Mutual Information-driven Triple Interaction Network for Efficient Image Dehazing

Hao Shen
Hefei University of Technology, Hefei, China
haoshenhs@gmail.com

Zhong-Qiu Zhao*†‡
Hefei University of Technology, Hefei, China
z.zhao@hfut.edu.cn

Yulun Zhang
ETH Zürich, Zürich, Switzerland
yulun100@gmail.com

Zhao Zhang*
Hefei University of Technology, Hefei, China
cszzhang@gmail.com

ABSTRACT

Multi-stage architectures have exhibited efficacy in image dehazing, which usually decomposes a challenging task into multiple more tractable sub-tasks and progressively estimates latent hazy-free images. Despite the remarkable progress, existing methods still suffer from the following shortcomings: (1) limited exploration of frequency domain information; (2) insufficient information interaction; (3) severe feature redundancy. To remedy these issues, we propose a novel Mutual Information-driven Triple interaction Network (MITNet) based on spatial-frequency dual domain information and two-stage architecture. To be specific, the first stage, named amplitude-guided haze removal, aims to recover the amplitude spectrum of the hazy images for haze removal. And the second stage, named phase-guided structure refined, devotes to learning the transformation and refinement of the phase spectrum. To facilitate the information exchange between two stages, an Adaptive Triple Interaction Module (ATIM) is developed to simultaneously aggregate cross-domain, cross-scale, and cross-stage features, where the fused features are further used to generate content-adaptive dynamic filters so that applying them to enhance global context representation. In addition, we impose the mutual information minimization constraint on paired scale encoder and decoder features from both stages. Such an operation can effectively reduce information redundancy and enhance cross-stage feature complementarity. Extensive experiments on multiple public datasets exhibit that our MITNet performs superior performance with lower model complexity. The code and models are available at https://github.com/it-hao/MITNet.

CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**; **Computer vision**; **Computer vision tasks**.

---

*Corresponding authors.
†Also with Intelligent Interconnected Systems Laboratory of Anhui Province.
‡Also with Guangxi Academy of Sciences.

KEYWORDS

Image dehazing, mutual information, adaptive triple interaction block, spatial-frequency domain information

1 INTRODUCTION

Hazy images usually cause the degradation of visual quality [16, 43] in object appearance, contrast, and color distortion, leading to significant performance drops in subsequent high-level vision tasks such as scene understanding [38], semantic segmentation [36], and object detection [29]. Since these tasks urgently require clean haze-free images, academic and industry communities, thereby, have focused on single image dehazing, which attempts to extract latent clean images from hazy ones.

Benefit from the important breakthrough of deep learning in computer vision tasks [26, 44, 45, 48, 52, 56], data-driven image restoration [10, 11, 60, 61] and image dehazing methods [9, 12, 14, 17, 33, 47, 50, 54] have achieved superior performance. For the image dehazing task, one can divide these methods into two categories: physics-free methods and physics-aware methods. The former first calculates the transmission map and atmospheric light independently before employing the atmospheric scattering model (ASM) [31] to produce haze-free images. The latter tries to explore a mapping from hazy images to clean counterparts directly in an end-to-end manner. Among them, multi-stage architectures [8, 35, 54, 55] play a significant role in performance improvement. The basic idea is to decompose a challenging task into multiple easy sub-tasks and adopt progressive learning. Despite the remarkable advancement, there still exist some issues.

(1) *Limited exploration of frequency domain information.* Most existing image dehazing methods [8, 35, 53, 55] focus exclusively on exploiting spatial features but fail to sufficiently leverage frequency discrepancies, such as the intrinsic prior of the physical properties of hazy images, thus limiting performance gains. As revealed in [54], the degraded haze information is almost displayed in the amplitude component, while the disparity between phase components of the corresponding hazy and clean images is small.
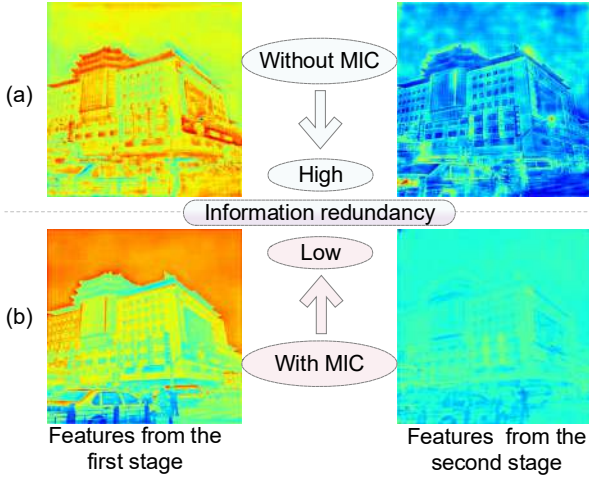
Figure 1: The feature comparisons before and after applying mutual information constraint (MIC).

