SCANet: Self-Paced Semi-Curricular Attention Network for Non-Homogeneous Image Dehazing

Yu Guo[1] Yuan Gao[1] Ryan Wen Liu[1*] Yuxu Lu[1]
Jingxiang Qu[1] Shengfeng He[2] Wenqi Ren[3]
[1]Wuhan University of Technology [2]Singapore Management University
[3]Sun Yat-sen University

Abstract

The presence of non-homogeneous haze can cause scene blurring, color distortion, low contrast, and other degradations that obscure texture details. Existing homogeneous dehazing methods struggle to handle the non-uniform distribution of haze in a robust manner. The crucial challenge of non-homogeneous dehazing is to effectively extract the non-uniform distribution features and reconstruct the details of hazy areas with high quality. In this paper, we propose a novel self-paced semi-curricular attention network, called SCANet, for non-homogeneous image dehazing that focuses on enhancing haze-occluded regions. Our approach consists of an attention generator network and a scene reconstruction network. We use the luminance differences of images to restrict the attention map and introduce a self-paced semi-curricular learning strategy to reduce learning ambiguity in the early stages of training. Extensive quantitative and qualitative experiments demonstrate that our SCANet outperforms many state-of-the-art methods. The code is publicly available at https://github.com/gy65896/SCANet.

Figure 1. Dehazing results of the proposed SCANet on the NTIRE2023 test set. Our method can reconstruct high-quality haze-free images.

1. Introduction

The existence of turbid media in the atmosphere can lead to the absorption and scattering of light, resulting in degraded hazy scenes that adversely affect the performance of vision-driven scene understanding and object detection methods [36, 42]. To tackle this issue, many physical prior-based image dehazing models have been proposed [5, 13, 17, 20, 32, 44, 45]. These models typically represent the imaging process using an atmospheric scattering model, which can be expressed as follows

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

where x is the pixel index, I, J, t, and A represent the hazy image, clear image, transmission map, and global atmospheric light, respectively. However, it is critical for the success of physical prior-based dehazing methods to estimate t and A. When hazy scenes are complex, the estimation of t and A may be inaccurate, leading to unsatisfactory dehazing performance. To achieve superior dehazing performance, numerous learning-based single image dehazing methods [6, 7, 10, 12, 14, 19, 22–25, 27, 28, 39] have been proposed by leveraging the powerful nonlinear feature representation capacity of deep neural networks. However, haze may be spatially variable and non-uniform in the realistic scenes, making many physical prior- and learning-based

*Corresponding author (wenliu@whut.edu.cn).

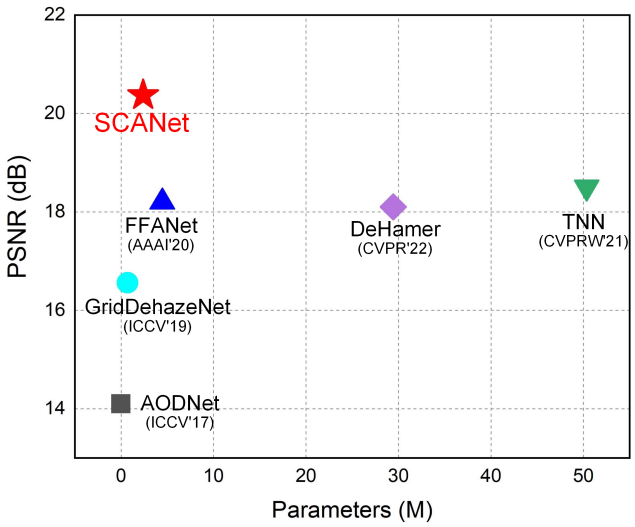

Figure 2. Comparisons of PSNR and parameters of several state-of-the-art dehazing methods on 15 non-homogeneous images from NTIRE2020, NTIRE2021, and NTIRE2023 datasets.

methods designed for homogeneous haze inapplicable.

In recent years, many methods have been proposed to address the challenge of non-homogeneous image dehazing [21, 31, 37, 41]. However, modeling the complex interactions between non-homogeneous haze and the underlying scene remains a challenging task. The key challenge is accurately perceiving the distribution of haze and reconstructing the texture detail of haze-dense areas with high quality. To address this issue, we propose a self-paced semi-curricular attention network (SCANet) for non-homogeneous image dehazing, which consists of an attention generation network and a scene reconstruction network. To better restore areas with significant luminance changes, we design a self-paced semi-curricular learning strategy to control the generation of attention maps. Figure 1 displays three dehazing cases on the NTIRE2023 test set. The proposed SCANet can adaptively extract non-homogeneous haze features and effectively suppress its interference. Furthermore, Figure 2 compares the peak signal-to-noise ratio (PSNR) and parameters of our method with state-of-the-art methods, demonstrating the competitive performance of our SCANet.

Overall, our main contributions are as follows

- To address the challenging problem of non-homogeneous image dehazing, we propose an attention network that learns complex interaction features between non-homogeneous haze and the underlying scene. The proposed method employs a novel "attention generation-scene reconstruction" paradigm specifically designed for non-homogeneous image dehazing.
- To enhance the haze removal ability in areas with significant luminance differences, we introduce a self-paced semi-curricular learning-driven attention map generation strategy. This approach improves model convergence and reduces the learning ambiguity caused by multi-objective prediction in the early stages of training.
- We extensively evaluate the proposed SCANet through qualitative and quantitative experiments, demonstrating its superior performance compared to state-of-the-art methods. We conduct an ablation analysis to confirm the effectiveness of our method, highlighting the contribution of each component to the overall performance of SCANet.

2. Related Works

Physical Prior-Based Dehazing. Physical prior-based methods depend on the physical scattering model. Some methods treat empirical observation as the prior knowledge to restore a hazy image, such as dark channel prior (DCP) [17], color attenuation prior [43], and non-local prior [5]. He *et al.* [17] proposed a dark channel prior (DCP) of clean outdoor images in terms of pixel intensities and achieved a nice dehazing performance. Zhu *et al.* [43] discovered the brightness and saturation of the pixels in hazy images are different and proposed the color attenuation prior. Berman *et al.* [5] proposed an effective non-local path prior based on the observation that the pixel are usually non-local in a given RGB space. While these priors can yield impressive results in certain scenarios, they may not always be practically applicable. In the real world, haze is often influenced by a variety of complex factors, making these priors unsuitable and resulting in suboptimal dehazing outcomes. For instance, the DCP [17] fails to dehaze the sky regions properly due to the inapplicable prior assumption.

