Journal of KIIT. Vol. 19, No. 5, pp. 93-100, May 31, 2021. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2021.19.5.93

An End-to-End Single Image Dehazing Method Using U-Net Architecture

Quoc Hieu Nguyen*, Bongsoon Kang**

This research was supported by research funds from Dong-A University, Busan, South Korea

Abstract

Images captured outdoor always suffer from degradation due to light transmission medium. This problem is even worse in bad weather conditions such as fog or haze. Accordingly, for various vision-based applications, image dehazing or haze removal becomes a crucial pre-processing procedure. This paper presents an end-to-end deep learning-based method for hazy image restoration using U-Net architecture. The loss function is designed to keep the model more invariant to different kinds of input. Training data is created by small pieces of a modest-size real dataset that reduces the dependence on the limitation of real data sources while still ensuring consistent performance. The superior benchmarking results on outdoor and indoor real datasets have shown that the proposed model achieved better performance than existing methods.

요 약

실외에서 촬영된 대부분의 이미지는 전송 매체에 의해 성능이 저하된다. 이러한 문제는 안개와 같은 악조건 날씨에서 더욱 심각하다. 따라서 다양한 vision-based applications의 전처리 과정에서 정확한 검출을 위해 안개를 제거하는 것은 중요하다. 본 논문은 U-Net architecture를 사용한 End-to-End 딥러닝 기반의 안개 제거 방법을 제안한다. 알고리즘에 사용되는 Loss-function은 다른 입력 이미지에 대해 성능이 바뀌지 않도록 설계되었다. 원본 이미지에 대한 의존도를 줄이면서 동일한 성능을 갖기 위해 실제 이미지를 분할하여 학습 데이터로 사용했다. 실내 및 실외의 실제 이미지에 대한 결과를 통해 제안한 모델이 기존 방법보다 우수한 성능임을 증명한다.

Keywords
dehazing, deep learning, U-net, loss function, image restoration

* Dept. of Electronic Engineering, Dong-A University
- ORCID ID: https://orcid.org/0000-0003-0250-5996
** Professor, Dept. of Electronic Engineering, Dong-A University
- ORCID ID: https://orcid.org/0000-0001-6716-5799

· Received: Feb. 19, 2021, Revised : Mar. 24, 2021, Accepted: Mar. 27, 2021
· Corresponding Author: Bongsoon Kang
Depart. of Electronic Engineering, Dong-A University, 37 Nakdong-Daero 550 beon-gil, Saha-gu, Busan, Korea,
Tel.: +82-51-200-7703, Email: bongsoon@dau.ac.kr

Ⅰ. Introduction

Haze is an atmospheric phenomenon in which the density of microscopic particles suspending in the air becomes higher than usual. Due to the absorption and scattering of these particles, image quality deteriorates and thus, negatively affects the performance of several computer vision applications such as object detection, classification, and tracking. Image dehazing, or haze removal, focuses on both visibility restoration and performance improvement for successive processes. Among the two categories of haze removal algorithms based on single- and multiple-image, the first approach is worthy of research since it is more practical for real-time applications.

Image dehazing has been tackled from various aspects. Solving the Koschmieder atmospheric scattering model[1], which expresses the relationship between hazy and haze-free images, gains massive attention at the beginning. He et al.[2] introduced the Dark Channel Prior (DCP), stating that pixels in non-sky patches usually possess very-low intensity in at least one color channel. With this assumption, the corresponding atmospheric light and transmission map could be estimated to derive the dehazed result. To overcome the computational drawbacks of the DCP, Galdran[3] utilized the Laplacian pyramid decomposition scheme and multi-exposed image fusion for haze removal, thus avoiding the need for depth estimation. For real-time application, some lightweight designs for hardware implementation have also been proposed in [4]-[5]. In the machine learning category, the use of Color Attenuation Prior (CAP) proposed by Zhu et al.[6] to build a linear model of scene depth is a typical example. Besides, some studies[7]-[8] have been conducted to address the shortcoming of CAP. In the deep learning field, DehazeNet[9] and Multi-scale CNN[10] were proposed to learn the mapping between a hazy image and its medium transmission map. These approaches yield better results than some traditional methods but suffer from the lack of real datasets for training.

