Depth Information Assisted Collaborative Mutual Promotion Network for Single Image Dehazing

Yafei Zhang[*]   Shen Zhou[*]   Huafeng Li[†]

Faculty of Information Engineering and Automation, Kunming University of Science and Technology

Abstract

Recovering a clear image from a single hazy image is an open inverse problem. Although significant research progress has been made, most existing methods ignore the effect that downstream tasks play in promoting upstream dehazing. From the perspective of the haze generation mechanism, there is a potential relationship between the depth information of the scene and the hazy image. Based on this, we propose a dual-task collaborative mutual promotion framework to achieve the dehazing of a single image. This framework integrates depth estimation and dehazing by a dual-task interaction mechanism and achieves mutual enhancement of their performance. To realize the joint optimization of the two tasks, an alternative implementation mechanism with the difference perception is developed. On the one hand, the difference perception between the depth maps of the dehazing result and the ideal image is proposed to promote the dehazing network to pay attention to the non-ideal areas of the dehazing. On the other hand, by improving the depth estimation performance in the difficult-to-recover areas of the hazy image, the dehazing network can explicitly use the depth information of the hazy image to assist the clear image recovery. To promote the depth estimation, we propose to use the difference between the dehazed image and the ground truth to guide the depth estimation network to focus on the dehazed unideal areas. It allows dehazing and depth estimation to leverage their strengths in a mutually reinforcing manner. Experimental results show that the proposed method can achieve better performance than that of the state-of-the-art approaches. The source code is released at https://github.com/zhoushen1/DCMPNet.

1. Introduction

Single image dehazing refers to restoring a clear image from a given hazy image. This technology has attracted wide attention due to its critical role in downstream com-
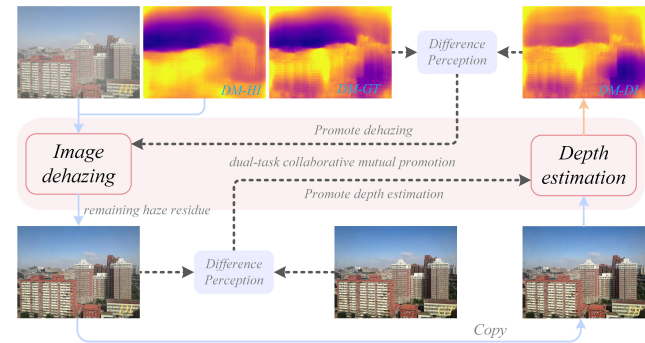
---

[*]Yafei Zhang and Shen Zhou contribute equally to this work.
[†]Huafeng Li (lhfchina99@kust.edu.cn) is the corresponding author.

Figure 1. Idea of dual task collaboration and mutual promotion. HI denotes the hazy image. *DM-HI*, *DM-GT* and *DM-DI* denote the depth maps of hazy image, ground truth and dehazed image, respectively.

puter vision tasks [12, 14, 19, 38, 46]. Mathematically, the atomization process is commonly simulated by employing the atmospheric scattering model [25, 27, 28]:

$$\boldsymbol{I}(x) = \boldsymbol{J}(x)\boldsymbol{T}(x) + (1 - \boldsymbol{T}(x))\boldsymbol{A}(x) \qquad (1)$$

where x represents the pixel position, $\boldsymbol{I}(x)$ denotes the hazy image, $\boldsymbol{J}(x)$ denotes the clear image, $\boldsymbol{T}(x)$ indicates the transmission map, and $\boldsymbol{A}(x)$ represents the atmospheric light. The transmission map $\boldsymbol{T}(x)$ can also be expressed as:

$$\boldsymbol{T}(x) = e^{-\beta \boldsymbol{d}(x)} \qquad (2)$$

where β is the scattering coefficient, and $\boldsymbol{d}(x)$ is the scene depth.

From Eq.(1), we can see that image dehazing methods based on the atmospheric scattering model necessitate the estimation of both $\boldsymbol{T}(x)$ and $\boldsymbol{A}(x)$. However, most studies claim that $\boldsymbol{T}(x)$ primarily contributes to generating haze within images. Consequently, the most existing Eq.(1)-based methods focus on estimating the transmission map $\boldsymbol{T}(x)$. For $\boldsymbol{A}(x)$, the maximum pixel value is taken as its value [39]. In practice, the maximum pixel value in an image may originate from the brightest object rather than representing atmospheric light. To address this issue, the methods in [2, 17, 37] consider the estimation of both $\boldsymbol{T}(x)$ and

$\boldsymbol{A}(x)$, achieving an improvement in dehazing performance. However, such atmospheric scattering model-based methods significantly depend on $\boldsymbol{T}(x)$ and $\boldsymbol{A}(x)$. In reality, the uneven distribution of haze poses a challenge in accurately estimating $\boldsymbol{T}(x)$ in each local region of the image, which limits the improvement of dehazing performance.

End-to-end image dehazing methods recover clear images directly from hazy images without the aid of atmospheric scattering models [6, 8, 22, 30, 44], thereby getting rid of the dependence on $\boldsymbol{T}(x)$ and $\boldsymbol{A}(x)$. However, the absence of guidance from the atmospheric scattering model also poses challenges for restoring clear images. To solve this problem, image dehazing based on prior information was proposed [1, 4, 24, 45]. Specifically, PSD [4] establishes a loss committee consisting of dark channel prior [10], bright channel prior [50] and histogram equalization to guide the restoration of clear images. NHFormer [24] uses dark channel priors and bright channel priors to guide the generalization of the dehazing model from synthetic domains to real-world applications. RIDCP [45] obtains high-quality codebook priors through pre-trained VQ-GAN. These priors are used to facilitate controllable high-quality priors matching, thereby achieving feature recovery for hazy images. In addition, through the two-stage network structure design, MITNet [34] simultaneously realizes the joint recovery of features in both the time and frequency domains and achieves the restoration of both amplitude and phase spectrum, ensuring the quality of dehazed images.

Although the above methods are effective, they ignore the correlation between the depth information of the hazy image and itself. According to Eq.(2), the depth information $\boldsymbol{d}(x)$ of the scene is directly related to $\boldsymbol{T}(x)$, while $\boldsymbol{T}(x)$ is one of the key factors that lead to the appearance of haze in image. Suppose $\boldsymbol{d}(x)$ can be accurately estimated from its hazy image, and the calculated result can be used to reconstruct the hazy-free image. In that case, it will help improve the performance of the dehazing model. Based on this idea, this paper proposes a dual-task collaborative mutual promotion network for single image dehazing within the framework of end-to-end deep learning. In the proposed method, we embed depth estimation and image dehazing within a unified framework. To effectively use the depth information of the hazy image for the dehazing model, a difference perception-based dual-task interaction mechanism is designed to serve as a bridge between the two tasks. This mechanism seamlessly integrates depth estimation and image dehazing to form a dual-task-driven dehazing method.

