APSIPA Transactions on Signal and Information Processing, 2023, 12, e101

Original Paper

FOANet: A Feedback Operation-Attention Network for Single Image Haze Removal

Chia-Lin Liu[1], Lei Chen[2], Ling Lo[2], Pin-Jui Huang[2], Hong-Han Shuai[2], Wen-Huang Cheng[2*], Ching-Hsuan Wang[3] and Fan Chou[3]

[1] *University of Washington, Seattle, USA*
[2] *National Yang Ming Chiao Tung University, Hsinchu City, Taiwan*
[3] *Chunghwa Telecom Laboratories, Taiwan*

ABSTRACT

Single image dehazing has become an important vision task for prevailing image degradation caused by detrimental atmosphere transmission conditions. Attention mechanism has been widely utilized in learning based methods to assist the model to discard redundant information and hence boost the performance. However, existing methods are mainly dealing with either channel-wise or pixel-wise attention, which requires much more parameters when the size of feature maps increases. In this paper, we introduce the operation-wise attention in the proposed Feedback Operation-Attention Network (FOANet) to focus on attaining optimal combination of network operations for image dehazing. Specifically, our model consists of two main steps. First, the extracted features of hazy input image are fed to the novel operation-attention block which can adjust the weight of different operations dynamically to produce the optimal processed features. The operation space of an operation-attention block comprises vanilla and dilated convolutions with different kernel sizes along with max pooling and average pooling. Second, we adopt curriculum learning with feedback mechanism to continually refine the features in a recurrent fashion and generate the haze-free image. The

*Corresponding Author: Wen-Huang Cheng, whcheng@nycu.edu.tw.

Received 13 July 2022; Revised 03 January 2023
ISSN 2048-7703; DOI 10.1561/116.00000144

experimental results on both synthetic and realistic subsets of RESIDE dataset have demonstrated our method can perform dehazing favorably against other dehazing algorithms.

1 Introduction

The visibility of outdoor images are often deficient due to some bad weather factors, such as fog, rain and haze. These weather phenomenons usually consist of visible aerosol of some tiny water droplets or ice crystals suspended in the air. The presence of these tiny particles often cause multiple light reflections between objects and the camera, resulting in noticeable degradation of visibility and visual contrast. For computer vision tasks like outdoor object detection [47], autonomous navigation [25] and remote sensing [29], the quality of the input images will highly affect the model performance since most of the methodologies proposed assume images with clear scenes under good weather conditions as input. As a result, several research works have focused on developing dehazing algorithms to eliminate the haze effects in images. However, in the real world scenario, the compositions of haze are often floating particles in the air, and so atmospheric scattering adds nonlinear and data-dependent noise to the outdoor images, which makes haze removal a more complicated process.

Currently, most of the previous dehazing works are based on the physical model of propagation of the light through the atmosphere proposed by Koschmieder [32]. The atmosphere scattering model formulates a simple linear model to deconstruct the haze formation of an image. Several researchers have followed the simple model and then focused on generating the transmission map intermediately to help diminish the hazy effect of images. These approaches can be roughly divided into those requiring multiple images to leverage auxiliary information from the scene [30, 48, 49], and those with only a single degraded image [8, 22, 26]. For the algorithms using multiple images [30, 48, 49], they usually require additional modalities like depth and texture to accomplish the task. Though they could achieve rather fine results sometimes, the additional information is usually unavailable and thus may not be practical when it comes to real-world data. On the other hand, using only a single hazy image is way more challenge owing to the fact that the atmosphere scattering model implies an under-constraint equation. The uncertainty makes the dehazing process more tricky. In order to estimate the uncertain transmission map, a variety of priors have been explored to enhance the visibility of the hazy images [10, 20, 21, 26, 51, 62]. Nevertheless, these approaches assume that the depth map of an image is a local constant. Therefore the obtained hazy-free images may sometimes suffer from artifacts and tend to be over-enhanced.

With the rapid development of deep learning, Convolutional Neural Networks (CNNs) has achieved great success on capturing visual features in image. Recently, a lot of efforts have been devoted to develop data-driven solutions for image dehazing based on learning algorithms [8, 13, 19, 35, 36, 40, 53–55, 57, 60, 68, 71, 72]. Instead of exploiting background knowledge as priors, some approaches utilize CNNs to learn the mapping between a hazy and clean images by predicting the parameters in the atmospheric scattering model, while some approaches seek for a direct mapping without the atmospheric scattering model by employing the generative models. Though powerful learning-based approaches seldom generate noises like color distortion when recovering the hazy images, the visual results may still have limitations and image details may be missing since the learned features may not be bound in the haze-related features.

Therefore, in this paper, we proposed a novel end-to-end solution, namely Feedback Operation-Attention Network (FOANet), which aims recovering a single hazy image using the feedback mechanism and novel operation-attention blocks. The feedback mechanism is introduced to eventually obtain the combination of features on different levels, to fuse low-level and high-level features and to remove the haze while retaining as more image details as possible. The operation-attention blocks in our proposed model can weight specific operations. Despite the non-homogeneous nature of haze, our proposed feedback operation-attention blocks can still eliminate haze in different level iteration by iteration. Moreover, a curriculum learning strategy is applied to improve the performance step by step and enforce a better output in pace with each iteration. Since the haze of an image is not always homogeneous in the real world, instead of employing the physical model of haze formation, our method can be more straightforward which can directly produce the estimated haze-free image from the original degraded image.

The rest of this paper is organized as follows. Section 2 reviews the related works, and Section 3 presents the proposed method. Section 4 presents the evaluations and conclusions are offered in Section 5.

2 Related Work

2.1 Single Image Dehazing

To formulate the process of image dehazing, one of the most important atmospheric scattering model of haze formation is proposed by Koschmieder [32] and has been widely used in the previous dehazing works. The model can be written as:

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

where I is the observed hazy image, J is the scene radiance (hazy-free image) to be recovered, A is the global atmospheric light which indicates the luminance of

the light source from infinite distance away, and t is the medium transmission map, and x denotes the pixel location. When the atmospheric light A is homogeneous, the transmission map t, which describes the light portion that is not scattered and reaches the camera:

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

where d is the distance from the scene point to the camera, and β is the scattering coefficient of the atmosphere. By solving the formulations (1) and (2), we will be able to restore haze-free images. However, this process becomes an ill-posed problem due to the existence of multiple valid solutions. Therefore a number of methods have been proposed with different basis to constraint the uncertainty.

2.1.1 Prior Based

Many of the previous methods for image restoration applied priors on deprecated images to better estimate the recovery function [10, 20, 21, 26, 51, 62]. Since the haze formation model involves the transmission map and the atmosphere light, the majority of the existing methods employ priors on scene depth to infer the surrounding depth information. Carr *et al.* [10] assume most of the hazy images are captured from outdoor cameras, allowing to conclude that objects appear around the top of images are usually further away. Based on this observed characteristic, their work improves the robustness of image dehazing techniques. Some other approaches assume that the images and the corresponding depths are piece-wise constant and use those priors based on statistical properties to estimate the original hazy-free image [21, 26]. Among them, the dark channel methodology proposed by He *et al.* [26] deserves a special mention. The novel and efficient dark channel priors are based on the observation that clear images are colorful with textures and shadows, and thus contain at least one channel with low intensity. They further discover the relations between the dark channel values and the transmission of each pixel, providing an estimate of the depth information to better obtain high-quality results. The main disadvantage of the methodologies using priors is that the assumptions may not apply under all circumstances and thus introduce some errors to the transmission estimation.

