBAM-enhanced network. The equations for CBAM are as follows:

$$CA(F) = \text{sigmoid}(\text{MLP}(\text{AvgPool}(F)) + \text{MLP}(\text{MaxPool}(F))), \quad (1)$$

$$SA(F) = \text{sigmoid}(\text{conv}^{7 \times 7}([\text{AvgPool}(F); \text{MaxPool}(F)])), \quad (2)$$

where $CA(\cdot)$ and $SA(\cdot)$ represent the channel- and spatial-attention modules, respectively; F represents a feature map; $\text{sigmoid}(\cdot)$ represents a sigmoid function; $\text{MLP}(\cdot)$ represents a multilayer perceptron; and $\text{conv}^{7 \times 7}(\cdot)$ represents a convolution operation with a filter size of 7×7 .

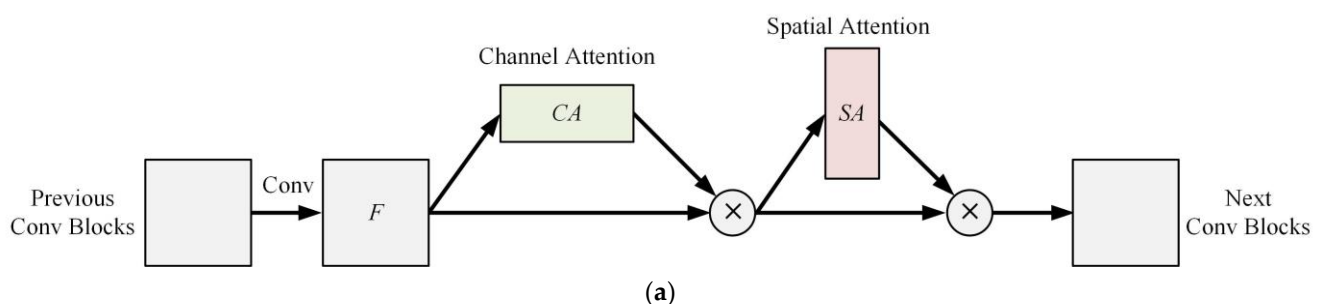


Figure 2. Cont.

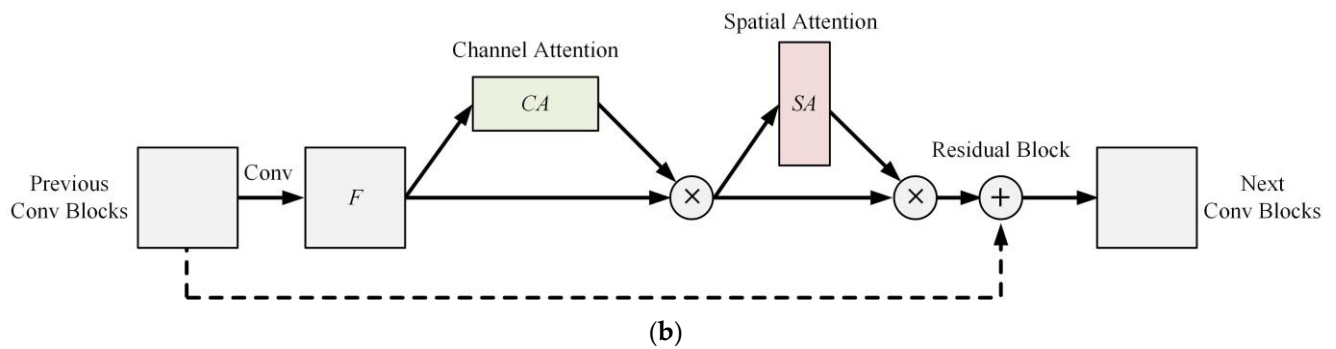


Figure 2. Diagrams of convolution block attention modules (CBAMs): (a) CBAMs; (b) residual convolution block attention module (RCBAM).

3. Proposed Method

In this study, we propose an image-translation model for effectively removing raindrops present in an image. The proposed model has three main components: data processing, raindrop-mask-generation network, and raindrop-removal network. Figure 3 illustrates the flow chart of the proposed model. The data-processing component preprocesses the input images, while the raindrop-mask-generation network generates masks to identify the raindrop regions. Furthermore, the raindrop-removal network utilizes the generated mask and real raindrop image to remove the raindrops from the image.

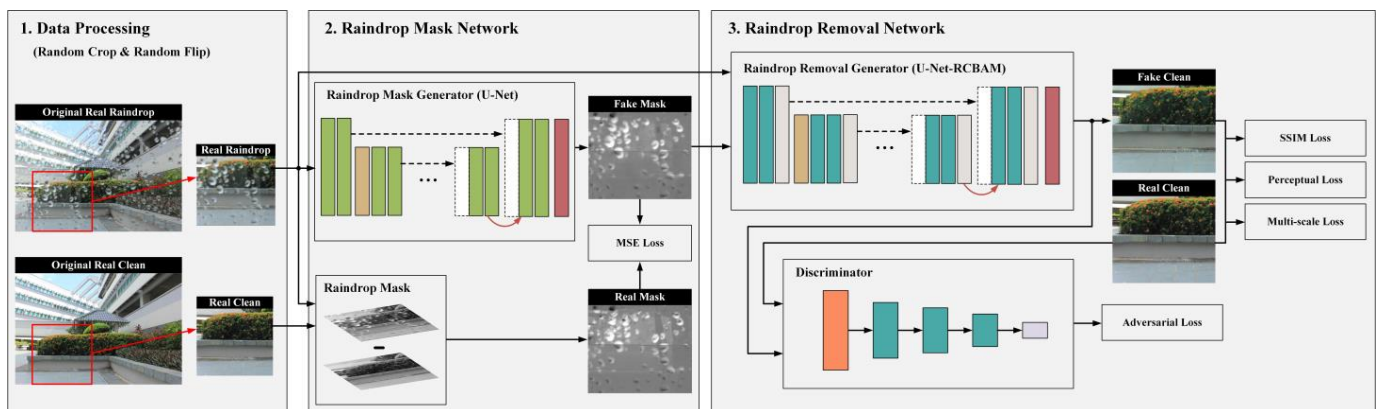


Figure 3. Flow chart of the proposed method.

3.1. Data Processing

Data processing is a preprocessing step that transforms the input images into a format that is useful for training. Deep-learning models require large amounts of data for effective training. However, obtaining many paired images for raindrop removal is difficult. Therefore, data-augmentation techniques are required to increase the amount of training data. In this study, data augmentation is performed using random cropping and random vertical- and horizontal-flip methods. Random cropping involves randomly selecting a 256×256 region of the original resolution image and creating various images from different positions. The copped images are flipped horizontally and vertically with a 50% probability. Data augmentation is applied during each batch of training. These methods generate a large number of datasets from a limited number of the datasets.

3.2. Raindrop-Mask Network

Visual-attention models have proven to be highly effective in localizing specific regions within an image to capture their distinctive features. This concept has found applications in visual recognition and classification tasks. Visual attention plays a crucial role in removing raindrops from images and generating clean versions of images. By incorporating visual

attention mechanisms into the network, the visual attention models gain the ability to identify areas that require the removal or restoration of raindrops. This not only enhances the overall performance of the model but also allows it to focus its efforts where they are most needed.

The raindrop-mask network serves as the second stage. To generate the raindrop mask, we made modifications to U-Net architecture. During the training phase, the input images comprise raindrop images, while the output images correspond to raindrop masks. These raindrop masks are utilized in the attention map for the raindrop-removal network. Existing models employ binary masks or absolute difference as the attention maps in the raindrop-removal process [4,20]. However, binary images often fall short in accurately representing the brightness and size variations of raindrops. This is because binary images are generated by applying a threshold to the absolute difference between raindrop and clean images. Important details regarding the intensity, size, and shape of raindrops are affected by the choice of the thresholding level. The absolute difference changes the intensity level of the raindrop region owing to the absolute value of the negative difference, which can impact the effectiveness of the raindrop removal task. To address this limitation, we propose using the raindrop mask, which is derived from the difference between the raindrop and clean images. This approach enables us to accurately represent the size and intensity levels of each individual raindrop.

Figure 4 shows a comparison between the proposed raindrop mask and the binary mask for an input raindrop image. The proposed mask is biased to display the negative intensity level. The red box in the figure indicates the cropped region of the image. In Figure 4a, the input image displays raindrops with varying intensity levels, including an intensity region that is darker than the surrounding area. In Figure 4b, the binary mask fails to capture the dark intensity region present in the input image due to thresholding and absolute difference representation. In the absolute difference mask of Figure 4c, the brightness of the dark raindrop region is reversed, resulting in the dark region appearing bright. However, in Figure 4d, the proposed raindrop mask accurately represents the shape and intensity variations of the raindrops, including the dark intensity region. The equations for the raindrop mask are as follows:

$$f(I) = 0.299 \times r + 0.587 \times g + 0.114 \times b, \quad (3)$$

$$M = f(I_r) - f(I_c), \quad (4)$$

where $f(\cdot)$ represents the conversion function that transforms a color image I into a grayscale image; r , g , and b represent the red, green, and blue channels of the color image, respectively; I_r and I_c represent the raindrop image and the clean image, respectively; and M represents the raindrop mask.