The main idea of the proposed approach is shown in Figure 1. The proposed method improves the depth estimation on hazy images by perceiving the difference between the output results of the dehazing network and the expected results so that the dehazing network can receive high-quality depth estimation information as guidance in the dehazing process. Moreover, the design of the dual-task collaborative mutual promotion framework will be conducive to the dehazing network to learn optimal network parameters. In terms of depth estimation improving the dehazing performance, it makes the dehazing network pay attention to areas where the dehazing effect could be unsatisfactory by perceiving the depth information difference between the dehazing result and the ideal image. In terms of the dehazing network promoting depth estimation, the robustness of the depth estimation is improved by making the depth estimation network pay attention to the non-ideal dehazing areas and obtaining more accurate prediction results on the hazy image. We have verified the effectiveness of the proposed method on three dehazing benchmarks and achieved compatible results of the state-of-the-art.

2. Related Work

2.1. Model-based Image Dehazing

Single image dehazing based on the atmospheric scattering model is a popular method. In this kind of method, the accuracy of the transmission map and atmospheric light estimation are pivotal factors affecting the quality of the dehazing results. To this end, a large number of dehazing methods based on transmission map estimation have emerged [3, 29, 32, 42]. These methods ignore the influence of atmospheric light on the dehazing results. To solve this problem, a joint estimation model for transmission map and atmospheric light is proposed. Particularly, Zhang *et al*. [49] proposed a joint discriminator based on GAN to assess the accuracy of the dehazed image and the estimated transmission map. Additionally, they employed a U-Net to predict the atmospheric light. Li *et al*. [18] proposed a multi-stage progressive learning approach to estimate the transmission map. Additionally, they used a global average pooling layer to normalize the features from different layers and superimpose them together, and thus achieved estimating of the atmospheric light. Guo *et al*. [9] used a shared DensetNet encoder and two non-shared DensetNet decoders to estimate the transmission map and atmospheric light jointly. Lee *et al*. [15] proposed a feature extraction network with a joint constraint of transmission map, atmospheric light and dehazing result. The network was used to predict more comprehensive information in the transmission map and atmospheric light to improve the quality of dehazing results.

Due to the absence of prior information, the dehazing results from the above methods still need improvement. To this end, prior knowledge was introduced to estimate transmission maps and atmospheric light [7, 21, 23, 51, 53]. Commonly used prior knowledge includes dark channel prior [10], bright channel prior [50] and color attenuation prior [54], etc. Although the above methods have proven ef-

fective, their performance depends on the estimation accuracy of the transmission map and atmospheric light. Compared with the physical model-driven approaches, end-to-end dehazing methods based on deep learning have attracted researchers' attention because they eliminate the constraints of transmission maps and atmospheric light.

2.2. End-to-End Image Dehazing

End-to-end dehazing methods usually restore a clear image directly from the hazy image. To improve the visual quality of dehazing results, Ren *et al.* [33] used white balance, image contrast enhancement and gamma transformation to preprocess the original image. They then determined the weights of the three processed images using a gating mechanism, ultimately achieving the restoration of hazy images. Wu *et al.* [44] proposed a regularization method based on contrast learning, which achieved the recovery of a clean image by pushing away the distance between the dehazing result and the original image and simultaneously narrowing the distance between the dehazing result and the label image. Given its proficiency in highlighting useful information, attention-based techniques have been widely used in image dehazing [22, 30]. Moreover, Guo *et al.* [8] proposed an image dehazing network combining CNN and Transformer, which leveraged the complementary characteristics of CNN and Transformer in feature extraction. Due to the high computational complexity of softmax-attention in transformer, Qiu *et al.* [31] proposed to approximate it with Taylor expansion, effectively reducing the complexity of traditional attention mechanisms in image dehazing. The above methods only consider the information recovery from the time domain, ignoring the influence of hazy information on the frequency domain. Therefore, Shen *et al.* [34] proposed to restore hazy-free image information in both the time and frequency domain.

Although the end-to-end dehazing methods do not rely on the physical model of the hazy image, they exhibit a high dependency on the training dataset. If we find a more effective way to leverage the information in the training samples, it will help improve the model performance. The distribution of haze has a relationship with the scene depth, Yang *et al.* [47] proposed a depth perception method to estimate the depth map of the hazy image and provide depth features for dehazing in a unified framework. Considering the challenge posed by the absence of ground truth of hazy images in real scenarios, Yang *et al.* [48] proposed a paired sample construction method for synthesizing authentic hazy images from clear haze-free images, thereby realizing the supervised training of the dehazing model. Wang *et al.* [43] proposed a self-improving depth-consistent dehazing network according to the differences in depth features between hazy image and its ground truth. Nevertheless, the method fails to achieve collaborative mutual promotion between depth
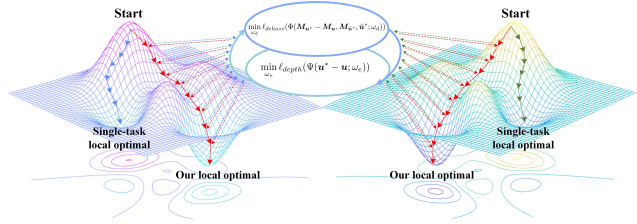

Figure 2. Dual-task Collaborative optimization formulation for image dehazing and depth estimation.

estimation and image dehazing, which limits the model performance. In this work, we fully consider the beneficial impact of depth information in hazy images on dehazing and propose to treat depth estimation and image dehazing as two independent tasks. The two tasks are seamlessly integrated into a unified learning framework through a dual-task interaction mechanism, which realizes collaborative mutual promotion.

3. The Proposed Method

3.1. Idea formulation

In this paper, we propose a single image dehazing method assisted by depth information. Unlike existing image dehazing methods that rely on elaborately designed algorithms to yield satisfactory results, our method improves the dehazing performance by leveraging auxiliary task sensitive to the quality of dehazing results. Assuming that the hazy image is $\tilde{\boldsymbol{u}}^*$, the dehazed image is $\boldsymbol{u}^*$, and the ground truth (GT) is $\boldsymbol{u}$, the idea of depth estimation based on the dehazed image can be formulated as follows:

$$\min_{\omega_e} \ell_{depth}(\Psi(\boldsymbol{u}^* - \boldsymbol{u}; \omega_e)), \quad (3)$$

where Ψ is the depth estimation network with parameter ω_e, ℓ_{depth} is specific depth estimation loss. As demonstrated in Eq.(3), the difference between $\boldsymbol{u}^*$ and $\boldsymbol{u}$ is used to optimize the parameter ω_e. This enables Ψ to improve its performance on $\boldsymbol{u}^*$ by focusing on the differences between $\boldsymbol{u}^*$ and $\boldsymbol{u}$. Since one of the contributors to the difference between $\boldsymbol{u}^*$ and $\boldsymbol{u}$ stems from the presence of residual haze information and regions where the dehazing effect is non-ideal. If the depth estimation network can improve its performance in these regions, the dehazing result can be also improved in the corresponding regions of $\tilde{\boldsymbol{u}}^*$.