To improve performance and simplify the dehazing process, we proposed an end-to-end trainable model based on the U-Net architecture[11] that directly maps a hazy image to the corresponding dehazed one. Besides, the dependence on training data has been reduced through dividing images in a modest-size real dataset into smaller pieces for learning procedure. In this paper, Section Ⅱ will address the proposed design, Section Ⅲ refers to the training procedure, Section Ⅳ evaluates the model, and Section Ⅴ concludes the study.

Ⅱ. Proposed Method

2.1 Architecture

The U-net, which is famous in biomedical image segmentation, is an autoencoder-like network with skip connections. It consists of three sections: the contraction, the bottleneck, and the expansion. The input image is progressively down-sampled until a bottleneck layer, after which the process is reversed to obtain the output image.

Fig. 1 shows the flow chart of the proposed dehazing method with a five-level U-net. The contraction section is a chain of multi-scale blocks where each of them includes two convolutional layers with activation function, followed by a pooling layer. The feature numbers are doubled after each down-sampling step to learn the complex structures effectively. The bottleneck section situates at the center with two convolutional layers with activation function after each and an up-scaling layer. The expansion part is a repetition of blocks with a similar structure to the contraction but following an upsampling direction. At each level, the sizes of convolutional layers in each block are the same for both contractive and expansive paths. Typically, the output and input share much structural information;

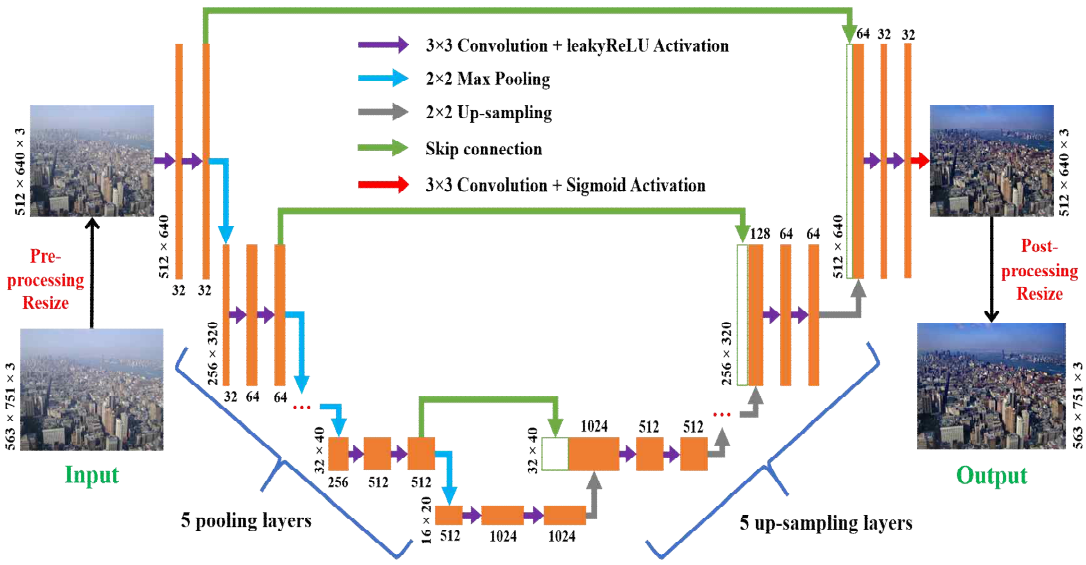

Fig. 1. The proposed dehazing method with a five-level U-net

thus, skip connections are used to directly combine data from the contracting section with the upsampled result at each level of the expansive path through concatenation. In addition, dropout layers, which randomly drop some connections, have also been inserted between two consecutive convolutional layers to avoid overfitting. In this design, all convolutional layers have a kernel size of 3×3, followed by a Leaky Rectified Linear Unit (LeakyReLU) activation function to avoid the gradient-vanishing problem, except the last layer that utilizes sigmoid activation function for a brighter image. The popular max-pooling method is employed for down-sampling while up-sampling is performed by the PixelShuffle[12] layer, which shifts the feature channels into the spatial domain to boost the detail of the output.

