

Y-NET: MULTI-SCALE FEATURE AGGREGATION NETWORK WITH WAVELET STRUCTURE SIMILARITY LOSS FUNCTION FOR SINGLE IMAGE DEHAZING

Hao-Hsiang Yang¹

Chao-Han Huck Yang²

Yi-Chang James Tsai²

¹ASUS Intelligent Cloud Services, Taiwan

²School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

Single image dehazing is the ill-posed two-dimensional signal reconstruction problem. Recently, deep convolutional neural networks (CNN) have been successfully used in many computer vision problems. In this paper, we propose a Y-net that is named for its structure. This network reconstructs clear images by aggregating multi-scale features maps. Additionally, we propose a Wavelet Structure SIMilarity (W-SSIM) loss function in the training step. In the proposed loss function, discrete wavelet transforms are applied repeatedly to divide the image into differently sized patches with different frequencies and scales. The proposed loss function is the accumulation of SSIM loss of various patches with respective ratios. Extensive experimental results demonstrate that the proposed Y-net with the W-SSIM loss function restores high-quality clear images and outperforms state-of-the-art algorithms. Code and models are available at <https://github.com/dectrfov/Y-net>

Index Terms— Single image dehazing, Y-net, discrete wavelet transform, structure similarity, multi-scale feature aggregation

1. INTRODUCTION

Single image dehazing is the two-dimensional signal reconstruction problem and intends to restore the unknown clear image given a hazy or foggy image. The image dehazing model [1] is formulated as follows:

$$I(x) = J(x)t(x) + A(1 - t(x)) \quad (1)$$

where $I(x)$ is the observed image, $J(x)$ is the clear image we need to estimate, A is the global atmospheric light, and $t(x)$ is the medium transmission. Assuming that the atmosphere is homogeneous, we express the transmission as $t(x) = e^{-\beta d(x)}$, where β is the scattering coefficient of atmosphere, and $d(x)$ is the scene depth. Since all variables in Equation (1) are unknown except the hazy image $I(x)$, image dehazing is an ill-posed problem.

Previously, many efforts on developing visual priors capture deterministic and statistical properties of hazy images.

In [2], color attenuation prior is proposed to recover the depth information. Berman *et al.* [3] find colors in haze-free images can be well approximated by a few hundred distinct colors that form tight clusters in RGB space. Garldran *et al.* [4] prove that Retinex on inverted intensities is a solution to the image dehazing problem. Cho *et al.* [5] propose the multi-band fusion approach using several inputs derived from the original image, and present balanced image enhancement while elaborating image details.

Deep learning frameworks also show competitive results in the dehazing problems. In [6], the multi-scale deep neural network is proposed to learn a non-linear mapping. Li *et al.* [7] reformulate the atmospheric scattering model and present a lightweight model to produce clear images from hazy images. Recently, several deep-learning-based U-Nets methods [8, 9, 10] have also been proposed. In [11], authors replace down-sampling with the discrete wavelet transform in U-net [8] to estimate dehazed images.

Similarly, we consider neural networks for image dehazing, since the dehazing model is only a crude approximation and CNN can capture more detailed features from hazy images. We think reconstructed images are composed of different scale feature maps and propose the Y-net. The Y-net aggregates various size feature maps that are up-sampled to identical size and convolves them by 1×1 kernels to restore clear images. Thus, all feature maps are comprised of the output images, and weights of all layers are updated effectively. In this paper, we show that applying existing encoder-decoder structures like U-net [8] or cascaded refinement networks [12] cannot produce optimal results. Our Y-net, on the contrary, amplifies the merit of various CNN structures and yields the feasibility for single image dehazing.

