Perceiving and Modeling Density
for Image Dehazing

Tian Ye[1], Yunchen Zhang[2], Mingchao Jiang[3], Liang Chen[4], Yun Liu[5],
Sixiang Chen[1], and Erkang Chen[1(✉)]

[1] School of Ocean Information Engineering, Jimei University, Xiamen, China
{201921114031,201921114013,ekchen}@jmu.edu.cn
[2] China Design Group Co.,Ltd., Nanjing, China
cydiachen@cydiachen.tech
[3] Joyy Ai Group, Guangzhou, China
[4] Fujian Provincial Key Laboratory of Photonics Technology,
Fujian Normal University, Fuzhou, China
[5] College of Artificial Intelligence, Southwest University, Chongqing, China
yunliu@swu.edu.cn

Abstract. In the real world, the degradation of images taken under haze can be quite complex, where the spatial distribution of haze varies from image to image. Recent methods adopt deep neural networks to recover clean scenes from hazy images directly. However, due to the generic design of network architectures and the failure in estimating an accurate haze degradation model, the generalization ability of recent dehazing methods on real-world hazy images is not ideal. To address the problem of modeling real-world haze degradation, we propose a novel Separable Hybrid Attention (SHA) module to perceive haze density by capturing positional-sensitive features in the orthogonal directions to achieve this goal. Moreover, a density encoding matrix is proposed to model the uneven distribution of the haze explicitly. The density encoding matrix generates positional encoding in a semi-supervised way – such a haze density perceiving and modeling strategy captures the unevenly distributed degeneration at the feature-level effectively. Through a suitable combination of SHA and density encoding matrix, we design a novel dehazing network architecture, which achieves a good complexity-performance trade-off. Comprehensive evaluation on both synthetic datasets and real-world datasets demonstrates that the proposed method surpasses all the state-of-the-art approaches with a large margin both quantitatively and qualitatively. The code is released in https://github.com/Owen718/ECCV22-Perceiving-and-Modeling-Density-for-Image-Dehazing.

Keywords: Image dehazing · Image restoration · Deep learning

T. Ye, Y. Zhang and M. Jiang—Equal contribution.

1   Introduction

Single image dehazing aims to generate a haze-free image from a hazy image. It is a classical image processing problem, which has been an important research topic in the vision communities within the last decade [3, 4, 21, 23]. Numerous real-world vision tasks (e.g., object detection and auto drive) require high-quality clean images, while the fog and haze usually lead to degraded images. Therefore, it is of great interest to develop an effective algorithm to recover haze-free images.
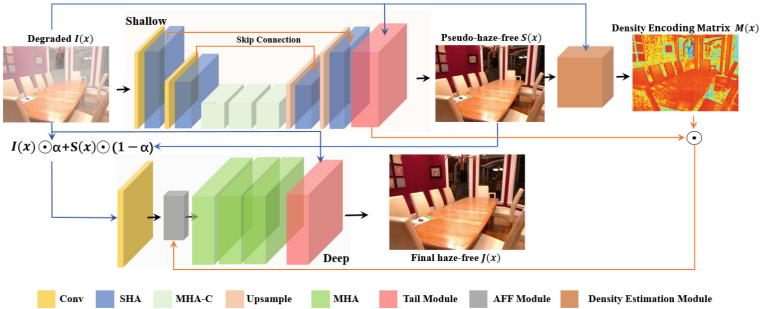

Fig. 1. Overview of network architecture. The Tail Module is formulated by a convolution layer and an activation layer. The shallow layers, consisting of a stack of Multi-branch Hybrid Attention modules with Contextual Transformer (MHAC) blocks and a Tail Module, are used to generate a pseudo-haze-free image. Then, the density encoding matrix is generated by a density estimation module. The deep layers emphasize on detailed reconstruction by using the Adaptive Features Fusion (AFF) module and Multi-branch Hybrid Attention Block (MHA). α is a learnable factor.

Haze is a common atmospheric phenomenon in our daily life. Images with haze and fog lose details and color fidelity. Mathematically, the image degradation caused by haze can be formulated by the following model:

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

where $\mathbf{I}(\mathbf{x})$ is the hazy image, $\mathbf{J}(\mathbf{x})$ is the clear image, $t(\mathbf{x})$ is the transmission map and $\mathbf{A}(\mathbf{x})$ stands for the global atmospheric light.

Methods [5, 9, 11, 14] based on priors first estimate the above two unknown parameters $t(\mathbf{x})$ and $\mathbf{A}(\mathbf{x})$, then infer $\mathbf{J}(\mathbf{x})$ according the Eq. 1 inversely. Unfortunately, these hand-crafted priors do not always hold in diverse real-world hazy scenes, resulting in inaccurate estimation of $t(\mathbf{x})$. Therefore, in recent years, more and more researchers have begun to pay attention to data-driven learning algorithms [8, 12, 16, 21, 23]. Different from traditional methods, deep learning methods achieve superior performance via training on large-scale datasets. However, methods introduce generic network architectures by simply stacking more layers or using wider layers, ignoring the uneven distribution that causes the unknown depth information which varies at different positions. Some approaches aim at introducing complex loss functions and designing fancy training strategies, leads

to high training-costs and poor convergence. Due to the airlight-albedo ambiguity and the limited number of datasets with ground-truth data of transmission maps, these methods result in poor performance degradation in real-world scenes. Moreover, inaccurate estimation of transmission map can significantly deteriorate the clean image restoration.