Given this discovery, how to embed frequency domain prior information into spatial features and effectively incorporate them into a unified dehazing network is an open issue.

(2) ***Insufficient information interaction.*** Current multi-stage architectures mainly consist of "early fusion" and "cross-stage fusion" designs. The first category [35] attempts to directly fuse the previous stage's output and the current stage's input at the beginning layer, which ignores the propagation of intermediate features from earlier to later stages. Thus, the complementary information will be gradually weakened with the increasing network depths. The second category [8, 54] fully integrates the contextually-enriched features from one stage to the next, easing the information loss caused by the encoder-decoder's repetitive up- and down-sampling operations. However, the fusion scheme only considers the interaction between paired scale features of two stages, ignores the exchange of cross-scale information of in-stage and cross-stage, and thereby fails to provide more precise and contextually enriched feature representations. Besides, recent research [22, 27, 41, 59] indicates the function of neurons should be modified adaptively in response to behavioral context. Thereby, it is vital to modify the convolution operation based on context information dynamically.

(3) ***Severe feature redundancy.*** Although performance can be improved by effective fusion strategies, existing dehazing methods do not explicitly enforce complementary information learning between different stage features, resulting in redundant features. We can observe from Figure 1 (a) that the learned features from the two stages have similar textures without any constraints. Conversely, information redundancy is reduced after endowing the mutual information constraint for two-stage features, as shown in Figure 1 (b). Thus, it is desirable to enforce complementary feature learning by adopting explicit constrain.

In this work, firstly, a simple yet effective two-stage network is designed by simultaneously processing the spatial-frequency dual domain information. The first stage is designed to recover the amplitude spectrum of the hazy images for haze removal, and the second stage learns the transformation and refinement of the

phase spectrum. To implement the network, we combine spatial-frequency information to construct the customized feature extraction block, *e.g.,* residual amplitude block for the first stage and residual phase block for the second stage, respectively. Secondly, we propose an adaptive triple interaction module (ATIM) to enhance information exchange. Specifically, we design a triple interaction solution to achieve cross-domain, cross-scale, and cross-stage feature integration, and then an adaptive dynamic filter block is developed to generate dynamic filters based on the fused features. The produced filters next are convolved with the decoder features of the second stage for representation capability enhancement. Since the generated filters are conditioned on the input features guided by Fourier prior and spatial information, the network can flexibly adapt to image contents. After being equipped with the two-stage architecture and the proposed ATIM, diverse frequency domain information, and rich spatial contextual information are exploited. To learn complementary information and alleviate feature redundancy from two stages, we introduce the mutual information minimization constraints on paired scale encoder and decoder features from two stages. Overall, the main contributions of this paper can conclude as follows:

(1) We propose a novel dehazing network termed MITNet, based on spatial-frequency dual domain information and two-stage architecture, which simplifies the challenging dehazing problem into two more manageable sub-tasks and embraces the advantages of high performance and lower model complexity.

(2) We design an adaptive triple interaction module (ATIM) capable of aggregating cross-domain, cross-scale, and cross-stage features and generating content-adaptive filters to enhance global context representation.

(3) We introduce the mutual information minimization constraint on the paired scale encoder and decoder features from two stages to reduce information redundancy and enhance cross-stage feature complementarity.

(4) Comprehensive experiments on several public benchmarks verify the superiority and strong generalization. Further, our method achieves an excellent model complexity versus performance trade-off.

2 RELATED WORK

2.1 Single Image Dehazing

Existing single image dehazing methods can be classified into prior-based and data-driven deep learning methods. The former [4, 7, 15, 18, 57, 66, 67] usually rely on the ASM or handcraft priors. However, these priors only work well in the scene that satisfies their assumptions.