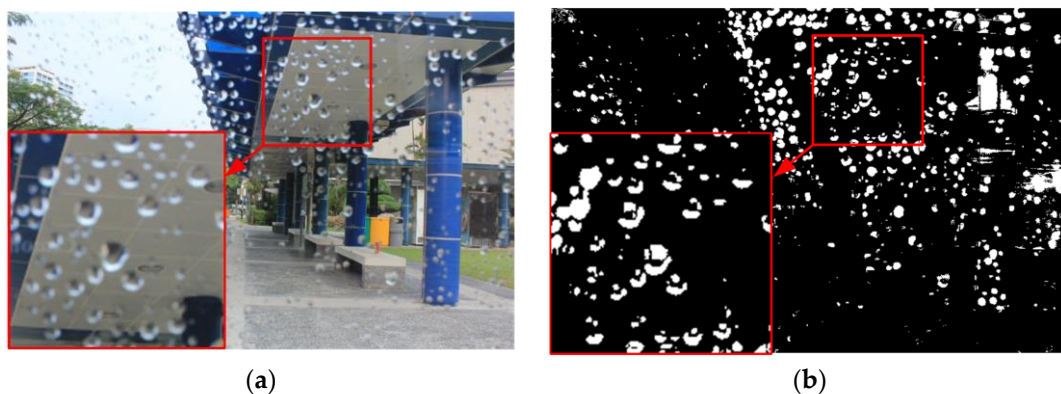


Figure 4. Cont.

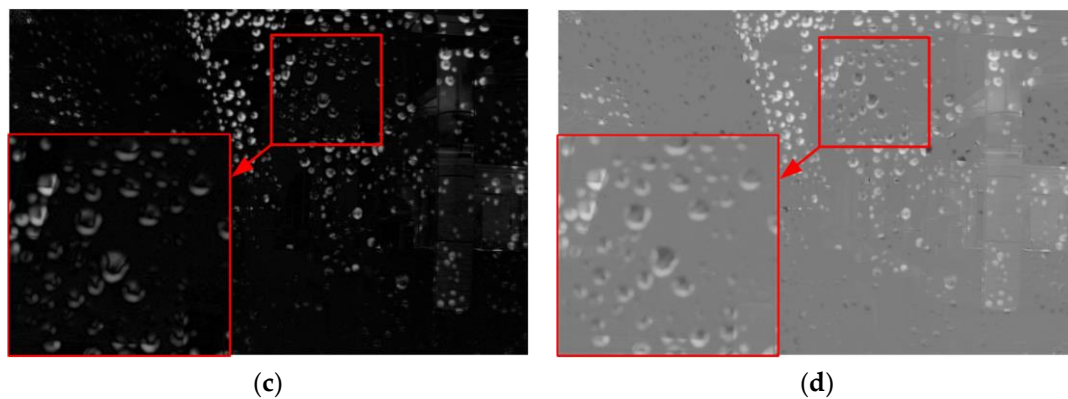


Figure 4. Comparison between binary mask and the raindrop mask: (a) real raindrop image, (b) binary image with thresholding level 30, (c) absolute difference, and (d) proposed raindrop mask. The red rectangle represents a magnified image of the raindrop area.

Figure 5 shows the architecture of the raindrop-mask network. This model comprises several components: the CBR block, the skip connection, and concatenation as well as the downsampling, upsampling, and convolution layers. The CBR block comprises the convolution layer, batch normalization, and the ReLU activation function. The downsampling and upsampling layers use max-pooling and two-dimensional (2D) transposed convolution layers, respectively. The skip connection connects the previous layers to the upsampled layers. This connection allows the network to learn the low-level and high-level features and facilitates the integration of detailed information with abstract representation. The skip-connected layers concatenate the previous layers and upsampled layer channel-wise. The convolution layer is applied without the ReLU activation function. All convolution layers in the rain-mask network use 3×3 spatial filters with a stride of 1 and padding of 1. The max-pooling and transposed convolution layers use a 2×2 spatial filter with a stride of 2. The input raindrop images are a color image with three channels and output raindrop masks are a grayscale image with a single channel. Each image is normalized from -1 to 1 . The loss is calculated using mean square error:

$$\mathcal{L}_m = \mathbb{E}[(M - M')^2], \quad (5)$$

where \mathcal{L}_m represents the raindrop mask loss and M and M' represent real raindrop mask and generated raindrop mask, respectively.

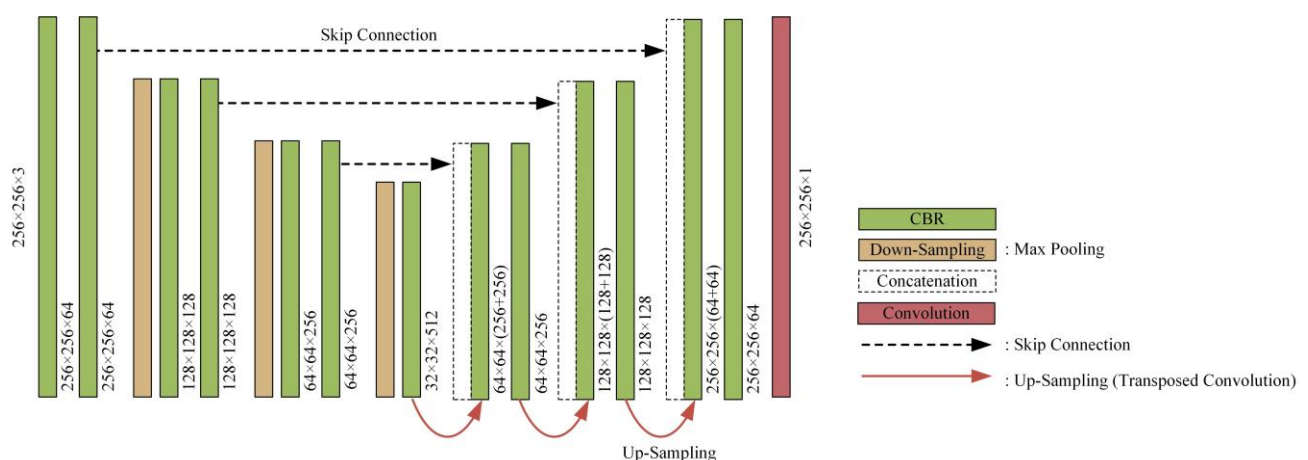


Figure 5. Architecture of the raindrop-mask network.

Figure 6 shows an example of the generated raindrop mask. Figure 6a,b depict the raindrop image and the clean image, respectively. Figure 6c shows the raindrop mask

obtained using Equation (4). Figure 6d shows the generated raindrop mask using the raindrop-mask network. The result demonstrates that the real mask and generated mask are almost identical.

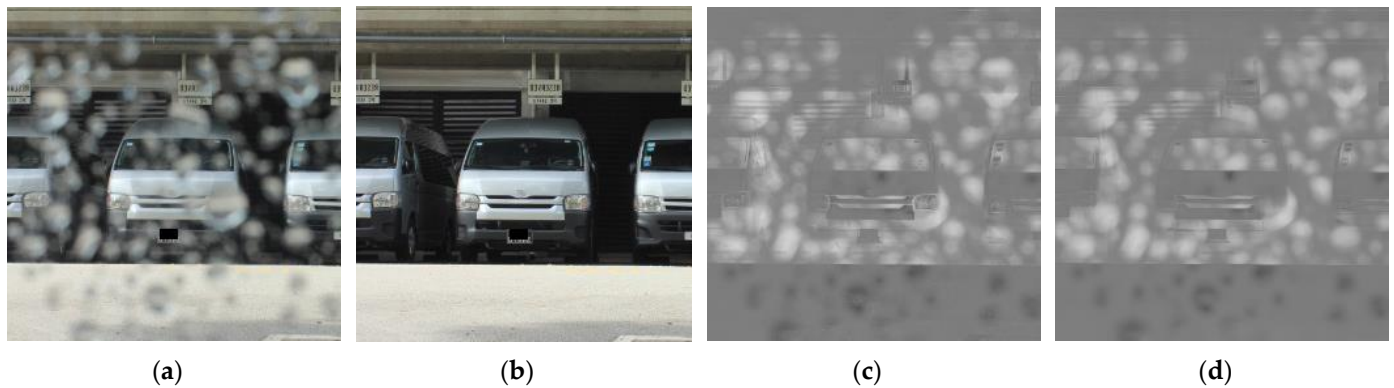


Figure 6. Example of the raindrop mask: (a) real raindrop image, (b) real clean image, (c) real raindrop mask, (d) generated raindrop mask.

3.3. Raindrop-Removal Network

The raindrop-removal network is based on a GAN, which comprises a generator and discriminator. The generator is constructed using U-Net and incorporates the self-attention module. The discriminator adopts the PatchGAN architecture used in Pix2Pix [2]. Figure 7 shows the architecture of the proposed model.