Not only network structures but loss functions impact accuracy. In [13], the loss function based on the Structure SIMilarity (SSIM) index [14] is proposed for image restoration. The authors demonstrate that when the network architecture is unchanged, the quality of the results improves significantly with better loss functions. Therefore, we extend the SSIM loss by combining it with the discrete wavelet transform (DWT) and propose the wavelet SSIM (W-SSIM) loss. Images are divided into many patches by DWT with various frequencies. Then, the SSIM loss of each patch is

calculated, and the weights of each loss are adjustable to preserve high-frequency details and prevent halo artifacts. Experimental results show that the proposed loss function achieves improved reconstructions with an identical network.

The contributions of this paper are two-fold. First, we propose a deep end-to-end trainable Y-net to reconstruct clear images without estimating any priors on atmospheric light and scene transmission. We demonstrate the utility and effectiveness of the Y-net for single image dehazing on the synthetic hazy image dataset. Second, the W-SSIM loss is proposed to improve accuracy. We show that the proposed dehazing model performs favorably against the related methods.

2. METHODOLOGY

2.1. Y-net: Multi-scale feature aggregation network

The proposed Y-net architecture mainly consists of three modules: down-sampling, up-sampling, and aggregation parts. The first two modules are similar to the classical U-net [8]. Our Y-net uses convolutions and down-sampling to extract multi-scale features. After bottom features are computed, convolutions and up-sampling are adopted to magnify the size of feature maps. The original U-net adopts the final up-sampled feature map to represent reconstructed images. Different from the U-net, we think reconstructed images are composed of different scale feature maps, and propose the aggregation module to combine each feature map to reconstruct clear images. Each feature map is adjusted to identical size by up-sampling and convolved by 1×1 kernels to restore clear images. The 1×1 convolution is seen as the operation of the weighted sum of various feature maps. Thus, all feature maps are comprised of the output images, and weights of all layers are updated effectively, which also avoids vanishing gradient. Our input is the 3-channel image, convolutions, whose size are 3×3 with a stride of two are adopted to down-sampling, the whole network architecture is shown in Fig. 1.

2.2. Wavelet SSIM loss

In the two dimensional discrete wavelet transform (DWT), for instance, four filters, f_{LL} , f_{HL} , f_{LH} and f_{HH} , are used to convolve with an image. The convolution results are then down-sampled. The 2-D DWT has separable property, so four filters are obtained by multiplication of the scaling function and the wavelet function. Scaling function is seen as the low-pass filter, and the wavelet function is seen as the high-pass filter. The block diagram of the DWT and the results are showing in Fig. 2, and the formula is represented as follows:

$$I^{LL}, I^{LH}, I^{HL}, I^{HH} = \text{DWT}(I) \quad (2)$$

where superscripts mean the output from respective filters. I^{LL} is the down-sampling image, I^{HL} and I^{LH} are horizontal and vertical edge detection images, and I^{HH} is the corner

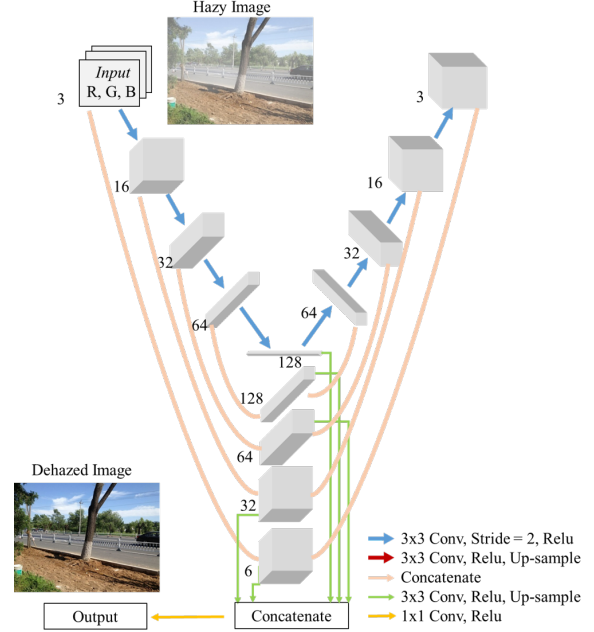


Fig. 1. The overview of our proposed Y-net. The clear image is composed of multi-scale feature maps from the hazy image. The digits under the blocks mean the numbers of channels.