With the introduction of sizeable hazy datasets [25, 30], a plethora of data-driven approaches have been proposed rapidly in the past years. Early methods [6, 24] usually design a network to estimate the transmission map and atmospheric light. However, inaccurate prediction for the atmospheric light and transmission map may lead to performance degradation. Recent research [14, 17, 28, 33–35, 50, 51, 54] has concentrated on recovering the clean image directly from the hazy image without the need for a physical model. GFN [35] proposes a revolutionary fusion-based strategy for generating haze-free images. FFA-Net [33] introduces a channel attention
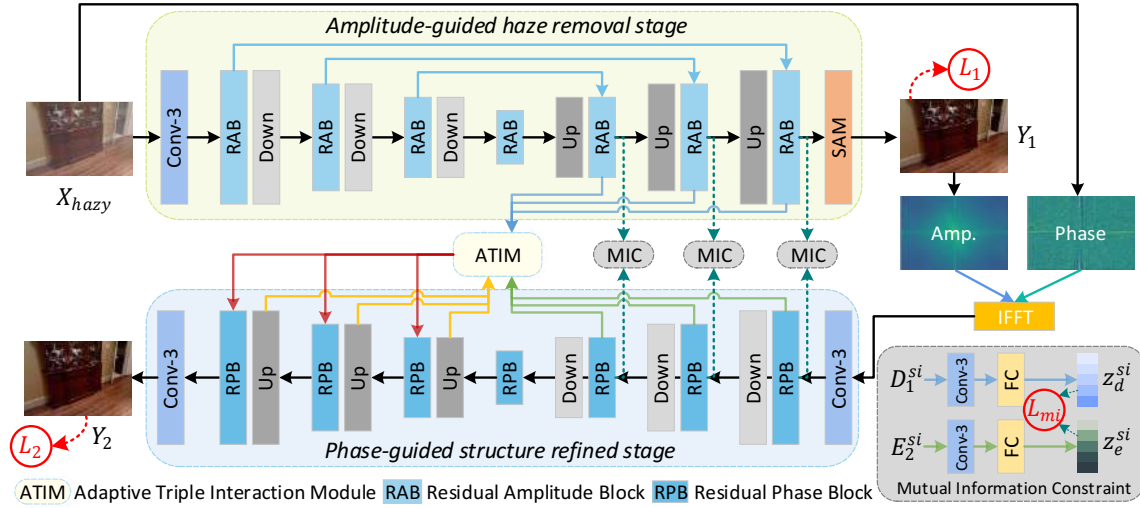
Figure 2: The detailed architecture of our Mutual Information-driven Triple interaction Network.

and pixel attention mechanism for constructing a feature attention block that can handle many forms of input. MSBDN [14] utilizes boosting and error feedback mechanisms to enhance the feature fusion. AECR-Net [50] exploits negative and positive image information via contrastive learning, achieving great network parameters versus performance trade-offs. Dehamer [17] combines CNN and Transformer to fuse local representation and global context modelling capability perfectly to achieve haze removal. However, these methods only utilize the spatial domain features but fail to take advantage of the frequency domain information. FSDGN [54] uses spatial-frequency information to achieve fast inference speed but fails to achieve the complementarity of cross-stage features.

In addition, current mainstream image restoration methods, including image dehazing [54], image denoising [41], image inpainting [63, 64], and image deblurring [62] adopt U-Net [37] architecture to pursue the model's efficiency. However, these methods neglect the interaction of cross-scale features. In contrast, full-resolution methods [33, 40] can obtain fine-grained features but fail to exploit contextual information. This paper inherits the benefits of U-Net and full-resolution architectures to develop a two-stage network.

2.2 Mutual Information Learning

Mutual information has deep connections with representation learning [3], aiming to capture the most important features similar to the input ideally. By extracting global features from the full image and local features from image patches, DIM [19] estimates and maximizes mutual information. CPC [32] enables mutual information maximization between the context and features extracted from different sequential data components. CMC [46] employs contrastive learning in a multiview context, seeking to maximize mutual information across representations of different views of the same scene. Based on the empirical of multiview features, Info3d [39] enables mutual information maximization between 3D objects and the corresponding geometric converted equivalents based on empirical multiview features to improve representations. CMINet [58] specifically uses multi-modal features in color images and depth data through mutual information learning, removing redundant features

for effective multi-modal learning. To learn mutual information on diverse image modalities, [65] implements mutual information minimization across PAN and MS features. Our proposed two-stage network explores different frequency domain characteristics in each stage and extracts rich spatial contextual features in both stages. Thus, common and diverse features exist in separate stages. To this end, we also apply mutual information as a constraint to alleviate the feature redundancy.

3 METHODS

3.1 Network Architecture

It is well known that the degradation information in low-quality images predominantly displays in the amplitude spectrum. Thus, we utilize the insight wisely and disentangle the dehazing procedure into a two-stage process so as to achieve haze degradation removal and preserve image structure information. Next, we will go through the operation and properties of the Fourier transformation.

For a image $X_{img} \in R^{H \times W \times 1}$ with single channel, its Fourier transformation operation [54] $\mathcal{F}$ can be expressed as:

$$\mathcal{F}(X_{img})(i, j) = \frac{1}{\sqrt{HW}} \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} X_{img}(h, w) e^{-k2\pi \left(\frac{h}{H} i + \frac{w}{W} j \right)}. \quad (1)$$

