FEATURE AGGREGATION ATTENTION NETWORK FOR SINGLE IMAGE DEHAZING

Lan Yan[1,2], *Wenbo Zheng*[3,1], *Chao Gou*[4], *Fei-Yue Wang*[1]

[1] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[4] School of Intelligent Systems Engineering, Sun Yat-sen University, Guangzhou 510275, China
{*yanlan2017@ia.ac.cn; zwb2017@stu.xjtu.edu.cn; gouchao@mail.sysu.edu.cn; feiyue.wang@ia.ac.cn*}

ABSTRACT

Due to its ill-posed nature, single image dehazing is a challenging problem. In this paper, we propose an end-to-end feature aggregation attention network (FAAN) for single image dehazing. It incorporates the idea of attention mechanism and residual learning and can adaptively aggregate different level features. In particular, in the proposed FANN, we design a novel block structure consisting of feature attention module, smoothed dilated convolution and local residual learning. The local residual learning allows the less useful information to be bypassed through multiple skip connections. The feature attention module is designed to assign more weight to important features. The smoothed dilated convolution is adopted to enlarge the receptive field without the negative influence of gridding artifacts. The experiments on the RESIDE dataset show that the proposed approach acquires state-of-the-art performance in both qualitative and quantitative measures.

Index Terms— Dehazing, image restoration, deep CNN.

1. INTRODUCTION

Due to the effect of light scattering through floating atmospheric particles such as mist, fumes, dust and smoke in the atmosphere, images taken in hazy conditions are easily subject to low contrast, saturation loss, blurring and other visible quality degradation. These image quality degradations make many succeeding high-level visual tasks, e.g., object detection and tracking for video surveilance and autonomous driving, become more challengeable. Hence, image dehazing, which aims to recover the clear version of a hazy image, has attracted much attention in computer vision community.

The atmospheric scattering model [1, 2, 3] has been a classical description for the hazy image generation process, it can be mathematically formulated as:

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

where $I(x)$ is the observed hazy image, x denotes the pixel location, $t(x)$ is the medium transmission map, A stands for the global atmospheric light and $J(x)$ is the haze-free scene radiance. When the haze is homogeneous, the transmission map can be expressed as $t(x) = e^{-\beta d(x)}$, where β represents atmosphere scattering parameter and $d(x)$ is the scene depth. Since only the observed image $I(x)$ is known, single image dehazing is an ill-posed problem.

Many previous dehazing methods [4, 5] employ many image priors to estimate the transmission maps and atmospheric light, then recover the haze-free images according to the atmosphere scattering model. For instance, DCP [4] is a dark channel prior method and postulates that local patches in clear outdoor images have some pixels with low intensity in at least one color channel. However, this assumption dose not hold when the scene object is similar to atmosphere lights. That is, the priors used in the prior-based approaches are not always valid, and they may not be feasible in certain real scenarios.

Recently, with the advent of deep learning, numerous learning-based methods [6, 7, 8, 9, 10, 11, 12, 13] are proposed for image dehazing. Dehazenet[8] learns and regresses the medium transmission map in an end-to-end way. DCPDN [10] builds upon the atmosphere scattering model and presents two sub-networks to respectively estimate the transmission map and atmospheric light. Compared to the prior-based approaches, learning-based approaches generally regress intermediate transmission maps or final clear images directly, and achieve good performance as well as robustness.

In this work, we propose a novel end-to-end learning-based method termed as feature aggregation attention network (FAAN), for single image dehazing. Considering the uneven distribution of haze in the image, the weights of thin and thick haze area pixels should be significantly different. Moreover, the discovery of dark channel in DCP suggests that the features from different channels have different important infor-

This work is supported in part by the Key Research and Development Program 2020 of Guangzhou (202007050002) and the National Natural Science Foundation of China (61533019, 61806198, U1811463). *(Corresponding author: Chao Gou).*

mation. All of these imply that equally treating the channel-wise and pixel-wise features will makes the network expend a lot of efforts on less important information and thus damage the expressive power of the network. Therefore, inspired by the widely used attention mechanism [14, 15] in network design, we introduce a feature attention module containing the pixel attention and channel attention for pixel-wise and channel-wise features respectively. This makes our model pay more attention on the thick hazy pixels and the more important channel information.