Due to 5 pairs of down- and up- sampling layers in the architecture, it is needed to resize input images to multiples of $2^5 \times 4 = 128$ (4 is the minimum size of features allowed in the bottleneck). Dehazed photo is then converted back to its original size at the end.

2.2 Loss function

The proposed model is optimized by minimizing the sum of four loss functions including total variation loss L_{TV}, Huber loss L_H, SSIM loss L_{SSIM}, and perceptual loss L_P as:

$$L = L_{TV} + L_H + L_{SSIM} + L_P \tag{1}$$

Total variation (TV), which is a measure of the complexity of an image with respect to its spatial variation is calculated by the sum of the absolute differences for neighboring intensities. Using total variation as a loss function helps reduce noise generated in dark regions during the dehazing process. The formula of total variation loss is described as:

$$L_{TV} = \frac{TV}{H \times W \times C} = \frac{\| \nabla x(i,j) \|}{H \times W \times C} \tag{2}$$

where (i,j) denotes the coordinates of pixel x, H, W, C are dimensions of the image.

Two of the most popular loss functions used for regression models are Mean Absolute Error (MAE)

and Mean Square Error (MSE), which help preserve both colors and luminance. MAE is more robust to outliers, but it is not differentiable at zero. In contrast, MSE is sensitive to outliers due to squared differences but gives a more stable solution. To utilize the advantages of these two loss functions, we propose to use the Huber loss as below:

$$L_H = \begin{cases} \frac{1}{2}(y_i - y_i^p)^2 & \text{for } |y_i - y_i^p| \le \delta \\ \delta |y_i - y_i^p| - \frac{1}{2}\delta^2 & otherwise \end{cases} \quad (3)$$

where y_i is the input value, y_i^p is the predicted result, and δ is a parameter that determines the shift of function toward MSE or MAE. Residuals smaller than δ are minimized with MSE, while MAE deals with the rest. Here, we use $\delta = 0.5$, which is the midpoint of the difference from 0 to 1 in image regression.

The SSIM (Structural SIMilarity)[8], which is a measure of the similarity in structural information between two images, is computed as:

$$SSIM(X, Y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (4)$$

where (μ_x, μ_y) and (σ_x, σ_y) are the local average and standard deviation of image X and Y, which are calculated by a Gaussian filter G_{σ_G}; C_1 and C_2 are constants to avoid dividing by zero; and σ_{xy} is the correlation coefficient between $(X - \mu_x)$ and $(Y - \mu_y)$. The SSIM value ranges from 0 to 1, in which the higher the number, the more structurally similarity between two images. SSIM loss function, which helps preserve the contrast in high-frequency regions, is computed as:

$$L_{SSIM} = 1 - SSIM \quad (5)$$

When using SSIM loss function, the output quality depends on the selected Gaussian standard deviation σ_G. Large numbers may generate noise, while the low values will lead to the loss of local structure and produce artifact. Through experiments, the value of 8 was selected for σ_G.

Perceptual loss[13] calculates the error between features extracted from both hazy image I and its ground-truth J with a pre-trained model as:

$$L_P = \sum_{l \in L} \| F_l(I) - F_l(J) \|_2 \quad (6)$$

where F_l is the feature map of layer l among L selected layers. In this study, we used the layers conv1_1, conv2_1, and conv3_1 inside the VGG19 model with ImageNet pre-trained weight set.

III. Training

3.1 Training data

To ensure that the proposed model can work well in a vast range of semantic concepts, we created a combinational dataset that consists of synthetic and real images. The first part includes hazy images synthesized by 1000 outdoor ground truth images using the atmospheric scattering model. The real data is created by dividing high-resolution images in the DenseHaze dataset[14], which contains 55 pairs of hazy and haze-free images collected by a professional setup, into 1100 smaller ones with the size of 256×256. We then divided it into a 1680-image training set and a 420-image validation set.

3.2 Parameter Settings

The proposed model was designed using Python 3.6 and Tensorflow 1.14-gpu. Training was performed on a computer with an Intel Core I9-9900K CPU, 64 GB DDR4 RAM 2400MHz, and NVIDIA Titan RTX.