Accordingly, $\mathcal{F}^{-1}$ is the inverse Fourier transformation, k represents the imaginary unit, i and j denote the horizontal and vertical coordinates. Following with FFT algorithm [5], the amplitude $\mathcal{A}(X_{img})$ and phase $\mathcal{P}(X_{img})$ information can be defined as:

$$\mathcal{A}(X_{img})(i, j) = \sqrt{Real^2(X_{img})(i, j) + Imag^2(X_{img})(i, j)},$$
$$\mathcal{P}(X_{img})(i, j)) = \arctan \left[\frac{Imag(X_{img})(i, j)}{Real(X_{img})(i, j)} \right], \quad (2)$$

where $Imag(X_{img})$ and $Real(X_{img})$ denote the imaginary and real component of $\mathcal{F}(X_{img})$. Here, it is worth emphasizing that the Fourier operation is computed alone in each channel for feature maps or color images.
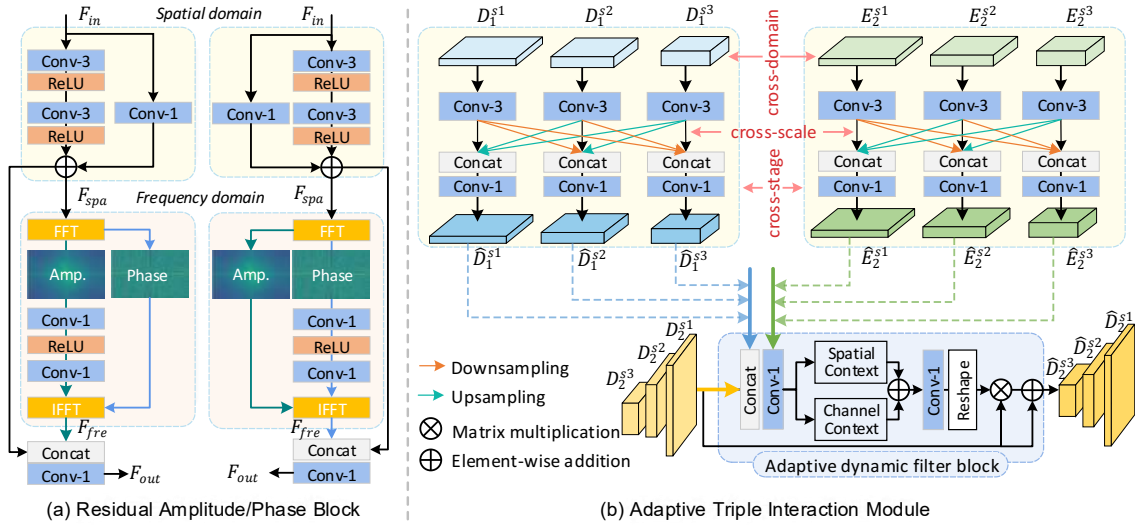
Figure 3: The detailed architecture of (a) residual amplitude/phase block and (b) adaptive triple interaction module.

The entire network, as shown in Figure 2, contains an ***amplitude-guided haze removal stage*** and a ***phase-guided structure refined stage***. The former restores the amplitude of hazy images to remove haze, and the latter restores the phase information to refine fine-grained background structures. Let X_{hazy} and X_{gt} denote hazy and ground truth images, Y_1 and Y_2 represent the outputs of two stages. Since each stage is designed by a task-specific sub-network, both the input and supervision signals should be different. For the first stage, the hazy images X_{hazy} are served as input, $\mathcal{F}^{-1}(\mathcal{A}(X_{gt}), \mathcal{P}(X_{hazy}))$ and $\mathcal{A}(X_{gt})$ are used to supervise the learning of amplitude representation. For the second stage, the X_{gt} are used as the supervision signal, but we do not directly use Y_1 as the input of the second stage. Instead, we use $\mathcal{F}^{-1}(\mathcal{A}(Y_1), \mathcal{P}(X_{hazy}))$ to ensure the retention of original phase information.

To achieve an efficient target, we use encoder-decoder-like (or U-Net-like) [37] design for each stage, consisting of seven basic units (*e.g.,* residual amplitude block (RAB) and residual phase block (RPB)), three downsampling blocks, and three upsampling blocks in each stage. In the first stage, except for the lowest scale, each scale contains a skip connection between the encoder and decoder. However, considering that the decoder of the second stage integrates features from the previous stage and the current encoder by adopting the proposed fusion module, we do not take a similar approach. In addition, we introduce a supervised attention module (SAM) [55] to stabilize the optimization procedure.

We show the structure of RAB and RPB in Figure 3 (a). Taking the former as an instance, which has a spatial domain branch and a frequency domain branch for processing dual domain representations. The spatial domain branch employs a residual block with two 3×3 convolutional layers to process spatial information. After obtaining the spatial features F_{spa}, we firstly utilize the fast Fourier transformation (FFT) to obtain the amplitude spectrum $\mathcal{A}(F_{spa})$ and phase spectrum $\mathcal{P}(F_{spa})$ information. Then, the $\mathcal{A}(F_{spa})$ is fed into two 1×1 convolutional layers to obtain $\mathcal{A}'(F_{spa})$. Subsequently, we use the inverse FFT (IFFT) algorithm to map $\mathcal{A}'(F_{spa})$ and $\mathcal{P}(F_{spa})$ back to their image space and obtain the frequency features F_{fre}.