Since the dilated convolution is very effective for the aggregation of context information and does not lose spatial resolution [16], we introduce it to cover more adjacent pixels and help get better dehazing effect. Nevertheless, the original dilated convolution suffers from the gridding artifacts [17] which hampers the performance, so we adopt the smoothed dilated convolution [18] to reduce the gridding artifacts and obtain better performance.

Considering that training a very deep network is feasible by using residual learning [19], we incorporate the idea of residual learning and the attention mechanism and make our network learn the different weights of different level feature information adaptively. Furthermore, we design a novel block structure composed by feature attention module, smoothed dilated convolution and multiple local residual learning skip connections. In this block structure, on the one hand, the local residual learning makes it possible to bypass the thin haze area and low frequency information, so that the main network can learn more useful features. On the other hand, the attention module and smoothed dilated convolution further improve the capability of our network.

To verify the effectiveness of our method, we conduct extensive qualitative and quantitative experiments on the RESIDE dataset[20], which is a large scale benchmark public available recently for single image dehazing. The results demonstrate that our method realizes the state-of-the-art dehazing performance. Moreover, we design several ablation experiments to show the effectiveness of key components.

In summary, our main contributions include:

1) We propose a novel feature aggregation attention network (FAAN) for single image dehazing, which incorporates attention mechanisms and residual learning and can adaptively aggregate different level features.

2) We design a block structure consisting of feature attention module, smoothed dilated convolution and local residual learning. Local residual learning permits the thin hazy area and low frequency information to be bypassed through multiple skip connections. Smoothed dilated convolution and attention module enhance the capability of our FAAN.

3) Extensive experiments show that our approach outperforms the state-of-the-art image dehazing methods and can suppress artifacts and color distortion.

2. THE PROPOSED APPROACH

In this section, our feature aggregation attention network is introduced. As illustrated in Figure 1, our FAAN uses a haze image as input and directly outputs a clear image. It adopts global residual learning and mainly contains a shallow feature extractor, three group structures with multiple skip connections and feature attention module. Furthermore, each group structure is mainly composed of N block structures with local residual learning. Each block structure includes the smoothed dilated convolution, local residual learning and the feature attention module which consist of a channel-wise and a pixel-wise attention mechanism.

2.1. Smoothed Dilated Convolution

Since dilated convolution can enlarge the receptive field without increasing the training parameters and will not decrease the spatial resolution of the response, some researchers have used it in image pixel-wise prediction tasks for efficient computation [21, 22]. In the one-dimensional case, the dilated convolution operation can be formulated as:

$$o[i] = \sum_{j=1}^{k} f[i + r \times k]w[j] \tag{2}$$

where f denotes the 1-D input, o is the output, k is the kernel size and r denotes the dilation rate. When we set r to 1, dilated convolutions degenerate to standard convolutions. Specially, the receptive field of the dilated convolution increases to $r \times (k-1)+1$ without sacrificing the spatial resolution.

However, when r of dilated convolution is greater than 1, adjacent units in the output are calculated from completely separate unit sets in the input. It leads to gridding artifacts [17], which hampers the performance. To alleviate this problem, we adopt the smoothed dilated convolution [18] which adds an additional convolution layer with kernel size $(2r-1)$ to increase the dependency among the input units before dilated convolutions (or the output units after dilated convolutions). More specially, we leverage separable and shared convolutions [18] as the additional convolutional layer to augment the interaction among input units.

2.2. Feature Attention

To assign different degrees of attention to the features which have different degrees of importance for the dehazing task and enhance the representational abilities of the network, we introduce a feature attention mechanism. This mechanism contains two part, including channel attention and pixel attention. The details are described as below.