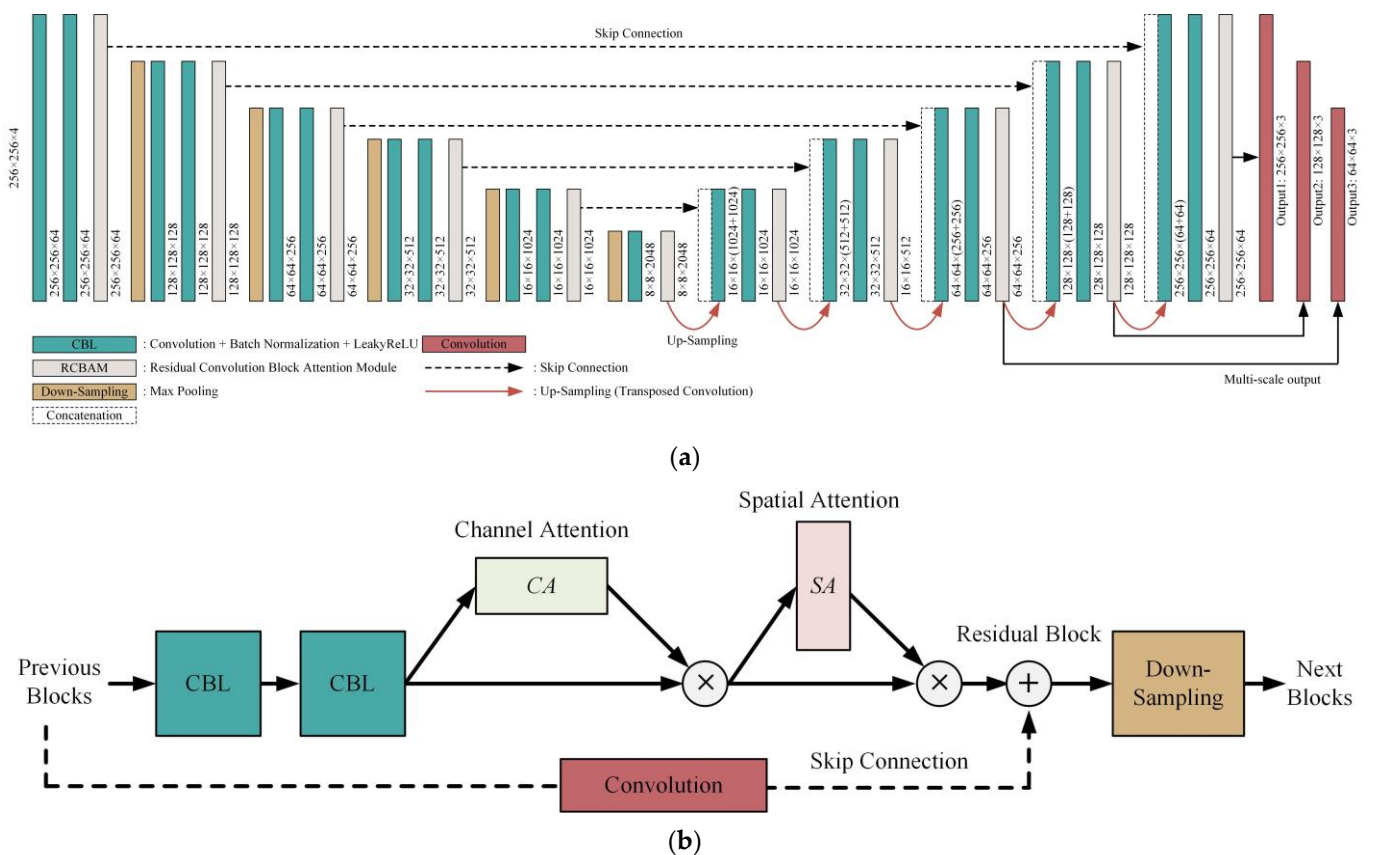


Figure 7. Cont.

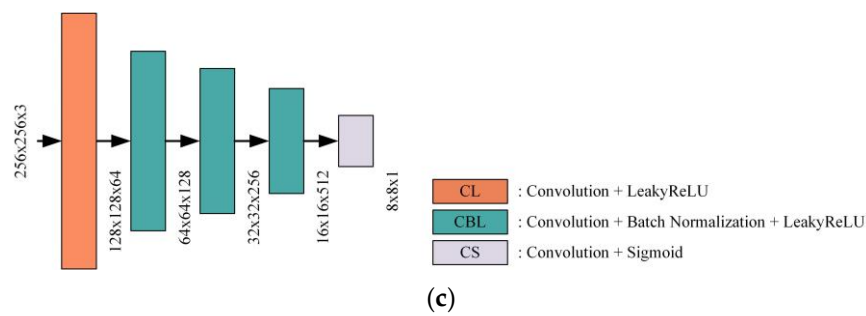


Figure 7. Architecture of the rain-removal network: (a) generator, (b) residual convolution block attention module, (c) discriminator.

According to GAN, the proposed adversarial loss can be expressed as follows:

$$\min_G \max_D V(D, G) = \mathbb{E}_{I_c \sim p_{\text{clean}}} [\log D(I_c)] + \mathbb{E}_{I_r \sim p_{\text{raindrop}}} [\log(1 - D(G(I_r, M)_1))], \quad (6)$$

where D and G represent the discriminator and generator, respectively; I_c represents the target clean image; I_r and M represent the input raindrop image and the generated raindrop mask, respectively; and subscript 1 of G indicates the first output image of the generator.

3.3.1. Generator

The generator generates a clean image from an input raindrop image. Figure 7a shows the proposed rain-removal generator. The generator of the raindrop-removal network uses two input images: the raindrop image (color image) and raindrop mask (grayscale image). These images are concatenated channel-wise, resulting in an input image with four channels. The output image is the generated clean image. The generator produces multiple output images, each with a different resolution. The input image scale is $256 \times 256 \times 3$ pixels. The output image scales are $256 \times 256 \times 3$, $128 \times 128 \times 3$, and $64 \times 64 \times 3$ pixels.

This generator comprises several components including the CBL block, the skip connection, concatenation, and RCBAM as well as the downsampling, upsampling, and convolution layers. The CBL block comprises the convolution layer, batch normalization, and the LeakyReLU activation function. RCBAM is the self-attention module. The downsampling and upsampling layers use max-pooling and 2D transposed convolution layers, respectively. The skip connection connects the previous layers to upsampled layers. RCBAM is used to improve model performance by effectively restoring the raindrop area with the input raindrop mask. In Figure 7b, RCBAM incorporates a convolution layer that matches the size of the input feature map with the size of the output feature map. The remaining elements of the network are used in the same way as in the raindrop-mask network. All convolution layers in the generator use 3×3 spatial filters with a stride of 1 and padding of 1. The max-pooling and transposed convolution layers use 2×2 spatial filter with a stride of 2. The negative slope of LeakyReLU is 0.2.

The generator uses four loss functions: the adversarial loss, perceptual loss, SSIM loss, and multiscale MSE loss.

The adversarial loss can be expressed as

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - G(I_r, M)_1)], \quad (7)$$

where $G(I_r, M)_1$ represents the first output of the raindrop-removal generator for the input raindrop image I_r and the generated raindrop mask M , which can also be expressed as I'_{c1} .

To measure the perceptual similarity between the two images, we used the perceptual loss, which was proposed in SRGAN [22] and measures the difference between features

obtained from a well-trained network. In the proposed model, we used the pretrained VGG19 [23], and it can be expressed as

$$\mathcal{L}_{\text{per}} = \mathbb{E}[(\phi(I_c) - \phi(I'_{c1}))^2], \quad (8)$$

where $\phi(\cdot)$ represents the feature map of the pretrained VGG19 network and I_c and I'_{c1} denote the real clean image and the first generated clean image, respectively.

SSIM loss is a loss function that utilizes the structural similarity of two images. The structural similarity measured by SSIM includes the luminance, contrast, and structure of the images and can be expressed as follows:

$$\mathcal{L}_{\text{ssim}} = \mathbb{E} \left[1 - \left(\frac{(2\mu_{I_c}\mu_{I'_{c1}} + \gamma_1)(2\sigma_{I_c I'_{c1}} + \gamma_2)}{(\mu_{I_c}^2 + \mu_{I'_{c1}}^2 + \gamma_1)(\sigma_{I_c}^2 + \sigma_{I'_{c1}}^2 + \gamma_2)} \right) \right], \quad (9)$$

where the right side of the minus term represents the SSIM; μ_{I_c} and $\mu_{I'_{c1}}$ represent the mean of the real clean image I_c and mean of the first generated clean image I'_{c1} , respectively; $\sigma_{I_c}^2$ and $\sigma_{I'_{c1}}^2$ represent the variance of the images I_c and I'_{c1} ; $\sigma_{I_c I'_{c1}}$ represents the covariance; and γ_1 and γ_2 are the constant values, 0.01^2 and 0.03^2 , respectively. Generally, SSIM has a value between 0 and 1; moreover, the closer the value is to 1, the better the match between the two images. The minus term is introduced to utilize SSIM as a loss function.

The multiscale MSE loss is a method for calculating loss by extracting features from different layers of the decoder in the generator. Each extracted feature map corresponds to a different scale. By considering multiple scales, multiscale losses can capture more contextual information from different scales [4]. They can be expressed as follows:

$$\mathcal{L}_{\text{mul}} = \frac{1}{N} \sum_{i=1}^N \mathbb{E}[(S(I_c)_i - I'_{ci})^2], \quad (10)$$

where N represents the number of output images of the rain-removal generator, which is set to 3, and $S(\cdot)_i$ represents the resize function that converts the scale of real clean image I_c to the scale of the i -th generated clean image I'_{ci} .

The total loss $\mathcal{L}_{\text{total}}$ is the sum of all the loss functions:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{per}} + \mathcal{L}_{\text{ssim}} + \mathcal{L}_{\text{mul}} \quad (11)$$

3.3.2. Discriminator

The discriminator is used to improve the performance of the generator by distinguishing the generated clean image from the real clean image. The proposed model uses a discriminator in the form of PatchGAN with a 70×70 receptive field. Unlike PixelGAN, where the existing discriminator distinguishes real/fake for a single value, PatchGAN distinguishes real and fake for local image patches. This is a form of texture/style loss and can generate more high-frequency components than PixelGAN. The size of the 70×70 receptive field results in higher sharpness in both the spatial and color dimensions [2]. Figure 7c shows the discriminator used in the proposed model. The first CL block comprises a convolution layer and the LeakyReLU raindrop-removal network. The middle block CBL includes a convolution layer, batch normalization, and a LeakyReLU activation function. The last CS block includes a convolution layer and sigmoid activation function. All convolution layers in the discriminator use 4×4 spatial filters with a stride of 2 and a padding of 1. The negative slope of LeakyReLU is 0.2.