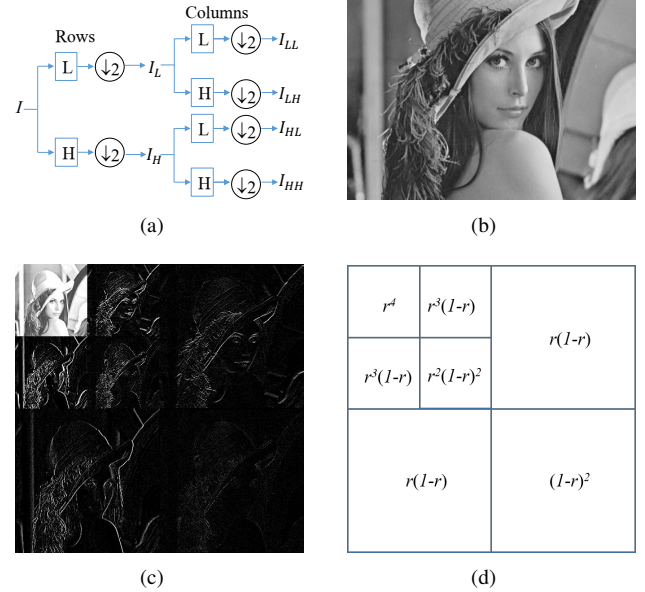


Fig. 2. The process of the DWT and the transformed example: (a) The process of the DWT, where downward arrows mean down-sampling. (b) The original image. (c) The result of the two-times DWT. (d) The ratios for different patches.

detection image. The DWT decomposes an image into four small patches with different frequency information. Further-

more, the low-frequency information of the image (I^{LL}) can be decomposed iteratively and divided into the multi-scale patches with various frequencies, as shown in Fig. 2(c). The iterative DWT can be formulated as follows:

$$I_{i+1}^{LL}, I_{i+1}^{LH}, I_{i+1}^{HL}, I_{i+1}^{HH} = \text{DWT}(I_i^{LL}) \quad (3)$$

where subscripts mean the output of the i_{th} time DWT. It is noted that I_0^{LL} means the original image.

We apply DWT to improve the original SSIM loss. The SSIM for two images x, y is defined as:

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

where μ and σ represent the means, standard deviation, and covariance of images. A higher SSIM indicates the two images are more similar to each other, and the SSIM equals 1 for identical images. The loss function for the SSIM can be then written as follows:

$$L_{\text{SSIM}} = -\text{SSIM}(x, y) \quad (5)$$

In practice, means and standard deviations are computed with a Gaussian filter. Therefore, the total function is differentiable and can be used in neural networks. The detailed derivation is analyzed in [13]. We extend the SSIM loss function and combine the DWT. The input is not only the whole image but the various patches computed by the DWT. Varying the image resolution in a coarse-to-fine manner in loss function prevents undesired artifacts. The proposed W-SSIM loss can be described as follows:

$$L_{\text{W-SSIM}}(x, y) = \sum_0^i r_i L_{\text{SSIM}}(x_i^w, y_i^w), \quad (6)$$

$$w \in \{LL, HL, LH, HH\}$$

where r_i controls weights of each patch. Because hazy images suffer from low contrast, faint color, and shifted luminance, the high-frequency patches own higher ratios [15]. We set ratios of four patches as:

$$I^{LL} : I^{LH} : I^{HL} : I^{HH} = r^2 : r(1-r) : r(1-r) : (1-r)^2 \quad (7)$$

Initially, the SSIM loss of the whole images is computed. Then, the DWT decomposes the images, and SSIM loss of three high-frequency parts are accumulated except the low-frequency part (I_i^{LL}). I_i^{LL} can be divided by the DWT to generate various patches, and accumulate new loss. The Algorithm 1 describes the weights of different patches. x, y and z are constructed as auxiliary numbers to record current weights in different scales. After the DWT is implemented, x, y and z are updated. The ratios for different patches are plotted in Fig. 2(d). The total loss function is $L_2 + L_{\text{W-SSIM}}$. Since ratios of low-frequency patches in Algorithm 1 are small, L_2 is helpful to reconstruct low-frequency parts.