For optimizing parameter ω_d of Φ, this paper employs the following optimization objective:

$$\min_{\omega_d} \ell_{dehaze}(\Phi(\boldsymbol{M}_{\boldsymbol{u}^*} - \boldsymbol{M}_{\boldsymbol{u}}, \boldsymbol{M}_{\tilde{\boldsymbol{u}}^*}, \tilde{\boldsymbol{u}}^*; \omega_d)), \quad (4)$$

where ℓ_{dehaze} is the specific dehazing loss, Φ is the dehazing network with parameter ω_d, $\boldsymbol{M}_{\boldsymbol{u}^*}$, $\boldsymbol{M}_{\boldsymbol{u}}$ and $\boldsymbol{M}_{\tilde{\boldsymbol{u}}^*}$ are the depth maps of $\boldsymbol{u}^*$, $\boldsymbol{u}$ and $\tilde{\boldsymbol{u}}^*$ respectively. In this
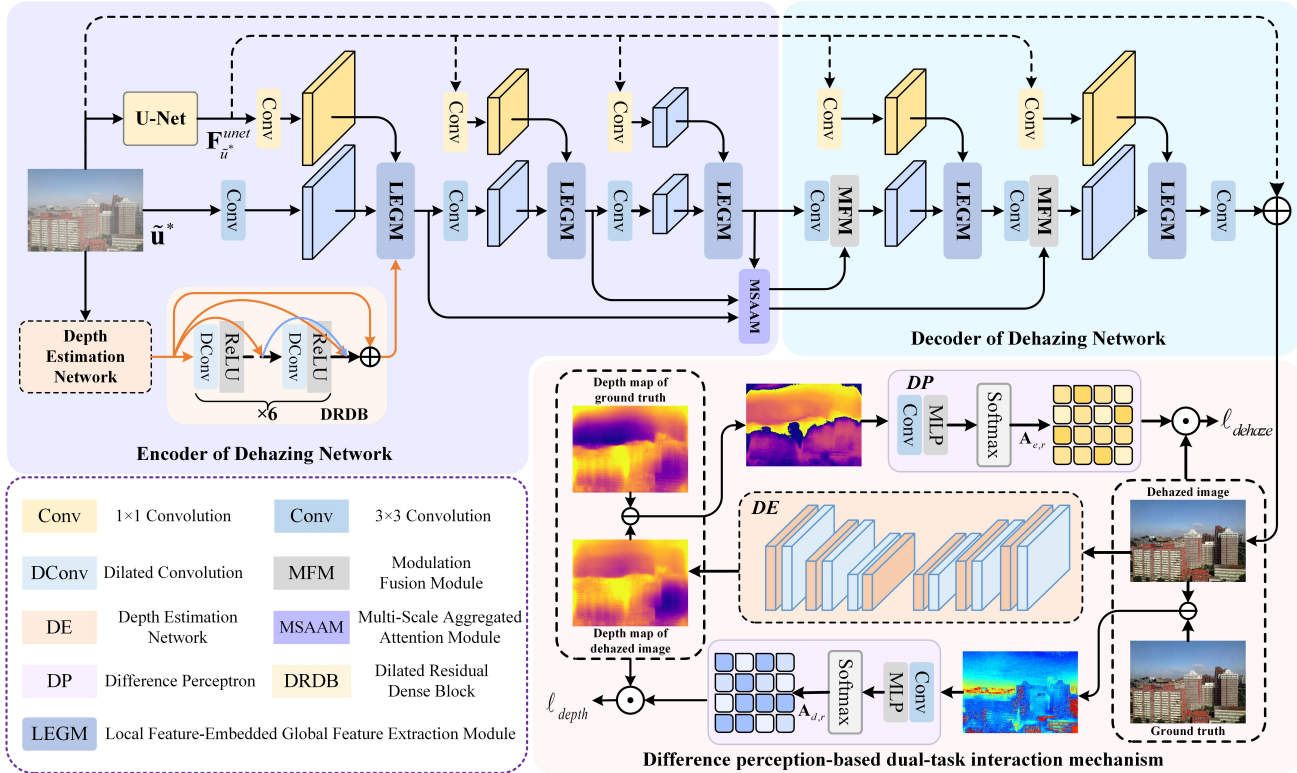
Figure 3. Architecture of the proposed method, which consists of the dehazing network, depth estimation network (DE) and difference perception (DP). DE and DP are the main components of difference perception-based dual-task interaction mechanism. This mechanism enables dehazing and depth estimation to be seamlessly integrated into a unified framework, and improves the model performance of two tasks through collaborative mutual promotion.

work, RA-Depth [11] is used to generate $\boldsymbol{M_u}$. In Eq.(4), $\boldsymbol{M_{u^*}} - \boldsymbol{M_u}$ is the difference between the depth maps of $\boldsymbol{u^*}$ and $\boldsymbol{u}$. Suppose this difference is fed back into the training process of dehazing network Φ, and used to prompt Φ to focus on areas where depth estimation is not ideal. In that case, the performance of Φ will be improved by Eq.(4). Driven by this idea, the depth estimation network Ψ will play a positive role in promoting the performance of the dehazing network Φ. Furthermore, as the performance of Ψ on $\tilde{\boldsymbol{u}}^*$ improves, the predicted depth information becomes more accurate, leading to a more comprehensive representation of structural details in the depth map. Consequently, injecting $\boldsymbol{M_{\tilde{u}^*}}$ into Φ will have a beneficial effect on the recovery of structural information in hazy images. This also reflects the promotion effect of Ψ on Φ. As shown in Figure 2, this design idea of dual-task collaboration can also effectively prevent the dehazing network parameter ω_d from falling into an undesirable local optimum [20].

3.2. Depth Information Assisted Image Dehazing

3.2.1 Encoder of Dehazing Network

As seen from Figure 3, the depth information-assisted image dehazing network consists of encoder and decoder.

