



## Research Article

# A Novel Attentive Generative Adversarial Network for Waterdrop Detection and Removal of Rubber Conveyor Belt Image

Xianguo Li <sup>1,2</sup>, Zongpeng Liu <sup>1,2</sup>, Bin Li<sup>1,2</sup>, Xinxin Feng<sup>1,2</sup>, Xiao Liu<sup>1,2</sup> and Debao Zhou<sup>1,3</sup>

<sup>1</sup>School of Electronic and Information Engineering, Tiangong University, Tianjin, China

<sup>2</sup>Tianjin Key Laboratory of Optoelectronic Detection Technology and System Tianjin, Tianjin, China

<sup>3</sup>Department of Mechanical and Industrial Engineering, University of Minnesota Duluth, Duluth, MN, USA

Correspondence should be addressed to Xianguo Li; [lixianguo@tjpu.edu.cn](mailto:lixianguo@tjpu.edu.cn)

Received 6 August 2019; Revised 4 October 2019; Accepted 16 November 2019; Published 22 February 2020

Guest Editor: Marco Perez-Cisneros

Copyright © 2020 Xianguo Li et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The lens for monitoring the rubber conveyor belt is easy to adhere to a large number of water droplets, which seriously affects the image quality and then affects the effect of fault monitoring. In this paper, a new method for detecting and removing water droplets on rubber conveyor belts based on the attentive generative adversarial network is proposed to solve this problem. First, the water droplet image of the rubber conveyor belt is input into the generative network composed of a cyclic visual attentive network and an autoencoder with skip connections, and an image of removing water droplets and an attention map for detecting the position of the water droplet are generated. Then, the generated image of removing water droplets is evaluated by the attentive discriminant network to assess the local consistency of the water droplet recovery area. In order to better learn the water droplet regions and the surrounding structures during the training, the image morphology is added to the precise water droplet regions. A dewatered rubber conveyor belt image is generated by increasing the number of circular visual attention network layers and the number of skip connection layers of the autoencoder. Finally, a large number of comparative experiments prove the effectiveness of the water droplet image removal algorithm proposed in this paper, which outperforms of Convolutional Neural Network (CNN), Discriminative Sparse Coding (DSC), Layer Prior (LP), and Attention Generative Adversarial Network (ATTGAN).

## 1. Introduction

Rubber conveyor belts [1] have been widely used in coal, mining, port, and other fields, mainly for the transportation of bulk, granular, and powdery solid materials. For a variety of reasons, rubber conveyor belts often show longitudinal tears [2] during operation, causing economic losses and even casualties. With the development of technology, the longitudinal tear detection system of the rubber conveyor belt based on machine vision has been gradually popularized. However, the rubber conveyor belt is prone to dust during transportation of some materials (such as coal and powder ore), which causes the camera lens to become dirty. In order to keep the lenses clean, it is common to spray water on the lenses and then wipe off water. This will leave water droplets on the lens. In addition, in order to reduce the environmental pollution caused by dust, many applications will

spray water on the materials. For example, when coal is transported in a coal mine, a large amount of coal dust is in the air. Spraying water is required to reduce the coal dust, which makes the monitoring of water droplets in the lens more common. Therefore, how to effectively remove the water droplets on the rubber conveyor belt monitoring image to ensure the sharpness of the image is an important issue to be solved.

The removal of image water droplets in a rubber conveyor belt is similar to the removal of raindrops in a natural image [3]. At present, the methods of water droplet removal at home and abroad are mainly divided into three categories: filter-based rain removal algorithms [4], sparse coding dictionary- and classifier-based rain removal algorithms [5], and deep learning-based rain removal algorithms [3].

In recent years, the deep convolutional neural network [6] with powerful feature learning ability has made a major

breakthrough and become the main method of water droplet removal.

In 2014, Eigen et al. [7] proposed a method for raindrop removing in a single image and trained a convolutional neural network with pairs of raindrop-degraded images and corresponding raindrop-free images to better affect relatively thin and small areas of raindrops or dust. But for larger and dense raindrops, it does not produce good results. In 2015, Luo et al. [5] used the discriminant dictionary to learn the sparse coding method, which improved the accuracy of background layer and rain layer separation. However, when the image contains image texture similar to the rain streak, the image details will be blurred. In 2016, Li [8] proposed single-image rain streak removal using layer prior algorithm based on the Gaussian mixture model, but it is easy to lead to smooth transition in nonrain areas. In 2018, Qian et al. [9] used a network of attentive generative adversarial training raindrop image, and visual attention was given into the generative and discriminative network, which had a good effect on raindrop removal, but there was a defect in losing image detail information. For the rubber belt longitudinal tear monitoring system, under the action of auxiliary light source illumination, the rubber conveyor belt image will have a certain amount of specular reflection effect, and the attention generation is not ideal for the network removal effect.

In this paper, we propose a new method for detecting and removing water droplets from rubber conveyor belts based on the attentive generative adversarial network [9]. The expected results are achieved.

## 2. Algorithm Implementation

Because the shape of the water droplets on the rubber conveyor belt images is different, the number is different, and the background information occluded by the water droplets is similar to the water droplets so that when the general algorithm detects the water droplets in the image, the area similar to the shape of the water droplet is mistakenly treated as a water droplet. The subsequent operation of removing water droplets will remove the background information in the image, and a clear, water-free background image cannot be restored. Even more difficult is that even if the position of the water drop area is correctly detected, it is impossible to restore a clear water-free background. Therefore, in our method, we utilize a GAN [10] as the backbone of our network, which is recently popular in dealing with the image inpainting or completion problem. Then, our main idea is to inject visual attention [9] into both the generative and discriminative networks. By increasing data preprocessing, improving network structure, and rationally designing network optimizer and hyperparameters, a new method for detecting and removing water droplets in the rubber conveyor belt based on attention generation against the network is designed. A block diagram of the water droplet detection and removal method for the rubber conveyor belt image based on the attentive generative adversarial network is shown in Figure 1.

Firstly, the rubber conveyor belt water droplet image datasets are subjected to data preprocessing, including input normalization, image cropping, image flipping, and image

morphology, and then the attention map is added in the generator to make the generation network focus on the area with water droplets and utilize the three loss (perceptual loss, multiscale losses, and attention map loss) functions in [9], and the independent design of the network optimizer make the generation of network training more stable and generate a clear image of the dewater droplets. The generated dewater droplet rubber conveyor belt image is input into the attentive discriminator together with the true clear background image to judge the true and false area of the water drop, and the optimizer and loss function which are most suitable for the discriminator are designed. The image data preprocessing and network optimizer sections designed in this paper are described in detail as follows. The improved network structure and loss function are described in Sections 3 and 4, respectively.