To address the above problems, we learn the implicit representation and estimate the explicit supervision information for modeling the uneven haze distribution. From Eq. 1, we can notice that the haze degradation model is highly associated with the absolute position of image pixels. The key to solve the dehazing problem lies in accurately encoding the haze intensity with its absolute position. As is demonstrated in Fig. 1, we propose a network to perceive the density of haze distribution with tailor designed modules that captures the positional-sensitive features. Moreover, we introduce a haze density encoding matrix to encode the image-level co-relationship between haze intensity and its absolute position.

We propose the method from three different levels: primary block of the network, architecture of the network and map of haze density information to refine features:

• *Primary Block:* We propose an efficient attention mechanism to perceive the uneven distribution of degradation of features among channel and spatial dimensions: Separable Hybrid Attention (SHA), which effectively samples the input features through a combination of different pooling operations, and strengthens the interaction of different kinds of dimensional information by scaling and shuffling channels. Our SHA based on horizontal and vertical encoding can obtain sufficient spatial clues from input features.

• *Density Encoding Matrix:* We design a coefficient matrix called as density encoding matrix, which encodes the co-relationship between haze intensity and absolute position. The density encoding matrix explicitly models the intensity of the haze degradation model at corresponding spatial locations. The density encoding matrix is obtained in an end-to-end manner, and semantic information of the scene is also introduced implicitly, which makes the density encoding matrix more consistent with the actual distribution.

• *Network Architecture:* We design a novel network architecture that restores hazy images with the coarse-to-fine strategy. The architecture of our network mainly consists of three parts: shallow layer, deep layer and density encoding matrix. We build the shallow layer and deep layer of our method based on SHA. The shallow layer will generate the coarse haze-free image, which we call as the pseudo-haze-free image. For modeling the uneven degradation of a hazy image to refine features explicitly, we utilize the pseudo-haze-free image and the input hazy sample to generate the density encoding matrix.

Our main contributions are summarized as follows:

– We propose the SHA as a task-specified attention mechanism, which perceives haze density effectively from its design of operation on orthogonal separated direction operations and enlarged perception fields.
– We propose the density encoding matrix as an explicit model of haze density, which enhances the coupling of our model. In addition, the proposed method

demonstrates the potential of dealing with the non-homogeneous haze distribution.

– We formulate a novel dehazing method that incorporates the implicit perception of haze features and explicit model of haze density in a unified framework. Our method confirms the necessity of perceiving and modeling of haze density and achieves the best performance compared with the state-of-the art approaches.

2   Related Works

2.1   Single Image Dehazing

Single image dehazing is mainly divided into two categories: a prior-based defogging method [11,15] and a data-driven method based on deep learning. With the introduction of large hazy datasets [17,20], image dehazing based on the deep neural network has developed rapidly. MSBDN [8] uses the classic Encoder-Decoder architecture, but repeated up-sampling and down-sampling operations result in texture information loss. The number of parameters of MSBDN [8] is large, and the model is complex. FFA-Net [21] proposes an Feature Attention (FA) block based on channel attention and pixel attention, obtains the final haze-free image by fusing features of different levels. With the help of two different-scale attention mechanisms, FFA-Net has obtained impressive PSNR and SSIM, but the entire model performs convolutional operations at the resolution of the original image, resulting in a large amount of calculation and slow speed. AECR-Net [23] reuses the FA block and proposes to use a novel loss function based on the contrast learning to make full use of hazy samples, and use a deformable convolution block to improve the expression ability of the model, pushing the index on SOTS [17] to a new height, but memory consumption of the loss function is so high. Compared with AECR-Net [23], our model only needs to utilize a simple Charbonnier [7] loss function to achieve higher PSNR and SSIM.

2.2   Attention Mechanism

Attention mechanisms have been proven essential in various computer vision tasks, such as image classification [13], segmentation [10], dehazing [12,21,23] and deblurring [24]. One of the classical attention mechanisms is SENet [13], which is widely used as a comparative baseline of the plugin of the backbone network. CBAM [22] introduces spatial information encoding via convolutions with large-size kernels, which sequentially infers attention maps along the channel and spatial dimension. The modern attention mechanism extends the idea of CBAM [22] by adopting different dimension attention mechanisms to design the advanced attention module. This shows that the key of the performance of the attention mechanism is to sample the original feature map completely. There is no effective information exchange of attention encoding across different dimensions in existing networks, thus limiting the promotion of networks. In response to the above problems, we utilize two types of pooling operations to sample the