The loss of discriminator \mathcal{L}_D can be expressed as follows:

$$\mathcal{L}_D = \mathbb{E}[\log D(I_c)] + \mathbb{E}[\log(1 - D(G(I_r, M)_1))] \quad (12)$$

4. Experimental Results

The computer system used in the experiment has an Intel Core i7-11700KF processor with 128 GB of RAM and an NVIDIA GeForce 3090 GPU. The proposed model utilized 861 pairs of training data and 307 pairs of testing data from the deraindrop dataset [4]. This dataset comprises pairs of images that have been degraded by raindrops, along with their corresponding clean images. The average image resolution of the deraindrop dataset is 720×480 , and the image format is a combination of Joint Photographic Experts Group and Portable Network Graphics. In the training phase, the proposed method is optimized by Adam, the parameters of which are set as $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and the learning rate = 0.0002. The input image size is 256×256 pixels and the batch size is 16. The number of epochs is 1800. Data augmentation was performed using random crop and random flip (vertical and horizontal). Random crop randomly selects a 256×256 -pixel area from the original image size. Images are normalized from -1 to 1 . The generator and discriminator are initialized from a Gaussian distribution with mean 0 and a standard deviation of 0.02. In the test phase, we used the original image size for testing, and images were normalized to a range from -1 to 1 .

4.1. Qualitative Evaluation

We compared the proposed model with state-of-the-art models, including Pix2Pix [2], ATGAN [4], R²Net [8], and TUM [24]. Pix2Pix was retrained using the same training data, whereas the other models used pretrained models to generate the output clean images. The parameters for all models were set to their default values. Figures 8–13 depict the input raindrop image and the generated image of each model. In the figures, the red box indicates the cropped region.

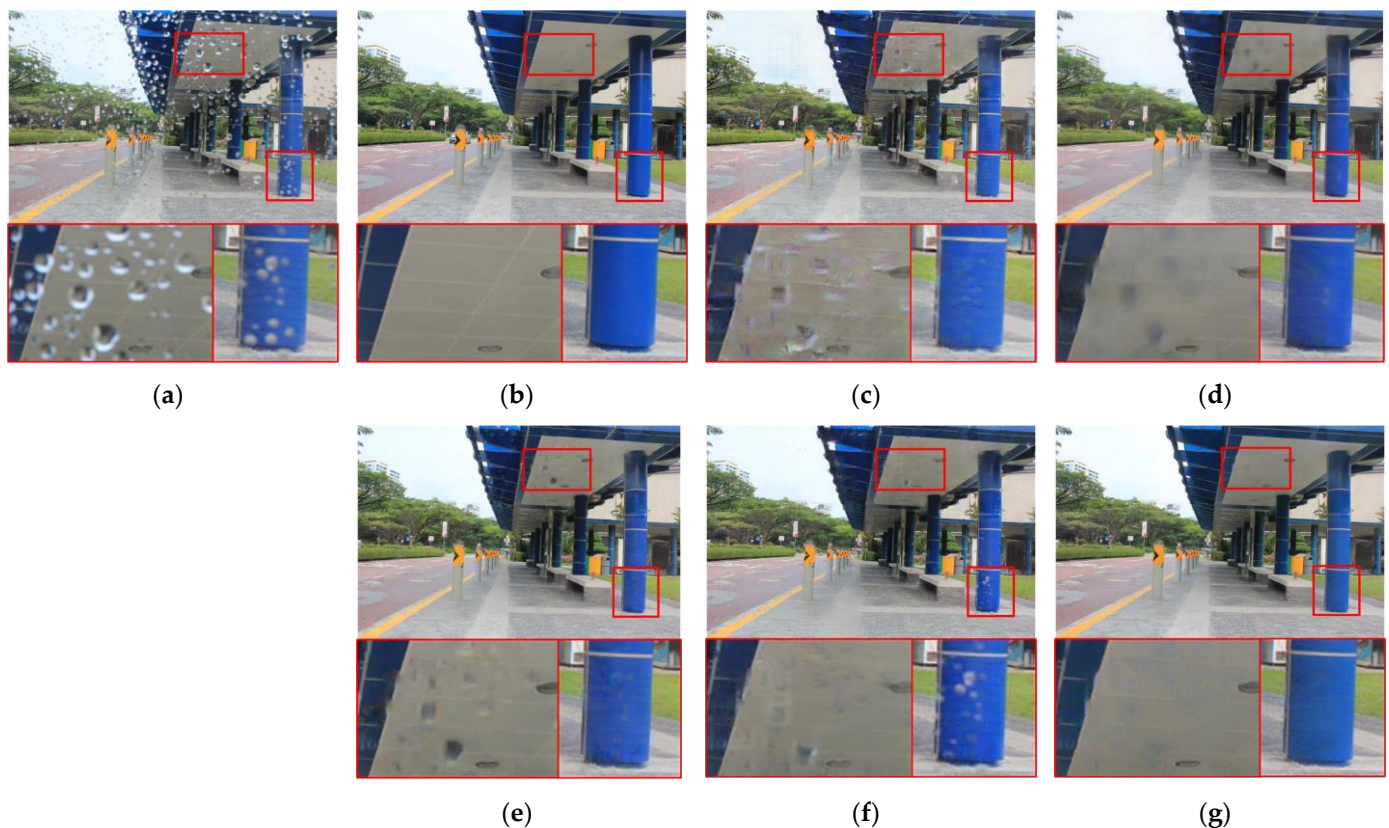


Figure 8. Rendering results of each model for Scene #1: (a) input, (b) ground truth, (c) Pix2Pix, (d) ATGAN, (e) R²Net, (f) TUM, and (g) proposed model.

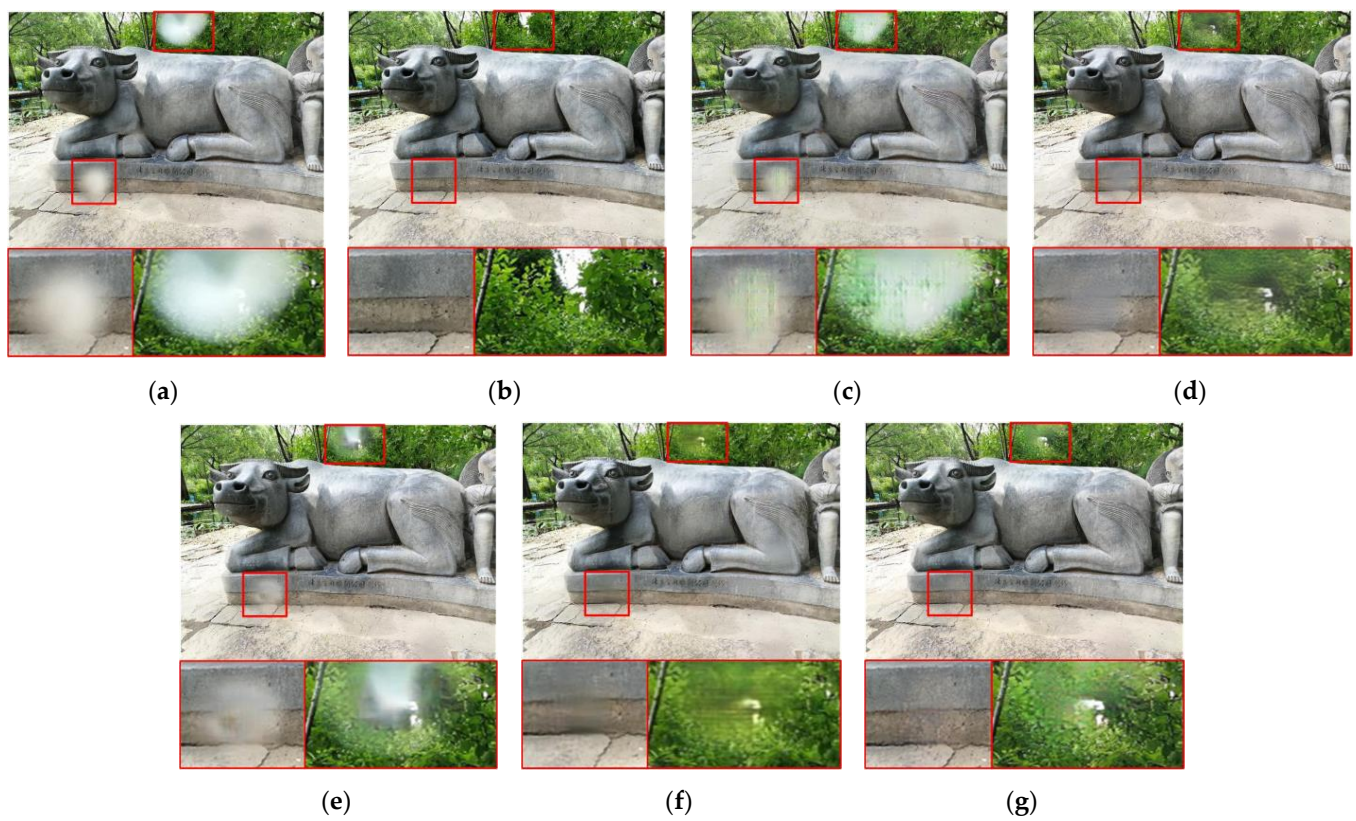


Figure 9. Rendering results of each model for Scene #2: (a) input, (b) ground truth, (c) Pix2Pix, (d) ATTGAN, (e) R²Net, (f) TUM, and (g) proposed model.