2.1.2 Learning Based

Motivated by the success of CNNs in other tasks and the availability of large-scale synthetic datasets, data-driven approaches for image restoration have received great attention in the last few years [8, 13, 19, 35, 36, 40, 53–55, 57, 60, 68, 71, 72]. Cai *et al.* [8] proposed an end-to-end haze removal architecture

based on CNNs, predicting the transmission map from a hazy image input and further perform the recovery. Ren *et al.* [54] developed a multi-scale network (MSCNN) as the learning framework to estimate a fine transmission map. Instead of computing the transmission map and the atmospheric light value separately, Zhang *et al.* [69] developed the densely connected pyramid dehaze network (DCPDN) to evaluate the transmission map, the atmospheric light, and the hazy image jointly. Some approaches [18, 19, 40, 72] adopted the generative adversarial network (GAN) to better recover the images with intense haze. Li *et al.* [40] proposed an end-to-end approach based on a conditional generative adversarial network (cGAN), which directly generates a hazy-free image from a hazy input without the use of haze formation model.

2.2 Attention Mechanism

Attention mechanism was originally proposed to introduce long-term relationship to the machine translation task in natural language processing field [61] and has demonstrated its power in boosting the performance of deep networks for a variety of computer vision tasks [15, 17, 64]. In the later sections, we will introduce some existing attention mechanisms.

2.2.1 Channel Attention

The main concept is to compress the channel information into low-dimensional representations to distinguish the importance of different channels and thus achieve channel attention. SE-Net [28] is the first attention method to learn channel information and achieves state-of-the-art performance.

2.2.2 Spatial Attention

Besides the aforementioned channel attention, some other researches also explore the attention of spatial information, which is to consider the distribution over the plane of 2D matrix and focus on the specific area of each channel. Representing works include [38, 41, 45].

2.2.3 Operation Attention

Different from channel and spatial attention, operation attention aims to consider attention over operations and give out the optimal combination of operations. Suganuma *et al.* [58] applied operation attention to restore images with unknown distortions. Our work is much inspired by [58] with additional learning strategy including curriculum learning and feedback loop to generate clearer images.

3 The Proposed FOANet

Most of the existing dehazing methods [8, 22, 26, 69] are based on a haze formation physical model proposed by Koschmieder [32]. The model assumes that the pollutants and the atmospheric light in the hazy images are equally distributed. Nevertheless, haze is not always homogeneous in the real world, and assuming that hazy scenes follow the atmospheric scattering model may be potentially hazardous to the model. In this paper, we propose an FOANet model and training scheme that directly recovers the hazy images without the above-mentioned assumption. In this section, we will first describe the overall architecture of the proposed model (Section 2.1.1). After that, the attention-based operation will be detailed (Section 3.2). Following, the feedback mechanism and the curriculum learning scheme will be covered (Section 3.1). Finally, the loss functions will be formulated (Section 3.3).

Our proposed FOANet (cf. Figure 1) consists of a feature extraction backbone (FEB) followed by a feedback operation-attention block (FOA) and a 1×1 convolution layer for feature map pooling. The FEB is composed of a stack of residual blocks and is designed to encode the important information from the input images. The extracted features are passed to the FOA as the input. In the FOA, we introduce the feedback mechanism along with the curriculum learning strategy to break down the complex dehazing procedure into multiple iterations. The feedback connection reroutes the output of the FOA back to the input. The learned high-level features can thus correct the low-level input in the next iteration to learn the hazy removal model

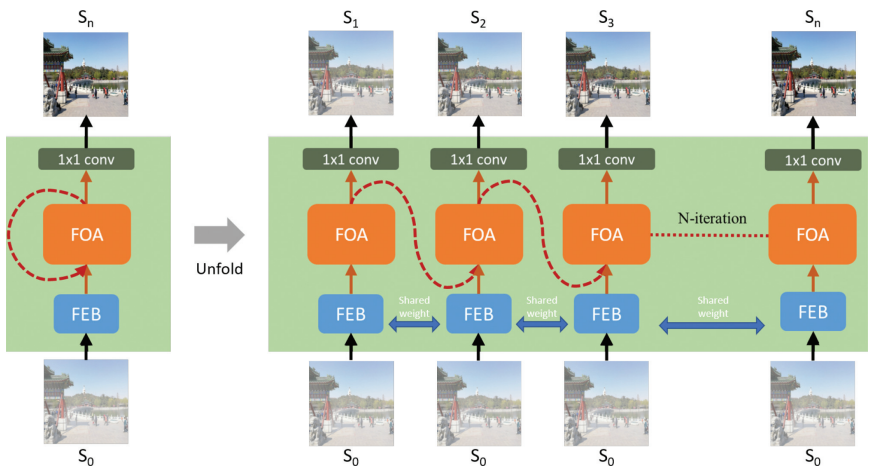


Figure 1: Overview of our proposed feedback operation-attention network (FOANet) for haze removal. The red dotted arrows indicate the feedback connections.

while preserving as many details as possible. Furthermore, our FOA takes advantages of attention mechanism and automatically searches the suitable model architecture for each specific input image with the selective capacity on different operation. After a few iterations, the output of FOA will be fed into an 1×1 convolution layer to perform pooling on feature maps that decrease the number of feature maps while retaining the salient features.

3.1 Feedback Mechanism with Curriculum Learning

The feedback mechanism can be considered as the feature fusion between iterations. Such concepts can be adopted to different computer vision tasks, including image super resolution [24, 43], visual attention [9], human pose estimation [11] and crowd counting [46, 56]. In our approach, we introduce the feedback mechanism to accomplish feature fusion within different iterations of the same network and to generate more informative representations while preserving as many details as possible. Furthermore, in order to make each iteration of FOA carry a notion of output that can correct the input, we adopt the curriculum learning strategy while training to dehaze the image progressively.

To implement the concept of feedback, there are two requirements in the system: (1) iterativeness and (2) redirecting the output of the subsystem, so that the collected losses of each iteration can be used to amend the final result. The recurrent process makes the original hazy image undergo our FOA repeatedly, and less hazy output is generated per iteration. Our recurrent operation-attention networks are trained to produce less and less hazy output at each iteration and converged to a hazy-free image in the end.

As shown in Figure 1, our proposed FOANet can be unrolled into N iterations. The network can be expressed as:

$$S_0 = FEB(I_{hazy}) \tag{3}$$

$$S_n = \Omega^n(S_0, S_{n-1}),\ n \geq 1\ for\ n \in \mathbb{Z} \tag{4}$$

$$I^n_{output} = \Phi(S_n) \tag{5}$$

where S_n means the feature extracted in the $n-th$ iteration, FEB is the feature extraction backbone, Ω indicates the main FOA block, and Φ represents the output layer which compresses the features into the output images.