The encoder is mainly composed of U-Net, local feature-embedded global feature extraction module (LEGM), depth estimation network (DE), dilated residual dense block (DRDB) and multi-scale aggregation attention module (MSAAM). To ensure the quality of U-Net output, the parameters of U-Net are optimized by l_1-loss:

$$\ell_{unet} = ||\boldsymbol{F}_{\tilde{\boldsymbol{u}}^*}^{unet} - \boldsymbol{F}_{\boldsymbol{u}}^{unet}||_1 \qquad (5)$$

where $\boldsymbol{F}_{\tilde{\boldsymbol{u}}^*}^{unet}$ and $\boldsymbol{F}_{\boldsymbol{u}}^{unet}$ denote the features of $\tilde{\boldsymbol{u}}^*$ and $\boldsymbol{u}$ of U-Net output respectively. As shown in Figure 4 (a), the self-attention block is the main component of LEGM. Its inputs include the features output by the 1×1 convolution after the U-Net, the features output by the 3×3 convolution and the features output by the DRDB after the DE. Since the features extracted by the convolutional network contain a large number of local information, we name the self-attention block combined with the convolution layer LEGM. In the depth information-assisted dehazing, only the first LEGM receives the depth information of the hazy image. The encoder of the dehazing network contains three LEGMs, and the outputs of these LEGMs are integrated through the MSAAM shown in Figure 4 (b) to prevent the loss of shallow features.
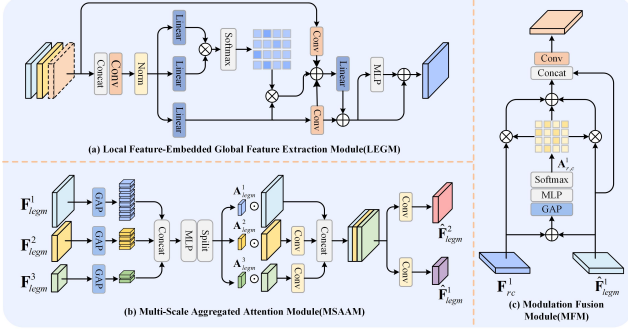
Figure 4. Architectures of the LEGM, MSAAM, MFM.

Let the outputs of three LEGMs be $\boldsymbol{F}_{legm}^1$, $\boldsymbol{F}_{legm}^2$ and $\boldsymbol{F}_{legm}^3$ respectively. The three outputs are then processed through GAP, Concat, MLP and split operations, the weight matrices $\boldsymbol{A}_{legm}^1$, $\boldsymbol{A}_{legm}^2$ and $\boldsymbol{A}_{legm}^3$ are obtained and used for modulating them. The modulated features can be expressed as:

$$\tilde{\boldsymbol{F}}_{legm}^i = \boldsymbol{A}_{legm}^i \odot \boldsymbol{F}_{legm}^i \quad (i = 1, 2, 3) \tag{6}$$

where $\odot$ denotes dot product. After modulation by Eq. (6), the features related to the clear image in $\boldsymbol{F}_{legm}^i$ $(i = 1, 2, 3)$ will be highlighted, which is conducive to the protection of detailed information. Since $\tilde{\boldsymbol{F}}_{legm}^1$, $\tilde{\boldsymbol{F}}_{legm}^2$ and $\tilde{\boldsymbol{F}}_{legm}^3$ have different dimensions, to mine the potential relationship between them, the convolution operation is respectively employed on $\tilde{\boldsymbol{F}}_{legm}^2$ and $\tilde{\boldsymbol{F}}_{legm}^3$ to facilitate their dimensional alignment with $\tilde{\boldsymbol{F}}_{legm}^1$. Subsequently, $\tilde{\boldsymbol{F}}_{legm}^1$, $\tilde{\boldsymbol{F}}_{legm}^2$ and $\tilde{\boldsymbol{F}}_{legm}^3$ are concatenated and processed through two convolutional layers with non-shared parameters, yielding the features $\hat{\boldsymbol{F}}_{legm}^1$ and $\hat{\boldsymbol{F}}_{legm}^2$.

3.2.2 Decoder of Dehazing Network

As can be seen from Figure 3, the decoder of the dehazed image mainly consists of two LEGMs with feature modulation implantation (FMI). Each of them consists of two convolution layers, a modulation fusion module (MFM) and an LEGM, where the structure of the MFM is shown in Figure 4 (c). The inputs of the first MFM are $\hat{\boldsymbol{F}}_{legm}^1$ and the feature $\boldsymbol{F}_{rc}^1$ derived after $\tilde{\boldsymbol{F}}_{legm}^3$ undergoes 3×3 convolution. The result obtained after the sum of $\hat{\boldsymbol{F}}_{legm}^1$ and $\boldsymbol{F}_{rc}^1$ is processed through GAP, MLP, and Softmax, is denoted as $\boldsymbol{A}_{r,c}^1$. The values in $\boldsymbol{A}_{r,c}^1$ indicate the significance of the features in $\hat{\boldsymbol{F}}_{legm}^1$ and $\boldsymbol{F}_{rc}^1$ in the reconstruction of the dehazed image. Their contribution to the image dehazing can be highlighted by adjusting with $\boldsymbol{A}_{r,c}^1$. The specific process can be expressed as follows:

$$\tilde{\boldsymbol{F}}_{rc}^1 = \boldsymbol{A}_{r,c}^1 \odot \hat{\boldsymbol{F}}_{legm}^1 + \boldsymbol{A}_{r,c}^1 \odot \boldsymbol{F}_{rc}^1 \tag{7}$$

$\tilde{\boldsymbol{F}}_{rc}^1$ and $\hat{\boldsymbol{F}}_{legm}^1$ are concatenated to enhance the shared information between them. Subsequently, the concatenation result is processed through a convolution layer to yield the output of the first LEGM with FMI. In the second LEGM with FMI, the first output of LEGM with FMI and $\hat{\boldsymbol{F}}_{legm}^2$ are used as its inputs, and the final output is sent to a convolution layer to reconstruct the final dehazed image $\boldsymbol{u}^*$.

3.3. Depth Estimation Driven by Difference Perception

The proposed dehazing method not only addresses the image dehazing but also incorporates depth estimation based on the dehazed images, thereby presenting a dual-task collaborative promotion network framework. Regarding network architecture, DE follows a U-Net structure, adopting a DRDB-based encoder and decoder design. The encoder and decoder comprises 4 DRDBs, respectively. To improve the performance of the DE on hazy images, we introduce the difference perception between dehazing images and their GTs. The process is shown in the difference perception-based dual-task interaction of Figure 3.