original feature map and smartly insert the channel shuffle block in our atten-
tion module. Experiments demonstrate that the performance of our attention
mechanism has been dramatically improved due to sufficient feature sampling
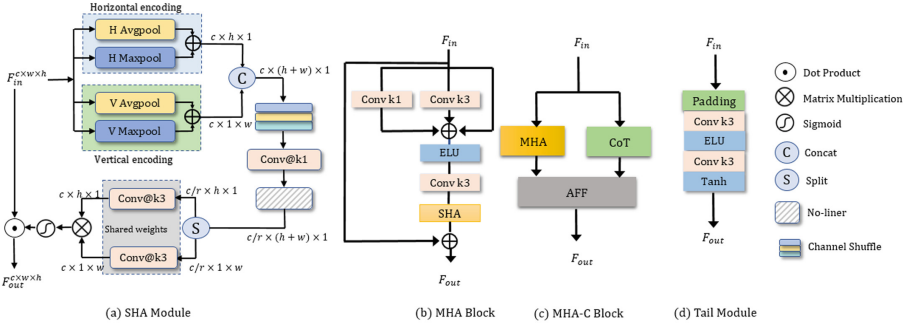and efficient feature exchange among different dimensions.

Fig. 2. Illustrations of the Separable Hybrid Attention (SHA) Module and our basic
blocks. (a) The Separable Hybrid Attention Module. (b) The Multi-branch Hybrid
Attention (MHA) Block. (c) The MHAC block of shallow layers. (d) The tail module
of the shallow and deep layers. Our SHA focuses on directional embedding, which
consists of unilateral directional pooling and convolution. **@k** denotes the kernel size
of the convolution layer.

3   Proposed Methods

We formulate the image dehazing task from two aspects: designing basic blocks
to capture positional-sensitive features from orthogonal directions, and modeling
haze density from the image-level to the feature-level explicitly.

3.1   Implicit Perception of Haze Density - Separable Hybrid Attention Mechanism and Its Variants

Previous image dehazing methods [12,21,23] follow the paradigm of subsequently
calculating attention weights on spatial dimensions and channel dimensions,
which lacks useful information exchange among dimensions and leads to the
sub-optimal perceptual ability in capturing uneven distributions.

In order to introduce the effective cross-dimension information exchange, we
propose a novel attention mechanism called Separable Hybrid Attention (SHA).
Different from previous methods that treat attention across the dimension indi-
vidually, the SHA mechanism acts in a unified way: It harvests contextual infor-
mation by compressing features from the horizontal and vertical axes, which
guarantees that a pixel at any position perceives contextual information from all
pixels. Specifically, it first compresses spatial information in separated directions,
exchanges information across channel dimensions, reduces the channel dimen-
sion and then recovers channel information and spatial information sequentially.

Specifically, a hybrid design of the pooling layer is proposed to preserve abundant spatial information. It combines the maximum pooling layer and the average pooling layer to minimize the loss on capturing high-frequency information and maximize noise resistances among spatial dimensions.

The details and formulas are as follows:

$$v_{avg}^h = \mathbf{AvgPool_h}(F_{in}), v_{avg}^v = \mathbf{AvgPool_v}(F_{in}), \tag{2}$$

$$v_{max}^h = \mathbf{MaxPool_h}(F_{in}), v_{max}^v = \mathbf{MaxPool_v}(F_{in}), \tag{3}$$

We get the direction encode features v^h and v^v from Eq. 2 and Eq. 3 respectively, so we can make distribution of features like a normal distribution, which can better preserve the important information of input features:

$$v^h = v_{avg}^h + v_{max}^h, v^v = v_{avg}^v + v_{max}^v, \tag{4}$$

We concatenate the encoded features and utilize the channel shuffle to interaction channel information, aiming to exchange the encoding information between different channels. Then, we utilize the 1x1 convolution to reduce the dimension of features, which pass through the nonlinear activation function, so that the features of different channels are fully interactive. The formulas are as follows:

$$[y_{c/r}^h, y_{c/r}^v] = \delta(\mathbf{Conv}(\mathbf{cat}([v_c^h, v_c^v]))), \tag{5}$$

wherein Eq. 5, δ is $ReLU6$ activation function, c is the number of dimensions and r is the channel scaling factor, r is usually 4. We use the shared 3x3 convolution to restore the number of channel dimensions for encoded features of different directions, making the isotropic features get similar attention weight values. The final weights can be determined by the larger receptive field of the input features.

$$y_c^h = \mathbf{Conv}(y_{c/r}^h), y_c^v = \mathbf{Conv}(y_{c/r}^v), \tag{6}$$

After restoring the channel of the feature, multiply y_c^h and y_c^w to obtain the attention weight matrix with the same size as the input feature. Finally, a *Sigmoid* function works to get the attention map $W_{c \times h \times w}$:

$$W_{c \times h \times w} = \mathbf{Sigmoid}(y_c^h \times y_c^v), \tag{7}$$

We multiply the attention weight matrix $W_{c \times h \times w}$ with the input feature F_{in} to get the output feature F_{out}:

$$F_{out} = W_{c \times h \times w} \otimes F_{in}. \tag{8}$$

Multi-branch Hybrid Attention. We design the Multi-branch Hybrid Attention (MHA) Block, mainly consisting of the SHA module with parallel convolution. The multi-branch design will improve the expressive ability of the network by introducing multi-scale receptive-filed. The multi-branch block comprises parallel 3x3 convolution, 1x1 convolution, and a residual connection, as shown in Fig. 2

(a). The degradation is often uneven in space in degraded images, such as hazy images and rainy images, and the spatial structure and color of some areas in the picture are not affected by the degradation of the scene, so we set the local residual learning to let the feature pass the current block directly without any process, which also avoids the disappearance of the gradient.