Channel Attention. Channel attention focuses on giving different weights to different channel features. As illustrated in Figure 1, we first use global average pooling to ac-

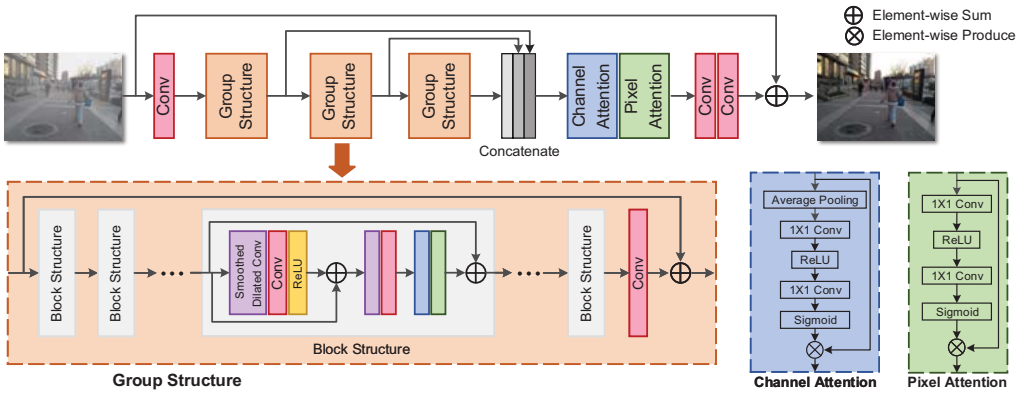

Fig. 1. The architecture of the proposed feature aggregation attention network.

quire channel-wise global spatial information, which can be expressed as:

$$G_c = P(F_c) = \frac{1}{H \times W} \sum_{a=1}^{H} \sum_{b=1}^{W} X_c(a, b) \tag{3}$$

where $X_c(a, b)$ denotes the value of c-th channel X_c at position (a, b), P stands for the global average pooling function. Then, the shape of the feature map becomes $C \times 1 \times 1$.

Subsequently, the feature passes through two convolutional layers with kernel size 1×1, ReLU and sigmoid activation layer. It can be represented as:

$$CA_c = Sigmoid(Conv(ReLU(Conv(G_c)))). \tag{4}$$

Finally, we use element-wise multiplication for the weights of channel CA_c and the input F_c, and get the output:

$$F_c^{'} = CA_c \otimes F_c. \tag{5}$$

Pixel Attention. We introduce the pixel attention mechanism to attach greater importance to informative features. As is evident in Figure 1, similar to the channel attention module, the output of channel attention $F^{'}$ is fed into two convolutional layers with kernel size 1×1, ReLU along with sigmoid activation layer. It can be formulated as:

$$PA = Sigmoid(Conv(ReLU(Conv(F^{'})))). \tag{6}$$

Then we get the pixel attention weight feature map whose shape is $1 \times H \times W$. In the end, we can obtain the output of the feature attention module by element-wise multiplying PA and $F^{'}$:

$$F^{*} = PA \otimes F^{'}. \tag{7}$$

2.3. Network Structure

Our network architecture is demonstrated in Figure 1. We use three group structures and each one outputs 64 filters and contains several block structures. In our experiment, we set the number of block structures in each group as 19. In addition to the feature attention module whose kernel size is 1×1, the kernel size of all convolutional layers is 3×3. Besides channel attention module all feature maps keep fixed size. The dilation rate of dilated convolutions is set as 2.

2.4. Loss Function

Considering that L1 loss can achieve better performances than L2 loss in many image restoration tasks [23], we adopt the L1 loss instead of the widely used mean square error (MSE) or L2 loss. Therefore, the loss function can be defined as:

$$\mathcal{L} = \|I_{gt} - FAAN(I_{haze})\|_1 \tag{8}$$

where I_{haze} stands for the hazy input and I_{gt} denotes the clean ground truth.

3. EXPERIMENTS

In this section, we quantitatively and qualitatively evaluate the proposed approach on the large-scale dehazing benchmark RESIDE [20]. We train our model with 313950 synthetic hazy outdoor images (OTS) and test it on 500 outdoor images from the synthetic objective testing set (SOTS) of RESIDE. Peak signal to noise ratio (PSNR) and structural similarity index (SSIM) are adopted for quantitative measurement. Experiments are conducted on two NVIDIA TITAN XP GPUs.

Training Details. We adopt the Adam solver for network training, with a batch size of 2 and momentum parameters 0.9 and 0.999. The input of our network is a cropped hazy image patch with a size of 240×240. We train the network for 8×10^5 steps and use the cosine annealing strategy [24] to decay the learning rate from the initial value of 1×10^{-4} to 0.