On the other hand, curriculum learning [6] is well known as an efficient learning strategy which gradually increases the difficulty of the learned target just as the procedure when human learned. Early research on curriculum learning is dedicated to a single task. Later on, more approaches proved that it can be extended to multiple tasks sequentially and resolve fixation problem in image restoration [23, 52].

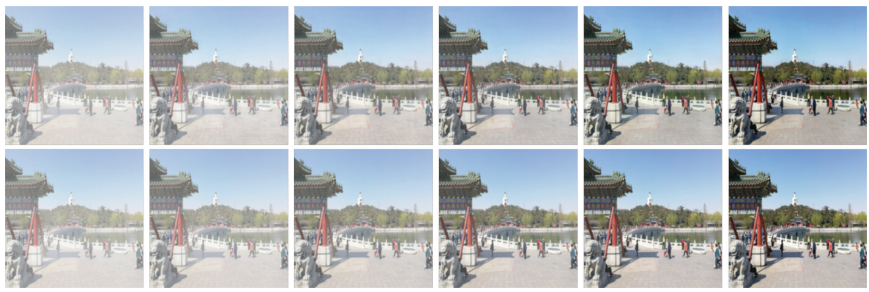


Figure 2: The upper row is the sample example of curriculum learning outputs of our network, and the bottom row is the target images. From left to right: input, $n = 2$, $n = 4$, $n = 6$, $n = 8$, and the final output (hazy-free image), respectively.

Technically, it is difficult to obtain a hazy-free image directly without the assistance of the atmospheric model and transmission map. The penalty of objective back-propagation may be affected by abundant factors. To alleviate the error correction process of each iteration and fully exploit the feedback iteration, we apply the curriculum learning strategy during training to make sure the network improves step by step. Moreover. the probability of divergence is also reduced by this way. During training, we interpolate between the hazy-free (ground truth) and hazy image to represent less density of haze as the target images $(I^1_{target}, I^2_{target}, \ldots, I^N_{target})$ for the FOA block to learn on different stages. We aim to compromise the complex dehaze problem into multiple stages, so that the model can learn the dehazing procedure stage by stage by setting the intermediate milestones for the model. Sample examples of the target outputs and the network outputs from all the stages are shown in Figure 2.

3.2 Attention-Based Operations

Attention mechanism has been utilized in many computer vision problems [27, 28, 34, 50, 61]. It has been proven as an efficient way to disregard the noise and focus on what is relevant by providing weights to specific items. Taking advantage of the attention model, which can be referred as a learning filter, [58] first combines it with parallel operations layer to restore an image. Here we introduce the idea of attention-based operations into our FOA. As mentioned, the haze in real-world is not always equally distributed over an image. Our model aims to take different operations according to individual attention value in each iteration, and therefore diverse combinations of multiple operations in each iteration can be performed for different density level of haze reduction. The FOA is adopted to transfer various operation layers to different level features so that each feature can be more adaptively used and delivered. Moreover, the learnable attention value and the operation selection are trained

to form the best FOA architecture tailored for each input image respectively. The selective ability of the attention-based operation makes the model more robust when there are multiple levels of haze covering the input image unevenly. The concept of attention-based operation is illustrated in Figure 3.

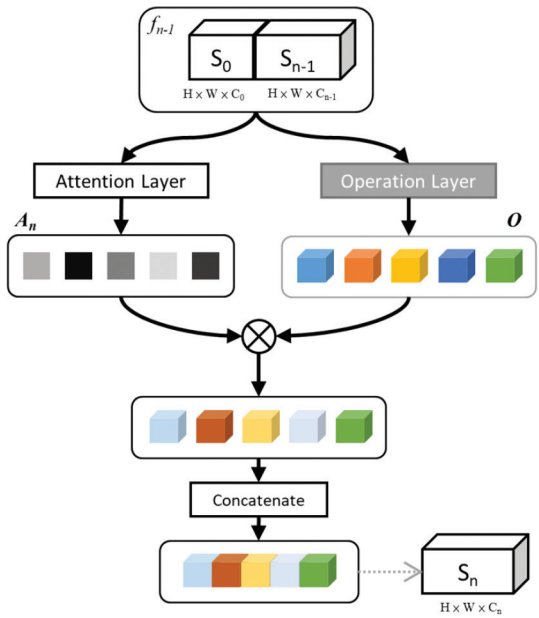


Figure 3: Structure of the attention-based operation layer, comprising operation layer, attention layer, and a concatenation operation. Each layer generates different weights of attention.

We denote the input feature of the operation-wise attention layer in the n-th iteration as $f_n = (S_0,\ S_{n-1})$. S_0 is the first input deduced by the feature extraction block. In the end of an iteration, the n-th output feature of the FOA S_n will be channel-wised concatenated with S_0 as $f_{n+1} = (S_0,\ S_n)$ to be the input of the operation-wise attention layer in the $(n+1)$ iteration. The size of f_n will be $H \times W \times (C_0 + C_{n-1})$, where H, W, and C are the height, width, and the number of channels, respectively. A set of I operations in operation layer is expressed as $O = \{o_1,\ o_2, \ldots,\ o_I\}$. Given input feature f_n, the attention weights used in each iteration n on operation set O is $A_n = \{a_n^{o_1},\ a_n^{o_2}, \ldots,\ a_n^{o_I}\}$, and can be computed as:

$$a_n^{o_i} = \frac{exp(L_n^i(f_n))}{\sum_{i=1}^{I} exp(L_n^i(f_n))} \tag{6}$$

where L is a mapping learned by the attention layer,

$$L(f) = W_2 \cdot \mathcal{R}(W_1 \overline{z}) \tag{7}$$

while $W_1 \in \mathbb{R}^{K \times C}$ and $W_2 \in \mathbb{R}^{|O| \times K}$ are learnable weight matrices and $\mathcal{R}(\cdot)$ is a $ReLU$ activation. $\overline{z} \in \mathbb{R}^C$ is the channel-wise averages vector of the input f defined as:

$$\overline{z}_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_{i,j,c} \tag{8}$$

To put it simply, A_n is generated channel-wisely to deduct computational complexity. Here we introduce depth-wise separable convolutions [16] with filter sizes of 1×1, 3×3, 5×5 and 7×7 for computation efficiency as like in [58]. To get the structure feature of the hazy image, we raise the receptive field using dilated convolution [66] with filter sizes of 3×3, 5×5 and 7×7 along with dilation rate of 2. We select 3×3 max pooling along with 3×3 average pooling for effective color retention. The operations are performed in parallel and the output of the operation are concatenated along channel, as shown in Figure 4.

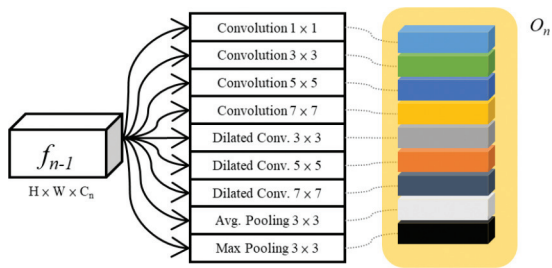


Figure 4: The detailed illustration of parallel operation layer in the attention-based operation framework.

3.3 Loss Functions

There are three loss functions we import here to guide our model training: 1) Curriculum Loss 2) Gradient Loss and 3) Maximizing Contrast Loss to refine the final model outputs.