*2.1. Data Normalization.* Data normalization mainly includes generating TFRecord format files and input normalization [11]. The TFRecord file stores binary data and label data (rubber conveyor belt with water droplets and no water droplet images) in the same folder, without compressing the data and quickly loading them into memory, improving network training efficiency. The data normalization classifies the input color image pixel values from  $[0, 255]$  to  $[-1, 1]$ , which match the pretraining model VGG16 [12] of the network to avoid the training loss explosion and accelerate the gradient descent to improve the convergence speed of the generative adversarial network (GAN) [10] model.

*2.2. Data Augmentation.* Data enhancements include random image cropping and image flipping. Image random cropping not only increases the amount of data but also weakens the data noise and increases the stability of the model. Assume that the water droplet regions of the rubber conveyor belt are  $C_1$ , the nonwater droplet regions are  $C_2$ , the main features of  $C_1$  are  $\{C_1, F_1, G_1\}$ , and  $C_2$  is  $\{E_2, F_2, G_2\}$ . Assume that background noise is added:  $C_1, C_2$  randomly add  $N_1, N_2, N_3$ , and the image we randomly cropped at this time is as follows:

$$\begin{aligned} I1 &= \{E1, F1, G1\}, \\ I2 &= \{E1, F1, G1, N1\}, \\ I3 &= \{E1, F1, N2\}, \\ I4 &= \{F1, G1, N3\}, \\ &\dots \end{aligned} \tag{1}$$

Since  $N_1, N_2$ , and  $N_3$  are random and  $\{E_1, F_1, G_1\}$  can always produce  $\{E_1, F_1, G_1\} \rightarrow C_1$  mapping with high probability, which is the identification of water droplet features, then any factor in  $\{E_1, F_1, G_1\}$  has a higher information gain or weight relative to  $N_1, N_2, N_3$ . If  $N_1, N_2$ , and  $N_3$  also have corresponding distributions in category  $C_2$ , then  $N1, N2$ , and  $N3$  have information gains close to zero for classification discriminant. So, the water droplet recognition effect is more accurate. It has a better effect on the network to remove water droplets. Image flip enriches the training set,

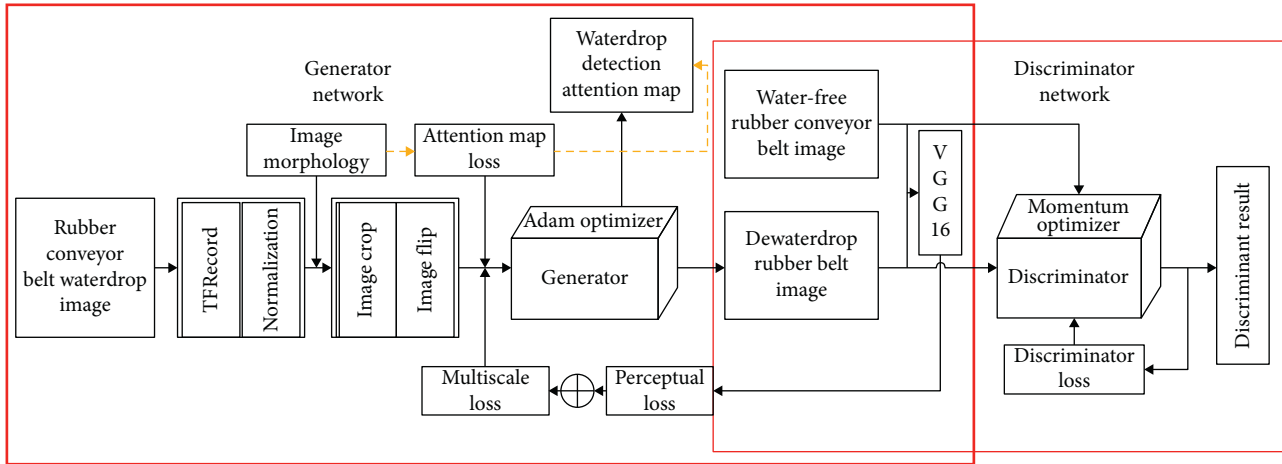


FIGURE 1: Block diagram of the method in the paper.

improves image features, generalizes the GAN model, and prevents overfitting.

**2.3. Image Morphology.** Image morphology [13] mainly performs closed and open operations on the water droplet binary mask. In order to detect the specific position of the water drop, the difference value between the water drop image and the original image is used to obtain a mask image. The closed operation of the mask map is to first etch and re-expand the image so that the white areas of the water droplet image are connected to each other and the small black holes isolated in the water droplet area are filled to highlight the entire raindrop area. Then, the opening operation of the corrosion after the first expansion is used to eliminate the background bright noise outside the raindrops, selectively retain the water droplets, and obtain the main object water droplets in the image. The image morphology module is placed after the process of obtaining the water droplet mask maps (water droplet binary mask) and image morphology operation can generate sharper water droplet attention map, and the difference between the network learning result and the water droplet mask is generated as small as possible. It is important to generate a water-free image later in the generative network section.

**2.4. The Optimizer.** After a lot of experiments, it is proved that the Adam optimizer is selected during generative network training, and the momentum optimizer is selected discriminative network. The Adam optimizer is able to calculate the adaptive learning rate for each parameter and solve the ill-posed problems caused by backpropagation. Experiments show that, with the decrease of the learning rate, the generation network can start to converge in the 500 epochs of training, and finally, the model is stable and convergent. Discriminative network selects momentum optimizer, which can achieve the momentum gradient descent algorithm. Momentum optimizer cannot be trapped in local minimum values and make the discriminant network converge faster and reduce oscillation. So, the model effect is better.

### 3. Network Design

**3.1. Single Water Drop Image Formation.** Single water drop image formation:

$$I = (1 - M) \odot B + W, \quad (2)$$

where  $I$  is the water droplet image;  $B$  is the background image; and  $M$  is the binary mask image, which is obtained by subtracting the background image  $B$  from the water droplet image  $I$ . In the mask,  $M(x) = 1$  indicates that the pixel  $x$  is part of the water droplet region, and  $M(x) = 0$  means that the pixel  $x$  is part of the background area;  $W$  means the effect of the water droplets, including the complex mixture of background information and the reflection of the lens through the rubber conveyor belt imaging monitoring device or the impact of illumination light. The operator represents the multiplication of the elements.