The difference between the dehazing result $\boldsymbol{u}^*$ and $\boldsymbol{u}$ is:

$$\boldsymbol{R}_{\boldsymbol{u}^*/\boldsymbol{u}} = \boldsymbol{u}^* - \boldsymbol{u} \tag{8}$$

where $\boldsymbol{R}_{\boldsymbol{u}^*/\boldsymbol{u}}$ reflects the difference between the dehazing result and the label. It also indirectly indicates the location of regions in the dehazing result where the dehazing effect is not ideal. Suppose the DE can focus on those regions and yield accurate depth estimation results in these regions. In that case, it will provide more accurate depth information for image dehazing in the mutual promotion of dual tasks to assist the dehazing network in obtaining ideal dehazing results. In this process, the difference $\boldsymbol{R}_{\boldsymbol{u}^*/\boldsymbol{u}}$ is processed through a difference perceptron consisting of convolutional layers, MLP and Softmax to predict the coefficient matrix $\boldsymbol{A}_{d,r}$ that reflects the difference between $\boldsymbol{u}^*$ and $\boldsymbol{u}$. DE is optimized by minimizing a tailored loss function as follows:

$$\ell_{depth} = \|\boldsymbol{A}_{d,r} \odot (\boldsymbol{M}_{\boldsymbol{u}^*} - \boldsymbol{M}_{\boldsymbol{u}})\|_1 + \|\boldsymbol{M}_{\tilde{\boldsymbol{u}}^*} - \boldsymbol{M}_{\boldsymbol{u}}\|_1 \tag{9}$$

It is worth noting that optimizing the DE guided by Eq. (9) embodies an integrated optimization strategy of dual tasks. On the one hand, the DE ensures consistency between the estimated depth maps of dehazed images and their corresponding GTs, particularly in regions that exhibit suboptimal dehazing results. This promotes the DE to allocate increased attention to the regions where the dehazed image differs from the GT. The difference is mainly from the residual haze in the dehazed image. Therefore, the optimization in Eq. (9) can improve the dehazing network's perceptiveness to the haze information in the hazy image. On the other hand, the constraint on hazy image depth estimation is introduced into the loss function in Eq. (9), which is equivalent

to optimizing the DE via a dual-task strategy. As depicted in Figure 2, this strategy can effectively mitigate the limitation of a single-task DE that may converge to a non-ideal local optimal solution, enhancing overall performance robustness.

3.4. Dual-task Collaborative Mutual Promotion Learning

DE and the image dehazing network are trained in a mutually promoting manner. When optimizing the DE, the main loss function is ℓ_{depth}. Since $M_{u^*} = \Psi(u^*, \omega_e)$, $M_{\tilde{u}^*} = \Psi(\tilde{u}^*, \omega_e)$ and $u^* = \Phi(\tilde{u}^*, \omega_d)$, we have $M_{u^*} = \Psi(\Phi(\tilde{u}^*, \omega_d), \omega_e)$, and then the formula for updating parameter ω_e can be expressed as:

$$
\begin{aligned}
\omega_e \leftarrow \omega_e - \\
\eta_e \nabla_{\omega_e} \ell_{depth}(\Psi(\Phi(\tilde{u}^*, \omega_d), \omega_e); \Psi(\tilde{u}^*, \omega_e); M_u)
\end{aligned}
\tag{10}
$$

where η_e denotes the learning rate. $\Psi(\Phi(\tilde{u}^*, \omega_d), \omega_e)$, $\Psi(\tilde{u}^*, \omega_e)$ and M_u represent the inputs of the loss function ℓ_{depth}. It can be seen that the output result $u^* = \Phi(\tilde{u}^*, \omega_d)$ is used when updating the parameter ω_e. In the iterative optimization of Φ and Ψ, the performance of the dehazing network Φ increases from weak to strong. Under the constraint of loss function $\|A_{d,r} \odot (M_{u^*} - M_u)\|_1$ in Eq. (9), the DE can improve its accuracy by focusing on the difference between the dehazing result and the GT. This demonstrates that the dehazing network plays a positive role in promoting the performance of the DE.

During the optimization of the dehazing network, $M_{u^*} - M_u$ is fed to the difference perceptron to yield the coefficient matrix $A_{e,r}$. With $A_{e,r}$, we use the following loss function to update the parameters of the dehazing network:

$$
\begin{aligned}
\ell_{dehaze} = \|A_{e,r} \odot (u^* - u)\|_1 \\
+ \sum_{i=1}^{n} \lambda_i \frac{\|VGG_i(u) - VGG_i(\Phi(\tilde{u}^*, \omega_d))\|_1}{\|VGG_i(\tilde{u}^*) - VGG_i(\Phi(\tilde{u}^*, \omega_d))\|_1}
\end{aligned}
\tag{11}
$$

where VGG_i represents the output feature of the i-th layer of VGG19 [35], and λ_i is the weight coefficient [44]. Therefore, the update process of the dehazing network can be described by:

$$
\begin{aligned}
\omega_d \leftarrow \omega_d - \\
\eta_d \nabla_{\omega_d} \ell_{dehaze}(\Phi(\tilde{u}^*, \omega_d); \Psi(\Phi(\tilde{u}^*, \omega_d), \omega_e); u; M_u)
\end{aligned}
\tag{12}
$$

where η_d is the learning rate. It can be seen that when updating ω_d, the depth estimation result of the dehazing image $\Psi(\Phi(\tilde{u}^*, \omega_d), \omega_e)$ is used. In this process, the difference between $\Psi(\Phi(\tilde{u}^*, \omega_d), \omega_e)$ and M_u is directly conveyed to the optimization of ω_d through the loss function, which promotes the dehazing network to improve its performance by updating ω_d, to reduce the difference between u^* and u.

Therefore, it is evident from the above analysis that the DE plays a positive role in promoting the performance of the dehazing network. On the other hand, as illustrated in Figure 3, the depth estimation result of $\tilde{u}^*$ is fed into the LEGM and participates in the feature extraction of hazy images. This facilitates the dehazing network to obtain useful auxiliary information from $M_{\tilde{u}^*}$ to improve the quality of the dehazing results.

4. Experiments

4.1. Experimental Settings

Datasets To ensure an unbiased comparison with current dehazing methods, we use the Indoor Training Set (ITS) and Outdoor Training Set (OTS) from the RESIDE dataset [16] as the training data. For evaluation, we employ the Synthetic Objective Test Set (SOTS), which contains 500 indoor and 500 outdoor hazy images for testing. The real-world images are collected to verify the model generalization.

Implementation Details. All experiments are conducted on an NVIDIA GeForce RTX 3090 with 24GB GPU, and the model is implemented in the Pytorch 1.12.0 framework. During the training phase, Adam optimizer [13] is used to optimize the network. We set the initial learning rate to 0.001 and use a cosine annealing strategy to adjust the learning rate. Moreover, we randomly crop the images into 256×256 patches for training. In each mini-batch, the patches are augmented through horizontal or vertical flipping to enlarge the training samples. The entire training process lasted for a total of 600 epochs on the indoor dataset and 60 epochs on the outdoor dataset.

Evaluation Metrics. The proposed method is compared with the state-of-the-art deep learning-based dehazing methods. Their performance in terms of fidelity and perceptual quality is evaluated by five metrics: PSNR, SSIM, NIQE [26], PIQE [41] and FADE [5].