3.3.1 Curriculum Loss

To ensure a progressively better result in each iteration of FOA, we adopt curriculum learning strategy during training. For that purpose, we interpolate between the hazy-free (ground truth) and hazy image to represent intermediate

density level of haze as our target images $(I^1_{target}, I^2_{target}, \ldots, I^N_{target})$ on different stages. Here we choose the $L1$ loss to optimize our model and output the result images $(I^1_{output}, I^2_{output}, \ldots, I^N_{output})$. The loss function can be expressed as below:

$$L(\theta) = \sum_{n=1}^{N} W^n ||I^n_{target} - I^n_{output}||_1 \tag{9}$$

where θ is the parameters of our model and W means the weight of loss in n-th iteration. The last output I^N_{output} of output sequence $(I^1_{output}, I^2_{output}, \ldots, I^N_{output})$ is expected to be I^N_{target}, which is the desired hazy-free image. As [43, 67], we take all the outputs into equal consideration and set all the W^n value to 1 in the curriculum learning.

3.3.2 Gradient Loss

Although L_1 loss is intuitive and effective, it tends to blur the final output. To deal with blurring and halo effect, gradient loss is used to preserve the edge details and be defined as:

$$\begin{aligned} L_{grad} = \lambda_{grad} \Bigg(& \sum_{W,H} ||\mathcal{G}_x(I_{output})_{(w,h)} - \mathcal{G}_x(I_{gt})_{(w,h)}|| + \\ & \sum_{W,H} ||\mathcal{G}_y(I_{output})_{(w,h)} - \mathcal{G}_y(I_{gt})_{(w,h)}|| \Bigg) \end{aligned} \tag{10}$$

where $\mathcal{G}_x$ and $\mathcal{G}_y$ are operators computing the horizontal and vertical derivative approximations of images I, i.e. gradients of two directions. Then, $W \times H$ indicates the width and height of the image. λ_{grad} is constant weight set to be 0.5 in our implementation.

3.3.3 Maximizing Contrast Loss

Images suffer from low contrast, faint color, and shifted luminance under bad weather conditions. The idea of increasing contrast has been adopted by many previous haze-removal papers [3, 5, 7, 21, 59]. And we also introduce the concept of high contrast to improve our result here by adding the maximizing contrast loss.

$$L_{contrast} = \lambda_{contrast} \cdot \left(-log \sqrt{\frac{1}{WH} \sum_{W,H} (I_{(w,h)} - \overline{I})^2} \right) \tag{11}$$

where $I_{w,h}$ is the intensity of w-th h-th pixel of output image of size W by H. $\overline{I}$ is the average intensity of all pixel values in the output image. The constant factor $\lambda_{contrast}$ is set to 0.005 in our experiment.

4 Experimental Results

In this section, we first present the details of the datasets we used and our implementation configuration. Afterward, qualitative and quantitative comparison with the state-of-the-art methods are presented. Finally, ablation studies on the losses and curriculum strategy are discussed to validate the improvement of our method.

4.1 Configuration

4.1.1 Dataset

RESIDE [37] is an well-known image dehazing benchmark, which contains both synthetic and real-word hazy images and is divided into several subsets for training, testing, indoor and outdoor data. We trained our module on the Indoor Training Dataset (ITS) and Outdoor Training Dataset (OTS) separately. In addition, we randomly picked each 8000 paired training images that are shuffled and divided them into a training set of 7500 images and a validation set of 500 images for training. After that, we tested our model on Synthetic Objective Testing Set (SOTS) for synthetic images as well as unannotated real-world hazy images from RESIDE [37]. The results are reported in Section 4.2 and Section 4.3 respectively. Moreover, we use O-HAZE [2], I-HAZE [4], and DENSE HAZE [1] in order to verify our performance on a wide variety of datasets.

4.1.2 Training Setting

Our network is trained on a workstation with NVIDIA GEFORCE RTX 2080 Ti GPU and Intel i7-8700 CPU at 3.20 GHz. The proposed FOANet is built by PyTorch, and all the tests are conducted with same environment. As for the training configuration, it is optimized by Adam method with learning rate of 0.001, where β_1 and β_2 take the default values of 0.9 and 0.999. The batch size is set to 1, total epoch is 300 and it costs approximately 200 epochs to converge. We train our network with three-channels RGB image patches of size 200×200. The recurrent step of our feedback mechanism is set to 10 iterations. In feature extraction block, all Conv layers and Resblocks are with kernel sizes of 3×3, strides of 1, and zero padding of 1.

4.2 Comparisons on Synthetic Images

For synthetic data, we used the well-known benchmark dataset, Synthetic Objective Testing Set (SOTS) from RESIDE [37], that contains 50 indoor images and 492 outdoor images to evaluate our module. To further compare our proposed architecture with single image dehaze methodologies, we apply two metrics to evaluate the quantitative results respectively. Table 1 shows the

Table 1: Quantitative comparison on synthetic images.

Dataset	SOTS-indoor		SOTS-outdoor	
Metrics	PSNR	SSIM	PSNR	SSIM
Zhang *et al.* [68]†	15.85	0.8175	19.93	0.8449
Ren *et al.* [55]†	22.30	0.8800	21.55	0.8444
Liu *et al.* [44]	22.46/0.8844			
Dong *et al.* [19]	22.81	0.8889	22.82	0.8886
Qu *et al.* [53]	25.06	0.9232	22.57	0.8630
Shao *et al.* [57]	27.76/0.93			
Li *et al.* [42]	–		–	0.934
Wu *et al.* [63]	23.85/0.91			
Chen *et al.* [14]	25.8079/0.9266			
Zhang *et al.* [70]	25.00	0.9172	29.03	0.9570
Li *et al.* [39]	23.93/0.936			
Kim *et al.* [31]	19.93	0.8633	24.96	0.9421
Ours	**29.37**	**0.9783**	**29.06**	**0.9749**

Note: For † we report the performance from the re-implemented result of [53].

comparison results of our proposed method and other twelve different existing dehazing approaches using Peak Signal to Noise Ratio (PSNR), Structural Similarity index (SSIM) on synthetic data. The table illustrates that our results can achieve a better performance of haze removal on both indoor and outdoor images among mentioned methods, proving that the dehazed images we can attain are much similar to the ground truth images. We also demonstrate the qualitative result by comparing the recovered images generated by our method and other approaches. Figures 5 and 6 shows the visual results with the haze removed images obtained by each approaches, containing outdoor and indoor images with different degrees of haze. Apparently, most of the dehazed results suffer from color distortion. Zhang *et al.* [68] tend to brighten light areas of the image and thus produce overexposed effect, making it difficult to identify objects. Results obtained by Dong *et al.* [19] and Ren *et al.* [55] usually have faded color and have unnatural image gradient in area containing only single color. Chen *et al.* [14] often generate color plaques. As for the results of Liu *et al.* [44], Qu *et al.* [53] and Shao *et al.* [57], the lighting condition is usually severely affected. Our proposed methods can obtain results more similar to the haze-free ground truth images without shifting the color of the images.