Based on the model (equation (1)), our goal is to obtain the background image  $B$  from a given input  $I$ . In order to realize the detection and removal of the water droplet image of the rubber conveyor belt, we create the attention map guided by the binary mask  $M$ . The threshold is used to determine whether a pixel is part of a water droplet region, and we set the threshold to 50 for all images in our training dataset.

**3.2. Network Structure.** Figure 2 shows the generative network structure, and Figure 3 shows the discriminant network structure, which together constitute the overall architecture of the generative adversarial network proposed in this paper. Input the water drop image of the rubber conveyor belt, and generate a network to generate as realistic a water drop-free image as possible, and generate a water drop attention map for detecting the position of the water drop. The discriminant network will verify that the image generated by the generator network is authentic.

**3.2.1. Generator Network.** As shown in Figure 2 above, the generator network function is divided into two parts: detecting water droplets and generating real images without water droplets. The detection of water droplets mainly uses

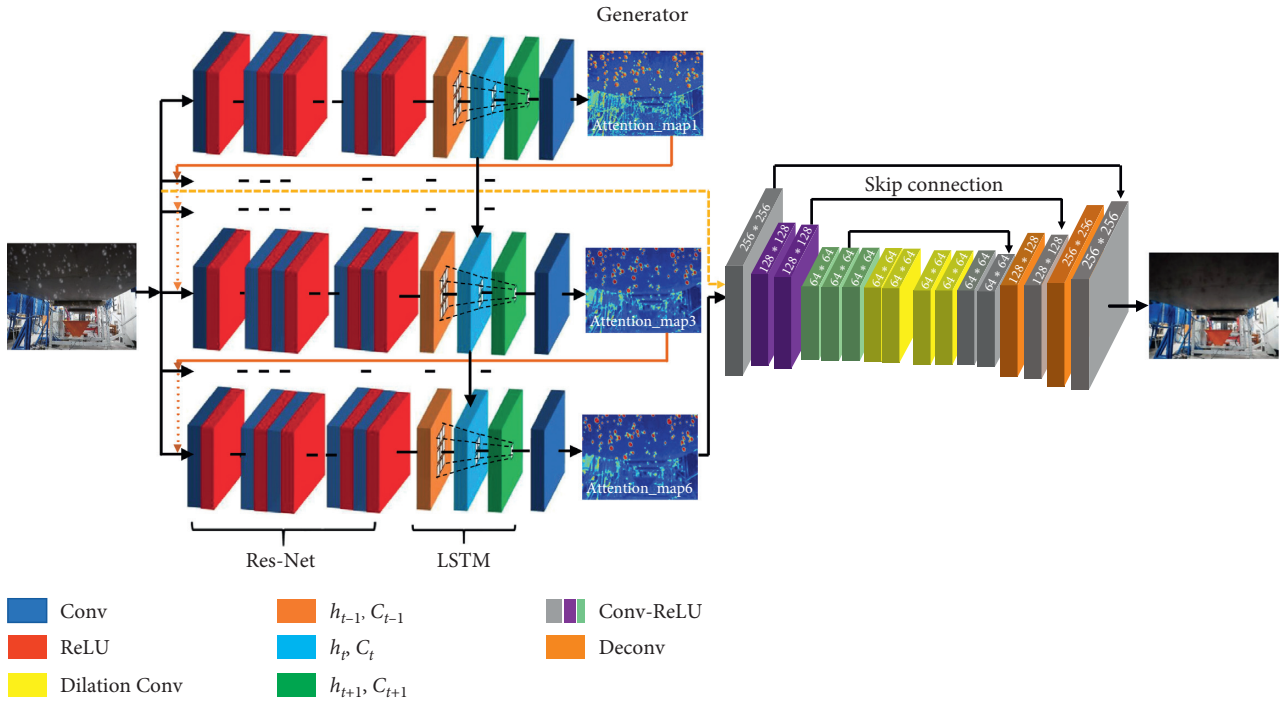


FIGURE 2: The generative network structure.

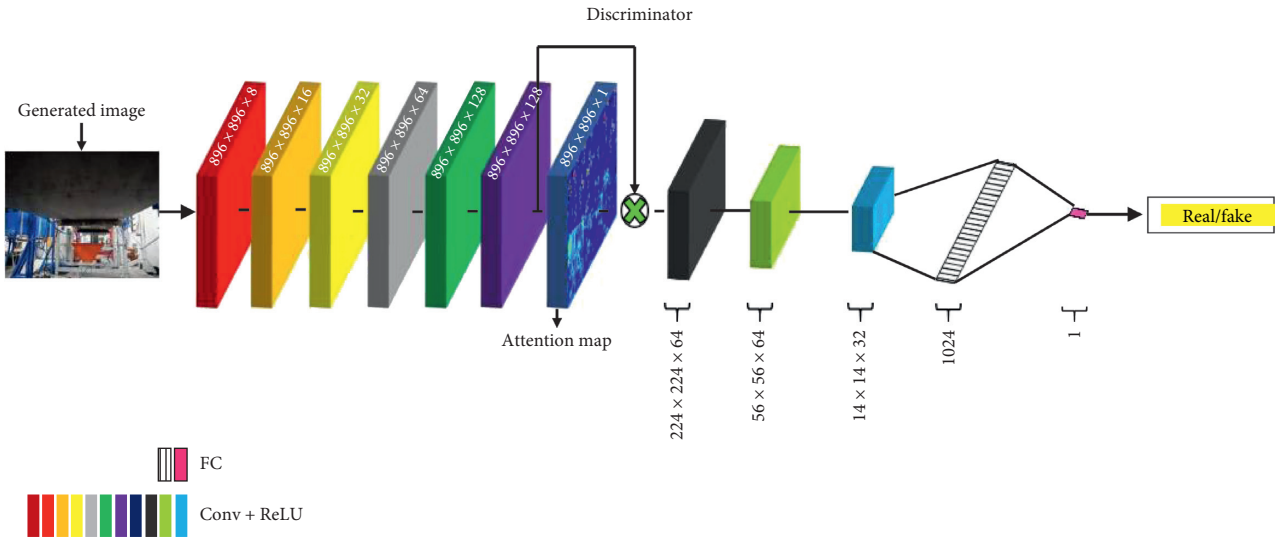


FIGURE 3: The discriminant network structure.

the mask attention model to extract the characteristics of water droplet regions and their surroundings. The addition of image morphology can more accurately display the water droplets and their surrounding areas. The part of the visual attention network designed in this paper is to learn mask and then generate attention map to detect the position of water droplets. The attention map and the input image are spliced into the automatic encoder of the generator network, and the final dewaterdrop rubber conveyor belt image is obtained by encoding and decoding.