Input: Two images I, J , the ratio for multi-frequency r and iterative times n

Output: $\text{loss} = L_{\text{W-SSIM}}(I, J)$

```

1  $I_0^{LL}, J_0^{LL} = I, J;$ 
2  $\text{Tensor } \text{loss} = 0;$ 
3  $x = r^2, y = r(1-r), z = (1-r)^2$  for
   $i = 1; i \leq n; i++$  do
4    $I_i^{LL}, I_i^{LH}, I_i^{HL}, I_i^{HH} = \text{DWT}(I_{i-1}^{LL})$ 
5    $J_i^{LL}, J_i^{LH}, J_i^{HL}, J_i^{HH} = \text{DWT}(J_{i-1}^{LL})$ 
6    $\text{loss} += L_{\text{SSIM}}(I_i^{LH}, J_i^{LH}) \cdot y +$ 
      $L_{\text{SSIM}}(I_i^{HL}, J_i^{HL}) \cdot y + L_{\text{SSIM}}(I_i^{HH}, J_i^{HH}) \cdot z$ 
      $[x, y, z] = x \cdot [x, y, z]$ 
7 end
8  $\text{loss} += L_{\text{SSIM}}(I_i^{LL}, J_i^{LL}) \cdot x$ 
9 return  $\text{loss}$ 
```

Algorithm 1: W-SSIM Loss

3. EXPERIMENTAL RESULTS

3.1. Datasets and training details

In this work, we adopt the RESIDE[16] dataset as training data. This dataset contains indoor images and outdoor images with corresponding hazy synthetic images. We mainly select 1400 outdoor images and the corresponding synthetic images as training data and select 500 images to evaluate. We set $r = 0.4$ in Algorithm 1. The 2D Haar wavelet [17] is adopted, and the W-SSIM loss implements the DWT three times. During training, Adam is used [18] as the optimization algorithm with a learning rate of 0.0001, and a batch size of 32. All training images are resized to 480×480 , and the network is trained for 400 iterations.

3.2. Image Dehazing Results

We adopt PSNR and SSIM for the quantitative evaluation and compare the proposed model with several state-of-the-art dehazing methods: Color Attenuation Prior (CAP) [2], AOD-Net [7], Multi-band enhancement [5], and Wavelet U-net (W U-net) [11]. CAP and MBE are prior-based methods, and the other two methods belong to deep-learning-based methods. The networks of these two methods are the cascaded refinement network and the U-net, respectively. The comparison results are shown in Table 1. It is obvious that the PSNR and SSIM from prior-based methods are lower than other deep-learning-based methods. Furthermore, our method performs favorably against the other two deep-learning-based competitors in this dataset, which shows that the proposed Y-net can reconstruct clear and detailed images.

Besides quantitative analysis, a few dehazed images by all methods are depicted in Fig. 3. All images of River contain artifacts or noises on the river surface except ours. Restored images from MBE own higher contrast but are vulnerable to