Deep Learning-Based Dehazing. With the rapid advancement of deep learning, numerous learning-based dehazing methods have been proposed. Cai *et al.* [6] introduced an end-to-end network (DehazeNet), which generates the transmission map of the hazy image and recovers a clear image via the atmospheric scattering model. Li *et al.* [22] proposed an all-in-one dehazing network (AOD-Net) that jointly estimates the atmospheric light and transmittance to recover the hazy image. Ren *et al.* [30] applied a fusion-based strategy using a multi-scale structure in their haze-free image generation framework. Zhang *et al.* [38] proposed a densely connected pyramid dehazing network (DCPDN), which estimates the transmission map using an edge-preserving densely connected en-decoder structure with a multilevel pyramid pooling module. Qu *et al.* [29] proposed an enhanced pix2pix dehazing network (EPDN)

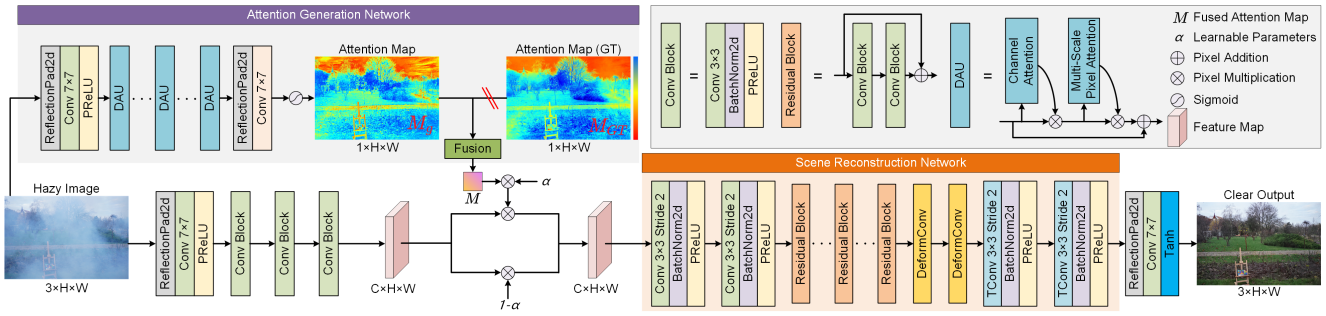

Figure 3. The network structure of our SCANet. The proposed method comprises an attention generation network and a scene reconstruction network. The red slash means we only use M_{GT} during the training phase.

that uses a generative adversarial network and an enhancer to accomplish the dehazing task. Chen *et al.* [7] introduced a gated context aggregation network (GCANet) that uses the smoothed dilation technique to efficiently generate a haze-free image. Liu *et al.* [26] proposed an attention-based multi-scale network (GridDehazeNet), which learns the feature map directly instead of estimating the transmission map. Recently, some studies [7, 9, 28, 35] tend to estimate the haze-free image or the residual between the hazy image and the corresponding clear image. Hong *et al.* [18] proposed an uncertainty-driven dehazing network (UDN) that improves the dehazing results by using the relationship between uncertain and confident representations. Although significant progress has been made by these methods in dehazing tasks, they tend to overlook the issue of non-homogeneous haze suppression. In recent years, several methods [21, 31, 37, 41] have been proposed to address this challenge. However, researchers are still struggling with the difficulty of learning haze distribution features and the poor quality of detail recovery in heavily hazy regions.

3. Proposed Method

In this section, we first introduce the network architecture of our SCANet. Then, we describe the proposed self-paced semi-curricular learning-driven attention map generation method. Finally, the loss functions employed in model training are mentioned.

3.1. Network Architecture

As illustrated in Figure 3, our method comprises two sub-networks: the attention generation network (AGN) and the scene reconstruction network (SRN). The AGN is composed of multiple dual-attention basic units (DAUs) to generate attention feature maps, while the SRN is an encoder-decoder network to reconstruct haze-free images.

Attention Generator Network. Our first sub-network (AGN) is designed to produce the attention feature map. Essentially, the AGN is stacked by multiple dual-attention units (DAUs), as shown in Figure 4. The input feature map

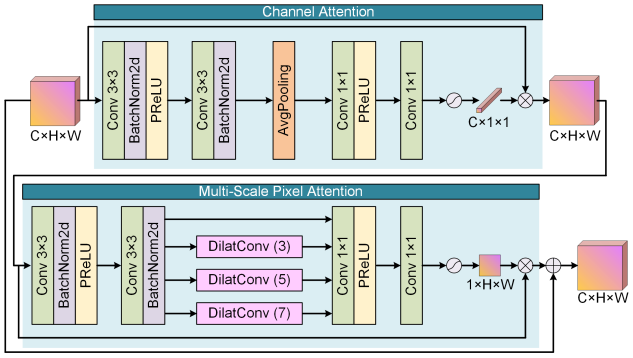

Figure 4. The pipeline of the dual attention unit (DAU). The DAU contains channel attention and multi-scale pixel attention.

will be sequentially processed by channel attention (CA) and multi-scale pixel attention (MSPA) to obtain the output feature map. The CA comprises two 3×3 convolutional layers, a global average pooling layer, two 1×1 convolutional layers, and a sigmoid function. The obtained weights of each channel by CA will be multiplied by the input feature map. The MSPA includes two 3×3 convolutional layers, three dilated convolutional layers with different dilated ratios $\in \{3, 5, 7\}$, two 1×1 convolutional layers, and a sigmoid function. To improve the perception of the spatial distribution of haze, three dilated convolutions are specially used to obtain feature information of multiple receptive fields. Finally, a 7×7 convolutional layer and a sigmoid function are employed to obtain the attention map M_g.