Adaptive Features Fusion Module. We hope that the network can adjust the proportion of feature fusion adaptively according to the importance of different kinds of information. Different from the MixUP [23] operation that fuses the information from different layers for feature preserving in AECR-Net, we utilize the Adaptive Features Fusion Module to combine two different blocks. The formula is as follows:

$$\begin{aligned} F_{out} &= \mathbf{AFF}(\mathbf{block1}, \mathbf{block2}), \\ &= \sigma(\theta) * \mathbf{block1} + (1 - \sigma(\theta)) * \mathbf{block2}, \end{aligned} \tag{9}$$

wherein, *block* denotes the block module, which has the same output size, σ is the *Sigmoid* activation function, and θ is a learnable factor.

Multi-branch Hybrid Attention with Contextual Transformer. Long-range dependence is essential for feature representation, so we introduce an improved Contextual Transformer (CoT) [18] block combined with the MHA block in the shallow layers to mine the long-distance dependence of sample features that further expand the receptive field. Specifically, we design a parallel block that uses the MHA block and the improved CoT block to capture local features and global dependencies simultaneously. To fuse the attention result, an Adaptive Features Fusion module is followed back as shown in Fig. 2 (b), we call this as MHAC block. The formulas are as follows:

$$\begin{aligned} F_{out} &= \mathbf{AFF}(\mathbf{MHAB}(F_{in}), \mathbf{CoT}(F_{in})), \\ &= \sigma(\theta) * (\mathbf{MHAB}(F_{in}) + (1 - \sigma(\theta)) * \mathbf{CoT}(F_{in}), \end{aligned} \tag{10}$$

F_{in} denotes the input features, F_{out} is the adaptive mixing results from MHA and CoT block. Considering that the BN layer will destroy the internal features of the sample, we use the IN layer to replace the BN layer in the improved CoT Block and use ELU as the activation function.

3.2   Shallow Layers

The shallow layers are stacked by several Multi-branch Hybrid Attention modules with CoT [18] blocks. We use the shallow layers to generate the pseudo-haze-free image, which has high-level semantic information.

Tail Module. As shown in Fig. 2 (c), we design the Tail module, which fuses the extracted features and restores the hazy image. The *tanh* activation function is often used on degradation reconstructions. Therefore, we use that as the activation of the output after a stack of 3x3 convolution.

Architecture of Shallow Layers. As shown in Fig. 1, we use the shallow layers to reconstruct the degraded image context content. In order to effectively reduce the amount of calculation and expand the receptive field of the convolution of MHAC, we utilize 2 convolutions with a stride of 2 to reduce the resolution of the feature map to 1/4 of the original input firstly; each convolution follows a SHA module. Then, we use a stack of 8 MHAC blocks with 256 channels, which has 2 skip connections to introduce shallow features before up-sampling, and utilize the Tail module for obtaining the residual of the restored image of the shallow layers:

$$S(x) = \mathbf{Shallowlayers}(x) + x, \tag{11}$$

wherein $S(x)$ denotes the pseudo-haze-free image, x denotes the hazy input image.

3.3 Explicit Model of Haze Density - Haze Density Encoding Matrix

Previous methods mainly focus on modeling haze density in an implicit manner. Methods based on physical model focus on obtaining an approximate estimation of haze distribution with direct supervision (e,g. transmission maps, K estimation) , which is limited with their inherent incompleteness in haze density modeling and their vulnerability in error that is introduced by degraded sub-tasks. Methods based on implicit constraints mainly rely on a complicated regularization term of loss functions, which only boosts the perception ability of the feature-level, ignoring the hidden spatial clue in the image-level. Some approaches based on explicit modeling haze density from haze images, lacks careful consideration of the mismatching between explicit image-level information and implicit feature-level distribution, resulting in sub-optimal dehazing results.

The proposed haze density encoding matrix addresses these problems from these aspects: Firstly, the haze density encoding matrix is an attention map that shares the shape prior of haze degradation. Secondly, the haze density encoding matrix, which is derived from haze image, encodes spatial clue from image-level to feature-level, bridging the gap between the channel mismatching between image space and feature space . Thirdly, the haze density encoding matrix is fully optimized with the network, which does not require any direct supervision, avoiding the incompleteness of hand-crafted priors.