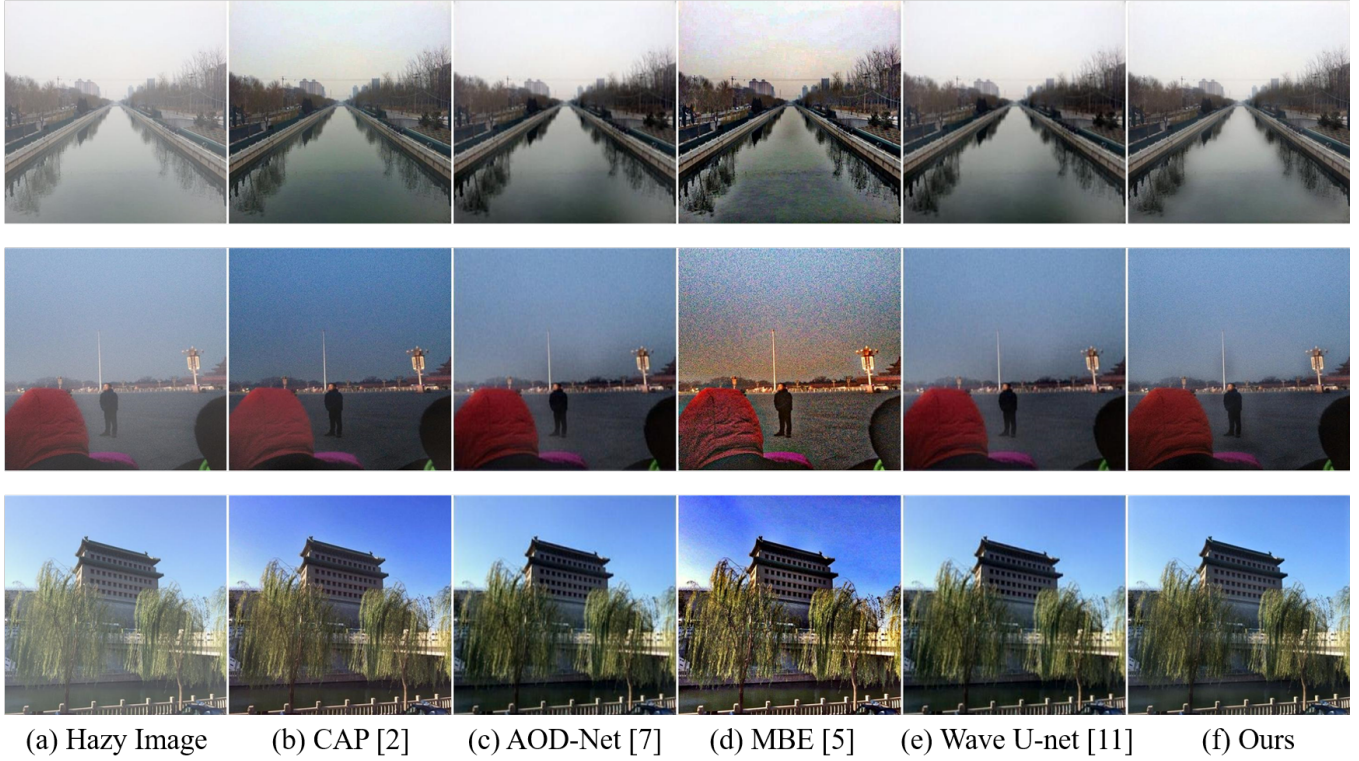


Fig. 3. Dehazed results in River, People and Willow.

noise. For example, People is filled with speckle noise. We calculate fog aware density evaluation (FADE) [19] scores and show them in Table 2. FADE not only provides a perceptual density score of the entire image but reports the local perceptual density of each patch. The remaining hazes, noises, and artifacts on images decrease scores. Our method produces good results on all images, which means the proposed network restores high-quality clear images.

Table 1. Quantitative SSIM and PSNR on the synthetic RE-SIDE dataset.

	PNSR	SSIM
CAP [2] (prior-based)	23.02	0.865
AOD-Net [7] (learning-based)	23.92	0.875
MBE [5] (prior-based)	18.83	0.790
W U-net [11] (learning-based)	24.81	0.910
Ours	26.61	0.947

Table 2. Quantitative FADE on restored images.