Scene Reconstruction Network. To improve the haze-free image reconstruction quality, an encoder-decoder network is employed. As illustrated in Figure 3, the SRN first adopts two 3×3 convolutional layers with the stride of 2 to extract $4\times$ downsampled features. Then, multiple residual blocks and two deformable convolutional layers are used to learn the hazy feature representations in the low-resolution. In particular, the deformable convolution [8] can adjust the kernel shape to focus on the features of interest by using offsets. Subsequently, two transposed convolutional layers

with a stride of 2 are used to restore the features to the original resolution. Finally, the haze-free results are produced by a tail block, which contains a reflection padding, a 7×7 convolutional layer, and a tanh function.

3.2. Self-Paced Semi-Curricular Attention

Why Supervise Attention Map. In non-homogeneous image dehazing, attention mechanisms can enable the network to flexibly focus on haze features to reconstruct high-quality haze-free images. However, attention maps are often unsupervised, which can lead to low-importance regions being assigned higher weights and generating low-quality reconstruction results. Figure 5 (b) and (e) display the attention map directly generated by AGN and the haze-free outputs generated by SRN. Obviously, the attention map has excessively high weights in the sky area, resulting in obvious block artifacts in the reconstruction result. According to our observations, non-homogeneous haze can significantly increase the luminance of occluded areas (except for the sky area). Theoretically, paying more attention to the restoration of areas with significant luminance changes can avoid the over-enhancement issue to improve the overall image reconstruction performance. Therefore, we transform the hazy and clear images into the YCbCr color space and calculate the Y channel-based luminance deviation as the ground truth of the attention map M_{GT}.

Self-Paced Semi-Curricular Learning. Note that multi-objective prediction tasks (i.e., obtaining both haze-free image and attention map) tend to increase the learning ambiguity. To make the model converge better, inspired by [11], we adopt a self-paced semi-curricular learning strategy to train the network from easy to hard. During training, the attention map M_g generated by AGN and the ground truth M_{GT} are fused to generate the final attention map M. Let λ be the trade-off parameter, M can be expressed mathematically as

$$M = \lambda \cdot M_g + (1 - \lambda) \cdot M_{GT}. \tag{2}$$

In particular, the trade-off parameter can be dynamically adjusted through the smooth L1 loss $\mathcal{L}^a_{sl1}$ of the attention map, i.e.,

$$\lambda = \begin{cases} 0, & \text{if } \mathcal{L}^a_{sl1} > 0.1, \\ \frac{\mathcal{L}^a_{sl1} - 0.1}{0.1 - 0.05}, & \text{if } 0.1 \geq \mathcal{L}^a_{sl1} > 0.05, \\ 1, & \text{if } \mathcal{L}^a_{sl1} \leq 0.05. \end{cases} \tag{3}$$

Eq. (3) is used to adjust the specific gravity of M_g and M_{GT}. In the initial stage, M mainly consists of M_{GT} to alleviate the learning ambiguity due to the large value of $\mathcal{L}^a_{sl1}$. As $\mathcal{L}^a_{sl1}$ decreases, the proportion of the attention map M_g generated by the network will continue to increase. When $\mathcal{L}^a_{sl1}$ is less than 0.05, M will only consist of M_g. Meanwhile, we only adopt the semi-curricular learning strategy in the first 25% epochs to avoid the model's over-reliance on M_{GT}.

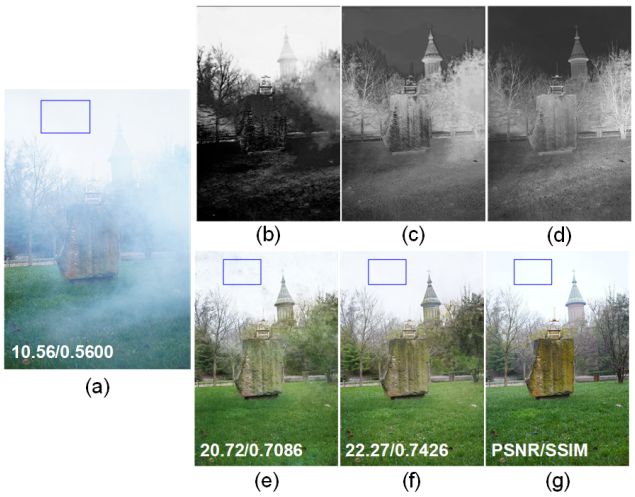

Figure 5. Visual comparisons of images generated by different strategies. From top-left to bottom-right: (a) hazy image, (b) attention map directly generated by AGN, (c) attention map generated by self-paced semi-curricular learning-driven AGN, (d) ground truth of the attention map, (e) dehazing result generated based on (b), (f) dehazing result generated based on (c), and (g) haze-free image. Note that the dehazing result (e) appears over-enhanced and exhibits noticeable artifacts, which can be attributed to the significant weight placed on the sky area by the attention map (b).

After obtaining the attention map M, we adaptively weight the feature map through a learnable parameter α. Let F_{in} be the input feature map, the feature map F_{out} weighted by the attention map can be given by

$$F_{out} = (1 - \alpha) \cdot F_{in} + \alpha \cdot M \otimes F_{in}, \tag{4}$$

with $\otimes$ being the operator of pixel-wise multiplication.

3.3. Loss function

In this section, we introduce the joint loss function of the proposed SCANet. Specifically, this joint loss function $\mathcal{L}_{\text{joint}}$ mainly consists of smooth L1 loss (including $\mathcal{L}_{sl1}$ and $\mathcal{L}^a_{sl1}$), multi-scale structural similarity (MS-SSIM) loss $\mathcal{L}_{\text{MS-SSIM}}$, perceptual loss $\mathcal{L}_{\text{p}}$, and adversarial loss $\mathcal{L}_{\text{a}}$, which can be expressed as follows

$$\mathcal{L}_{\text{joint}} = \gamma_1 \mathcal{L}_{sl1} + \gamma_2 \mathcal{L}^a_{sl1} + \gamma_3 \mathcal{L}_p + \gamma_4 \mathcal{L}_{\text{MS-SSIM}} + \gamma_5 \mathcal{L}_a, \tag{5}$$

where γ_1, γ_2, γ_3, γ_4, and γ_5 are the hyper-parameters. The best performance is achieved when we assign them the values of 1, 0.3, 0.01, 0.5, and 0.0005, respectively.