As is mentioned in Fig. 2, the SHA mechanism perceives haze density effectively with its spatial-channel operation. It is obvious that shallow layers have the ability to generate sub-optimal results. The proposed density estimation module encodes spatial clues in the image-level information by a concatenation of input haze image I(x) and sub-optimal haze-free image $\bar{J}(x)$. The formulas are as follows:

$$F(x) = G(I(x); \bar{J}(x)) \tag{12}$$

The simple convolution network encodes the spatial clues in image-level to a 64-channel feature map with a 3x3 convolution after utilizing the Reflected Padding
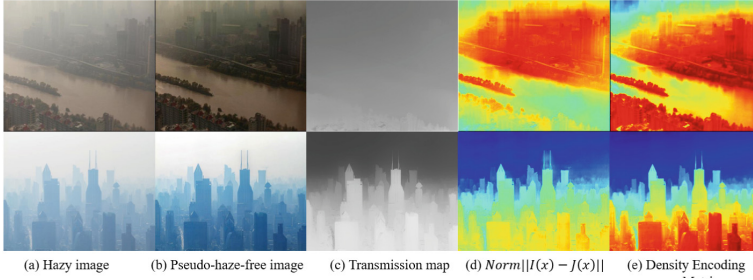
(a) Hazy image   (b) Pseudo-haze-free image   (c) Transmission map   (d) $Norm\|I(x) - J(x)\|$ (e) Density Encoding

Fig. 3. Visualization of different representations of haze density. The transmission map in (c) are predicted by DehazeNet [6]. The $Norm\|I(x) - J(x)\|$ [12] in (d) is the normalized result of the difference map. Our Density Encoding Matrix is shown in (e).
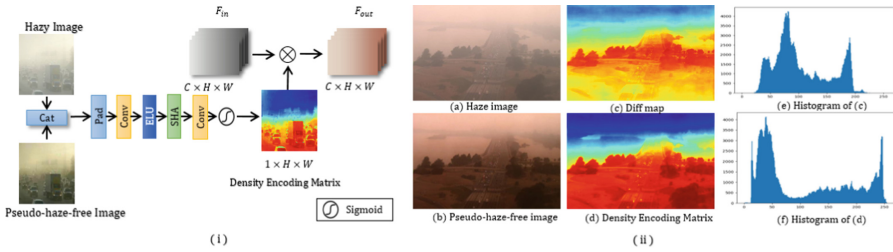

Fig. 4. (i) An illustration of the pipeline that generates the density encoding matrix. (ii) An illustration of the histogram of the difference map and density encoding matrix. The difference map is the numerical difference between the pseudo-haze-free images and haze images. The value of the density encoding matrix is mapped from [0, 1] to [0, 255]. The difference map and density encoding matrix are visualized by ColorJet for observation and comparison. Note that the histograms of the difference map and density encoding matrix have similar intensity distributions.

to avoid the detail loss of the feature edge. Afterwards, we utilize the SHA module to explore perceiving the uneven degeneration of input features fully, and finally use a convolution operation to compress the shape of the feature. The sigmoid function is used to get the density encoding matrix $M \in \mathbb{R}^{1 \times H \times W}$. After getting the density encoding matrix M, we multiply M by the input feature $F_{in} \in \mathbb{R}^{C \times H \times W}$ to get the final output $F_{out} \in \mathbb{R}^{C \times H \times W}$:

$$F_{out} = F_{in} \otimes M, \tag{13}$$

3.4 Deep Layers

To preserve abundant pixel-level texture details, we utilize the deep layers with the supervised signal from density encoding matrix. As shown in Fig. 1, we utilize 10 MHA blocks with 16 channels to extract features at the resolution of the original input. In order to avoid unnecessary calculations caused by repeated

extraction of features, we use the AFF module to introduce and mix features refined by our density encoding matrix from the shallow layers adaptively.

4   Experiments

4.1   Datasets and Metrics

We choose the PSNR and SSIM as experimental metrics to measure the performance of our network. We perform comprehensive evaluation on both synthetic datasets and real datasets.

To evaluate performance on synthetic datasets, we train on two large synthetic datasets, RESIDE [17] and Haze4K [20], and testing on SOTS [17] and Haze4K [20] testing sets, respectively. The indoor training set of RESIDE [17] contains 1,399 clean images and 13,990 hazy images generated by corresponding clean images. The indoor testing set of SOTS contains 500 indoor images. The outdoor testing set of SOTS contains 500 outdoor images. The training set of Haze4K [20] contains 3,000 hazy images with ground truth images and the testing set of Haze4K [20] contains 1,000 hazy images with ground truth images.

To evaluate performance on real-world datasets, i.e., Dense-Haze [1], NH-HAZE [3] and O-HAZE [2], we train the network following the official train-val split for fair comparison with other methods.

4.2   Loss Function

We only use Charbonnier loss [7] as our optimization objective:

$$\mathcal{L}(\Theta) = \mathcal{L}_{\text{char}}\left(S(x), J_{gt}(x)\right) + \mathcal{L}_{\text{char}}\left(D(x), J_{gt}(x)\right)), \tag{14}$$

where Θ denotes the parameters of our network, the $S(x)$ denotes pseudo-haze-free image, $D(x)$ denotes output of deep layers, which is the final output image, J_{gt} stands for ground truth, and $\mathcal{L}_{\text{char}}$ is the Charbonnier loss [7]:

$$\mathcal{L}_{\text{char}} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\left\|X^i - Y^i\right\|^2 + \epsilon^2}, \tag{15}$$

with constant ϵ emiprically set to $1e^{-3}$ for all experiments.