	River	People	Willow
CAP [2]	1.41	0.410	0.496
AOD-Net [7]	1.19	0.373	0.391
MBE [5]	0.440	0.184	0.184
W U-net [11]	1.51	0.647	0.562
Ours	1.77	2.37	0.592

We also consider four loss functions, L_2 , L_{SSIM} , L_{W-SSIM} and $L_{W-SSIM} + L_2$ to train the proposed Y-net; results are listed in Table 3. Experimental results show that using the W-SSIM loss improves both PSNR and SSIM, even though it is an SSIM-based function. The reason is that the proposed loss function is similar to perceptual loss [20]. Based on the proposed loss function, our model not only learns whole images but the different frequencies and scale features. Moreover, $L_{W-SSIM} + L_2$ is slightly better than original L_{W-SSIM} .

Table 3. SSIM and PSNR results of all loss functions applied for the proposed network.

	L_2	L_{SSIM}	L_{W-SSIM}	$L_{W-SSIM} + L_2$
PSNR	26.31	26.27	26.50	26.61
SSIM	0.925	0.929	0.939	0.947

4. CONCLUSION

In this paper, we address the single image dehazing problem via the end-to-end Y-net that aggregates multi-scale features to reconstruct clear images. The experimental results demonstrate the proposed network outperforms the other existing methods, both qualitatively and quantitatively. Furthermore, the proposed W-SSIM loss is able to improve accuracy when the network architecture is unchanged.

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] Qingsong Zhu, Jiaming Mai, Ling Shao, et al., "A fast single image haze removal algorithm using color attenuation prior.," *IEEE Trans. Image Processing*, vol. 24, no. 11, pp. 3522–3533, 2015.
- [3] Dana Berman, Shai Avidan, et al., "Non-local image dehazing," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1674–1682.
- [4] Adrian Galdran, Aitor Alvarez-Gila, Alessandro Bria, Javier Vazquez-Corral, and Marcelo Bertalmío, "On the duality between retinex and image dehazing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8212–8221.
- [5] Younggun Cho, Jinyong Jeong, and Ayoung Kim, "Model-assisted multiband fusion for single image enhancement and applications to robot vision," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 2822–2829, 2018.
- [6] Wenqi Ren, Si Liu, Hua Zhang, Jinshan Pan, Xiaochun Cao, and Ming-Hsuan Yang, "Single image dehazing via multi-scale convolutional neural networks," in *European conference on computer vision*. Springer, 2016, pp. 154–169.
- [7] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng, "Aod-net: All-in-one dehazing network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 1, p. 7.
- [8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [9] C-H Huck Yang, Fangyu Liu, Jia-Hong Huang, Meng Tian, MD I-Hung Lin, Yi Chieh Liu, Hiromasa Morikawa, Hao-Hsiang Yang, and Jesper Tegner, "Auto-classification of retinal diseases in the limit of sparse data using a two-streams machine learning model," in *Asian Conference on Computer Vision*. Springer, 2018, pp. 323–338.
- [10] C-H Huck Yang, Jia-Hong Huang, Fangyu Liu, Fang-Yi Chiu, Mengya Gao, Weifeng Lyu, Jesper Tegner, et al., "A novel hybrid machine learning model for auto-classification of retinal diseases," *arXiv preprint arXiv:1806.06423*, 2018.
- [11] Hao-Hsiang Yang and Yanwei Fu, "Wavelet u-net and the chromatic adaptation transform for single image dehazing," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 2736–2740.
- [12] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3883–3891.
- [13] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on Computational Imaging*, vol. 3, no. 1, pp. 47–57, 2016.
- [14] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [15] Xin Liu, He Zhang, Yiu-ming Cheung, Xinge You, and Yuan Yan Tang, "Efficient single image dehazing and denoising: An efficient multi-scale correlated wavelet approach," *Computer Vision and Image Understanding*, vol. 162, pp. 23–33, 2017.
- [16] 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, 2019.
- [17] Stéphane Mallat, *A wavelet tour of signal processing*, Elsevier, 1999.
- [18] D Kinga and J Ba Adam, "A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015, vol. 5.
- [19] Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik, "Referenceless prediction of perceptual fog density and perceptual image defogging," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3888–3901, 2015.
- [20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711.