Finally, we adopt channel-wise concatenation followed with one 1×1 convolution to fully integrate the cross-domain features F_{spa} and F_{fre}. The RPB is similar to RAB, just swapping the operations between $\mathcal{A}$ and $\mathcal{P}$.

3.2 Adaptive Triple Interaction Module

The information exchange is an important ingredient in the two-stage design because simply passing the previous stage's output to the next stage neglects the potential intermediate features, but the original, useful shallow features will be gradually weakened with the increasing network depths. In our framework, we construct the ATIM to associate two stages by adequately incorporating cross-domain, cross-scale, and cross-stage features.

As for the cross-domain interaction, the proposed RAB and RPB employed in two stages fully utilize the spatial and frequency domain information to reconstruct the representation of low-frequency (contrast, illumination, and color) and structural components.

As for the cross-scale interaction, in the first stage's decoder and second stage's encoder, the ATIM integrates information from all scales, allowing for both top-down and bottom-up information flow. Further, this procedure combines features with different receptive fields, which enriches contextual representations. As shown in Figure 3 (b), the features $\{D_1^{si}\}_{i=1}^{3}$ in the decoder of the first stage firstly will be extracted by three independent 3×3 convolutions. Upsampling and downsampling operations will further resize the obtained resolution features. To achieve the combination of precisely spatial high-resolution features with low-resolution features containing rich contextual information, a simple channel concatenation operation followed by 1×1 convolution is used to fuse them. The resultant multi-scale features can be denoted as $\{\widehat{D}_1^{si}\}_{i=1}^{3}$. Similarly, one can obtain the fused features $\{\widehat{E}_2^{si}\}_{i=1}^{3}$ in the encoder of the second stage.

As for the cross-stage interaction, the ATIM is equipped with an adaptive dynamic filter block (ADFB), which generates weight filters with flexibility based on the fused cross-domain and cross-scale features and applies them to the decoder's features of the

second stage for representation capability enhancement. Taking the specific-scale feature set $\left\{\widehat{D}_1^{s1}, \widehat{E}_2^{s1}, D_2^{s1}\right\} \in R^{c \times h \times w}$ as an example. The ADFB first uses one 1×1 convolution to fuse three inputs and then embeds resulting features into lightweight spatial and channel context branches to enhance filter representations. The spatial context branch includes a 3×3 depthwise convolution [21], whereas the channel context branch consists of a pooling layer and a convolutional layer. To take full advantage of them, the element-wise addition operation is used to integrate them. In order to formulate spatially-varying filters, a 1×1 convolutional layer is utilized to produce feature maps whose dimension is $(k \times k \times c) \times h \times w$; we then reshape them into a series of per-pixel kernels $W_{i,j} \in \mathbb{R}^{k^2 \times c}$, where $i \in \{1, 2, \cdots, h\}$, $j \in \{1, 2, \cdots, w\}$. In this way, a series of content-adaptive filters for each location is learned. Finally, given the scale-specific features D_2^{s1}, the dynamic filter results can be written as:

$$\widehat{D}_2^{s1} = D_2^{s1} * W + D_2^{s1}, \tag{3}$$

where "$*$" indicates the convolution operation, $\widehat{D}_2^{s1}$ is the enhanced features in the decoder of the second stage. Here, we adopt residual learning to preserve more low-level features.

Here, we mainly emphasize two points. (1) The generated filters of ADFB are conditioned on the triple interaction features, thus being guided by the enriched spatial-frequency features (cross-domain), multi-scale contextual features (cross-scale), and deep-shallow features (cross-stage). Such a mechanism allows the ADFB to adapt flexibly to image contents. (2) The two stages explore different frequency domain characteristics and extract rich spatial contextual features, so redundant features still exist.

3.3 Mutual Information Constraint

Given the multi-scale features $\left\{D_1^{si}\right\}_{i=1}^3$ from the first stage's decoder and $\left\{E_2^{si}\right\}_{i=1}^3$ from the second stage's encoder, the first step is to transform them into low-dimensional feature vectors for feature embedding. In detail, the paired features of the same scale are inputted into two 3×3 convolutional layers. Then the obtained features are mapped using two fully connected layers ("FC" in Figure 2) to produce a low-dimensional feature vector set $\left\{v_d^{si}\right\}_{i=1}^3$ and $\left\{v_e^{si}\right\}_{i=1}^3$. After obtaining the embedded vectors v_d and v_e, the next step is to introduce the mutual information minimization constraint to mitigate feature redundancy and achieve the complementary learning of diverse features.