4.3   Ablation Study

To demonstrate the effectiveness of separated hybrid attention mechanism, we design a baseline network with the minimized design. The network consists of one convolution with kernel size of 3 and stride of 2, followed by four residual blocks, one upsample layer and one tail module. The baseline network is trained and evaluated on the Haze4K [20] dataset. We employ the Charbonnier loss [7] as the training loss function for ablation study, and utilize Haze4K [20] dataset for both training and testing. Detailed implementation of the baseline model is demonstrated in the supplementary material.

Effectiveness of Separated Hybrid Attention Module. In order to make a fair comparison between attention modules on implicit perceiving haze density, we replace the residual blocks in our baseline model with different kinds of attention blocks. The performance of above models are summarized in Table 1. We provide visualization results of high-level feature maps in the supplementary material.

Table 1. Comparisons on Haze4K [20] testset for different configurations of Shallow Layers. The number of parameters and flops is calculated only on the attention modules.

Settings	Model	# of Params	# of Flops	PSNR
1	Baseline	–	–	24.40
2	SE	512	520	24.71
3	ECA	3	114	24.78
4	CBA	4.9 K	78.69 M	25.36
5	FA	8.25 K	4.6 M	25.30
6	SWRCA	10.3 K	302 M	25.87
7	SHA (Ours)	2.6 K	368 K	26.39
8	MHAB (Ours)	22.2 K	321 M	27.02
9	MHA-C (Ours)	38.3 K	587.3 M	27.58

As is depicted in Table 1, we can see that attention modules boost the performance of image dehazing network. Comparing with the models of Setting 1, Setting 2, Setting 3, haze specific attention mechanisms gain a noticeable performance boost. Unfortunately, the performance of Setting 5 is worse than the performance of Setting 4, indicate the brute design of Setting 5 is not optimal. Compared with Setting 5 and Setting 6, the proposed modules obtain higher PSNR scores while maintaining the appropriate efficiency. The key to the performance lies in the mutual interaction between spatial and channel dimensions, and the extracted positional-sensitive features is effective in perceiving haze density. Furthermore, We find that MHAC and MHAB can achieve better performance, but the number of parameters also increases. Therefore, the careful design of shallow layers and deep layers is essential.

Effectiveness of Density Encoding Matrix. Compared with previous methods that explores haze density from a variety of forms, we perform an experiment on validating the effectiveness of the proposed Density Encoding Matrix. Since the haze density representation can be obtained by attaching direct supervisions on the density estimation module, we compare different forms of haze density. First, we replace the haze density encoding matrix (as is denoted with doted orange line in Fig. 1) with the identity matrix as the baseline. Second, We replace the haze density encoding matrix with a ground-truth transmission

map provided with Haze4K [20] as the validation of the effectiveness of classical model derives from the physical model. Third, we modify the Density Estimation Module with an extra transmission map supervision, enabling the output of online estimation of transmission map. As is depicted in KDDN, a normed difference map is introduced as another unsupervised haze density representations. Finally, the proposed Density Encoding Matrix is optimized without any direction supervision.

As is depicted in Table 2, we can observe that different kinds of density representations boost the performance of the dehazing network than the identity matrix, which indicates that the ability of the implicit perception of haze distribution is limited and it is necessary to introduce external image-level guidance to the dehazing network. The comparison between Setting 2 and Setting 3 indicates that methods with direct supervisions are vulnerable to error introduced by the auxiliary task, which suffer from more severe performance degradation in real-world scenes. Compared with Settings 2, 3 and Setting 4, the online estimation of the normalized difference map achieves a higher score than online estimation of the transmission map, which indicates no direct constraint or joint optimization allows better performance in feature-level. Unfortunately, it suffers from serious performance degradation than ground-truth transmission map, which indicates that image-level spatial distribution is mismatched with feature-level distribution. The proposed density encoding matrix achieves the highest score than its competitors, which adopts a fully end-to-end manner in training, shares a shape prior with haze degradation model and aligns the uneven image-level distribution with feature-level distribution. In addition, we provide visualization results of different representations of haze density.

Table 2. Comparison with different forms of haze representations.