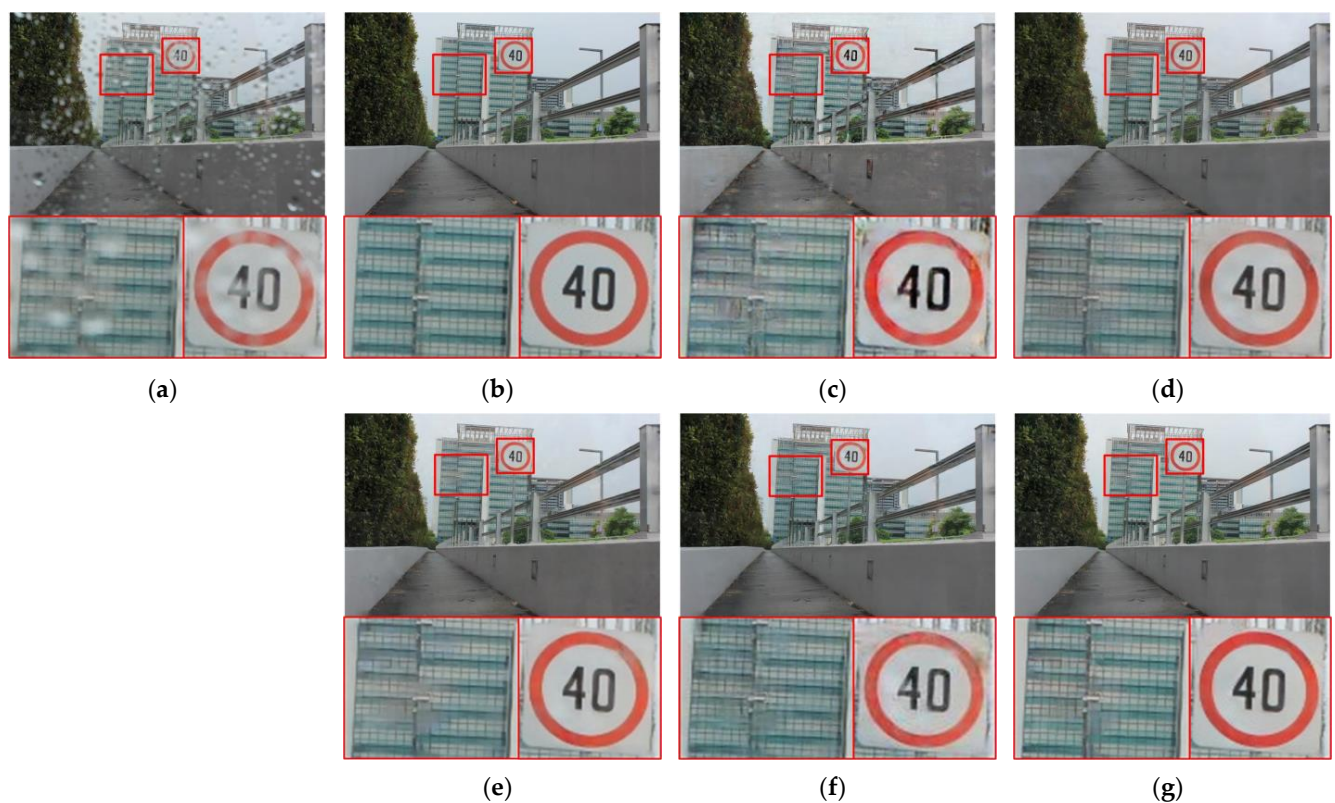


Figure 10. Rendering results of each model for Scene #3: (a) input, (b) ground truth, (c) Pix2Pix, (d) ATTGAN, (e) R²Net, (f) TUM, and (g) proposed model.

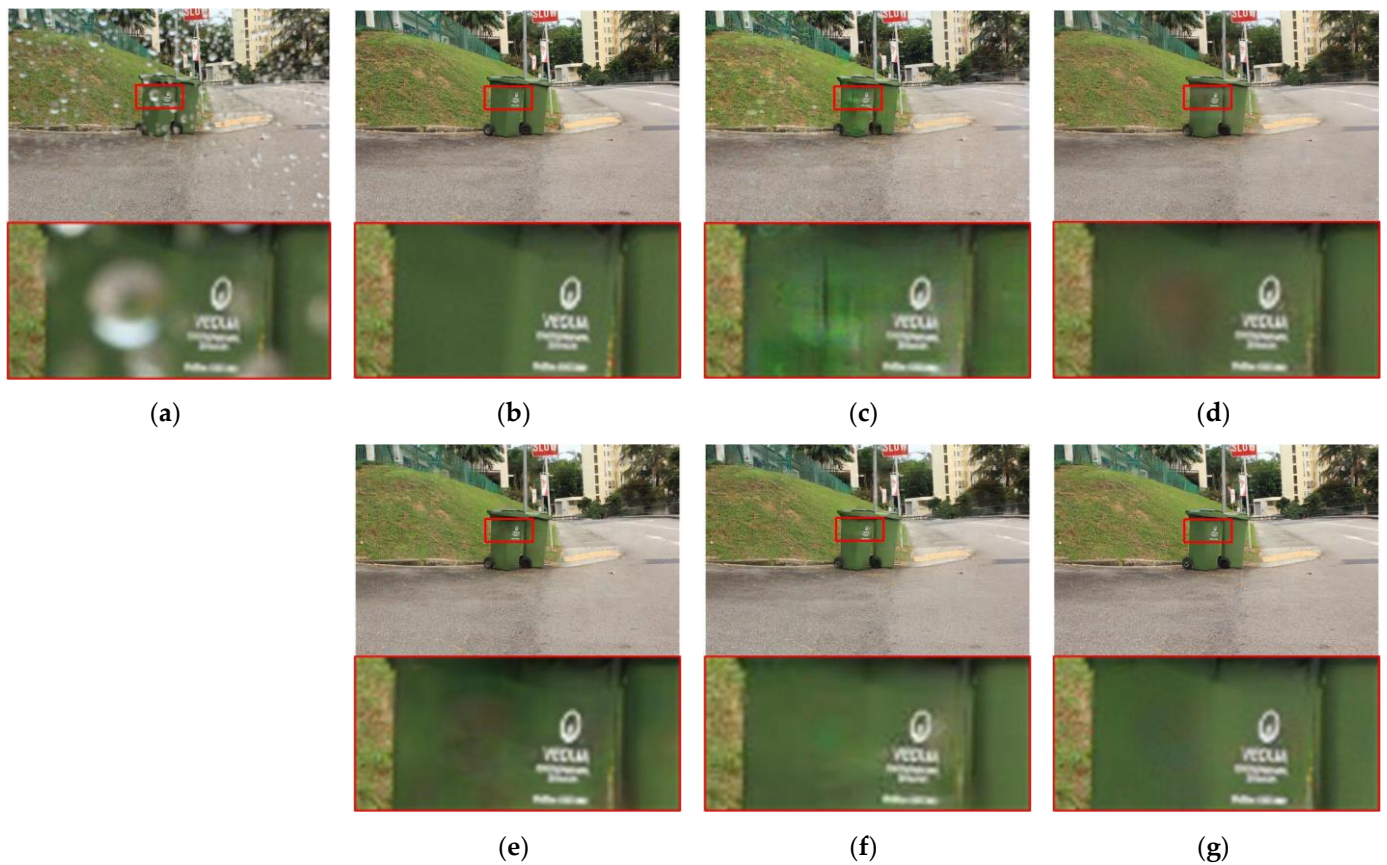


Figure 11. Rendering results of each model for Scene #4: (a) input, (b) ground truth, (c) Pix2Pix, (d) ATTGAN, (e) R²Net, (f) TUM, and (g) proposed model.

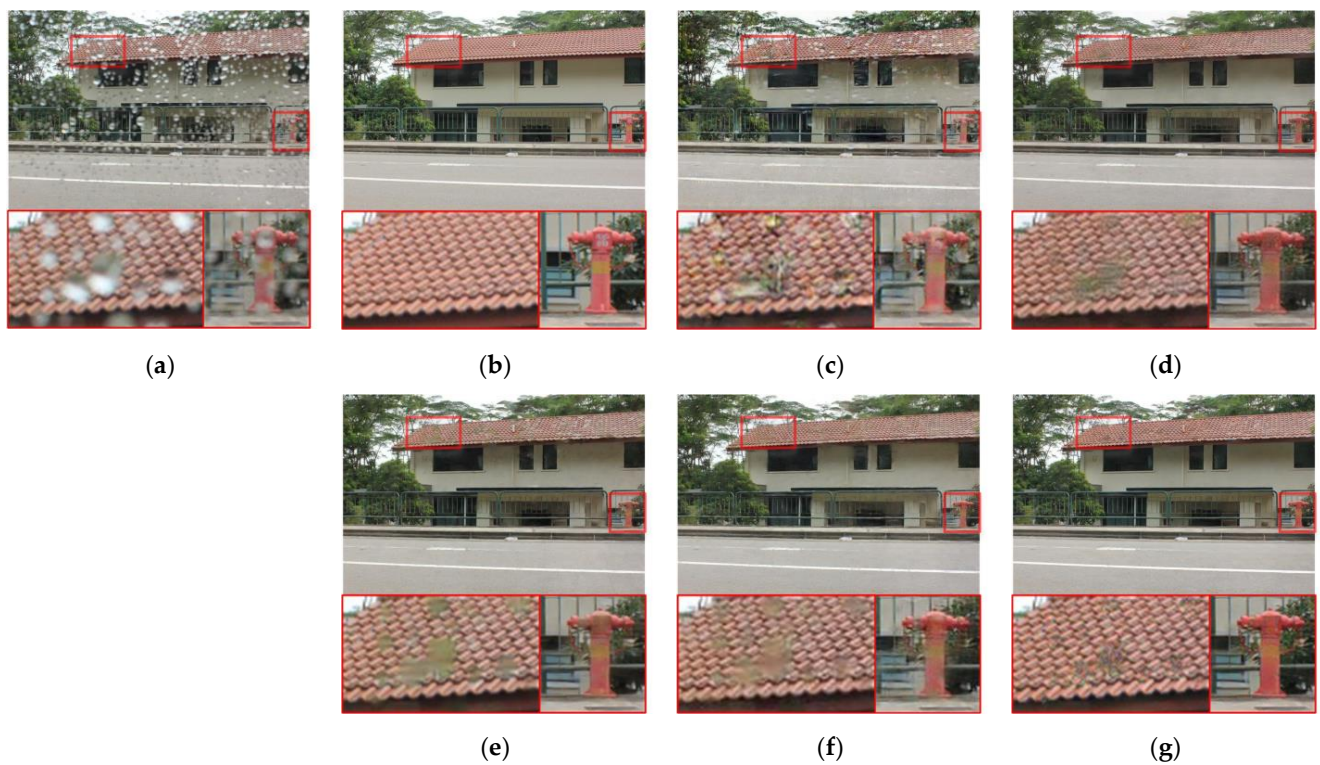


Figure 12. Rendering results of each model for Scene #5: (a) input, (b) ground truth, (c) Pix2Pix, (d) ATTGAN, (e) R²Net, (f) TUM, and (g) proposed model.