Smooth L1 Loss. In the image restoration task, Zhao *et al.* [40] have demonstrated that the L1 loss function has better effects compared with L2 loss. Therefore, we use smooth L1 loss [15] to supervise the final output $\hat{J}$ and the

predicted attention map M_g, which can be expressed as follows

$$\mathcal{L}_{sl1} = L_1(\hat{J} - J), \tag{6}$$

$$\mathcal{L}^a_{sl1} = L_1(M_g - M_{GT}), \tag{7}$$

where $L_1(\cdot)$ represents the smooth L1 loss function, $\mathcal{L}_{sl1}$ is the loss between the network's output $\hat{J}$ and the ground truth J, $\mathcal{L}^a_{sl1}$ is the loss between the predicted attention map M_g and the ground truth of attention map M_{GT}.

Let Q denotes the input, the L_1 operation can be expressed as follows

$$L_1(Q) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{D}_{l1}(Q(i)), \tag{8}$$

where i is the index pixel, N denotes the sum of pixels. Finally, the smooth L1 operator $\mathcal{D}_{l1}$ can be given by

$$\mathcal{D}_{l1}(Q(i)) = \begin{cases} 0.5 \cdot Q^2(i), & \text{if } |Q(i)| < 1, \\ |Q(i)| - 0.5, & \text{otherwise.} \end{cases} \tag{9}$$

Perceptual Loss. To improve the similarity between the output and ground truth in feature space, we add the perceptual loss $\mathcal{L}_{\text{p}}$, which can be written as follows

$$\mathcal{L}_{\text{p}} = \frac{1}{3}\sum_{r}\frac{\left\|\phi^v_J\left(\hat{J}\right) - \phi^v_k(J)\right\|^2_2}{C_k H_k W_k}, \tag{10}$$

where $\phi^v_k(\cdot)$ represents the feature map of VGG16 in k-layer, and (C_k, H_k, W_k) denotes the shape of the feature map in the corresponding layer. In this paper, $r \in \{$ relu1_2, relu2_2, relu3_3$\}$.

MS-SSIM Loss. To improve the contrast of high-frequency regions in the image, we adopt MS-SSIM loss $\mathcal{L}_{\text{MS-SSIM}}$, which can be defined as follows

$$\mathcal{L}_{\text{MS-SSIM}} = L_{\text{MS-SSIM}}(J, \hat{J}), \tag{11}$$

where $L_{\text{MS-SSIM}}(\cdot)$ represents the multi-scale structure similarity function. The SSIM value can be written as follows

$$\begin{aligned} \text{SSIM}(x) &= \frac{2\mu_J\mu_{\hat{J}} + c}{\mu^2_J + \mu^2_{\hat{J}} + c} \cdot \frac{2\sigma_{J\hat{J}} + c_*}{\sigma^2_J + \sigma^2_{\hat{J}} + c_*} \\ &= l(x) \cdot cs(x), \end{aligned} \tag{12}$$

where x demotes the pixel index, c and c_* are two constants to avoid the denominator becoming zero. The means μ_J , $\mu_{\hat{J}}$, standard deviations σ_J, $\sigma_{\hat{J}}$, and covariance $\sigma_{J\hat{J}}$ are computed by a Gaussian filter. Finally, the operation of MS-SSIM can be defined as follows

$$L_{\text{MS-SSIM}} = 1 - l^{\alpha}_{\mathcal{P}} \cdot \prod_{j=1}^{\mathcal{P}}\left[cs_j\right]^{\beta_j}, \tag{13}$$

Table 1. The details of the datasets used in our experiments. (w/o GT) represents the lack of public ground truth for this set.

Datasets	Train	Validation	Test	Image Size
NTIRE2020	45	5	5	1200 × 1600
NTIRE2021	25	5 (w/o GT)	5 (w/o GT)	1200 × 1600
NTIRE2023	40	5 (w/o GT)	5 (w/o GT)	4000 × 6000

where $\mathcal{P}$ denotes the default parameter of scales.

Adversarial Loss. To improve the generalization ability of the proposed network, we add the additional adversarial loss, i.e.,

$$\mathcal{L}_a = -\frac{1}{S}\sum_{n=1}^{S}\log\left(D\left(J - \hat{J}\right)\right), \tag{14}$$

where $D(\cdot)$ represents the discriminator, S represents the number of training data.

4. Experiments

In this section, we first describe the datasets, implementation details, evaluation metrics, and competitors. Then, we compare the proposed SCANet with other state-of-the-art dehazing methods. Finally, we conduct the ablation study to demonstrate the rationality of each module in the proposed SCANet.

4.1. Experiment Settings

Datasets. We choose NTIRE2020 [1, 2], NTIRE2021 [3], and NTIRE2023 [4] datasets to train and evaluate the proposed SCANet. The haze patterns in all three datasets are non-uniformly distributed. Specifically, NTIRE2020 dataset (termed NH-Haze) contains 45 training, 5 validation, and 5 test image pairs. NTIRE2021 dataset (termed NH-Haze2) contains 25 training image pairs, 5 validation hazy images, and 5 test hazy images. NTIRE2023 dataset contains 40 training image pairs, 5 validation hazy images, and 5 test hazy images. Note that only the validation and test sets of NTIRE2020 dataset contain the corresponding ground truth. More details about these datasets can be found in Table 1.

Implementation Details. The proposed SCANet is implemented by PyTorch 1.9.1 and trained on a PC with an Intel(R) Core(TM) i9-13900K CPU @5.80GHz and Nvidia GeForce RTX 3080 GPU. We use the Adam with exponential decay rates being $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for optimization. The initial learning rate and batchsize are set to 0.0001 and 2, respectively. During the training stage, we resize the images into 0.5, 0.7, and 1 scales and randomly crop them to several image patches of size 512 × 512 with a stride of 400. Meanwhile, these image patches are randomly flipped 0, 90, 180, and 270 degrees. In addition, we train two models for NTIRE2023 validation and

Table 2. Quantitative comparisons for non-homogeneous dehazing on NHIRE2020, NHIRE2021, and NHIRE2023 datasets. The best results are in **bold**, and the second best are with underline.