The generator network consists of a circular visual attentive network and an automatic encoder with skip

connections. In order to detect water droplets more accurately, we designed a six-layer cyclic visual attentive network. Each layer (each time step) consists of five-layer ResNet (Deep Residual Network) [14], a ConvLSTM unit (Convolutional LSTM Network) [15], and a standard convolutional layer [16]. The ResNet is mainly used to extract features from the input image and mask of the previous block. Each residual block includes a two-layer convolution kernel with  $3 \times 3$  convolution with ReLU nonlinear activation function; it is used for image feature extraction. The extracted feature map and the initialized attention map are spliced and transferred to ConvLSTM for learning. Through

the updating of the cell state in ConvLSTM, the nondroplet part information is lost through the forgetting gate  $f_t$ ; the new water droplet feature information is determined through the input gate  $i_t$ . Update the cell state for better network learning. Firstly, the input gate passes the sigmoid activation function to determine the water droplet information which needs to be updated, and the tanh function generates a new vector candidate to update the water droplet information. The two steps are combined to discard the nonwater droplet information, adding new water droplet information for learning update; finally, the water droplet area feature is determined by the output gate  $o_t$ . The output characteristics of the ConvLSTM are input into the convolutional layer to generate 2D attention maps. Our ConvLSTM unit consists of an input gate  $i_t$ , a forget gate  $f_t$ , and an output gate  $o_t$ , as well as a cell state  $C_t$ . The interaction between states and gates along time dimension is defined as

$$\begin{aligned} i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i), \\ f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f), \\ C_t &= f_t \circ C_{t-1} + i_t \circ \tan h(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c), \\ o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o), \\ H_t &= o_t \circ \tan h(C_t), \end{aligned} \quad (3)$$

where  $X_t$  is a feature generated by ResNet;  $C_t$  is encoded in the state to be forwarded to the next ConvLSTM;  $H_t$  represents the output characteristics of the ConvLSTM unit; and operator  $*$  represents a convolution operation.

The autoencoder of the skip connections is composed of 16 Conv-ReLU blocks. The first layer of convolution uses a  $5 \times 5$  kernel, which obtains a large image receptive field and extracts more information about the water droplet image of the rubber conveyor belt. Convolution 2<sup>nd</sup>–6<sup>th</sup> layers use  $3 \times 3$  kernel stacks; compared with  $7 \times 7$  or  $5 \times 5$  kernels directly, the convolutional layer parameters are reduced by half when the effect is the same, which makes the network training convergence faster. In order to make the features of water droplets extracted by the convolutional layer more comprehensive, it is necessary to increase the receptive field of the convolutional layers but also to avoid too many weights between the convolutional layers, so we use different rates of dilated convolution to replace the traditional convolutional layer. This operation can extract features more comprehensively, the number of network weights is smaller, and the calculation efficiency is higher. The decoding part of the automatic encoder uses  $4 \times 4$  deconvolution and adds the average pooling layer. In order to better generate the waterdrop image and prevent the fuzzy output, the outputs of last 1<sup>st</sup>, 3<sup>rd</sup>, 5<sup>th</sup>, and 7<sup>th</sup> layers are added to skip connections [17]. Compared with [9], the experimental results show that the method is more effective.

**3.2.2. Discriminator Network.** To differentiate fake images from real ones, a few GAN-based methods adopt global and local image-content consistency in the discriminative part (e.g., [18, 19]). Like in the case of image inpainting, where the

regions to be restored are given, the local discriminator strategy for directly discriminating is useful. So, using an attentive discriminator to directly determine whether the water droplet area of our generated image is true is the most efficient method.

The function of discriminating the network is equivalent to the second classifier. The discriminant network consists of 9 convolutional layers. Each layer is connected to the ReLU activation function. The  $5 \times 5$  convolution kernel is used to extract and fuse the texture features. The first six output channels are 8, 16, 32, 64, 128, and 128, respectively. Extracting features from the 6<sup>th</sup> layer of the discriminator, outputting an attention mask through a convolution layer, and then multiplying the attention mask by the 6<sup>th</sup> volume of the discriminant network before inputting to the next layer cause the discriminator to focus on the area specified by the attention map. The latter 3-layer convolution uses a (stride=4) convolutional layer, and the lack of a pooling layer mainly draws on the techniques mentioned in the deep convolution generative adversarial network (DCGAN) [20], which not only extracts high-dimensional texture features but also makes input features smaller and more controllable. Finally, through two layers of fully connected layers, the image features extracted after dimension reduction are weighted, and a specific value is output to represent the probability that the input image is a dewaterdrop rubber conveyor belt image. The sigmoid is used for activation, and the output value is limited to [0, 1] for the decision of the second classifier.

Compared with [9], in generative networks, we increase the visual cycle attention network to the 6<sup>th</sup> layer and customize the convolution kernel size and the optimal activation function in each convolution network layer, deconvolution network layer, and dilated convolution layer. The autoencoder added 4-layer skip connections. In the discriminant network, the pooling layer is removed from DCGAN discriminator, and the convolution layer with stride larger more than 1 is adopted for downsampling to prevent gradient sparse so that the entire generation is more stable against the network without causing the network to not converge. After numerous experimental improvements, our experimental results have outperformed the state-of-the-art methods.

**3.2.3. Loss Function.** In this paper, the minimum and maximum game error expression in the generator network and discriminator network is the same as the original GAN definition [10] as follows:

$$\begin{aligned} \min_G \max_D V(D, G) &= E_{W \sim P_{\text{clean}}} [\log(D(w))] \\ &+ E_{I \sim P_{\text{waterdrop}}} [\log(1 - D(G(I))), \end{aligned} \quad (4)$$

where  $W$  is the image of the water drop rubber conveyor belt generated by the generation network and  $I$  is a true no-waterdrop image.