(a) Hazy image (b) He et al.[2] (c) Galdran et al.[3] (d) Cai et al.[9] (e) Ren et al.[10] (f) Ours

Fig. 2. Qualitative comparison between different methods

Convolutional weights were initialized using Glorot initialization with all bias values were set to zero. The model was trained by using RMSprop optimizer with a fixed learning rate of 0.0001 and a batch-size of 1. We utilized early stopping mechanism to finish the training after the validation loss stopped decreasing for 7 consecutive epochs.

IV. Evaluation

In this section, we evaluate the proposed method through comparison with four approaches in [2], [3], [9], and [10]. The image-processing-based technique from He et al.[2] works well in most cases but gets difficulties dealing with sky regions due to limitations in the DCP. Image enhancement method by Galdran[3] can solve this problem but at the cost of leftover haze and color degradation due to histogram equalization. Cai et al.[9] and Ren et al.[10] only use deep architectures for medium transmission estimation in lieu of an end-to-end mapping between haze and haze-free images; thus cannot exploit the powerful potential of neural network for the whole system.

IVC[7], O-HAZE[7], and I-HAZE[7] real datasets are considered, in which the last two contain outdoor and indoor pairs of hazy and haze-free images while the former only consists of hazy scenes.

Fig. 2 presents real hazy photos with corresponding results from five benchmarking methods. The algorithm of He et al.[2] has a pleasant dehazing ability, but it suffers from strong artifacts at the horizon in the first two images and results in dark scenes after eliminating dense haze. Galdran[3] can moderately restore the overall structure of the images but at the cost of slightly dark scenes with pale color and remaining haze in most cases. The DehazeNet from Cai et al.[9] obtains similar results to Gardran[3], but shows lower performance with thick fog. Ren et al.[10] seems to be an over-exposed version of He et al.[2], except that it can reduce the visual artifacts in the sky regions. In contrast, our model possesses strong dehazing power through clear,

bright, and detailed images in both outdoor and indoor cases. For the first scene, it produces a cleared sky without any artifacts or leftover haze. In the second one, our method recovers vehicles that cannot be seen in other results and introduces less artifacts than [2]. It can also diminish the very thick haze, which is a big challenge to the methods in [3], [9], and [10]; while still keeping other regions from becoming too dark compared to [2], as shown in the last two images.

Next, we quantitatively assess the dehazing performance by utilizing four evaluation metrics, including MSE, SSIM[7], Tone Mapped Image Quality Index (TMQI)[7], and Feature SIMilarity Index extended to color image (FSIMc)[7]. MSE and SSIM have been mentioned in section 2.2. TMQI, ranging from 0 to 1 for the best, measures the structural fidelity and statistical naturalness. The first part works similarly to SSIM, while the second one is computed by the probability distribution of the own brightness and contrast of the evaluated image. FSIMc quantifies image quality based on salient low-level features and chromatic similarity.

It is the only measurement, among the four metrics, that assesses images in terms of color with value varies between 0 and 1, where the higher the better.

Table 1. Average results of O-HAZE datasets

Method	MSE	SSIM	TMQI	FSIMc
He et al.[2]	0.0200	0.7709	0.8403	0.8423
Galdran et al.[5]	0.0168	0.7877	0.8410	0.8468
Cai et al.[11]	0.0266	0.6999	0.8413	0.7865
Ren et al.[12]	0.0155	0.7997	0.8737	0.8553
Proposed method	**0.0088**	**0.8610**	**0.9228**	**0.9076**

Table 2. Average results of I-HAZE datasets

Method	MSE	SSIM	TMQI	FSIMc
He et al.[2]	0.0535	0.6580	0.7319	0.8208
Galdran et al.[5]	0.0336	0.7547	0.7613	0.8558
Cai et al.[11]	0.0320	0.7115	0.7598	0.8482
Ren et al.[12]	0.0223	0.7786	0.7819	0.8634
Proposed method	**0.0157**	**0.8297**	**0.8649**	**0.9076**

The average scores for O-HAZE and I-HAZE datasets are provided in Tables 1 and 2, where the bold number represents the best value in each field. Our method obtains the best results in all metrics with considerable differences from the second place. DenseHaze images with thick fog in training data help the model eliminate dense haze in most test cases and successfully recover the original structure. Under the FSIMc metric, which assesses images concerning chrominance, the similarity in color between our dehazed results and ground-truths is the main reason for its highest score, while images from other methods are darkened due to the dehazing effect.