Methods	NTIRE2020		NTIRE2021		NTIRE2023		Average	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
Hazy	11.31	0.4160	11.24	0.5787	8.86	0.4702	10.47	0.4883
(TPAMI'10) DCP [17]	12.35	0.4480	10.57	0.6030	10.98	0.4777	11.30	0.5096
(ICCV'17) AODNet [22]	14.04	0.4450	14.52	0.6740	13.75	0.5619	14.10	0.5603
(ICCV'19) GridDehazeNet [26]	14.78	0.5074	18.05	0.7433	16.85	0.6075	16.56	0.6194
(AAAI'20) FFANet [28]	16.98	0.6105	19.75	0.7925	17.85	0.6485	18.20	0.6838
(CVPRW'21) TNN [37]	17.18	0.6114	20.13	**0.8019**	18.19	0.6426	18.50	0.6853
(CVPR'22) DeHamer [16]	18.53	0.6201	18.17	0.7677	17.61	0.6051	18.10	0.6693
SCANet	**19.52**	**0.6488**	**21.14**	0.7694	**20.44**	**0.6616**	**20.37**	**0.6933**

Table 3. FLOPs and Parameters comparisons of all methods.

Methods	FLOPs	Parameters
(ICCV'17) AODNet [22]	1.68G	1.76K
(ICCV'19) GridDehazeNet [26]	271.95G	702.47K
(AAAI'20) FFANet [28]	4211.91G	4.46M
(CVPRW'21) TNN [37]	1235.84G	50.35M
(CVPR'22) DeHamer [16]	866.96G	29.44M
SCANet	258.63G	2.39M

test sets and NTIRE2020/2021/2023 datasets, respectively. For NTIRE2023 validation and test sets, we only use 35 training pairs in NTIRE2023 for training. The epoch is set to 85, and the learning rate decays by 0.5 every 20 epochs. Due to the large size of the test images, we adopt the Nvidia A100 GPU for testing. For NTIRE2020, NTIRE2021, and NTIRE2023 datasets, we select 45 training pairs and 5 validation pairs in NTIRE2020, the first 20 training pairs in NTIRE2021, and the first 35 training pairs in NTIRE2023 as the train set. The test set is composed of 5 test pairs in NTIRE2020, the last 5 training pairs in NTIRE2021, and the last 5 training pairs in NTIRE2023. In this experiment, the images of NTIRE2023 are compressed to 1/4 (i.e., 1000 $\times$ 1500) to ensure a similar size with other datasets. In addition, the epoch is set to 500, and the learning rate decays by 0.5 every 150 epochs.

Evaluation Metrics and Competitors. To conduct an exhaustive analysis of the dehazing performance, we employ the peak signal-to-noise ratio (PSNR) [33] and structural similarity index (SSIM) [34] to quantitatively evaluate the restored images. Meanwhile, we compare the proposed SCANet with the state-of-the-art methods, including a prior-based method (i.e., DCP [17]), a physical model-based CNN method (i.e., AODNet [22]), three hazy-to-clear CNN methods (i.e., GridDehazeNet [26], FFANet [28], and TNN [37]), and a CNN-Transoformer combined method (i.e., DeHamer [16]).

4.2. Comparisons with the State-of-the-Arts

Results on NTIRE2020/2021/2023. Table 2 presents the PSNR and SSIM results of various dehazing methods on NTIRE2020, NTIRE2021, and NTIRE2023 datasets. Prior knowledge of DCP fails in the non-homogeneous dehazing task, resulting in relatively low values of PSNR and SSIM. Learning-based methods show better adaptability in generating haze-free images, with a significant boost in metrics. Among these methods, the proposed SCANet achieves satisfactory performance, ranking first in most cases. We also show the visual comparisons in Figure 6. The results generated by DCP have the issue of serious color distortion. AODNet, GridDehazeNet, and FFANet fail to remove haze completely. The performance of TNN on the NTIRE2023 benchmark falls short of expectations. Specifically, five images lack sufficient color saturation, while the first image exhibits overly darkened ground. DeHamer is effective in haze suppression. However, color restoration and detail preservation abilities still need improvement. Compared with other methods, the proposed SCANet exhibits superior visual performance.

Results on NTIRE2023 Validation and Test Sets. According to our submission on the NTIRE2023 website, our SCANet can achieve PSNR 21.13dB and SSIM 0.6907 on the validation set and PSNR 21.75dB and SSIM 0.6955 on the test set. Meanwhile, the visual comparison of our method and the state-of-the-art on 5 validation and 5 test images are shown in Figure 7. It can be observed that DCP, AODNet, and GridDehazeNet perform poorly in non-homogeneous image dehazing. Although FFANet, TNN, and DeHamer can partially remove the haze, residues still exist in dense haze regions. Compared to existing methods, the proposed SCANet have a more natural performance. However, our method still cannot fully restore the color and details of high haze concentration areas.