Settings	Density representations	Online inference/Supervision Type	PSNR
1	Identity matrix	Offline/-.	30.33
2	GT Transmission Map	Offline/Sup.	31.42
3	Est. Transmission Map	Online/Sup.	31.17
4	Norm $\|I(x) - J(x)\|$ [12]	Online / UnSup	32.21
5	Density Encoding Matrix	Online / UnSup	33.49

(a) Groud truth   (b) Hazy   (c) AOD-Net   (d) MSBDN   (e) FFA-Net   (f) Ours   (g) Density Encoding Matrix
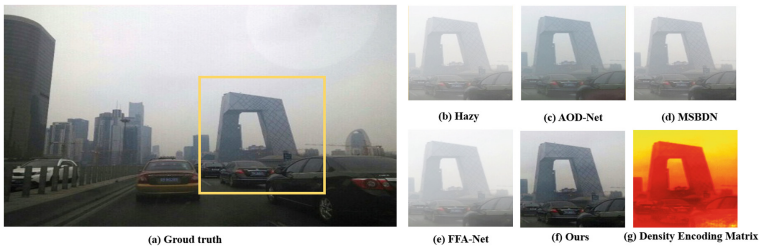
Fig. 5. Visual comparison of dehazing results of one image from Haze4k [20] dataset.

5   Compare with SOTA Methods

5.1   Implementation Details

We augment the training dataset with randomly rotated by 90, 180, 270 degrees and horizontal flip. The training image patches with the size 256×256 are extracted as input I_{in} of our network. The network is trained for 7.5×10^5, 1.5×10^6 steps on Haze4K [20] and RESIDE [17] respectively. We use the Adam optimizer with initial learning rate of 2×10^{-4}, and adopt the CyclicLR to adjust the learning rate, where on the triangular mode, the value of gamma is 1.0, base momentum is 0.8, max momentum is 0.9, base learning rate is initial learning rate and max learning rate is 3×10^{-4}. PyTorch is used to implement our model with 4 RTX 3080 GPUs with total batchsize of 40.

(a) hazy   (b) DehazeNet   (c) AOD-Net   (d) MSBDN   (e) FFA-Net   (f) AECR-Net   (g) Ours   (h) ground truth (i) Density Encoding Matrix

Fig. 6. Visual comparison of dehazing results on Haze4k [20] dataset.

(a) hazy   (b) DehazeNet   (c) AOD-Net   (d) MSBDN   (e) FFA-Net   (f) AECR-Net   (g) Ours   (h) Density Encoding Matrix
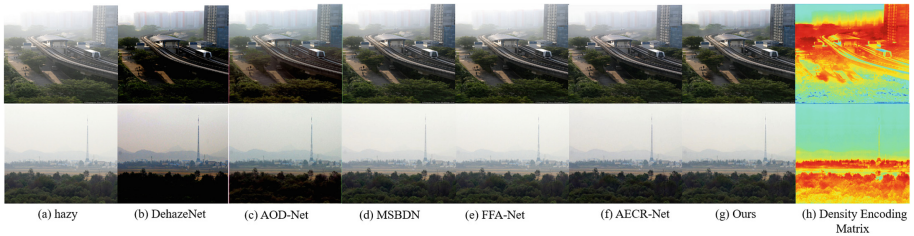
Fig. 7. Visual comparision of dehazing results on real-world images.

5.2   Qualitative and Quantitative Results on Benchmarks

Visual Comparisons. To validate the superiority of our method, as shown in Fig. 5, 6, 7, firstly we compare the visual results of our method with previous SOTA methods on synthetic hazy images from Haze4K [20] and real-world hazy images. It can be seen that other methods are not able to remove the haze in all the cases, while the proposed method produced results close to the real clean scenes visually. Additionally, the visualization results of the density encoding matrix demonstrates the ability of capturing uneven haze distribution. Our method is superior in the recovery performance of image details and color fidelity. Please refer to the supplementary materials for more visual comparisons on the synthetic hazy images and real-world hazy images.

Quantitative Comparisons. We quantitatively compare the dehazing results of our method with SOTA single image dehazing methods on Haze4 K [20], SOTS [17] datasets, Dense-Haze [1], NH-Haze [3] and O-HAZE [2]. As shown in Table 3, our method outperforms all SOTA methods, achieving 33.49 dB PSNR and 0.98 SSIM on Haze4 K [20]. It increases the PSNR by 4.93 dB compared to the second-best method. On the SOTS [17] indoor test set, our method also outperforms all SOTA methods, achieving 38.41 dB PSNR and 0.99 SSIM. It increases the PSNR by 1.24 dB, compared to the second-best method. Our method also outperforms all SOTA methods on the SOTS [17] outdoor test set, achieving 34.74 dB PSNR and 0.97 SSIM. In conclusion, our method achieves the best performance on the 6 synthetic and real-world benchmarks compared to previous methods.

Table 3. Quantitative comparisons of our method with the state-of-the-art dehazing methods on Haze4K [20] and SOTS [17] datasets (PSNR(dB)/SSIM). Best results are underlined.