The generator network loss function is expressed as

$$L_G = 0.01 \times (\log(1 - D(O))) + L_{\text{ATT}}(\{A\}, M) + L_M(\{S\}, \{T\}) + L_P(O, T), \quad (5)$$

where  $L_{\text{ATT}}(\{A\}, M)$  is the mean square error (MSE) between the water droplet binary mask diagram of ResNet and ConvLSTM learning and the generated attention map;  $L_M(\{S\}, \{T\})$  represents the multiscale loss, which is MSE between the output extracted from the decoding layer and the true value of the same ratio; and  $L_P(O, T)$  represents the perceived loss and is used to measure the difference in global characteristics between the image of the dewaterdrop and the image of the real rubber belt.

$$L_{\text{ATT}}(\{A\}, M) = \sum_{t=1}^6 0.8^{6-t} L_{\text{MSE}}(A_t, M), \quad (6)$$

$$A_t = \text{ATT}_t(F_{t-1}, H_{t-1}, C_{t-1}),$$

where  $A_t$  represents the attention map of the cell state generated by the ConvLSTM. The initial attention map value is 0.5 and  $F_{t-1}$  is the splicing of the input image and the attention map.

$$L_M(\{S\}, \{T\}) = \sum_{i=1}^3 \lambda_i L_{\text{MSE}}(S_i, T_i), \quad (7)$$

where  $S_i$  represents the output extracted from the decoding layer and  $T_i$  represents the true value labeling in the same proportion as  $S_i$ . We use the output of the last 1<sup>st</sup>, 3<sup>rd</sup>, 5<sup>th</sup>, and 7<sup>th</sup> layers of the decoding layer, corresponding to the original size of 0.125, 0.25, 0.5, and 1, and the weights  $\lambda_i$  of the 3 layers are set to 0.4, 0.6, 0.8, and 1.0, respectively.

$$L_P(O, T) = L_{\text{MSE}}(\text{VGG}(O), \text{VGG}(T)). \quad (8)$$

Among them,  $\text{VGG}(O)$  and  $\text{VGG}(T)$  are the spatial features of the generated dewaterdrop image and the real water-free rubber conveyor belt image extracted by the pretrained VGG16.

The discriminator loss function is expressed as

$$L_D(O, W, A_N) = -\log(D(W)) - \log(1 - D(O)) + 0.05 \times L_{\text{map}}(O, W, A_N), \quad (9)$$

where  $L_{\text{map}}$  is the difference between the attention mask generated from a layer in the middle of the discriminator and the real attention map:

$$L_{\text{map}}(O, W, A_N) = L_{\text{MSE}}(D_{\text{map}}(O), A_N) + L_{\text{MSE}}(D_{\text{map}}(W), 0), \quad (10)$$

where  $D_{\text{map}}$  represents the process of producing a 2D map by the discriminative network.

The loss function of the generator visual attentive-recurrent network is designed in this paper. The loss function in each recurrent block is defined as the mean squared error (MSE) between the output attention map and the binary mask at time step  $t$ , and the cycle time step is increased to 6. The earlier attention maps have smaller values and get larger when approaching 6<sup>th</sup> time step indicating the increase in

confidence, the error of generative network learning is smaller, and a better attention map of the water droplets of the rubber conveyor belt is generated accordingly. In addition, to prevent the output of the autoencoder from blurring and increase the skip connections and to generate an image without water droplets, we extract the output image at the decoder end and then calculate the mean square error of the ground truth in the same proportion in order to reduce the error between the generated image and the ground truth to zero. We scale in the same size as the original size: 0.125, 0.25, 0.5, and 1, making the generator loss function error the most accurate and minimal. Discriminator loss function: we use an attentive discriminator loss function in [9]. Specifically, we extract the features from the interior layers of the discriminator and feed them to VGG16. We define a loss function based on the VGG16's output and the attention map. The loss function is defined as the mean square error calculated between the output and the attention map. Moreover, we use the VGG16's output and multiply it with the original features from the discriminative network before feeding them into the next layers. Our underlying idea of doing this is to guide our discriminator to focus on regions indicated by the attention map. Finally, at the end layer, we use a fully connected layer to decide whether the input image is fake or real.

## 4. Experiments

In this section, first we introduce the metrics for evaluating. Next, we describe datasets used for training and testing our method. Then, our training setup is given. Finally, we provide quantitative and qualitative comparisons with CNN, DSC, LP, and ATTGAN.

**4.1. Evaluation Indicators.** We use peak signal-to-noise ratio (PSNR) [21] and structural similarity (SSIM) [22] as primary image evaluation criteria. These two evaluation methods are the most common and widely used objective evaluation indicators for image quality. In addition, we add some newer full-reference objective image quality measures, such as MS-SSIM [23], IW-SSIM [24], VIF [25], and FSIM [26]; they could also be used to better compare the proposed method with existing methods. Finally, the average time (time) required for each water droplet image removal under the statistical test set and the time efficiency of water droplet image removal have great influence from the viewpoint of rubber conveyor engineering application.

**4.2. Dataset.** The dataset of this paper [9] consists of 861 pairs of natural image raindrop datasets in various scenes and 400 pairs of rubber conveyor belt water droplet images prepared. It is divided into training set (RCB-Train), test set (RCB-Test), and verification set (RCB-Val) according to 75%, 15%, and 1% of the dataset. We used Sony a6300 for the rubber conveyor belt water droplet image acquisition, and our glass slabs have the thickness of 3 mm and are attached to the camera lens. We set the distance between the glass and the camera varying from 5 cm to 10 cm to generate diverse

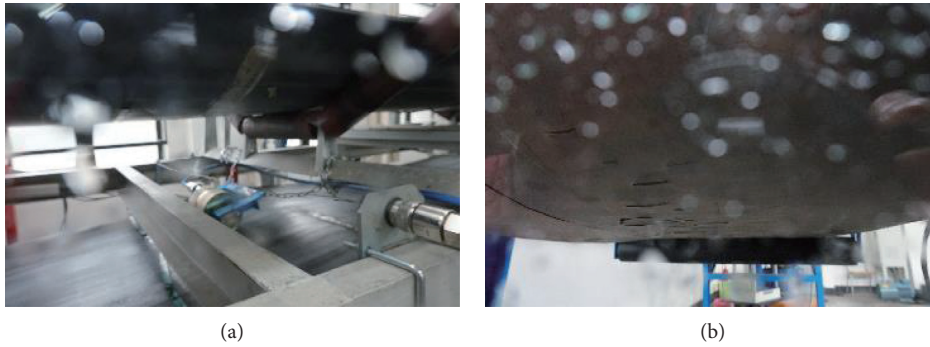


FIGURE 4: The shape of water droplets of the rubber conveyor belt. (a) Big white pixels. (b) Small white pixels.