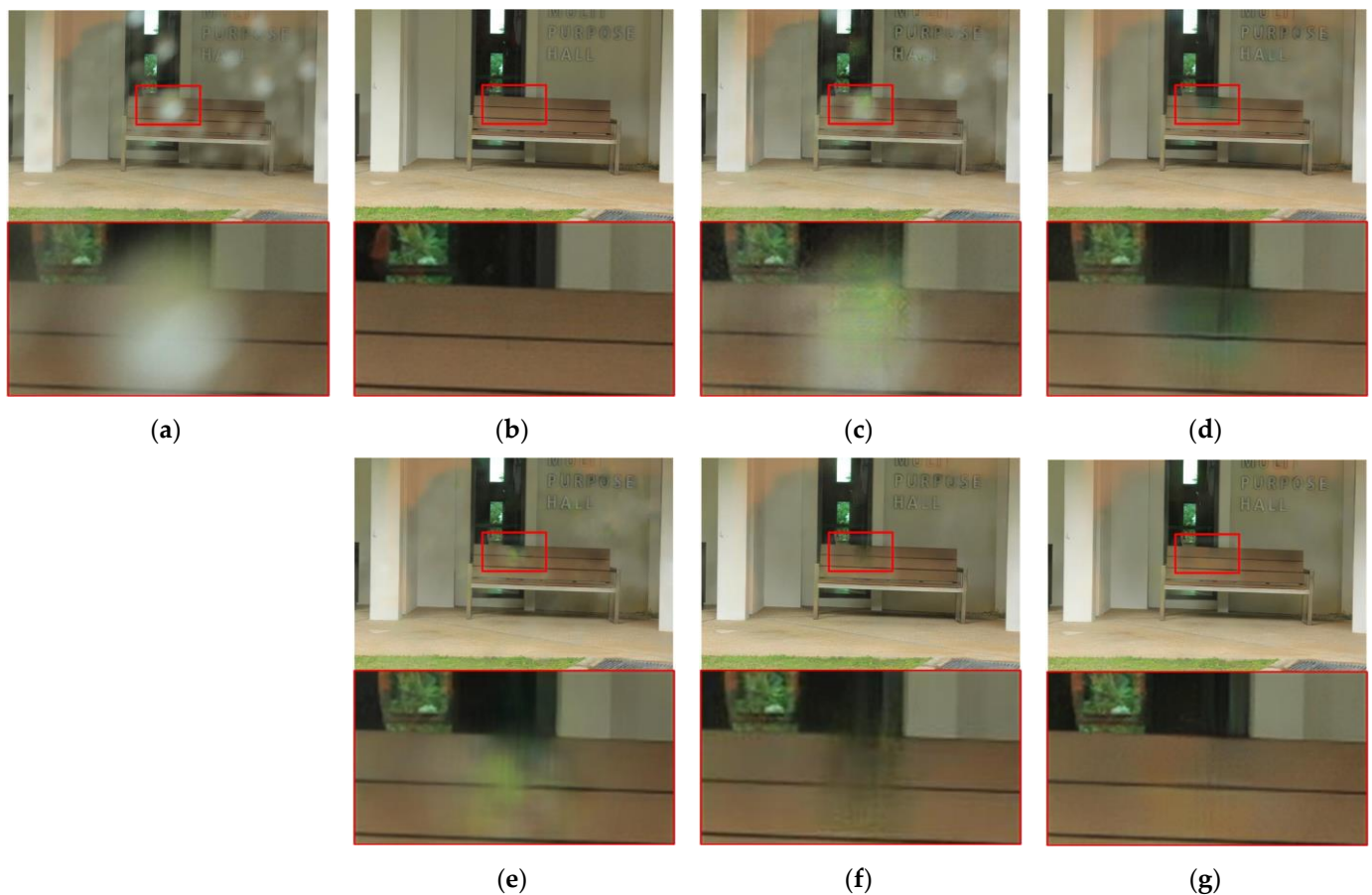


Figure 13. Rendering results of each model for Scene #6: (a) input, (b) ground truth, (c) Pix2Pix, (d) ATTGAN, (e) R²Net, (f) TUM, and (g) proposed model.

In Figure 8, small transparent circular raindrops are visible throughout the image. The raindrops within the red box, seen on the structure's ceiling, appear dark against the white color of the ceiling. In addition, several small gray raindrops are distributed within the red boxes on the pillars of the building. In Figure 8c (Pix2Pix), the raindrops are still visible, and noise is noticeable around the ceiling. In Figure 8d (ATTGAN), the raindrop shapes are replaced by black stains. In Figure 8e (R²Net), the raindrop shapes remain and appear as black spots. In Figure 8f (TUM), compared with the other models, a considerable number of raindrop shapes are erased, but noise is apparent in the image. Furthermore, unlike the other models, the raindrops present in the pillar area could not be removed. In the proposed model shown in Figure 8g, all raindrops in the ceiling and column area have been successfully removed.

In Figure 9, large raindrops are visible on the tree in the background and on the statue pedestal. These raindrops have low transparency, making it impossible to accurately identify objects behind them. Figure 9c (Pix2Pix) shows that the two raindrop regions are not successfully removed, while Figure 9d (ATTGAN) shows that two raindrop regions are removed. However, the restoration of the area behind the raindrops on the statue pedestal is unsuccessful, resulting in a dark-gray blur. In Figure 9e (R²Net), some raindrops are removed, but they appear as white blurred regions. In Figure 9f (TUM), raindrops are better removed overall compared with those in cases of other algorithms. However, the generated image is blurry, and the level of detail is low. In Figure 9g (proposed model), the two raindrop regions are better restored than in cases of other models.

In Figure 10, small raindrops are evenly distributed throughout the image. Each model is compared with respect to the areas of buildings and road signs in the background. In Figure 10c (Pix2Pix), the details of the windows of the building appear blurred, and the

numbers and red circular parts on the road sign appear blurry. In Figure 10d (ATTGAN), the details of the window area of the building are clear, but the restored area has a gray color. Additionally, the saturation of the road signs is low. In Figure 7e (R²Net), the details of the window area of the building are not clear. In Figure 10f (TUM), the restoration of the window area of the building is excellent, but traces of raindrops can still be observed on the road sign. In Figure 10g (proposed model), a clearer image is produced compared with the other models, with excellent restoration of the window areas of the building and road signs; moreover, the saturation of the road sign is high.

Figure 11 shows the distribution of raindrops throughout the image, and it can be seen that there are large raindrops present in the trashcan. In Figure 11c (Pix2Pix), the large raindrops are removed from the trashcan, but a light-green color distortion and noise could be observed. In Figure 11d (ATTGAN) and 11e (R²Net), the restored areas appear as brown stains. In Figure 11f (TUM) and the proposed model in Figure 11g, the raindrops are effectively removed, and the shape of the trashcan is clearly displayed.

Figure 12 shows that small raindrops with bright or dark brightness are distributed over the entire image. The restored roof area in Figure 12c (Pix2Pix) exhibits a significant amount of noise, and traces of raindrops in the fire hydrants are still visible. In Figure 12d–f, the shape of the raindrops remains in the roof area, and the inside of the raindrop regions appears blurred, resulting in a lack of detail in the roof. In Figure 12g (proposed model), although there is some noise in the roof area, the overall restoration is good. Additionally, the fire hydrant has been well restored, resulting in high color saturation and a clear image.

Figure 13 shows that the raindrops within the image are opaque. There are large raindrops on the bench. In the models shown in Figure 13c–f, the restoration of the bench area is not successful. As a result, the raindrop area remains and the color appears distorted. In Figure 13g (proposed model), the raindrop area is effectively removed and the color is restored to be similar to the surrounding color.

4.2. Quantitative Evaluation

To compare the quality of the generated images, we utilize several image-quality-assessment metrics (IQAMs), including SSIM [25], peak signal-to-noise ratio (PSNR), contrast-changed image-quality measure (CEIQ) [26], naturalness image-quality evaluator (NIQE) [27], Fréchet inception distance (FID) [28], and learned perceptual image patch similarity (LPIPS) [29]. SSIM, PSNR, FID, and LPIPS are full-reference IQAMs that quantify the image quality by evaluating the similarity between reference and generated images. Conversely, CEIQ and NIQE are no-reference IQAMs that analyze the image quality without a reference image.