4.2. Comparison with State-of-the-arts

Table 1 shows the quantitative results of different dehazing methods on the SOTS indoor and outdoor datasets, respectively. The evaluation results indicate that the proposed method has achieved the best values on SSIM, NIQE, PIQE, and FADE, while also obtaining comparable scores on PSNR. Figure 5 depicts the dehazing results obtained on the SOTS indoor and outdoor datasets by different methods. It can be seen from the local residual maps that the proposed method has better dehazing performance than other methods on synthetic data. Moreover, to assess the generalization of the proposed method, experiments are conducted on real-world hazy images. The experimental results are presented in Figure 6. It shows that the proposed

Table 1. Performance of the proposed method is compared with that of state-of-the-art methods on the synthetic datasets (SOTS-indoor and SOTS-outdoor). Best values are in bold.

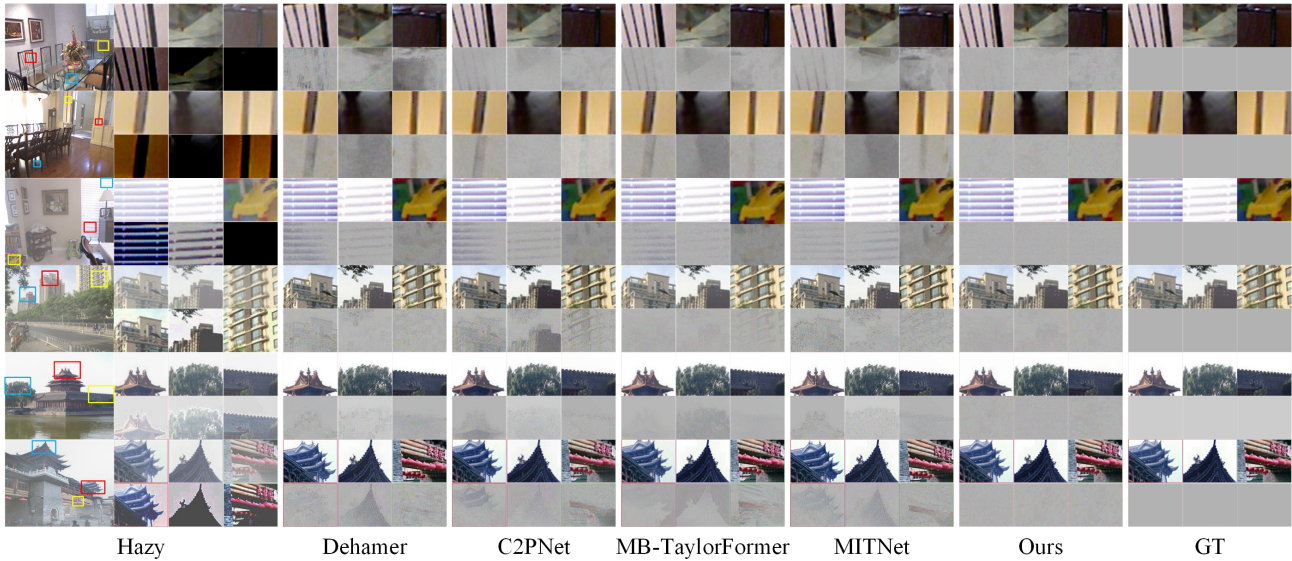| Methods	SOTS-indoor					SOTS-outdoor				
	PSNR↑	SSIM↑	NIQE↓	PIQE↓	FADE↓	PSNR↑	SSIM↑	NIQE↓	PIQE↓	FADE↓
AECR-Net [44]	37.17	0.9901	–	–	–	–	–	–	–	–
PSD [4]	12.50	0.7153	4.6350	35.1740	0.6023	15.51	0.7488	3.2619	11.7166	0.7235
MAXIM-2S [40]	38.11	0.9908	4.2962	31.8841	0.5209	34.19	0.9846	3.0964	7.7737	0.7697
Dehamer [8]	36.63	0.9881	4.6482	31.4427	0.5020	35.18	0.9860	3.0088	9.7072	0.8318
D4 [48]	25.24	0.9320	4.3261	31.5310	0.7084	25.83	0.9560	2.9474	7.6678	0.7340
Dehazeformer [36]	40.05	0.9958	4.3276	32.2755	0.4987	34.95	0.9840	2.9507	7.8495	0.8093
C2PNet [52]	42.56	0.9954	4.3978	31.7937	0.4983	36.68	0.9900	3.1437	11.7275	1.0630
MB-TaylorFormer [31]	**42.63**	0.9942	4.3047	31.9366	0.4959	**38.09**	0.9910	2.9952	7.9968	0.8378
MITNet [34]	40.23	0.9920	4.2529	31.3239	0.5019	35.18	0.9920	2.9454	7.8575	0.7563
Proposed	42.18	**0.9967**	**4.2403**	**31.3046**	**0.4907**	36.56	**0.9931**	**2.9377**	**7.6498**	**0.7112**

Figure 5. Visual comparisons on SOTS-indoor and SOTS-outdoor. Due to space limitations, we only show the visual effects of the images obtained by the methods with excellent performance of each year in Table 1. To facilitate the visual comparison, we display the visual effects of enclosed areas from the dehazed results and the differences between areas enclosed and their GTs. Less residual information in the difference map indicates better dehazing effect.

method achieves better dehazing effect than other methods compared on real-world data.

4.3. Ablation Study

The proposed method mainly consists of the LEGM, MFM, MSAAM, DE and DP. To verify their contribution to the full model, we conduct ablation experiments on SOTS-Indoor dataset. We exclude the MSAAM, DE, and DP from the model depicted in Figure 3, and replace the LEGM and MFM with a summation operation, then construct the Baseline. It is trained with constrains of loss in Eq.(11). The quantitative evaluation results of the ablation experiment are listed in Table 2.

Table 2. Ablation study of each module on SOTS-Indoor dataset.