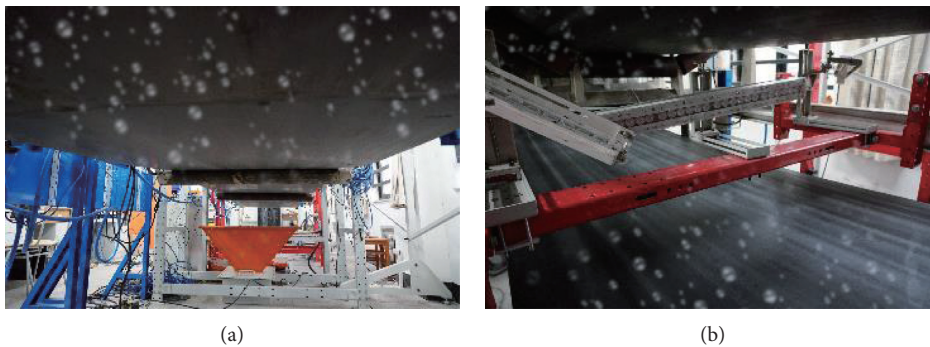


FIGURE 5: The shape of water droplets of the rubber conveyor belt. (a, b) Stilliform water droplets.

water droplet images of the rubber conveyor belt. In order to expand the dataset and improve the generalization ability of the model, Photoshop software was also used to synthesize the water droplets of the rubber conveyor belt image, and a total of 3,600 pairs of  $480 \times 720$  size with/without water droplets were enhanced by data augmentation. This paper deliberately selects 20 water droplets with a relatively uniform background pixel position as a test set (RCS-Waterdrop20) for measuring PSNR, SSIM, and so on. In the indoor and outdoor working environment, the rubber conveyor belt has illumination lamps for lighting to assist its work. This image of water droplets produced by the imaging monitoring device produces two shapes: the shape of white pixels and stilliform water droplets. Figures 4 and 5 show the specific shape of water droplets.

**4.3. Training Details and Implementation.** The experiment in this paper is carried out under the 64-bit Ubuntu18.04 system. The TensorFlow 1.10.0 deep learning framework is used for network training. The hardware configuration is Intel(R) Core(TM) i7-6800K CPU @ 3.40 GHz, 64 GB RAM, GeForce GTX1080 Ti GPU. The training initial learning rate is 0.002, and the learning rate exponential decay method is used to attenuate the learning rate by 0.1 times every 10,000 iterations. Thus, using a large learning rate, the optimal solution can be quickly obtained, and the model training is more stable. In addition, due to the limitation of GPU memory, the number of samples per training is 1 (batch

size = 1), the number of iterations is more stable, which is 400,000 times (epoch = 400k), the model converges, and the SSIM and PSNR value lines reach the highest point and continue steadily.

In order to highlight the significant effect of the water droplet image removal on the rubber conveyor belt, the average value of PSNR, SSIM, MS-SSIM, IW-SSIM, VIF, and FSIM and time of the rubber belt water droplet image were tested by using the code disclosed by the original author in the comparison method, and the uniform size of the test set image was  $240 \times 360$ . The results are shown in Table 1.

Through the evaluation data values in Table 1, it can be seen that the proposed algorithm has improved significantly. The values of SSIM of the algorithm tested in this paper are improved by 0.1326, 0.2532, 0.2474, and 0.0195 compared with those of CNN, DSC, LP, and ATTGAN. It shows that the distortion degree of the water droplet image after the rubber conveyor belt is removed is minimized, and the water droplet image and the original image have higher similarity. The values of PSNR were improved by 3.7047 dB, 4.7846 dB, 3.4821 dB, and 2.5983 dB, respectively. It shows that the image quality of the rubber belt is better after the water droplet is removed, and the image feature information is more abundant. In addition, the values of MS-SSIM, IW-SSIM, VIF, and FSIM are 0.9234, 0.9501, 0.9391, and 0.9676. According to Zhang et al. [27], the two new image quality evaluation algorithms, FSIM and IW-SSIM, have the highest accuracy, which proves the effectiveness of the proposed algorithm for water droplet image removal in rubber

TABLE 1: Test results on RCB-waterdrop20.

Name	CNN	DSC	LP	ATTGAN	Ours
SSIM	0.7961	0.6755	0.6813	0.9092	0.9287
PSNR/dB	26.6778	25.5979	26.9004	27.7842	30.3825
MS-SSIM <sup>1</sup>	0.7922	0.6770	0.6803	0.9056	0.9234
IW-SSIM <sup>2</sup>	0.8225	0.7003	0.7145	0.9307	0.9501
VIF <sup>3</sup>	0.8070	0.6890	0.6967	0.9183	0.9391
FSIM <sup>4</sup>	0.8337	0.7145	0.7215	0.9462	0.9676
Time/s	9.26	18.66	260.01	5.00	5.1