Complexity Analysis. Table 3 shows the number of network parameters and floating point operations (FLOPs) on 1200 $\times$ 1600 images of the proposed method and other

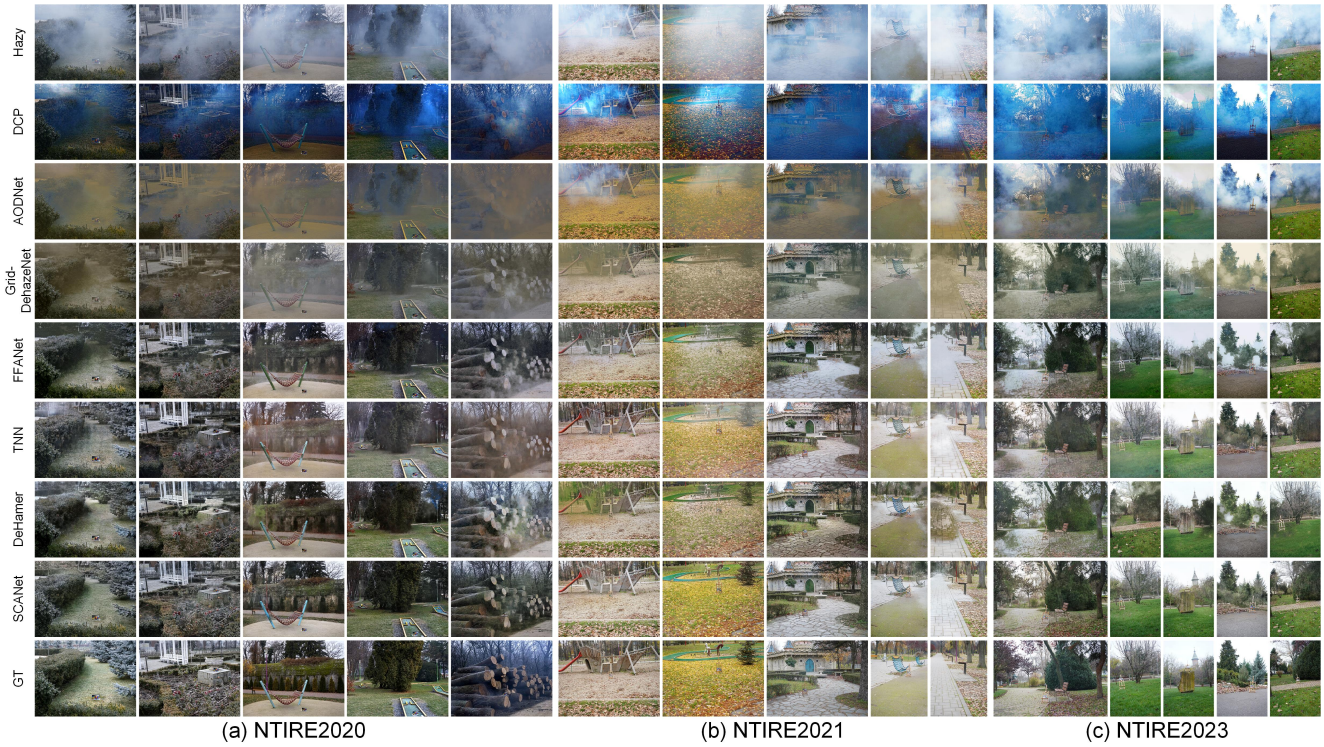

Figure 6. Visual comparisons of various methods on NTIRE2020, NTIRE2021, and NTIRE2023 datasets.

comparable methods. By comparison, our SCANet has lower FLOPs and fewer network parameters. To visually demonstrate the superiority of our method, we compare the PSNR and parameter amount of different methods in Figure 2. It is worth mentioning that our time complexity is also relatively modest. It takes an average of 0.1962 seconds to process a 1200 $\times$ 1600 image on the NVIDIA GeForce RTX 3080 GPU.

4.3. Ablation Analysis

We conduct a series of experiments as an ablation study to demonstrate the effectiveness of different components, including attention generation network (AGN), scene reconstruction network (SRN), self-paced semi-curricular learning strategy (SCL), and each loss function. As shown in Table 4, we design seven models with different configurations and employ NTIRE2020, NTIRE2021, and NTIRE2023 datasets as both training and test sets.

The quantitative results are presented in Table 4. By comparing Model (1) and (2), our method achieves performance improvement after adding the attention generator network (AGN) before the scene reconstruction network (SRN). This result demonstrates that unlike homogeneous image dehazing, restoring non-homogeneous images requires the network to be more sensitive to the haze regions. Moreover, we use $\mathcal{L}_{sl1}^{a}$ to supervise the attention feature map, resulting in satisfactory improvement in both PSNR and SSIM by observing Model (2) and (3). The supervision of the attention map avoids assigning higher weights to low-importance regions, which can provide better reconstruction results. Additionally, applying self-paced semi-curricular learning (SCL) during training leads to the further improvement of metrics, which indicates that SCL can reduce the network's convergence difficulty and improve its performance. By comparing the examples shown in Figure 5, we can find the change from Model (2) to Model (4) more intuitively. Obviously, our SCL strategy for attention map constraint can make the SRN more fully focus on the regions with significant luminance changes and avoid the distortion issue in the sky region. Furthermore, the usage of MS-SSIM loss, perceptual loss, and generative adversarial loss can further enhance the dehazing performance of our SCANet by comparing Model (5), (6), and (7) in Table 4.

5. Conclusion

In this paper, we provided a robust solution (termed SCANet) for non-homogeneous image dehazing by effectively extracting non-uniform haze distribution features and reconstructing the details with high quality. Our attention generator network and scene reconstruction network work together in a novel "attention generation-scene reconstruction" paradigm. Moreover, we proposed a self-paced semi-