Methods	PSNR↑	SSIM↑	NIQE↓	PIQE↓	FADE↓
Baseline	35.69	0.9900	4.3974	31.9224	0.5950
+LEGM	40.41	0.9947	4.2701	31.4971	0.5948
+MFM	40.60	0.9957	4.2601	31.3885	0.5717
+MSSAM	40.75	0.9957	4.2485	31.3657	0.5423
+DE	41.55	0.9958	4.2269	31.3438	0.5018
+DP	42.18	0.9967	4.2400	31.3050	0.4907

Effectiveness of LEGM. As shown in Table 2, compared to the Baseline, the PSNR of the Baseline+LEGM increases by $4.72dB$. Moreover, in terms of other metrics, notable enhancements can be also observed. Those improvements can be attributed to the inherent capability of the LEGM,
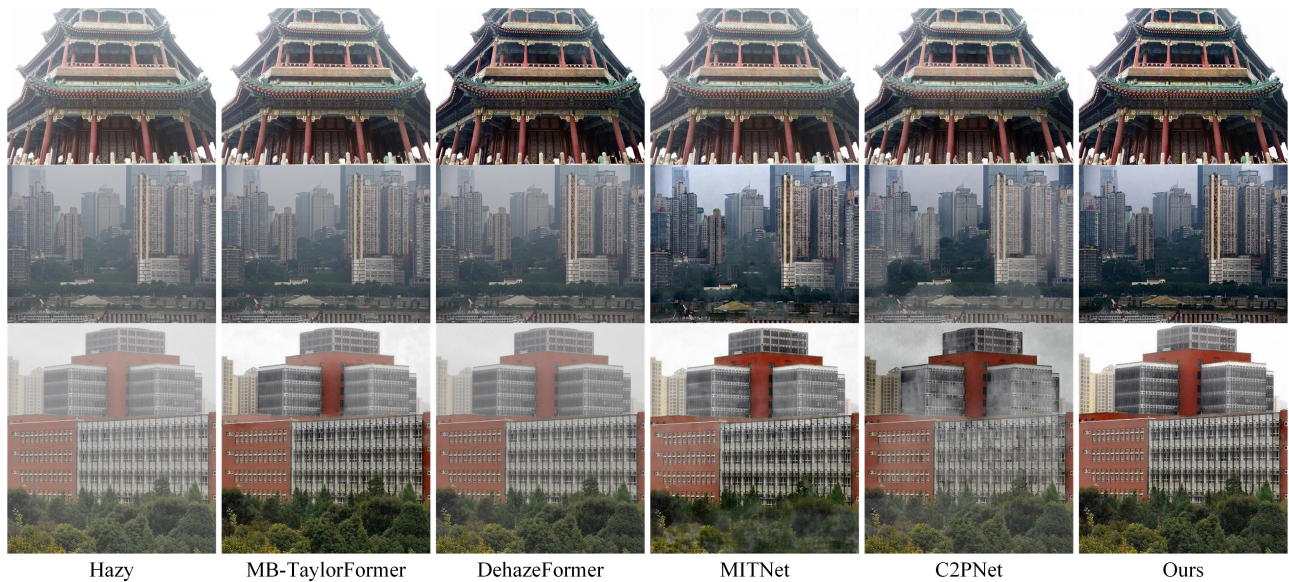
Figure 6. Visual comparisons on real-world images collected by ourselves. These images were taken casually with a mobile phone under foggy weather conditions.

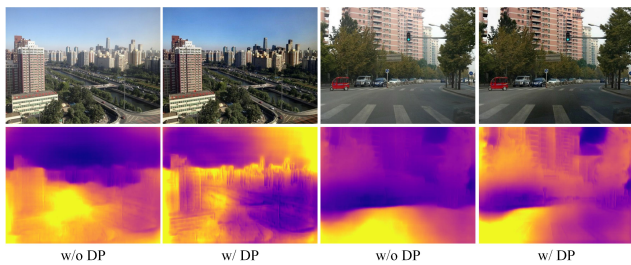| Hazy | MB-TaylorFormer | DehazeFormer | MITNet | C2PNet | Ours |

Figure 7. Depth estimation and Dehazing results of the proposed method without DP and with DP.

which integrates both global and local features.

Effectiveness of MFM. To enhance the feature representation ability of the network, MFM facilitates the feature interaction across diverse channels by dynamically adjusting fusion weights. As listed in Table 2, the addition of MFM to the Baseline+LEGM improves the model performance.

Effectiveness of MSAAM. As shown in Table 2, the performance of the model is further improved after MSAAM is added into Baseline+LEGM+MFM. The features of different scales aggregated through MSAAM are fused with the decoding layer features, which could alleviate the feature dilution problem in the encoding and decoding process.

Effectiveness of DE. DE is incorporated into the Baseline+LEGM+MFM+MSAAM to verify its effectiveness. As listed in Table 2, the DE effectively enhances the performance of the model. This can be attributed to the fact that depth information offers crucial insights, including object depth and spatial coherence. It contributes to the preser-

vation of image structure stability, which also proves the effectiveness of dual task mutual promotion.

Effectiveness of DP. DP is added into Baseline+LEGM+MFM+MSAAM+DE to verify its effectiveness. As compared in Table 2, the DP facilitates the enhancement of model performance. It is discernible from Figure 7 that the network with DP module acquires a noteworthy improvement in both depth estimation and dehazing results.

5. Conclusion

We propose a novel approach to treat depth estimation and image dehazing as two independent tasks, and integrate them into a joint learning framework through a dual task interaction mechanism for joint optimization. The proposed difference perception in the depth maps between the dehazing result and the ideal image allows the dehazing network to focus on the sub-optimal dehazing areas to improve the network performance. Concurrently, the difference between the dehazed image and its label guides the depth estimation network to pay attention to the sub-optimal dehazing areas, thereby improving the depth prediction accuracy. In this manner, the image dehazing and depth estimation network are mutually reinforced. Experimental results show that the proposed method exhibits superior performance on both synthetic and real-world hazy images.

References

[1] Dana Berman, Tali Treibitz, and Shai Avidan. Non-local image dehazing. In *CVPR*, pages 1674–1682, 2016.

[2] Dana Berman, Tali Treibitz, and Shai Avidan. Air-light estimation using haze-lines. In *ICCP*, pages 1–9, 2017.

[3] 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*, 25(11):5187–5198, 2016.

[4] Zeyuan Chen, Yangchao Wang, Yang Yang, and Dong Liu. Psd: Principled synthetic-to-real dehazing guided by physical priors. In *CVPR*, pages 7176–7185, 2021.

[5] Lark Kwon Choi, Jaehee You, and Alan Conrad Bovik. Referenceless prediction of perceptual fog density and perceptual image defogging. *IEEE Transactions on Image Processing*, 24(11):3888–3901, 2015.

[6] 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 *CVPR*, pages 2154–2164, 2020.

[7] Junkai Fan, Fei Guo, Jianjun Qian, Xiang Li, Jun Li, and Jian Yang. Non-aligned supervision for real image dehazing. *arXiv preprint arXiv:2303.04940*, 2023.

[8] Chunle Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. Image dehazing transformer with transmission-aware 3d position embedding. In *CVPR*, pages 5802–5810, 2022.

[9] Tiantong Guo, Xuelu Li, Venkateswararao Cherukuri, and Vishal Monga. Dense scene information estimation network for dehazing. In *CVPRW*, pages 2122–2130, 2019.

[10] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(12):2341–2353, 2011.

[11] Mu He, Le Hui, Yikai Bian, Jian Ren, Jin Xie, and Jian Yang. Ra-depth: Resolution adaptive self-supervised monocular depth estimation. In *ECCV*, pages 565–581, 2022.