According to the information theory [23], the mutual information between v_d and v_e, with the joint entropy $G(v_d|v_e)$, and marginal entropies $G(v_d)$ and $G(v_e)$, can be expressed as:

$$I(v_d, v_e) = G(v_d) + G(v_e) - G(v_d|v_e). \tag{4}$$

Formally, following [39], the representation of KL divergence of these two variables is also given as:

$$\begin{aligned} K(v_d\|v_e) &= G_{v_e}(v_d) - G(v_d), \\ K(v_e\|z_d) &= G_{v_d}(v_e) - G(v_e), \end{aligned} \tag{5}$$

where $G_{v_e}(v_d)$ and $G_{v_d}(v_e)$ represent the cross-entropy. Thereby, combining Eq. 4 and Eq. 5, we can obtain:

$$\begin{aligned} I(v_d, v_e) = &G_{v_d}(v_e) + G_{v_e}(v_d) - G(v_d|v_e) \\ &- K(v_d\|v_e) - K(v_e\|v_d). \end{aligned} \tag{6}$$

Due to the non-negative property of $G(v_d|v_e)$, minimizing the mutual information is equivalent to optimizing the following formula:

$$L_{mi} = G_{v_d}(v_e) + G_{v_e}(v_d) - K(v_d\|v_e) - K(v_e\|v_d). \tag{7}$$

To enforce the mutual information constraint on paired scale-specific features, we use a multi-scale joint learning manner. By minimizing $I(v_e, v_d)$, one can explore the complementary characteristics of two stages effectively. Additionally, the MIC is only used in the training phase and thus does not affect the model's efficiency.

3.4 Loss Function

Since there are two outputs, Y_1 and Y_2, from our two-stage network, then the loss function L_1 and L_2 for both stages is:

$$L_1 = \underbrace{\left\| Y_1 - \mathcal{F}^{-1}(\mathcal{A}(X_{gt}), \mathcal{P}(X_{hazy})) \right\|_1}_{spatial} + \alpha \underbrace{\left\| \mathcal{A}(Y_1) - \mathcal{A}(X_{gt}) \right\|_1}_{frequency},$$
$$\tag{8}$$

$$L_2 = \underbrace{\left\| Y_2 - X_{gt} \right\|_1}_{spatial} + \beta \underbrace{\left\| \mathcal{F}(Y_2) - \mathcal{F}(X_{gt}) \right\|_1}_{frequency}. \tag{9}$$

Both terms in each equation above are performed on spatial and frequency domains, respectively. $\|\cdot\|$ denotes the Mean Absolute Error (MAE), α and β are the trade-off factor, and we set them as 0.05. As a consequence, the overall loss after combining with mutual information minimization loss is:

$$L = L_1 + L_2 + \gamma \sum_{i=1}^{3} L_{mi}\left(v_d^{si}, v_e^{si}\right). \tag{10}$$

As the L_{mi} is much larger than other terms, we empirically set the weight $\gamma = 0.001$ for balanced learning.

4 EXPERIMENTS

We will first describe the experimental setup in this section. Then, experimental results and a detailed ablation analysis of our proposed methods will be illustrated.

4.1 Experiment Setting

Datasets. For fair comparisons with existing dehazing methods, we utilize **ITS** and **OTS** subsets of RESIDE [25] dataset as the training data and evaluate performance on **SOTS** subset for synthetic image dehazing. For real image dehazing, two common real-world datasets, **Dense-Haze** [1] and **NH-Haze** [2], are adopted to test the robustness of our MITNet. It should be emphasized that we execute all evaluations on these datasets separately and do not use extra data to boost results.

Implementation Details. From the first to the fourth scale, the encoder and decoder have 20, 40, 80, and 160 channels, respectively. The upsampling and downsampling layers are implemented by transposed and strided convolution, respectively. Our model is implemented using the *PyTorch* library on two NVIDIA GeForce GTX 2080Ti GPUs. ADAM optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ is used in the whole training phase. We set the initial learning rate as 0.0002 and then linearly decrease it to half every 200 epoch. In addition, 16 patches of size 256×256 are cropped randomly as input images for each training. And in each mini-batch, we enrich
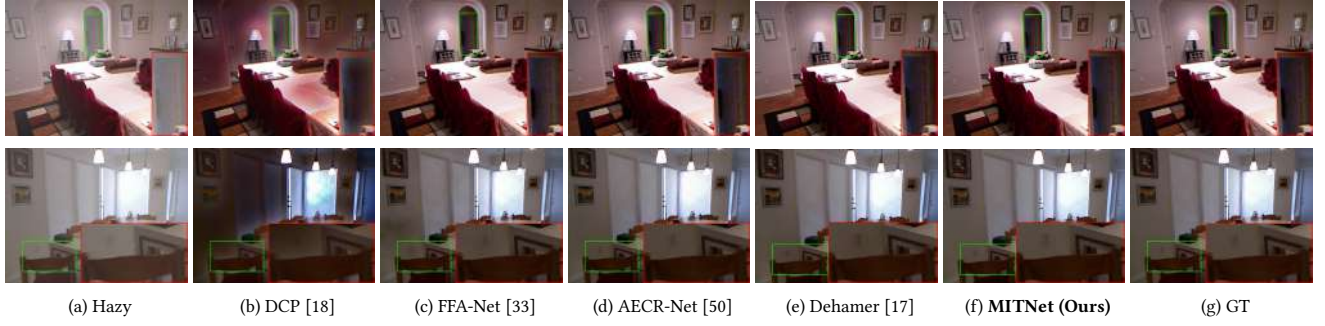
| (a) Hazy | (b) DCP [18] | (c) FFA-Net [33] | (d) AECR-Net [50] | (e) Dehamer [17] | (f) **MITNet (Ours)** | (g) GT |