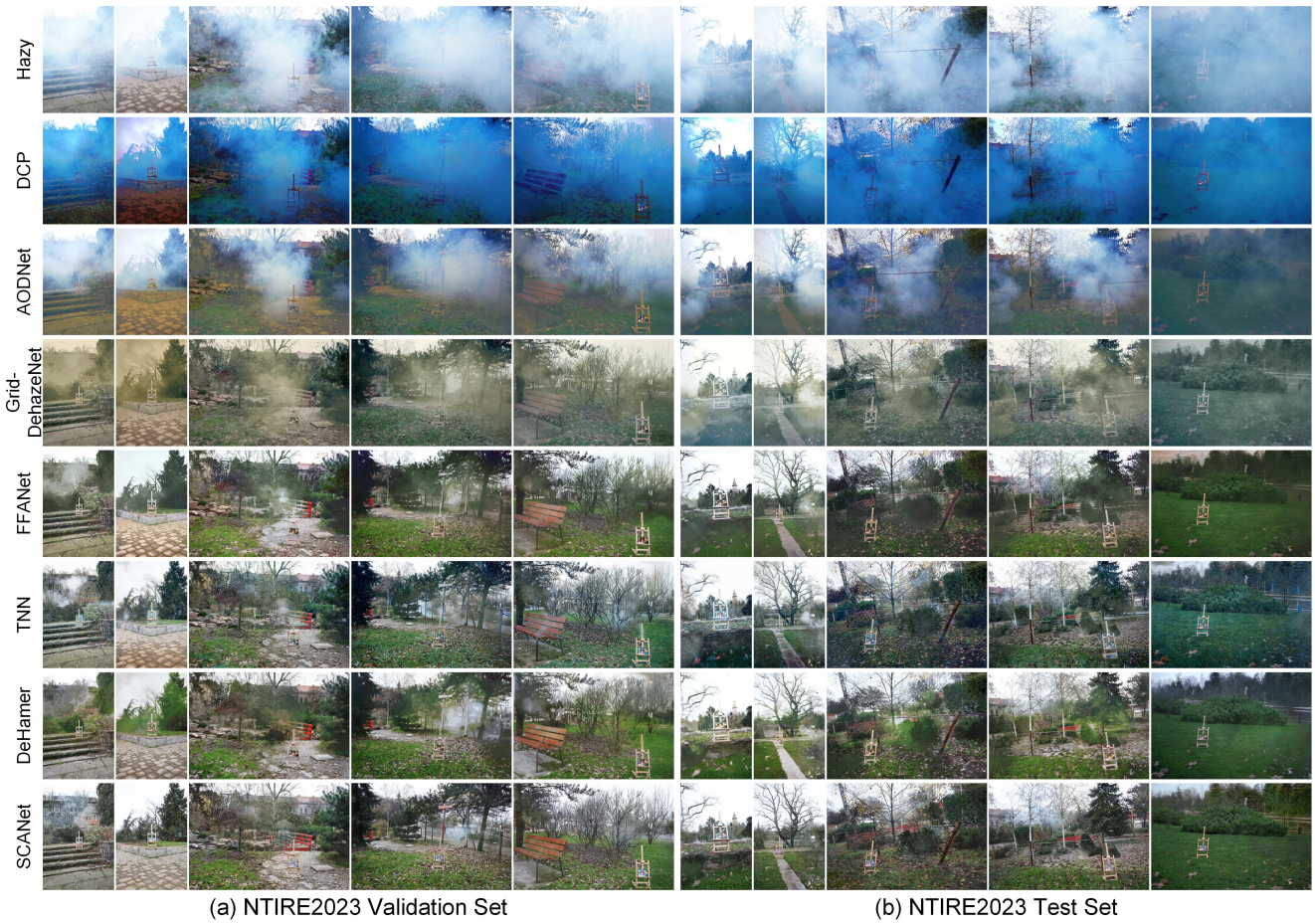

Figure 7. Visual comparisons of various methods on NTIRE2023 validation set ($\#41 \sim 45$) and test set ($\#46 \sim 50$).

Table 4. The ablation study of different configurations. The best results are in **bold**, and the second best are with underline.

Number	Methods	$\mathcal{L}_{sl1}^{f}$	$\mathcal{L}_{sl1}^{a}$	$\mathcal{L}_{p}$	$\mathcal{L}_{\text{MS-SSIM}}$	$\mathcal{L}_{a}$	PSNR ↑	SSIM ↑
(1)	SRN	✔					18.84	0.6634
(2)	AGN + SRN	✔					19.29	0.6714
(3)	SRN + AGN	✔	✔				19.71	0.6787
(4)	SRN + AGN + SCL	✔	✔				19.92	0.6881
(5)	SRN + AGN + SCL	✔	✔	✔			19.85	0.6890
(6)	SRN + AGN + SCL	✔	✔	✔	✔		20.02	**0.6957**
(7)	SRN + AGN + SCL	✔	✔	✔	✔	✔	**20.37**	0.6933

curricular learning-driven attention map generation strategy to improve the model convergence and reduce learning ambiguity during the early stage of training. Our proposed method outperforms many state-of-the-art methods in both quantitative and qualitative experiments, demonstrating the effectiveness of our approach. Additionally, ablation analysis confirms the contribution of each component in the overall performance of our SCANet. We believe that the proposed method can provide a promising solution for the applications of real-world non-homogeneous image dehazing. Future work can extend our method to handle more complex scenarios. Examples include handling multiple types of haze and integrating our method with other computer vision tasks.

6. Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.: 52271365). The authors would like to thank the three anonymous reviewers for their professional comments and constructive suggestions.

References

[1] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proc. IEEE CVPRW*, pages 444–445, 2020. 5

[2] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. NTIRE 2020 challenge on nonhomogeneous dehazing. In *Proc. IEEE CVPRW*, pages 490–491, 2020. 5

[3] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. NTIRE 2021 nonhomogeneous dehazing challenge report. In *Proc. IEEE CVPRW*, pages 627–646, 2021. 5

[4] Codruta O Ancuti, Cosmin Ancuti, Florin-Alexandru Vasluianu, and Radu Timofte. Ntire 2023 challenge on nonhomogeneous dehazing. In *Proc. IEEE CVPRW*, 2023. 5

[5] Dana Berman, Shai Avidan, et al. Non-local image dehazing. In *Proc. IEEE CVPR*, pages 1674–1682, 2016. 1, 2

[6] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. Dehazenet: An end-to-end system for single image haze removal. *IEEE Trans. Comput. Imaging*, 25(11):5187–5198, 2016. 1, 2

[7] 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 *Proc. IEEE WACV*, pages 1375–1383, 2019. 1, 3

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proc. IEEE ICCV*, pages 764–773, 2017. 3

[9] Qili Deng, Ziling Huang, Chung-Chi Tsai, and Chia-Wen Lin. Hardgan: A haze-aware representation distillation gan for single image dehazing. In *Proc. ECCV*, pages 722–738, 2020. 3

[10] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. Multi-scale boosted dehazing network with dense feature fusion. In *Proc. IEEE CVPR*, pages 2157–2167, 2020. 1

[11] Yong Du, Junjie Deng, Yulong Zheng, Junyu Dong, and Shengfeng He. DSDNet: Toward single image deraining with self-paced curricular dual stimulations. *Comput. Vision Image Understanding*, page 103657, 2023. 4