4.3 Comparisons on Real-world Images

Recovering real world hazy images is quite challenging since outdoor images usually have a large region of sky that comes in the color of white or gray which may confuse the dehazing model. Figure 7 shows the processed results

Figure 5: Comparison on SOTS indoor images. Methods starting from second row are DCPDN [68], GFN [55], FD-GAN [19], PMS-Net [14], LDP [44], EPDN [53], DA dehazing [57] and Ours.

of our proposed method along with other existing approaches. Zhang *et al.* [68] and Dong *et al.* [19] suffer from over enhancement on the sky region and the color of the images are severely affected. As for those approaches which can recognize sky region correctly such as Ren *et al.* [55] fail to remove the haze. The performance is obviously worse than ours, especially in regions with dense haze, such as the second and eighth image. Also Qu *et al.* [53], Shao *et al.* [57] have severe vignetting in the first column. On the other hand, though our proposed methods here can attain results with haze removed to a certain degree with less over-saturated color, we still cannot successfully separate the white or gray background and thus contain some artifacts due to the misjudge of the region of the objects and the background.

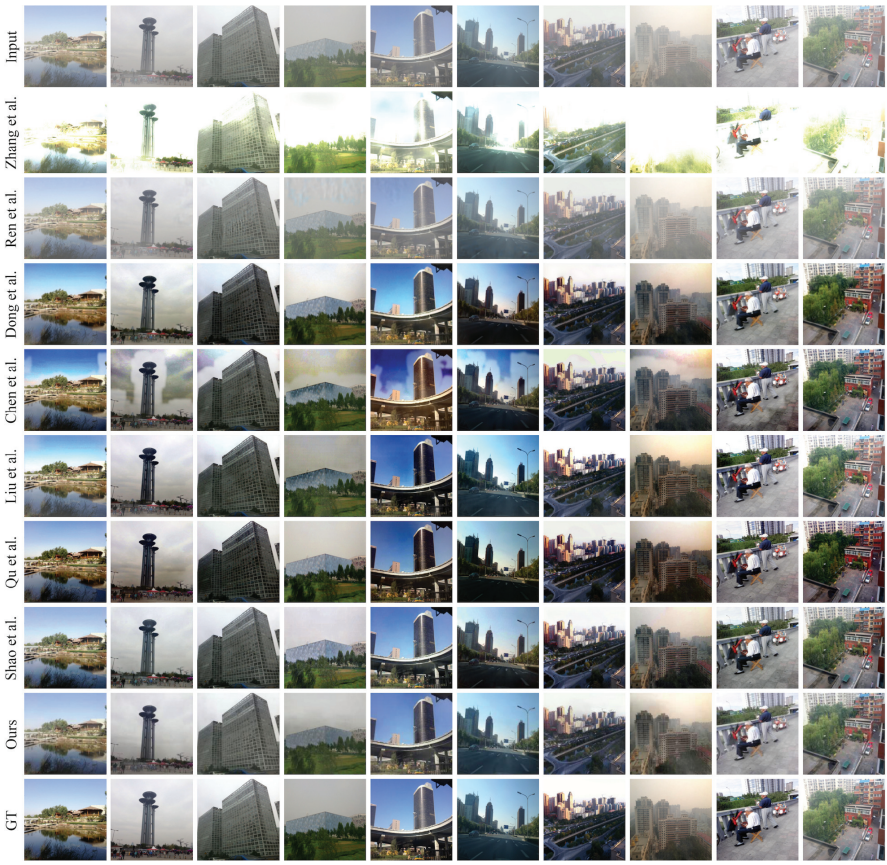


Figure 6: Comparison on SOTS outdoor images. Methods starting from second row are DCPDN [68], GFN [55], FD-GAN [19], PMS-Net [14], LDP [44], EPDN [53], DA dehazing [57] and Ours.

4.4 Ablation Study

We conduct ablation study on dense haze dataset to demonstrate the improvement of different components: Constraint Loss, Gradient Loss, and Curriculum Strategy. The results of ablation study are shown in Figure 8 and we will discuss in detail as follows.

4.4.1 Without Contrast Loss

The second image shows the result without contrast loss is overall darker than the others. The entire image is more like vintage style and the tone is closer

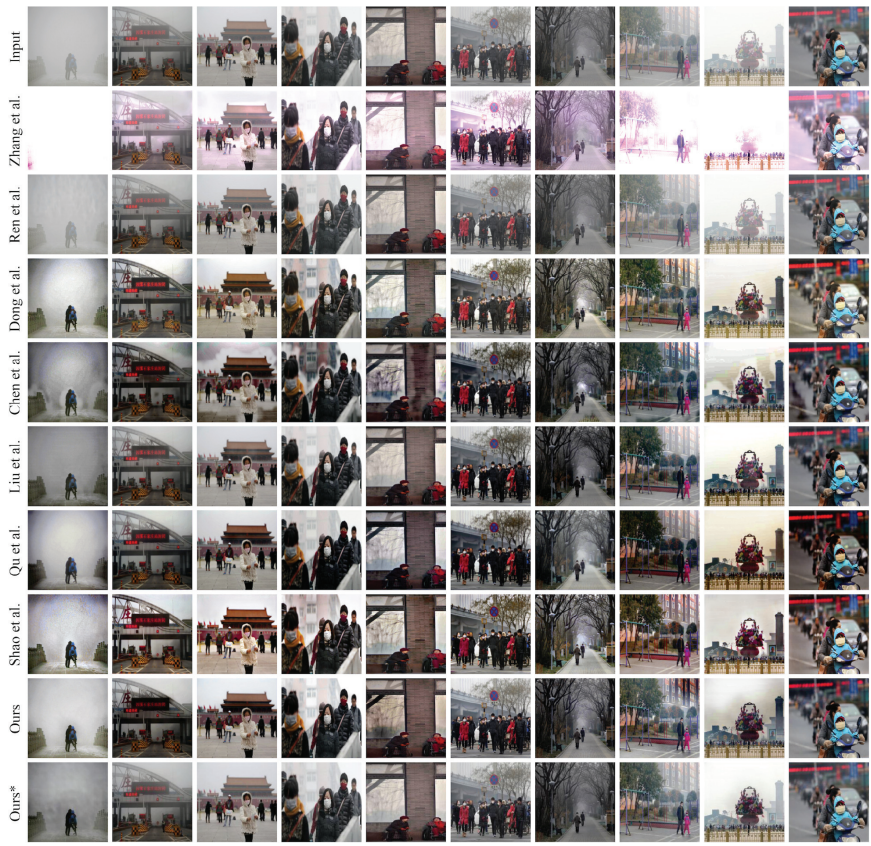


Figure 7: Comparison on realistic images. Methods starting from second row are DCPDN [68], GFN [55], FD-GAN [19], PMS-Net [14], LDP [44], EPDN [53], DA dehazing [57] and Ours. For * we train our model on RESIDE [37] OTS subset.

Figure 8: Results of ablation study.

to shade compared to the result of our proposed method. On the contrary, in our result, the colors compensate each other in a more harmonious way and demonstrate various color than just white-and-black scene.