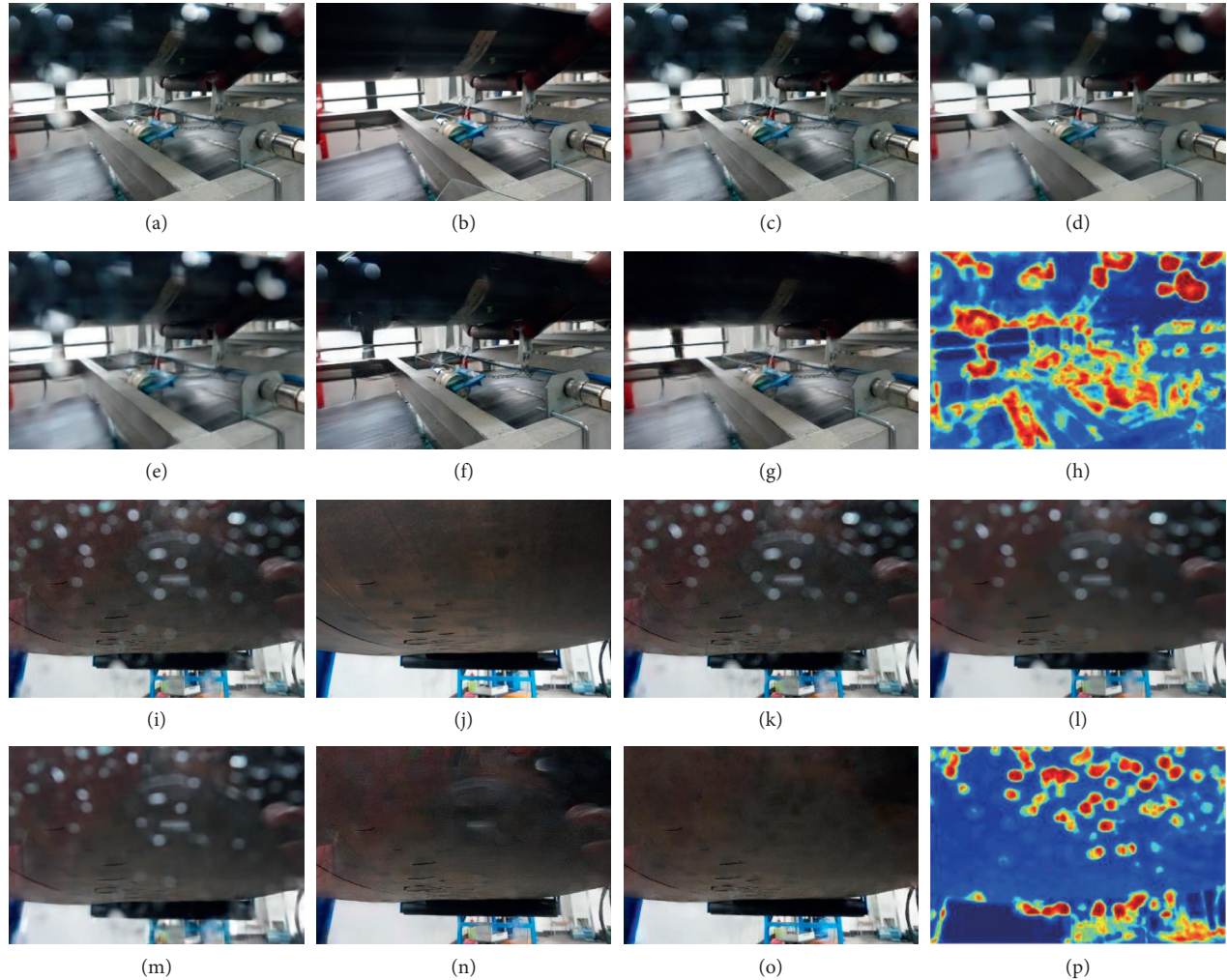


FIGURE 6: Results of comparing a few different methods on the rubber conveyor belt image of the water droplet shape of white pixels, and figures show in the sequence input, ground Truth, DSC results, LP results, CNN results, ATTGAN results, Our results detected waterdrop attention map. Nearly all water droplets are removed by our method despite the diversity of their colors, shapes, and transparency. (a) Input 1. (b) Ground truth. (c) DSC. (d) LP. (e) CNN. (f) ATTGAN. (g) Ours. (h) Attention map. (i) Input 2. (j) Ground truth. (k) DSC. (l) LP. (m) CNN. (n) ATTGAN. (o) Ours. (p) Attention map.

conveyor belts and also proves that the algorithm is currently the best. In this paper, the time consumed by our algorithm to remove water droplets in the image of the rubber conveyor belt is much less than that of the traditional machine vision algorithm, such as DSC and LP, and the other three deep learning algorithms take less time, especially ATTGAN and our algorithm take about 5 s, which greatly improves the efficiency of the model.

The above proves the superiority of the algorithm from the objective point of view. Figures 6 and 7 demonstrate the effectiveness of our proposed algorithm for water droplet removal in different shapes of rubber conveyor belt images.

Figure 6 represents the results on the rubber conveyor belt of the water droplet shape of white pixels, and Figure 7 represents the results on the rubber conveyor belt of the water droplet shape of stilliform water droplets. It can be



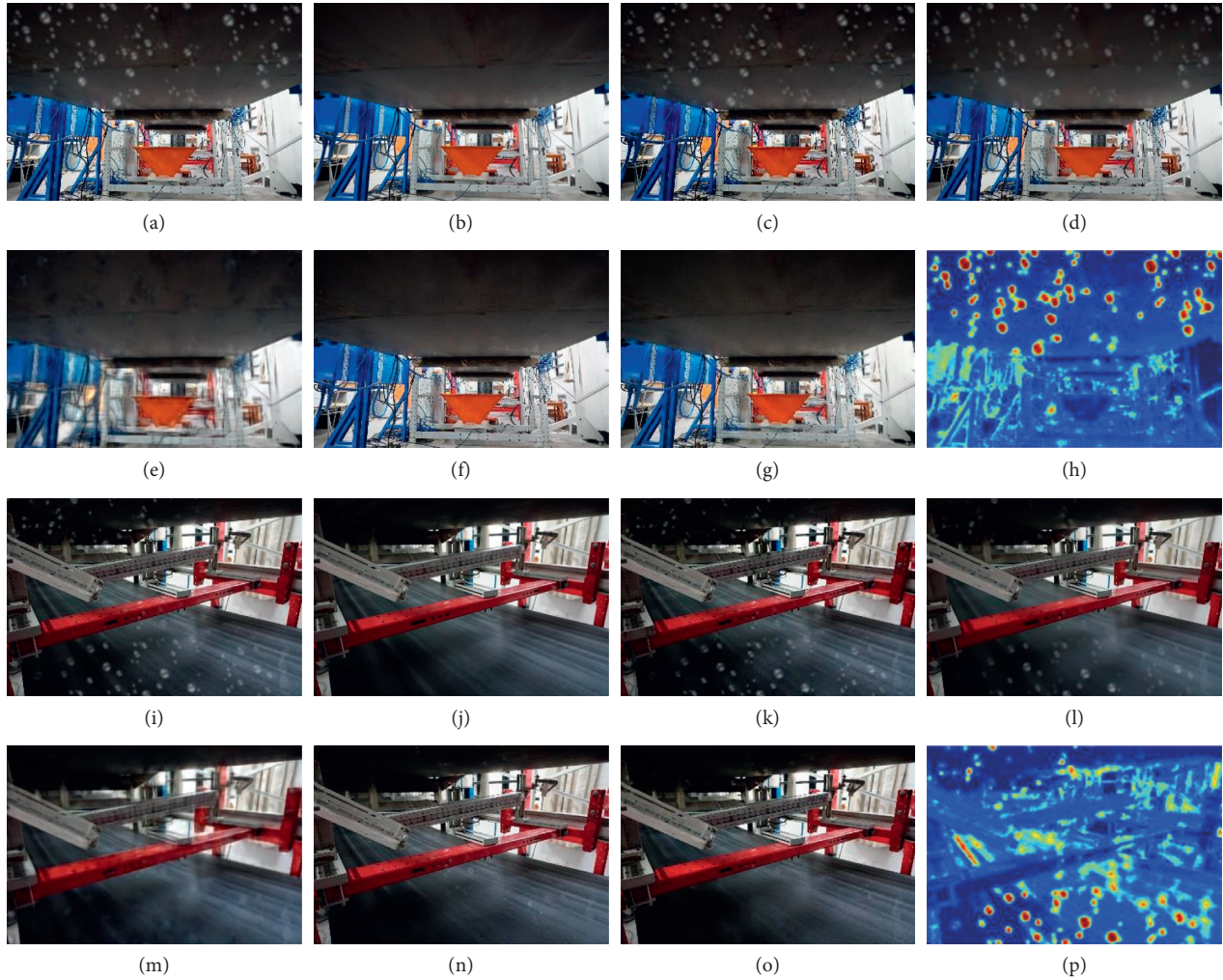


FIGURE 7: Results of comparing a few different methods on the rubber conveyor belt image of the water droplet shape of stillform water droplets, and figures show in the sequence input, ground truth, DSC results, LP results, CNN results, ATTGAN results, our results detected waterdrop attention map. Nearly all water droplets are removed by our method despite the diversity of their colors, shapes, and transparency. (a) Input 1. (b) Ground truth. (c) DSC. (d) LP. (e) CNN. (f) ATTGAN. (g) Ours. (h) Attention map. (i) Input 2. (j) Ground truth. (k) DSC. (l) LP. (m) CNN. (n) ATTGAN. (o) Ours. (p) Attention map.