Qualitative and Quantitative Evaluation. To demonstrate the superiority of the proposed FAAN, we design both qualitative and quantitative comparision experiments. We compare our model with seven previous image dehazing approaches quantitatively including DCP [4], AOD-Net [6],

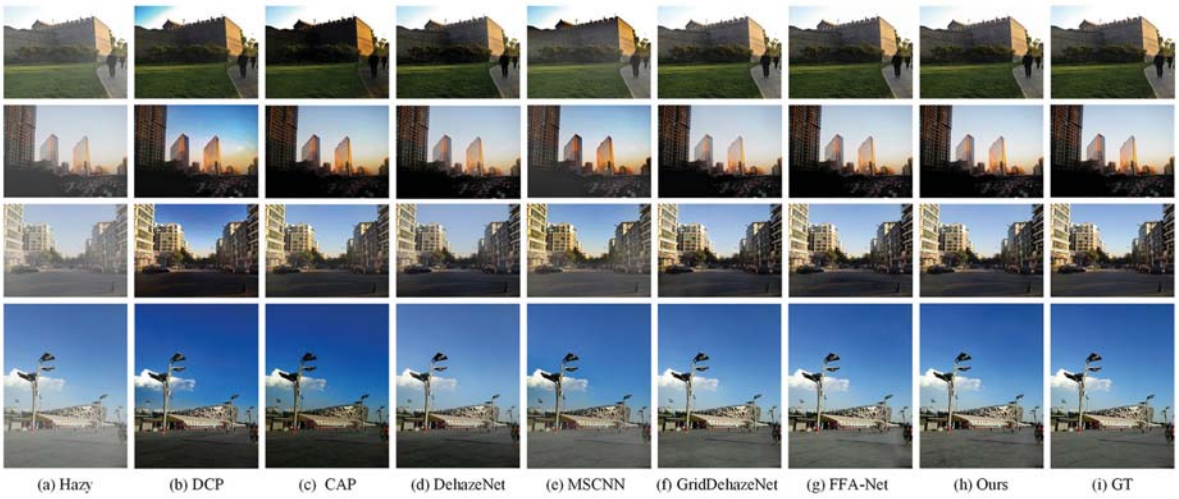

Fig. 2. Qualitative comparisons on the SOTS outdoor testing set.

Table 1. Quantitative evaluation results on the SOTS outdoor testing set for different dehazing methods.

Methods	PSNR	SSIM
DCP [4]	19.13	0.8148
AOD-Net [6]	20.29	0.8765
MSCNN [7]	22.06	0.9078
DehazeNet [8]	22.46	0.8514
GFN [9]	21.55	0.8444
GridDehazeNet [12]	30.86	0.9819
FFA-Net [13]	33.38	0.9804
Ours	**34.10**	**0.9850**

MSCNN [7], DehazeNet [8], GFN [9], GridDehazeNet [12] and FFA-Net [13]. The comparison results are reported in Table 1. As is evident in Table 1, our FAAN achieves the best performance in both PSNR and SSIM metrics. It indicates the superiority of our model over other state-of-the-art methods.

Furthermore, we display some dehazing results in Figure 2 for qualitative comparisons. In Figure 2, it can be observed that the result of DCP [4] is darker than the ground-truth which is caused by the inaccurate estimation of haze thickness. Besides, due to the underlying prior assumptions, the prior-based methods DCP and CAP [5] suffer from severe color distortion which dramatically damages the quality of their outputs. Dehazenet recovers images with excessive brightness relative to ground-truth. For DehazeNet, its dehazing results are also affected by color distortion. MSCNN has poor dehazing capacity and a mass of haze still remains unremoved. Although GridDehazeNet and FFA-Net produce quite good dehazing results, the proposed FAAN is better at suppressing artifacts and color distortion while removing haze as much as possible from input images.

Ablation Study. To further illustrate the effectiveness of key components in our FAAN, we consider different component configurations and conduct ablation studies. In particular, we mainly concern three components including attention mechanism, local residual learning and smoothed dilation. Correspondingly, we evaluate the three different configurations of our network on the RESIDE dataset and the results are listed in Table 2. It is clear that the performance keeps raising by incrementally adding each key component. This suggests that the designed components are effective and reasonable, and the maximum gain can be achieved by combing all these components together.

Table 2. Evaluation results on the SOTS outdoor testing set for different configurations.