Method	Haze4K [20]		SOTS Indoor [17]		SOTS Outdoor [17]		Dense-Haze [1]		NH-HAZE [3]		O-HAZE [2]	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DCP [11]	14.01	0.76	15.09	0.76	19.13	0.8148	10.06	0.39	10.57	0.52	16.78	0.653
DehazeNet [6]	19.12	0.84	20.64	0.80	20.29	0.88	13.84	0.43	16.62	0.52	17.57	0.77
AOD-Net [16]	17.15	0.83	19.82	0.82	24.14	0.92	13.14	0.41	15.40	0.57	15.03	0.54
GDN [19]	23.29	0.93	32.16	0.98	30.86	0.98	–	–	–	–	23.51	0.83
MSBDN [8]	22.99	0.85	33.79	0.98	23.36	0.88	15.37	0.49	19.23	0.71	24.36	0.75
FFA-Net [21]	26.96	0.95	36.39	0.98	33.57	0.98	14.39	0.45	19.87	0.69	22.12	0.77
AECR-Net [23]	–	–	37.17	0.99	–	–	15.80	0.47	19.88	0.72	–	–
DMT-Net [20]	28.53	0.96	–	–	–	–	–	–	–	–	–	–
Ours	33.49	0.98	38.41	0.99	34.74	0.99	16.79	0.51	20.42	0.73	24.64	0.83

6   Conclusion

In this paper, we propose a powerful image dehazing method to recover haze-free images directly. Specifically, the Separable Hybrid Attention is designed to better perceive the haze density, and a density encoding matrix is to further refine extracted features. Although our method is simple, it is superior to all the previous state-of-the-art methods with a very large margin on two large-scale hazy datasets. Our method has a powerful advantage in the restoration of image detail and color fidelity. We hope to further promote our method to other low-level vision tasks such as deraining, super-resolution, denoising and desnowing.

Limitations: The proposed method recovers high-fidelity haze-free images in common scenarios. However, as is shown in Fig. 7 (top), the dense hazy area with high-light scenes still has residues of haze. Following the main idea of this work, future research can be made in various aspects to generate high-quality images with pleasant visual perceptions.

References

1. Ancuti, C.O., Ancuti, C., Sbert, M., Timofte, R.: Dense haze: a benchmark for image dehazing with dense-haze and haze-free images. In: IEEE International Conference on Image Processing (ICIP). IEEE ICIP 2019 (2019)
2. Ancuti, C.O., Ancuti, C., Timofte, R., Vleeschouwer, C.D.: O-haze: a dehazing benchmark with real hazy and haze-free outdoor images. In: IEEE Conference on Computer Vision and Pattern Recognition, NTIRE Workshop. NTIRE CVPR'18 (2018)
3. Ancuti, C.O., Ancuti, C., Vasluianu, F.A., Timofte, R.: Ntire 2020 challenge on nonhomogeneous dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 490–491 (2020)
4. Ancuti, C.O., Ancuti, C., Vasluianu, F.A., Timofte, R.: Ntire 2021 nonhomogeneous dehazing challenge report. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 627–646 (2021)
5. Berman, D., Avidan, S., et al.: Non-local image dehazing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1674–1682 (2016)
6. Cai, B., Xu, X., Jia, K., Qing, C., Tao, D.: Dehazenet: an end-to-end system for single image haze removal. IEEE Trans. Image Process. **25**(11), 5187–5198 (2016)
7. Charbonnier, P., Blanc-Feraud, L., Aubert, G., Barlaud, M.: Two deterministic half-quadratic regularization algorithms for computed imaging. In: Proceedings of 1st International Conference on Image Processing, vol. 2, pp. 168–172. IEEE (1994)
8. Dong, H., et al.: Multi-scale boosted dehazing network with dense feature fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2157–2167 (2020)
9. Fattal, R.: Dehazing using color-lines. ACM Trans. Graph. **34**(1), 13:1–13:14 (2014)
10. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3146–3154 (2019)
11. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. **33**(12), 2341–2353 (2010)
12. Hong, M., Xie, Y., Li, C., Qu, Y.: Distilling image dehazing with heterogeneous task imitation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3462–3471 (2020)
13. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
14. Jiang, Y., Sun, C., Zhao, Y., Yang, L.: Image dehazing using adaptive bi-channel priors on superpixels. Comput. Vis. Image Underst. **165**, 17–32 (2017)
15. Katiyar, K., Verma, N.: Single image haze removal algorithm using color attenuation prior and multi-scale fusion. Int. J. Comput. Appl. **141**(10), 037–042 (2016). https://doi.org/10.5120/ijca2016909827

16. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: all-in-one dehazing network. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4770–4778 (2017)
17. Li, B.: Benchmarking single-image dehazing and beyond. IEEE Trans. Image Process. **28**(1), 492–505 (2018)
18. Li, Y., Yao, T., Pan, Y., Mei, T.: Contextual transformer networks for visual recognition. arXiv preprint arXiv:2107.12292 (2021)
19. Liu, X., Ma, Y., Shi, Z., Chen, J.: Griddehazenet: attention-based multi-scale network for image dehazing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7314–7323 (2019)
20. Liu, Y., et al.: From synthetic to real: image dehazing collaborating with unlabeled real data. arXiv preprint arXiv:2108.02934 (2021)
21. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: feature fusion attention network for single image dehazing. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11908–11915 (2020)
22. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
23. Wu, H., et al.: Contrastive learning for compact single image dehazing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10551–10560 (2021)
24. Zamir, S.W., et al.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14821–14831 (2021)