SSIM is utilized to evaluate the structural information and perceptual similarity between reference and distorted images.

$$\text{SSIM} = \frac{(2\mu_x\mu_y + \gamma_1)(2\sigma_{xy} + \gamma_2)}{(\mu_x^2 + \mu_y^2 + \gamma_1)(\sigma_x^2 + \sigma_y^2 + \gamma_2)}, \quad (13)$$

where μ_x and μ_y represent the pixel sample mean of the images x and y , respectively; σ_x^2 and σ_y^2 represent the pixel variance of the images x and y , respectively; and σ_{xy} represents their covariance. $\gamma_1 = (k_1L)^2$ and $\gamma_2 = (k_2L)^2$, where L represents the dynamic range of the pixel value and k_1 and k_2 represent the constant values, set to 0.01 and 0.03, respectively.

PSNR calculates the difference between a reference image and a distorted image by comparing their pixel values. It calculates the mean squared error between the corresponding pixels and converts it to a logarithmic scale.

$$\text{PSNR} = 20 \log_{10} \left(\frac{\text{MAX}_I}{\sqrt{\text{MSE}}} \right), \quad (14)$$

where MAX_I represents the maximum pixel value. MSE is the mean square error that quantifies the average squared difference between the corresponding pixels of a reference image and a distorted image.

FID is used to evaluate the quality and diversity of generated images in image-synthesis tasks such as GAN. It measures the distance between the feature representations of real and generated images using a pretrained deep neural network.

$$FID = \| \mu - \mu_w \|_2^2 + Tr(\rho + \rho_w - 2\sqrt{\rho\rho_w}), \quad (15)$$

where μ and ρ represent the mean and covariance matrix of the distribution of model samples and μ_w and ρ_w represent the distribution of the samples from the real world.

LPIPS is a perceptual-similarity metric that quantifies the perceptual difference between two images. It measures the similarity between image patches based on learned features from deep neural networks.

$$LPIPS(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \| \omega_l \odot (\hat{y}_{hw}^l - \hat{y}_{0hw}^l) \|_2^2, \quad (16)$$

where x and x_0 represent reference and distorted patches with networks; H and W represent the height and width of the layer l , respectively; \hat{y}^l and \hat{y}_0^l represent the feature stack extracted from layer l and unit-normalized in the channel dimension of the reference and distorted patches, respectively; and ω_l represents the scale factor of the layer l .

CEIQ is employed to measure the quality of contrast-altered images. It utilizes histogram-based entropy and cross-entropy between the original image and the histogram-equalized image to assess the quality of the image. The measurements are performed using a support-vector-machine regressor model.

$$E = - \sum_{i=0}^b h(i) \log h(i) \quad (17)$$

$$E_{ge} = - \sum_{i=0}^b h_g(i) \log h_e(i) \quad (18)$$

$$E_{eg} = - \sum_{i=0}^b h_e(i) \log h_g(i) \quad (19)$$

$$CEIQ = SVM(SSIM_{ge}, E_g, E_e, E_{ge}, E_{eg}), \quad (20)$$

where $h(\cdot)$ represents the histogram function and b represents the number of bins in the histogram. Subscripts g and e denote the gray image and histogram-equalized image, respectively. $SSIM_{ge}$ represents the SSIM for the gray image and histogram-equalized image. E_x, E_y and E_{xy}, E_{yx} denote the histogram-based entropy and histogram cross-entropy. $SVM(\cdot)$ represents the support vector regression.

NIQE utilizes a “quality aware” collection of statistical features. These features are constructed based on a successful space-domain natural-scene statistic model and derived from a dataset of natural, undistorted images.

$$NIQE = \sqrt{(\mu_1 - \mu_2)^T \left(\frac{\rho_1 + \rho_2}{2} \right)^{-1} (\mu_1 - \mu_2)}, \quad (21)$$

where μ_1, μ_2 and ρ_1, ρ_2 , are the mean vectors and covariance matrices of the natural multivariate Gaussian model and the multivariate Gaussian model of the distorted image. The NIQE model was retrained to assess the naturalness of the generated clean images. The training dataset of NIQE comprised all the ground-truth images available in the deraindrop dataset.

Table 1 and Figure 14 show the results of each model for the IQAMs. The best-performing model is indicated in bold, while the second-best model is underlined. The

proposed model demonstrates superior performance, achieving the highest scores in CEIQ, NIQE, FID, and LPIPS, as well as the second-highest scores in SSIM and PSNR. These results confirm the superiority of the proposed model over other models. FID and LPIPS scores show a significant improvement of 19.84% and 4.84%, respectively, compared to ATTGAN. This indicates that the perceptual similarity between the generated images and real images is higher in the proposed model compared to other models.

Table 1. Comparisons of image-quality-assessment metrics.

Model	SSIM \uparrow	PSNR \uparrow	CEIQ \uparrow	NIQE \downarrow	FID \downarrow	LPIPS \downarrow
Pix2Pix	0.770	23.621	3.332	2.499	47.490	0.114
ATTGAN	0.830	26.266	<u>3.344</u>	<u>2.442</u>	25.994	<u>0.062</u>
R ² Net	0.835	26.160	3.338	3.015	26.319	0.071
TUM	0.663	23.757	3.269	2.908	26.995	0.136
Proposed	<u>0.832</u>	<u>26.165</u>	3.351	2.224	20.837	0.059

Up arrow represents that a higher score is better, while down arrow represents that a lower score is better. The best-performing model is indicated in bold, while the second-best model is underlined.

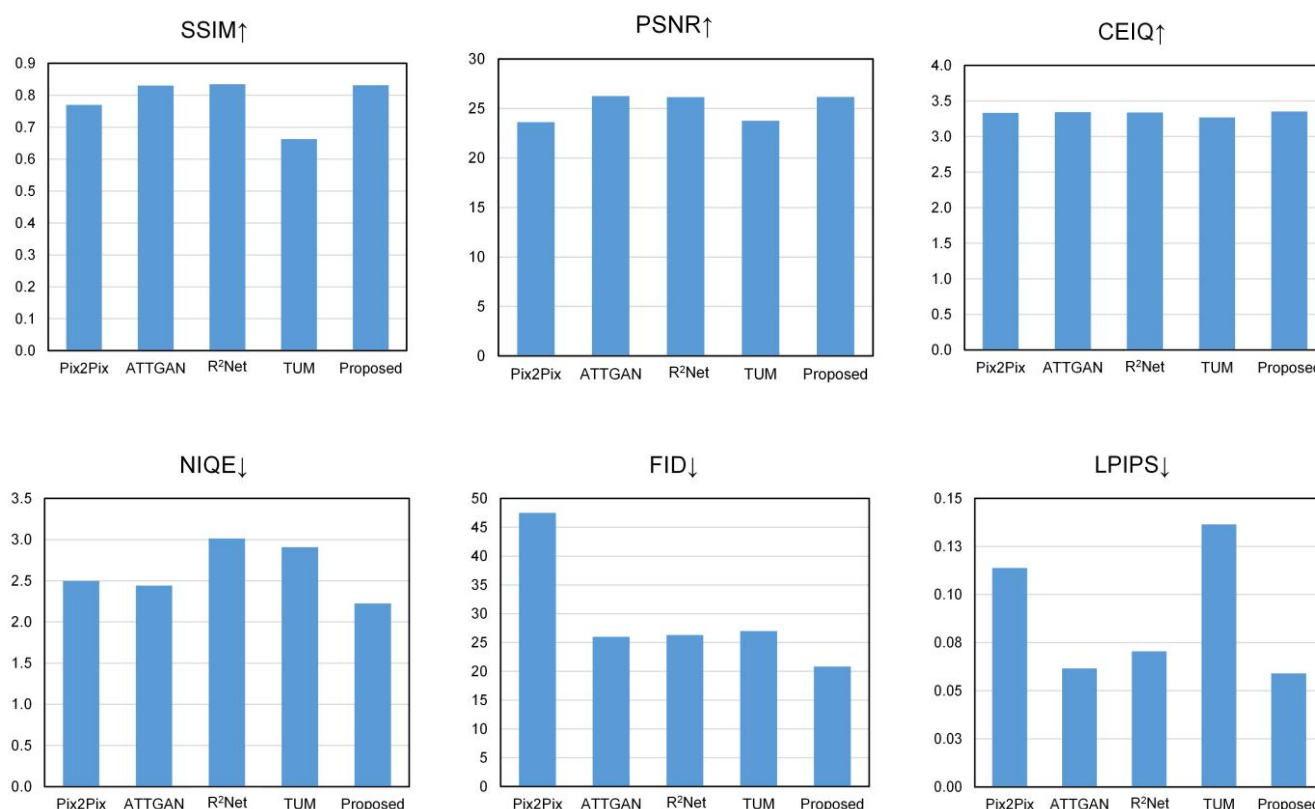


Figure 14. Image-quality-assessment scores for each model. Up arrow represents that a higher score is better, while down arrow represents that a lower score is better.

LPIPS is a distance metric used to evaluate visual similarity between images, learned from deep-learning models. FID is a metric used to measure the difference between generated and real images, assessing how similar the generated images are to real ones. Both metrics are widely used for evaluating deep-learning-model performance and quantifying image quality.

To assess whether the raindrop-removal model enhances object-detection performance, we conducted tests using YOLOv7 [30]. The results in Figure 15 show that the proposed model achieved higher object-detection rates compared to the object-detection rates from the raindrop images. Furthermore, in Figure 15b on the right image, it is evident that the proposed model successfully detected the car behind the truck, which was not detected

in the raindrop image. These results validate the effectiveness of the proposed raindrop-removal method in improving object-detection performance.