Figure 4: Visual comparisons of SOTS [25] dataset by different methods.

and augment these patches to expand training samples by flipping horizontally or vertically and rotating 90°.

Comparison Methods and Evaluation Metrics. We compare our MITNet with four earlier method including DCP [18], DehazeNet [6], AOD-Net [24], GFN [35] and eight recent competing methods including FFA-Net [33], MSBDN [14], AECR-Net [50], UDN [20], PMDNet [53], Dehamer [17], MAXIM [47], and FSDGN [54]. Additionally, all of the compared results are provided by the authors or retrained by their available codes(*e.g.,* the real-world dehazing results on FSDGN [54]). Note that **(1)** we do not compare recent methods (*e.g.,* DehazeFormer [42] and SFNet[13]) with very high computational cost and network parameters; **(2)** AECR-Net [50] uses the last five images in the real-world training set for evaluation, leading to the reported results being inconsistent with the testing set; we utilize their provided pre-trained models to obtain the results on two testing sets. In addition, Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [49] are utilized for dehazing performance evaluation.

4.2 Comparison with State-of-the-art Methods

In this subsection, the results of several compared methods with ours on synthetic and real-world dehazing datasets are provided.

Evaluation on Synthetic Dataset. The quantitative evaluation of our approach and other methods on the SOTS dataset are listed in Table 1. As observed, our MITNet acquires the highest PSNR on the SOTS-indoor dataset, the second highest PSNR on the SOTS-outdoor dataset, and outperforms all other methods on the SSIM index. Specifically, compared to the Transformer-based method Dehamer and spatial-frequency guided method FSDGN, we achieve performance gains by 3.6 dB and 1.6 dB on the SOTS-indoor dataset. Although there is no performance improvement on the SOTS-outdoor compared to Dehamer [17], our method is more lightweight and efficient. We also show the visualized comparisons with representative methods on the SOTS dataset in Figure 4, from which we can see that the compared methods retain haze or produce color deviations. In contrast, our method preserves more details and involves fewer color distortions, resulting in the closest match to the original clean images and accurate haze removal.

Evaluation on Real-world Datasets. We also perform the evaluation on real-world datasets (*e.g.,* the Dense-Haze and NH-Haze) and present the quantitative results in Table 2. We all know that removing haze from real-world photographs is significantly more

Table 1: Quantitative comparisons of various dehazing methods on SOTS (indoor and outdoor) [25].

Method	Venue	SOTS-indoor		SOTS-outdoor	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DCP [18]	TPAMI'10	16.61	0.8546	19.14	0.8605
DehazeNet [6]	TIP'16	19.82	0.8209	27.75	0.9269
AOD-Net [24]	ICCV'17	20.51	0.8162	24.14	0.9198
GFN [35]	CVPR'18	22.30	0.8800	21.55	0.8444
FFA-Net [33]	AAAI'20	36.39	0.9886	33.57	0.9840
MSBDN [14]	CVPR'20	32.77	0.9812	34.81	0.9857
AECR-Net [50]	CVPR'21	37.17	0.9901	-	-
UDN [20]	AAAI'22	38.62	0.9909	34.92	0.9871
PMDNet [53]	ECCV'22	38.41	0.9900	34.74	0.9850
Dehamer [17]	CVPR'22	36.63	0.9881	**35.18**	0.9860
MAXIM-2S [47]	CVPR'22	38.11	0.9908	34.19	0.9846
FSDGN [54]	ECCV'22	38.63	0.9903	-	-
MITNet (Ours)	-	**40.23**	**0.9920**	**35.18**	**0.9881**

Table 2: Quantitative comparisons of various dehazing methods on Dense-Haze [1] and NH-HAZE [2].