4.4.2 Without Gradient Loss

As Figure 8 shows, it is obvious that the one without gradient loss is the most blurred one, where the climbing slope in the center of the picture is almost disappeared. Also, it is hard to differentiate the playground facilities from the house in the background, because the slides at the left side are nearly blended into the house without the help of gradient loss. However, in our proposed method with gradient loss, not only the climbing slope is presented but also the window frames at the top-right corner of the image are depicted more clearly.

4.4.3 Without Curriculum Strategy

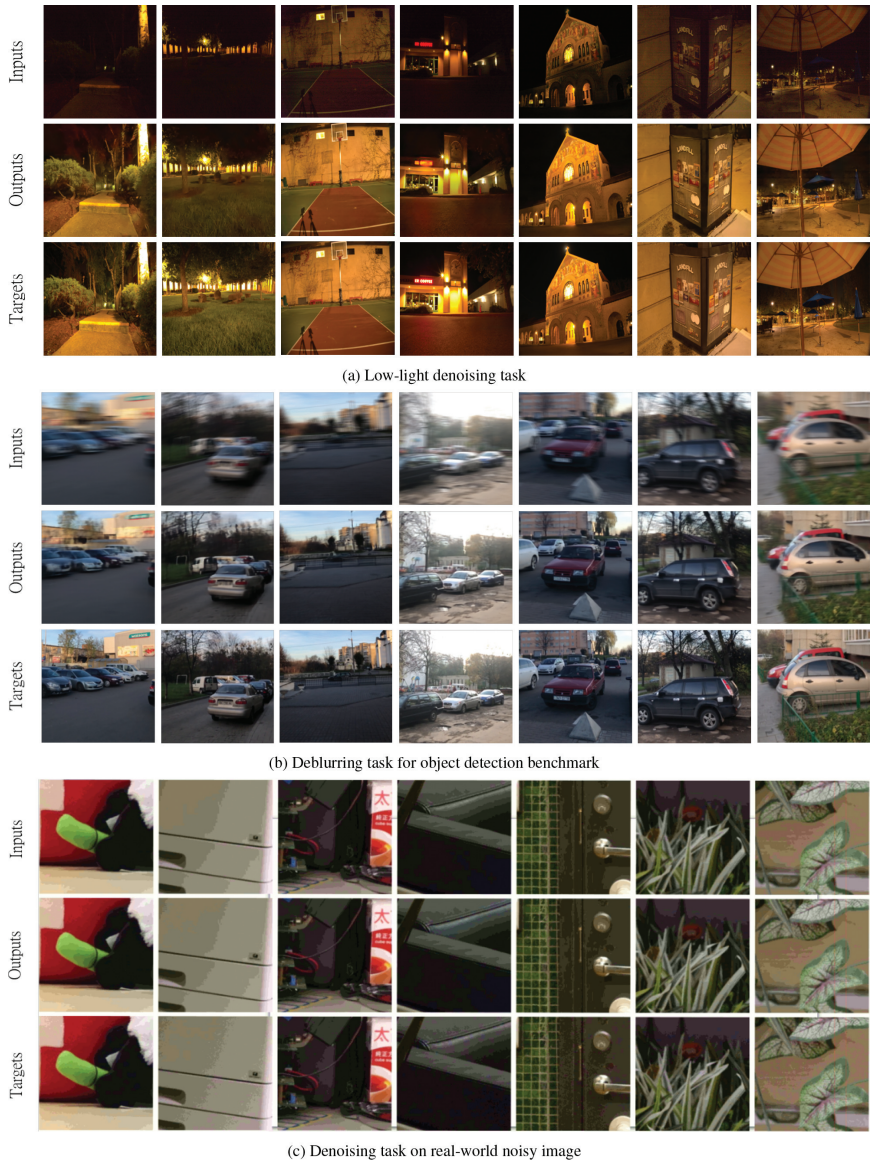
The result without curriculum strategy is the worst one amongst the ablation study. The dense haze is not removed well, causing low color saturation. We can conclude that the multi-stage neural network without the guide of curriculum strategy is more likely to be biased because of the computational error between different stage. However, introducing the curriculum strategy and considering the losses in between can reduce the distances from the input images to the target images in each stage, which makes each stage of the network can be amended more precisely and thus leads to the much better result shown in Figure 8. According to the ablation study, we therefore decide to retain all these three components in our network, which are dedicated to the color improvement, the detail retention and the haze removal respectively.

4.5 Extended Applications

Besides haze removal, we also conducted several experiments on various image restoration tasks, such as low-light denoising on SID dataset [12], deblurring task for object detection benchmark on YOLO dataset [33] and real-world noisy image denoising on PolyU dataset [65]. All the results are shown in Figure 9. Note that we use our model with no architectural changes or fine tuning for all applications.

5 Conclusions and Future Work

In this paper, we proposed a novel end-to-end method for single image dehazing problem by using feedback mechanism and operation-wise attention blocks. The feedback mechanism is applied to consider features on different levels. The design of the attention-based operations is to weight different operations and reduce the different density of haze each iteration. We employed the curriculum learning strategy in order to make sure the performance improve step by



(a) Low-light denoising task

(b) Deblurring task for object detection benchmark

(c) Denoising task on real-world noisy image

Figure 9: Results of various image restoration task.

step. Furthermore, We additionally import gradient loss and Contrast loss to refine the results. Experiments on various datasets with both synthetic and realistic data show that the proposed network has an outstanding performance

on image dehazing issue. Finally, we study the impact of the two losses and the curriculum learning strategy in ablation study section. Since our model is computationally expensive and memory intensive, we consider model compression an interesting direction of future work to further apply our method to real-time tasks.

Financial Support

This work was supported in part by Ministry of Science and Technology of Taiwan under the grant numbers: MOST-109-2223-E-009-002-MY3, MOST-110-2218-E-A49-018 and MOST-111-2634-F-007-002.

Author Biographies

Chia-Lin Liu received a B.S. degree from the Department of Electronics Engineering, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan, in 2019, and a M.S. degree from the Department of Electrical and Computer Engineering, University of Washington, Seattle, United States, with a specialty in data science track.

Lei Chen received the B.S. degree from Department of Computer Science and Information Engineering, Chang Gung University (CGU), Taoyuan, Taiwan, R.O.C., in 2019, and the M.S. degree in Artificial Intelligence Graduate Program, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan, R.O.C., in 2021. Her current research topics contain network anomaly detection, image dehazing, and atrial fibrillation detection.

Ling Lo received the B.S. degree from Department of Electronics Engineering, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan, R.O.C., in 2019, and now she is pursuing a Ph.D. degree in Institute of Electronics, NYCU. Her recent work includes facial expression recognition and micro-expression recognition.

Pin-Jui Huang received the B.S. degree from Department of Biomedical Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, R.O.C., in 2020, and now he is pursuing a master degree in Artificial Intelligence Graduate Program, NYCU. His current research topics contain image dehazing.

Hong-Han Shuai received the M.S. degree in computer science from NTU in 2009, and the Ph.D. degree from Graduate Institute of Communication Engineering, NTU, in 2015. He is now an Associate Professor in NYCU. His research interests are in the area of multimedia processing, machine learning, social network analysis, and data mining.