[12] Shih-Chia Huang, Trung-Hieu Le, and Da-Wei Jaw. Dsnet: Joint semantic learning for object detection in inclement weather conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2623–2633, 2021.

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[14] Ashutosh Kulkarni and Subrahmanyam Murala. Aerial image dehazing with attentive deformable transformers. In *WACV*, pages 6294–6303, 2023.

[15] Byeong-Uk Lee, Kyunghyun Lee, Jean Oh, and In So Kweon. Cnn-based simultaneous dehazing and depth estimation. In *ICRA*, pages 9722–9728, 2020.

[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*, 28(1):492–505, 2019.

[17] Huafeng Li, Jirui Gao, Yafei Zhang, Minghong Xie, and Zhengtao Yu. Haze transfer and feature aggregation network for real-world single image dehazing. *Knowledge-Based Systems*, 251:109309, 2022.

[18] Yunan Li, Qiguang Miao, Wanli Ouyang, Zhenxin Ma, Huijuan Fang, Chao Dong, and Yining Quan. Lap-net: Level-aware progressive network for image dehazing. In *ICCV*, pages 3275–3284, 2019.

[19] Yazhao Li, Yanwei Pang, Jianbing Shen, Jiale Cao, and Ling Shao. Netnet: Neighbor erasing and transferring network for better single shot object detection. In *CVPR*, pages 13346–13355, 2020.

[20] Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, and Xin Fan. Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation. In *ICCV*, pages 8081–8090, 2023.

[21] Risheng Liu, Xin Fan, Minjun Hou, Zhiying Jiang, Zhongxuan Luo, and Lei Zhang. Learning aggregated transmission propagation networks for haze removal and beyond. *IEEE Transactions on Neural Networks and Learning Systems*, 30 (10):2973–2986, 2019.

[22] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*, pages 7313–7322, 2019.

[23] Yang Liu, Jinshan Pan, Jimmy Ren, and Zhixun Su. Learning deep priors for image dehazing. In *ICCV*, pages 2492–2500, 2019.

[24] Yun Liu, Zhongsheng Yan, Sixiang Chen, Tian Ye, Wenqi Ren, and Erkang Chen. Nighthazeformer: Single nighttime haze removal using prior query transformer. In *ACM MM*, pages 4119–4128, 2023.

[25] Earl J McCartney. Optics of the atmosphere: scattering by molecules and particles. *New York, NY, USA: Wiley*, 1976.

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[27] Srinivasa G. Narasimhan and Shree K. Nayar. Chromatic framework for vision in bad weather. In *CVPR*, pages 598–605 vol.1, 2000.

[28] Srinivasa G. Narasimhan and Shree K. Nayar. Vision and the atmosphere. *International Journal of Computer Vision*, 48: 233–254, 2002.

[29] Yanwei Pang, Jing Nie, Jin Xie, Jungong Han, and Xuelong Li. Bidnet: Binocular image dehazing without explicit disparity estimation. In *CVPR*, pages 5930–5939, 2020.

[30] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. Ffa-net: Feature fusion attention network for single image dehazing. In *AAAI*, pages 11908–11915, 2020.

[31] Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, and Zhi Jin. Mb-taylorformer: Multi-branch efficient transformer expanded by taylor formula for image dehazing. In *ICCV*, pages 12756–12767, 2023.

[32] 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*, pages 154–169, 2016.

[33] 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*, pages 3253–3261, 2018.

[34] Hao Shen, Zhong-Qiu Zhao, Yulun Zhang, and Zhao Zhang. Mutual information-driven triple interaction network for efficient image dehazing. In *ACM MM*, pages 7–16, 2023.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[36] Yuda Song, Zhuqing He, Hui Qian, and Xin Du. Vision transformers for single image dehazing. *IEEE Transactions on Image Processing*, 32:1927–1941, 2023.

[37] Matan Sulami, Itamar Glatzer, Raanan Fattal, and Mike Werman. Automatic recovery of the atmospheric light in hazy images. In *ICCP*, pages 1–11, 2014.

[38] Shangquan Sun, Wenqi Ren, Tao Wang, and Xiaochun Cao. Rethinking image restoration for object detection. *NeurIPS*, 35:4461–4474, 2022.

[39] Robby T. Tan. Visibility in bad weather from a single image. In *CVPR*, pages 1–8, 2008.

[40] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxim: Multi-axis mlp for image processing. In *CVPR*, pages 5759–5770, 2022.

[41] N Venkatanath, D Praneeth, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *NCC*, pages 1–6, 2015.

[42] Cong Wang, Yuexian Zou, and Zehan Chen. Abc-net: Avoiding blocking effect & color shift network for single image dehazing via restraining transmission bias. In *ICIP*, pages 1053–1057, 2020.

[43] Cong Wang, Jinshan Pan, Wanyu Lin, Jiangxin Dong, and Xiao-Ming Wu. Selfpromer: Self-prompt dehazing transformers with depth-consistency. *arXiv preprint arXiv:2303.07033*, 2023.

[44] 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 *CVPR*, pages 10546–10555, 2021.

[45] Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. Ridcp: Revitalizing real image dehazing via high-quality codebook priors. In *CVPR*, pages 22282–22291, 2023.

[46] Jiaqi Xu, Xiaowei Hu, Lei Zhu, Qi Dou, Jifeng Dai, Yu Qiao, and Pheng-Ann Heng. Video dehazing via a multi-range temporal alignment network with physical prior. In *CVPR*, pages 18053–18062, 2023.

[47] Fei Yang and Qian Zhang. Depth aware image dehazing. *The Visual Computer*, 38:1579–1587, 2022.

[48] Yang Yang, Chaoyue Wang, Risheng Liu, Lin Zhang, Xiaojie Guo, and Dacheng Tao. Self-augmented unpaired image dehazing via density and depth decomposition. In *CVPR*, pages 2027–2036, 2022.

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

[50] Libao Zhang, Shan Wang, and Xiaohan Wang. Single image dehazing based on bright channel prior model and saliency analysis strategy. *IET Image Processing*, 15(5):1023–1031, 2021.

[51] Shiyu Zhao, Lin Zhang, Ying Shen, and Yicong Zhou. Refinednet: A weakly supervised refinement framework for single image dehazing. *IEEE Transactions on Image Processing*, 30:3391–3404, 2021.

[52] Yu Zheng, Jiahui Zhan, Shengfeng He, Junyu Dong, and Yong Du. Curricular contrastive regularization for physics-aware single image dehazing. In *CVPR*, pages 5785–5794, 2023.

[53] Hao Zhou, Zekai Chen, Yun Liu, Yongpan Sheng, Wenqi Ren, and Hailing Xiong. Physical-priors-guided dehazeformer. *Knowledge-Based Systems*, 266:110410, 2023.

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