seen that DSC and LP have poor removal effect on water droplets because of these two methods mainly has a certain effect on the removal of the rain streak. CNN has a certain effect on the removal of water droplets from the rubber conveyor belt, but it is obvious that it is limited to the removal of small water droplets, and the restored background image is blurred. The algorithm proposed by us has better effect on the removal of water droplets in different shapes for large and small water droplets, and the objective evaluation image index is much higher than other algorithms.

## 5. Conclusions

This paper proposes a new method for detecting and removing water droplets on rubber conveyor belts based on the attentive generative adversarial network. By increasing the number of visual attentive-recurrent network layers, skip connections are added on the automatic encoder, changing

the convolution kernel size and threshold that is used to determine if the pixels in the image are water droplets. The method utilizes a generative adversarial network, where the generative network produces the attention map via an attentive-recurrent network and applies this map along with the input image to generate a waterdrop-free image through an autoencoder. In the process of training, through data standardization, data enhancement, improved network optimizer, and learning rate changes, our algorithm is far superior to other advanced methods in objective quantitative evaluation criteria PSNR, SSIM, MS-SSIM, IW-SSIM, VIF, FSIM, and time, and the rubber conveyor belt background is seen from subjective visual effects, and the information of background is clearer and richer. This is of great significance for ensuring that the belt conveyor monitoring imaging system monitors and keeps the High Definition (HD) working state in real time. Moreover, the algorithm of this paper has a very good effect on the removal of natural water droplet images.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. 51504164), Key Scientific and Technological Support Projects of Tianjin Key R&D Program (grant no. 18YFZCGX00930), the Tianjin Natural Science Foundation (grant no. 17JCZDJC31 600), and Tianjin Key Research and Development Program (grant no. 18YFJLCG00060).

## Supplementary Materials

It is mainly to explain the removal effect of the network model of our training on the water droplets of the rubber conveyor belt from the visual subjective effect. The network model we tested was a model that was iterated 385,000 times. In the process of testing, the window displays a total of 8 pictures, where `src_image` is the image of the water droplets on the rubber conveyor belt and `derain_ret` is the image generated by the attentive generative adversarial network model after removing the water droplets, and `atte_map_4` is one of the attention maps for detecting the position of the water droplets (the generated network in this paper includes a six-layer cyclic visual attentive network; the main function is to find the regions of water droplets and also to find the surrounding regions of water droplets; therefore, the attention heat map is visualized during the test phase; with the increasing of time step, our network focuses more and more on the raindrop regions and relevant structures; the window displays a total of 6 attention heat maps). (*Supplementary Materials*)

## References

- [1] M. Peter, "Conveyor belt structure," US Patent No. 3,144,930, 1964.
- [2] M. Fukuda, "Method for detecting longitudinal tear in a conveyor belt," US Patent No. 5,168,266, 1992.
- [3] M. H. A. Wahab, C.-H. Su, N. Zakaria, and R. A. Salam, "Review on raindrop detection and removal for weather degraded image," in *Proceedings of the IEEE International Conference on Computer Science and Information Technology 2013*, IEEE, Amman, Jordan, March 2013.
- [4] J. Xu, W. Zhao, P. Liu, and X. Tang, "Removing rain and snow in a single image using guided filter," in *Proceedings of the 2012 IEEE International Conference on Computer Science and Automation Engineering (CSAE)*, Zhangjiajie, China, May 2012.
- [5] Y. Luo, Y. Xu, and H. Ji, "Removing rain from a single image via discriminative sparse coding," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, IEEE, Santiago, Chile, December 2015.
- [6] C. Szegedy et al., "Going deeper with convolutions," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2014.
- [7] D. Eigen, D. Krishnan, and R. Fergus, "Restoring an image taken through a window covered with dirt or rain," in *Proceedings of the 2013 IEEE International Conference on Computer Vision IEEE*, Sydney, Australia, December 2013.
- [8] Y. Li, R. T. Tan, X. Guo et al., "Rain streak removal using layer priors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2736–2744, Vegas, NV, USA, June 2016.
- [9] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu, "Attentive generative adversarial network for raindrop removal from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June 2018.
- [10] I. J. Goodfellow, "Generative adversarial nets," in *Proceedings of the International Conference on Neural Information Processing Systems*, Montreal, Canada, December 2014.
- [11] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the International Conference on International Conference on Machine Learning JMLR.Org*, Lille, France, 2015.
- [12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, <https://arxiv.org/abs/1409.1556>.
- [13] E. Dougherty, *Mathematical Morphology in Image processing*, CRC Press, Boca Raton, FL, USA, 2018.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2016.
- [15] X. Shi, "Convolutional LSTM network: a machine learning approach for precipitation nowcasting," 2015.
- [16] J. Gu, Z. Wang, J. Kuen et al., "Recent advances in convolutional neural networks," *Pattern Recognition*, vol. 77, pp. 354–377, 2018.
- [17] X. J. Mao, C. Shen, and Y. B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," 2016, <https://arxiv.org/abs/1603.09056>.
- [18] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1–14, 2017.
- [19] Y. Li, S. Liu, J. Yang, and M.-H. Yang, "Generative face completion," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI, USA, July 2017.
- [20] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, <https://arxiv.org/abs/1511.06434>.
- [21] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, p. 800, 2008.
- [22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004.
- [23] Z. Wang, E. Simoncelli, and A. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proceedings of the Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003, Pacific Grove, CA, USA, November 2003.

- [24] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, 2011.
- [25] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.
- [26] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
- [27] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proceedings of the 2012 19th IEEE International Conference on Image Processing IEEE*, Orlando, FL, USA, October 2012.