Ⅴ. Conclusion

This paper proposed an end-to-end convolutional network based on a five-level U-Net architecture for single-image haze removal. The powerful loss function plays a vital role in the dehazing ability. Besides, creating training dataset by splitting high-resolution images is useful when dealing with the lack of real databases. The model successfully recovered hazy images in detail, contract, and color, proving through qualitative evaluations. Moreover, our superior performance is shown by the best quantitative scores in four measurements as witnessed by the difference of at least 5% from the second-best algorithm. However, this design still leaves artifacts in flat regions, and we leave this issue for future research.

References

[1] Z. Lee and S. Shang, "Visibility: How Applicable is the Century-Old Koschmieder Model?", Journal of Atmospheris Science, Vol. 73, No. 11, pp. 4573-4581, Nov. 2016

[2] K. He, J. Sun, and X. Tang, "Single Image Haze Removal Using Dark Channel Prior", IEEE Transactions on Pattern Analysis and Machine

Intelligence, Vol. 33, No. 12, pp. 2341-2353, Dec. 2011.

[3] A. Galdran, "Image dehazing by artificial multiple-exposure image fusion", Signal Processing, Vol. 149, pp. 135-147, Aug. 2018.

[4] Q.-H. Nguyen, B. Kang, "FPGA-based Haze Removal Architecture using Multiple-exposure Fusion", The Journal of Korean Institute of Information Technology, Vol. 18, No. 5, pp. 85-90, May 2020.

[5] D. Ngo, S. Lee, Q.-H. Nguyen, T. M. Ngo, G.-D. Lee, and B. Kang, "Single Image Haze Removal from Image Enhancement Perspective for Real-Time Vision-Based Systems", Sensors, Vol. 20, No. 18, pp. 5170, Sep. 2020.

[6] Q. Zhu, J. Mai, and L. Shao, "A Fast Single Image Haze Removal Algorithm Using Color Attenuation Prior", IEEE Transactions on Image Processing, Vol. 24, No. 11, pp. 3522-3533, Nov. 2015.

[7] D. Ngo, G. D. Lee, and B. Kang, "Improved Color Attenuation Prior for Single-Image Haze Removal", Applied Sciences, Vol. 9, No. 19, pp. 4011, Jan. 2019.

[8] D. Ngo, S. Lee, G.-D. Lee, and B. Kang, "Single-Image Visibility Restoration: A Machine Learning Approach and Its 4K-Capable Hardware Accelerator," Sensors, Vol. 20, No. 20, pp. 5795, Oct. 2020.

[9] 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, Vol. 25, No. 11, pp. 5187-5198, Nov. 2016.

[10] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, M. Yang, "Single Image Dehazing via Multi-scale Convolutional Neural Networks", 2016 European Conference on Computer Vision, pp. 154-169, Nov. 2016.

[11] O. Ronneberger, P. Fischer, T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", 2015 Medical Image Computing and Computer-Assisted Intervention, pp. 234-241, Oct. 2015

[12] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, Z. Wang, "Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network", 2016 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874-1883, Jul. 2016.

[13] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution", 2016 European Conference on Computer Vision, pp. 694-711, Nov. 2016.

[14] C. O. Ancuti, C. Ancuti, M. Sbert and R. Timofte, "Dense-Haze: A Benchmark for Image Dehazing with Dense-Haze and Haze-Free Images," 2019 IEEE International Conference on Image Processing, pp. 1014-1018, Sep. 2019.

Authors

Quoc Hieu Nguyen

2016 : BS degree in Department of Electronic and Telecommunication Engineering, University of Danang.

2019 ~ present : Pursuing MS degree in Electronic Engineering, Dong-A University.

Research interests : VLSI architecture design, and image processing

Bongsoon Kang

1985 : BS degree in Electronic Engineering, Yonsei University.
1987 : MS degree in Electronic Engineering, Pennsylvania University.
1990 : PhD degree in Electrical and Computer Engineering, Drexel University.
1989 ~ 1999 : Senior Staff Researcher, Samsung Electronics.
1999 ~ present : Prof. of Dept. Electronic Engineering, Dong-A University.
Research interests : image/video processing, pattern recognition, and SoC designs