Wen-Huang Cheng is Distinguished Professor with the Institute of Electronics, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan. He is also Jointly Appointed Professor with the Artificial Intelligence and Data Science Program, National Chung Hsing University (NCHU), Taichung, Taiwan. His current research interests include multimedia, artificial intelligence, computer vision, and machine learning.

Ching-Hsuan Wang is currently a researcher in Chunghwa Telecom laboratories. He is responsible for the research on AI application in traffic law enforcement. He received a bachelor's degree from National Central University, Taiwan, and a master's degree from National Yang Ming Chiao Tung University, Taiwan, both in computer science. He is interested in multi-target tracking, deep learning, and traffic analysis.

Fan Chou is a deputy senior researcher of IoT lab of Chunghwa Telecom Laboratories. He received the M.S. degree in multimedia engineering from National Chiao Tung University, Taiwan in 2008. He received the Outstanding Young Electrical Engineer Award from Chinese Institute of Electrical Engineering in 2020. His research focus on applying AI image/video analysis technology to smart city application, such as intelligent video surveillance, advanced traffic management, etc.

References

[1] C. O. Ancuti, C. Ancuti, M. Sbert, and R. Timofte, "Dense-Haze: A Benchmark for Image Dehazing with Dense-Haze and Haze-Free images," in *2019 IEEE International Conference On Image Processing (ICIP)*, IEEE, 2019, 1014–8.

[2] C. O. Ancuti, C. Ancuti, R. Timofte, and C. De Vleeschouwer, "O-haze: A Dehazing Benchmark with Real Hazy and Haze-Free Outdoor Images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2018, 754–62.

[3] C. O. Ancuti and C. Ancuti, "Single Image Dehazing by Multi-Scale Fusion," *IEEE Transactions on Image Processing*, 22(8), 2013, 3271–82.

[4] C. Ancuti, C. O. Ancuti, R. Timofte, and C. De Vleeschouwer, "I-HAZE: A Dehazing Benchmark with Real Hazy and Haze-Free Indoor Images," in *International Conference on Advanced Concepts For Intelligent Vision Systems*, Springer, 2018, 620–31.

[5] L. Bao, Y. Song, Q. Yang, and N. Ahuja, "An Edge-Preserving Filtering Framework for Visibility Restoration," in *Proc. ICPR*, 2012, 384–7.

[6] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum Learning," in *Proc. ICML*, 2009, 41–8.

[7] L. K. C. J. Y. A. C. Bovik, "Referenceless Prediction of Perceptual Fog Density and Perceptual Image Defogging," *IEEE Transactions on Image Processing*, 24(11), 2015, 3888–901.

[8] B. Cai, X. Xu, K. Jia, C. Qing, and D. Tao, "Dehazenet: An End-to-End System for Single Image Haze Removal," *IEEE Transactions on Image Processing*, 25(11), 2016, 5187–98.

[9] C. Cao, X. Liu, Y. Yang, Y. Yu, J. Wang, Z. Wang, Y. Huang, L. Wang, C. Huang, W. Xu, *et al.*, "Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, 2956–64.

[10] P. Carr and R. Hartley, "Improved Single Image Dehazing Using Geometry," in *2009 Digital Image Computing: Techniques And Applications*, IEEE, 2009, 103–10.

[11] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human Pose Estimation with Iterative Error Feedback," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, 4733–42.

[12] C. Chen, Q. Chen, J. Xu, and V. Koltun, "Learning to See in the Dark," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 3291–300.

[13] D. Chen, M. He, Q. Fan, J. Liao, L. Zhang, D. Hou, L. Yuan, and G. Hua, "Gated Context Aggregation Network for Image Dehazing and Deraining," in *2019 IEEE Winter Conference On Applications Of Computer Vision (WACV)*, 2019, 1375–83.

[14] W.-T. Chen, J.-J. Ding, and S.-Y. Kuo, "PMS-net: Robust Haze Removal based on Patch Map for Single Images," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 11681–9.

[15] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion Meets Computer Vision: A Survey," *ACM Computing Surveys*, 2021.

[16] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 1251–8.

[17] C.-L. Chou, C.-Y. Chen, C.-W. Hsieh, H.-H. Shuai, J. Liu, and W.-H. Cheng, "Template-Free Try-on Image Synthesis viaSemantic-guided Optimization," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[18] Q. Deng, Z. Huang, C.-C. Tsai, and C.-W. Lin, "HardGAN: A Haze-Aware Representation Distillation GAN for Single Image Dehazing," in *Proceedings of European Conference on Computer Vision (ECCV)*, Springer, 2020, 722–38.

[19] Y. Dong, Y. Liu, H. Zhang, S. Chen, and Y. Qiao, "FD-GAN: Generative Adversarial Networks with Fusion-Discriminator for Single Image Dehazing," in *Proc. AAAI Conference On Artificial Intelligence*, Vol. 34, No. 07, 2020, 10729–36.

[20] X. Fang, Q. Zhou, J. Shen, C. Jacquemin, and L. Shao, "Text Image Deblurring Using Kernel Sparsity Prior," *IEEE Trans. Cybern.*, 50(3), 2020, 997–1008.

[21] R. Fattal, "Dehazing Using Color-Lines," *ACM Transactions on Graphics (TOG)*, 34(13)(11), 2014.

[22] R. Fattal, "Single Image Dehazing," *ACM Transactions on Graphics (TOG)*, 27(3), 2008, 1–9.

[23] R. Gao and K. Grauman, "On-demand Learning for Deep Image Restoration," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, 1086–95.

[24] W. Han, S. Chang, D. Liu, M. Yu, M. Witbrock, and T. S. Huang, "Image Super-resolution via Dual-state Recurrent Networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 1654–63.

[25] N. Hautiere, J.-P. Tarel, and D. Aubert, "Towards Fog-Free In-Vehicle Vision Systems through Contrast Restoration," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, 1–8.

[26] K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12), 2010, 2341–53.

[27] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng, "FashionOn: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information," in *Proceedings of ACM International Conference on Multimedia (MM)*, 2019, 275–83.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze-and-Excitation Networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 7132–41.

[29] W. T. Jiao Long Zhenwei Shi and C. Zhang, "Single Remote Sensing Image Dehazing," *IEEE Geoscience and Remote Sensing Letters*, 11(1), 2013, 59–63.

[30] N. Joshi and M. F. Cohen, "Seeing Mt. Rainier: Lucky Imaging for Multi-image Denoising, Sharpening, and Haze Removal," in *2010 IEEE International Conference on Computational Photography (ICCP)*, IEEE, 2010, 1–8.

[31] S. E. Kim, T. H. Park, and I. K. Eom, "Fast Single Image Dehazing using Saturation based Transmission Map Estimation," *IEEE Transactions on Image Processing*, 29, 2019, 1985–98.

[32] H. Koschmieder, *Theorie der horizontalen Sichtweite: Kontrast und Sichtweite, Keim & Nem-nich*, 1925.

[33] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "Deblurgan: Blind Motion Deblurring using Conditional Adversarial Networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 8183–92.

[34] W.-C. Lai, Z.-X. Xia, H.-S. Lin, L.-F. Hsu, H.-H. Shuai, I.-H. Jhuo, and W.-H. Cheng, "Trajectory Prediction in Heterogeneous Environment via Attended Ecology Embedding," in *Proceedings of ACM International Conference on Multimedia (MM)*, 2020, 202–10.