Attention meachanism	√	√	√
Local residual learning		√	√
Smoothed dilation			√
PSNR	31.78	33.21	34.10

4. CONCLUSION

In this work, we propose an end-to-end trainable feature aggregation attention network and its superior performance in single image dehazing is demonstrated by quantitatively and qualitatively comparison experiments. By ingenious design and application of attention mechanism, smoothed dilated convolution and local residual learning, our method can remove haze as much as possible while preserving image detail and color fidelity as much as possible. Considering the superiority of our network in suppressing color distortion and artifacts, it is expected to be used to address other low-level vision problems, such as de-raining and denoising.

5. REFERENCES

[1] Earl J McCartney, "Optics of the atmosphere: scattering by molecules and particles," *New York, John Wiley and Sons, Inc., 1976. 421 p.*, 1976.

[2] Srinivasa G Narasimhan and Shree K Nayar, "Chromatic framework for vision in bad weather," in *CVPR*. IEEE, 2000, vol. 1, pp. 598–605.

[3] Srinivasa G Narasimhan and Shree K Nayar, "Vision and the atmosphere," *International journal of computer vision*, vol. 48, no. 3, pp. 233–254, 2002.

[4] Kaiming He, Jian Sun, and Xiaoou Tang, "Single image haze removal using dark channel prior," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 12, pp. 2341–2353, 2010.

[5] Qingsong Zhu, Jiaming Mai, and Ling Shao, "A fast single image haze removal algorithm using color attenuation prior," *IEEE transactions on image processing*, vol. 24, no. 11, pp. 3522–3533, 2015.

[6] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng, "Aod-net: All-in-one dehazing network," in *ICCV*, 2017, pp. 4770–4778.

[7] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang, "Single image dehazing via multi-scale convolutional neural networks," in *ECCV*. Springer, 2016, pp. 154–169.

[8] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao, "Dehazenet: An end-to-end system for single image haze removal," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5187–5198, 2016.

[9] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang, "Gated fusion network for single image dehazing," in *CVPR*, 2018, pp. 3253–3261.

[10] He Zhang and Vishal M Patel, "Densely connected pyramid dehazing network," in *CVPR*, 2018, pp. 3194–3203.

[11] Dongdong Chen, Mingming He, Qingnan Fan, Jing Liao, Liheng Zhang, Dongdong Hou, Lu Yuan, and Gang Hua, "Gated context aggregation network for image dehazing and deraining," in *WACV*. IEEE, 2019, pp. 1375–1383.

[12] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen, "Griddehazenet: Attention-based multi-scale network for image dehazing," in *ICCV*, 2019, pp. 7314–7323.

[13] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia, "Ffa-net: Feature fusion attention network for single image dehazing," *arXiv preprint arXiv:1911.07559*, 2019.

[14] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[15] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, "Image super-resolution using very deep residual channel attention networks," in *ECCV*, 2018, pp. 286–301.

[16] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha, "Recurrent squeeze-and-excitation context aggregation net for single image deraining," in *ECCV*, 2018, pp. 254–269.

[17] Ryuhei Hamaguchi, Aito Fujita, Keisuke Nemoto, Tomoyuki Imaizumi, and Shuhei Hikosaka, "Effective use of dilated convolutions for segmenting small object instances in remote sensing imagery," in *WACV*. IEEE, 2018, pp. 1442–1450.

[18] Zhengyang Wang and Shuiwang Ji, "Smoothed dilated convolutions for improved dense prediction," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2486–2495.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[20] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang, "Benchmarking single-image dehazing and beyond," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 492–505, 2018.

[21] Alessandro Giusti, Dan C Cireşan, Jonathan Masci, Luca M Gambardella, and Jürgen Schmidhuber, "Fast image scanning with deep max-pooling convolutional neural networks," in *ICIP*. IEEE, 2013, pp. 4034–4038.

[22] George Papandreou, Iasonas Kokkinos, and Pierre-André Savalle, "Modeling local and global deformations in deep learning: Epitomic convolution, multiple instance learning, and sliding window detection," in *CVPR*, 2015, pp. 390–399.

[23] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[24] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li, "Bag of tricks for image classification with convolutional neural networks," in *CVPR*, 2019, pp. 558–567.