Figure 15. The results of the object detection (YOLOv7) for the raindrop and raindrop-removal images. (a) raindrop images; (b) proposed raindrop-removal images.

Table 2 shows the average processing speeds for each model. The test images had an average size of 720×480 , and a total of 307 images were used to evaluate the processing speeds. Based on the results, it is evident that ATTGAN demonstrated the fastest processing speed, while the proposed algorithm exhibited relatively slower performance. At present, the proposed model may not be suitable for real-time processing, but with future code optimization, it has the potential to achieve real-time processing capabilities.

Table 2. Processing speeds of each model.

Image Resolution	Pix2Pix	ATTGAN	TUM	Proposed
720×480	0.0473 s	0.0391 s	0.0466 s	0.1169 s

4.3. Ablation Study

We compare the proposed model against ablations of RCAM, raindrop-mask network (MASK), perceptual loss \mathcal{L}_{per} , and SSIM loss $\mathcal{L}_{\text{ssim}}$, while multiscale loss \mathcal{L}_{mul} and adversarial loss \mathcal{L}_{adv} are commonly applied. Table 3 shows the results of the ablation study. The best-performing model is indicated in bold, while the second-best model is underlined. In Table 3, we can observe a gradual increase in IQAM scores as the proposed modules or loss functions are added. Comparing Case 1 and Case 2, we can see that the Mask module

contributes to the improvement of SSIM, PSNR, CEIQ, and NIQE performance. In particular, in Case 3, the addition of the perceptual loss significantly enhances the performance of all IQAMs, except for CEIQ. By comparing Case 4 with the proposed model, it can be confirmed that RCBAM also contributes to improving model performance.

Table 3. Ablation study of the proposed model.

Model	Module		Loss				Metric					
	RCBAM	MASK	\mathcal{L}_{per}	$\mathcal{L}_{\text{ssim}}$	\mathcal{L}_{mul}	\mathcal{L}_{adv}	SSIM \uparrow	PSNR \uparrow	CEIQ \uparrow	NIQE \downarrow	FID \downarrow	LPIPS \downarrow
Case 1	✓				✓	✓	0.822	25.768	3.334	2.476	25.592	0.070
Case 2	✓	✓			✓	✓	0.824	25.794	<u>3.358</u>	2.426	27.857	0.073
Case 3	✓	✓	✓		✓	✓	<u>0.829</u>	<u>25.948</u>	3.354	2.355	<u>22.110</u>	<u>0.061</u>
Case 4		✓	✓	✓	✓	✓	0.828	25.796	3.361	<u>2.288</u>	23.102	0.063
Proposed	✓	✓	✓	✓	✓	✓	0.832	26.165	3.351	2.224	20.837	0.059

Up arrow represents that a higher score is better, while down arrow represents that a lower score is better. The best-performing model is indicated in bold, while the second-best model is underlined.

5. Conclusions

In this study, we proposed a deep-learning model for the removal of raindrops in images; this model incorporates an attention mechanism based on the GAN framework and comprises two networks designed to effectively remove the raindrops. The first network is the raindrop-mask-generation network. This network accurately identifies the raindrop regions in the input image based on information such as location, size, and brightness. Unlike conventional binary or absolute difference masks, this approach uses the difference image between the raindrop image and the corresponding clean image to generate precise raindrop masks. The second network is the raindrop-removal network, which is based on the GAN model. The raindrop-removal generator combines U-Net architecture with the RCBAM and produces multiscale outputs. The input to the raindrop-removal generator comprises four channels: the raindrop image and raindrop-mask image. The raindrop mask image and RCBAM effectively guide the removal of raindrops. The discriminator structure in the GAN framework is designed to distinguish between real and fake images at the local patch using PatchGAN. To further enhance model performance, we incorporated perceptual loss, SSIM loss, multiscale loss, and adversarial loss. Based on qualitative and quantitative evaluations, our proposed model exhibits superior performance in terms of the raindrop removal and enhancing details compared to existing models.

Author Contributions: Conceptualization, S.-H.L.; Data curation, H.-J.K.; Formal analysis, H.-J.K. and S.-H.L.; Funding acquisition, S.-H.L.; Investigation, H.-J.K. and S.-H.L.; Methodology, H.-J.K. and S.-H.L.; Project administration, S.-H.L.; Resources, H.-J.K.; Software, H.-J.K.; Supervision, S.-H.L.; Validation, H.-J.K. and S.-H.L.; Visualization, H.-J.K.; Writing—original draft, H.-J.K.; Writing—review & editing, S.-H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Korea (NRF-2021R1I1A3049604) and supported by the MSIT (Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program (IITP-2023-RS-2022-00156389) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation).

Data Availability Statement: Not applicable.

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

References

- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [\[CrossRef\]](#)
- Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 5967–5976.
- Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 2242–2251.
- Qian, R.; Tan, R.T.; Yang, W.; Su, J.; Liu, J. Attentive Generative Adversarial Network for Raindrop Removal from A Single Image. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 2482–2491.
- Alletto, S.; Carlin, C.; Rigazio, L.; Ishii, Y.; Tsukizawa, S. Adherent Raindrop Removal with Self-Supervised Attention Maps and Spatio-Temporal Generative Adversarial Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Long Beach, CA, USA, 16–17 June 2019; pp. 2329–2338. [\[CrossRef\]](#)
- Quan, Y.; Deng, S.; Chen, Y.; Ji, H. Deep Learning for Seeing through Window with Raindrops. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 2463–2471. [\[CrossRef\]](#)
- Shao, M.; Li, L.; Wang, H.; Meng, D. Selective generative adversarial network for raindrop removal from a single image. *Neurocomputing* **2021**, *426*, 265–273. [\[CrossRef\]](#)
- Anwar, S.; Barnes, N.; Petersson, L. Attention-Based Real Image Restoration. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–11. [\[CrossRef\]](#) [\[PubMed\]](#)
- Yang, H.D. Restoring Raindrops Using Attentive Generative Adversarial Networks. *Appl. Sci.* **2021**, *11*, 7034. [\[CrossRef\]](#)
- Xia, H.; Lan, Y.; Song, S.; Li, H. Raindrop Removal from a Single Image Using a Two-Step Generative Adversarial Network. *Signal Image Video Process.* **2022**, *16*, 677–684. [\[CrossRef\]](#)
- Chen, R.; Lai, Z.; Qian, Y. Image Raindrop Removal Method for Generative Adversarial Network Based on Difference Learning. *J. Phys. Conf. Ser.* **2020**, *1544*, 012099. [\[CrossRef\]](#)
- Xu, C.; Gao, J.; Wen, Q.; Wang, B. Generative Adversarial Network for Image Raindrop Removal of Transmission Line Based on Unmanned Aerial Vehicle Inspection. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 6668771. [\[CrossRef\]](#)
- Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18. Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
- Torbunov, D.; Huang, Y.; Yu, H.; Huang, J.; Yoo, S.; Lin, M.; Viren, B.; Ren, Y. UVCGAN: UNet Vision Transformer Cycle-Consistent GAN for Unpaired Image-to-Image Translation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 702–712. [\[CrossRef\]](#)
- Bai, J.; Chen, R.; Liu, M. Feature-Attention Module for Context-Aware Image-to-Image Translation. *Vis. Comput.* **2020**, *36*, 2145–2159. [\[CrossRef\]](#)
- Hu, X.; Naiel, M.A.; Wong, A.; Lamm, M.; Fieguth, P. RUNet: A Robust UNet Architecture for Image Super-Resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–20 June 2019; pp. 505–507. [\[CrossRef\]](#)
- Masutani, E.M.; Bahrami, N.; Hsiao, A. Deep Learning Single-Frame and Multiframe Super-Resolution for Cardiac MRI. *Radiology* **2020**, *295*, 552–561. [\[CrossRef\]](#) [\[PubMed\]](#)
- Huang, H.; Tao, H.; Wang, H. A Convolutional Neural Network Based Method for Low-Illumination Image Enhancement. In Proceedings of the 2nd International Conference on Artificial Intelligence and Pattern Recognition, Beijing, China, 16–18 August 2019; pp. 72–77. [\[CrossRef\]](#)
- Liu, F.; Hua, Z.; Li, J.; Fan, L. Dual UNet Low-Light Image Enhancement Network Based on Attention Mechanism. *Multimed. Tools Appl.* **2022**, *82*, 24707–24742. [\[CrossRef\]](#)
- Yan, W.; Xu, L.; Yang, W.; Tan, R.T. Feature-Aligned Video Raindrop Removal With Temporal Constraints. *IEEE Trans. Image Process.* **2022**, *31*, 3440–3448. [\[CrossRef\]](#) [\[PubMed\]](#)
- Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19. ISBN 978-3-030-01234-2.
- Ledig, C.; Theis, L.; Huszar, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 105–114.
- Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
- Chen, W.-T.; Huang, Z.-K.; Tsai, C.-C.; Yang, H.-H.; Ding, J.-J.; Kuo, S.-Y. Learning Multiple Adverse Weather Removal via Two-Stage Knowledge Learning and Multi-Contrastive Regularization: Toward a Unified Model. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 17632–17641.

25. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)]
26. Yan, J.; Li, J.; Fu, X. No-Reference Quality Assessment of Contrast-Distorted Images Using Contrast Enhancement. *arXiv* **2019**, arXiv:1904.08879.
27. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a “Completely Blind” Image Quality Analyzer. *IEEE Signal Process. Lett.* **2013**, *20*, 209–212. [[CrossRef](#)]
28. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
29. 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 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 18–23 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 586–595.
30. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *arXiv* **2022**, arXiv:2207.02696.

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