[35] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "MADNet: A Fast and Lightweight Network for Single-Image Super Resolution," *IEEE Transactions on Cybernetics*, 51(3), 2021, 1443–53.

[36] R. Lan, L. Sun, Z. Liu, H. Lu, Z. Su, C. Pang, and X. Luo, "Cascading and Enhanced Residual Networks for Accurate Single-Image Super-Resolution," *IEEE Transactions on Cybernetics*, 51(1), 2021, 115–25.

[37] B. Li, W. Ren, D. Fu, D. Tao, D. Feng, W. Zeng, and Z. Wang, "Benchmarking Single-image Dehazing and Beyond," *IEEE Transactions on Image Processing*, 28(1), 2018, 492–505.

[38] J. Li, Z. Pan, Q. Liu, Y. Cui, and Y. Sun, "Complementarity-Aware Attention Network for Salient Object Detection," *IEEE Transactions on Cybernetics*, 2020, 1–14.

[39] R. Li, J. Pan, M. He, Z. Li, and J. Tang, "Task-oriented Network for Image Dehazing," *IEEE Transactions on Image Processing*, 29, 2020, 6523–34.

[40] R. Li, J. Pan, Z. Li, and J. Tang, "Single Image Dehazing via Conditional Generative Adversarial Network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 8202–11.

[41] X. Li, Z. Zhao, and Q. Wang, "ABSSNet: Attention-Based Spatial Segmentation Network for Traffic Scene Understanding," *IEEE Transactions on Cybernetics*, 2021, 1–11.

[42] Y. Li, Q. Miao, W. Ouyang, Z. Ma, H. Fang, C. Dong, and Y. Quan, "LAP-Net: Level-Aware Progressive Network for Image Dehazing," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, 3275–84.

[43] Z. Li, J. Yang, Z. Liu, X. Yang, G. Jeon, and W. Wu, "Feedback Network for Image Super-resolution," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 3867–76.

[44] Y. Liu, J. Pan, J. Ren, and Z. Su, "Learning Deep Priors for Image Dehazing," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, 2492–500.

[45] K.-Y. Lung, C.-R. Chang, S.-E. Weng, H.-S. Lin, H.-H. Shuai, and W.-H. Cheng, "ROSNet: Robust One-Stage Network for CT Lesion Detection," *Pattern Recognition Letters*, 2021.

[46] Y.-J. Ma, H.-H. Shuai, and W.-H. Cheng, "Spatiotemporal Dilated Convolution with Uncertain Matching for Video-based Crowd Estimation," *IEEE Transactions on Multimedia*, 2021.

[47] B. T. Nalla, T. Sharma, N. K. Verma, and S. Sahoo, "Image Dehazing for Object Recognition using Faster RCNN," in *2018 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2018, 1–7.

[48] S. G. Narasimhan and S. K. Nayar, "Contrast Restoration of Weather Degraded Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8), 2003, 713–24.

[49] S. G. Narasimhan and S. K. Nayar, "Contrast Restoration of Weather Degraded Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6), 2003, 713–24.

[50] D.-K. Nguyen and T. Okatani, "Improved Fusion of Visual and Language Representations by Dense Symmetric Co-attention for Visual Question Answering," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 6087–96.

[51] K. Nishino, L. Kratz, and S. Lombardi, "Bayesian defogging," *International Journal of Computer Vision*, 98(3), 2012, 263–78.

[52] A. Pentina, V. Sharmanska, and C. H. Lampert, "Curriculum Learning of Multiple Tasks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, 5492–500.

[53] Y. Qu, Y. Chen, J. Huang, and Y. Xie, "Enhanced Pix2pix Dehazing Network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, 8152–60.

[54] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang, "Single Image Dehazing via Multi-scale Convolutional Neural Networks," in *Proceedings of European Conference on Computer Vision (ECCV)*, Springer, 2016, 154–69.

[55] W. Ren, L. Ma, J. Zhang, J. Pan, X. Cao, W. Liu, and M.-H. Yang, "Gated Fusion Network for Single Image Dehazing," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 3253–61.

[56] D. B. Sam and R. V. Babu, "Top-down Feedback for Crowd Counting Convolutional Neural Network," in *Proceedings of AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, 2018.

[57] Y. Shao, L. Li, W. Ren, C. Gao, and N. Sang, "Domain Adaptation for Image Dehazing," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, 2808–17.

[58] M. Suganuma, X. Liu, and T. Okatani, "Attention-based Adaptive Selection of Operations for Image Restoration in the Presence of Unknown Combined Distortions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, 9039–48.

[59] R. T. Tan, "Visibility in Bad Weather from a Single Image," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2008, 1–8.

[60] K. Tang, J. Yang, and J. Wang, "Investigating Haze-relevant Features in a Learning Framework for Image Dehazing," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, 2995–3000.

[61] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All You Need," *Proceedings of NIPS*, 2017.

[62] A. Wang, W. Wang, J. Liu, and N. Gu, "AIPNet: Image-to-image Single Image Dehazing with Atmospheric Illumination Prior," *IEEE Transactions on Image Processing*, 28(1), 2018, 381–93.

[63] Q. Wu, J. Zhang, W. Ren, W. Zuo, and X. Cao, "Accurate Transmission Estimation for Removing Haze and Noise from a Single Image," *IEEE Transactions on Image Processing*, 29, 2019, 2583–97.

[64] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "AU-assisted Graph Attention Convolutional Network for Micro-Expression Recognition," in *Proceedings of the ACM International Conference on Multimedia (MM)*, 2020, 2871–80.

[65] J. Xu, H. Li, Z. Liang, D. Zhang, and L. Zhang, "Real-world Noisy Image Denoising: A New Benchmark," 2018.

[66] F. Yu and V. Koltun, "Multi-scale Context Aggregation by Dilated Convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[67] A. R. Zamir, T.-L. Wu, L. Sun, W. B. Shen, B. E. Shi, J. Malik, and S. Savarese, "Feedback Networks," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, 1308–17.

[68] H. Zhang and V. M. Patel, "Densely Connected Pyramid Dehazing Network," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, 3194–203.

[69] H. Zhang, V. Sindagi, and V. M. Patel, "Joint Transmission Map Estimation and Dehazing using Deep Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019, 1–1.

[70] J. Zhang and D. Tao, "Famed-net: A Fast and Accurate Multi-scale End-to-End Dehazing Network," *IEEE Transactions on Image Processing*, 29, 2019, 72–84.

[71] Y. Zhou, X. Du, M. Wang, S. Huo, Y. Zhang, and S. .-. Kung, "Cross-Scale Residual Network: A General Framework for Image Super-Resolution, Denoising, and Deblocking," *IEEE Transactions on Cybernetics*, 2021, 1–13.

[72] H. Zhu, Y. Cheng, X. Peng, J. T. Zhou, Z. Kang, S. Lu, Z. Fang, L. Li, and J. .-. Lim, "Single-Image Dehazing via Compositional Adversarial Network," *IEEE Transactions on Cybernetics*, 51(2), 2021, 829–38.