[12] Akshay Dudhane and Subrahmanyam Murala. Cdnet: Single image de-hazing using unpaired adversarial training. In *Proc. IEEE WACV*, pages 1147–1155, 2019. 1

[13] Raanan Fattal. Dehazing using color-lines. *ACM Trans. Graphics*, 34(1):1–14, 2014. 1

[14] Minghan Fu, Huan Liu, Yankun Yu, Jun Chen, and Keyan Wang. Dw-gan: A discrete wavelet transform gan for non-homogeneous dehazing. In *Proc. IEEE CVPRW*, pages 203–212, 2021. 1

[15] Ross Girshick. Fast r-cnn. In *Proc. IEEE ICCV*, pages 1440–1448, 2015. 4

[16] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *Proc. IEEE CVPR*, pages 5812–5820, 2022. 6

[17] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(12):2341–2353, 2010. 1, 2, 6

[18] Ming Hong, Jianzhuang Liu, Cuihua Li, and Yanyun Qu. Uncertainty-driven dehazing network. In *Proc. AAAI*, pages 906–913, 2022. 3

[19] Ming Hong, Yuan Xie, Cuihua Li, and Yanyun Qu. Distilling image dehazing with heterogeneous task imitation. In *Proc. IEEE CVPR*, pages 3462–3471, 2020. 1

[20] Shih-Chia Huang, Bo-Hao Chen, and Yi-Jui Cheng. An efficient visibility enhancement algorithm for road scenes captured by intelligent transportation systems. *IEEE Trans. Intell. Transp. Syst.*, 15(5):2321–2332, 2014. 1

[21] Eunsung Jo and Jae-Young Sim. Multi-scale selective residual learning for non-homogeneous dehazing. In *Proc. IEEE CVPRW*, pages 507–515, 2021. 2, 3

[22] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proc. IEEE ICCV*, pages 4770–4778, 2017. 1, 2, 6

[23] Runde Li, Jinshan Pan, Zechao Li, and Jinhui Tang. Single image dehazing via conditional generative adversarial network. In *Proc. IEEE CVPR*, pages 8202–8211, 2018. 1

[24] Jing Liu, Haiyan Wu, Yuan Xie, Yanyun Qu, and Lizhuang Ma. Trident dehazing network. In *Proc. IEEE CVPRW*, pages 430–431, 2020. 1

[25] Ryan Wen Liu, Yu Guo, Yuxu Lu, Kwok Tai Chui, and Brij B Gupta. Deep network-enabled haze visibility enhancement for visual iot-driven intelligent transportation systems. *IEEE Trans. Ind. Inf.*, 19(2):1581–1591, 2022. 1

[26] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proc. IEEE ICCV*, pages 7314–7323, 2019. 3, 6

[27] Aditya Mehta, Harsh Sinha, Pratik Narang, and Murari Mandal. Hidegan: A hyperspectral-guided image dehazing gan. In *Proc. IEEE CVPRW*, pages 212–213, 2020. 1

[28] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. FFA-Net: Feature fusion attention network for single image dehazing. In *Proc. AAAI*, pages 11908–11915, 2020. 1, 3, 6

[29] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. Enhanced pix2pix dehazing network. In *Proc. IEEE CVPR*, pages 8160–8168, 2019. 2

[30] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. Gated fusion network for single image dehazing. In *Proc. IEEE CVPR*, pages 3253–3261, 2018. 2

[31] Lithesh Shetty et al. Non homogeneous realistic single image dehazing. In *Proc. IEEE WACV*, pages 548–555, 2023. 2, 3

[32] Qiaoling Shu, Chuansheng Wu, Zhe Xiao, and Ryan Wen Liu. Variational regularized transmission refinement for image dehazing. In *Proc. IEEE ICIP*, pages 2781–2785, 2019. 1

[33] Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Process Mag.*, 26(1):98–117, 2009. 6

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 6

[35] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proc. IEEE CVPR*, pages 10551–10560, 2021. 3

[36] Jinrong Yang, Songtao Liu, Zeming Li, Xiaoping Li, and Jian Sun. Real-time object detection for streaming perception. In *Proc. IEEE CVPR*, pages 5385–5395, 2022. 1

[37] Yankun Yu, Huan Liu, Minghan Fu, Jun Chen, Xiyao Wang, and Keyan Wang. A two-branch neural network for non-homogeneous dehazing via ensemble learning. In *Proc. IEEE CVPRW*, pages 193–202, 2021. 2, 3, 6

[38] He Zhang and Vishal M Patel. Densely connected pyramid dehazing network. In *Proc. IEEE CVPR*, pages 3194–3203, 2018. 2

[39] Jing Zhang and Dacheng Tao. FAMED-Net: A fast and accurate multi-scale end-to-end dehazing network. *IEEE Trans. Image Process.*, 29:72–84, 2019. 1

[40] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Trans. Comput. Imaging*, 3(1):47–57, 2016. 4

[41] Yu Zheng, Jiahui Zhan, Shengfeng He, Junyu Dong, and Yong Du. Curricular contrastive regularization for physics-aware single image dehazing. *arXiv preprint arXiv:2303.14218*, 2023. 2, 3

[42] Wujie Zhou, Shaohua Dong, Jingsheng Lei, and Lu Yu. MTANet: Multitask-aware network with hierarchical multimodal fusion for rgb-t urban scene understanding. *IEEE Trans. Intell. Veh.*, 8(1):48–58, 2022. 1

[43] Qingsong Zhu, Jiaming Mai, and Ling Shao. Single image dehazing using color attenuation prior. In *BMVC*, pages 1–10, 2014. 2

[44] Qingsong Zhu, Jiaming Mai, and Ling Shao. A fast single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Process.*, 24(11):3522–3533, 2015. 1

[45] Yingying Zhu, Gaoyang Tang, Xiaoyan Zhang, Jianmin Jiang, and Qi Tian. Haze removal method for natural restoration of images with sky. *Neurocomputing*, 275:499–510, 2018. 1