Method	Venue	Dense-Haze		NH-HAZE	
		PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
DCP [18]	TPAMI'10	10.06	0.3854	10.57	0.5196
DehazeNet [6]	TIP'16	13.84	0.4252	16.62	0.5238
AOD-Net [24]	ICCV'17	13.14	0.4144	15.40	0.5693
GDN [28]	ICCV'19	13.31	0.3681	13.80	0.5370
FFA-Net [33]	AAAI'20	14.39	0.4524	19.87	0.6915
MSBDN [33]	CVPR'20	15.37	0.4858	19.23	0.7056
AECR-Net [50]	CVPR'21	14.88	0.5049	19.92	0.6717
Dehamer [17]	CVPR'22	16.62	0.5602	20.66	0.6844
FSDGN [54]	ECCV'22	16.91	0.5806	19.99	0.7086
MITNet (Ours)	-	**16.97**	**0.6056**	**21.26**	**0.7122**

challenging than removing haze from synthetic photographs due to the dense and nonhomogeneous haze distribution. However, the proposed MITNet surpasses all compared methods on both datasets. The visually compared images are also presented in Figure 5. We can see although other competitors successfully remove most of the haze, their results lose the colorfulness of images. Our method effectively removes the homogeneous or nonhomogeneous haze and reconstructs vivid colors.
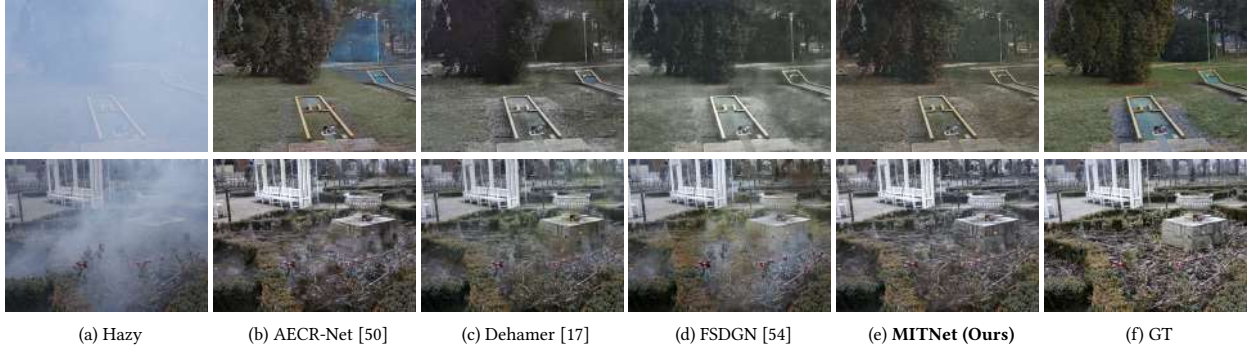
| (a) Hazy | (b) AECR-Net [50] | (c) Dehamer [17] | (d) FSDGN [54] | (e) **MITNet (Ours)** | (f) GT |

Figure 5: Visual comparisons of real-world image dehazing methods on Dense-Haze [1] and NH-Haze [2] datasets.
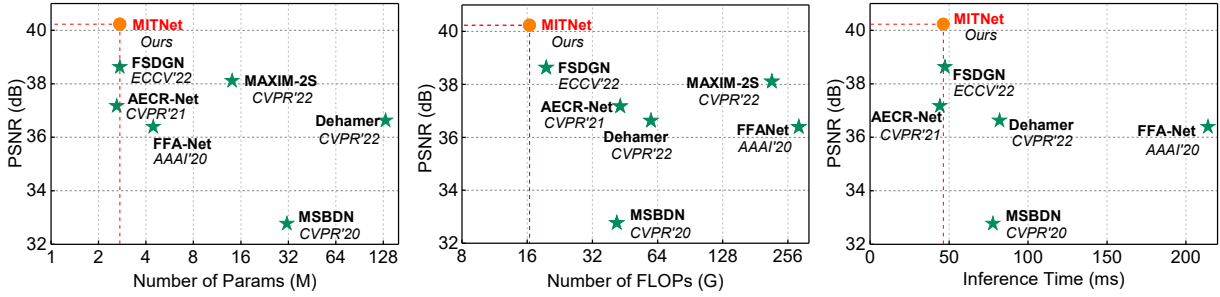

Figure 6: Comparisons of model complexity of different methods. FLOPs are calculated with an input size of 256×256, and the average inference time is evaluated on SOTS-indoor [25].

Table 3: Ablation study about the two-stage design and the combination of spatial-frequency domain features.

Model	M_1	M_2	M_3	M_4	M_5	M_6
#Parmas(M)	2.08	2.22	2.22	2.39	2.39	2.39
PSNR(dB)	35.43	35.98	36.11	34.17	37.43	**37.64**
SSIM	0.9870	0.9890	0.9899	0.9863	0.9905	**0.9908**

Table 4: Ablation investigation on the effectiveness of the proposed ATIM and MIC.

Model	Triple Interaction	ADFB	MIC	#Params (M)	PSNR (dB)	SSIM
M_a	✗	✗	✗	2.39	37.64	0.9908
M_b	✓	✗	✗	2.54	38.64	0.9913
M_c	✗	✓	✗	2.49	38.81	0.9910
M_d	✓	✓	✗	2.73	39.47	0.9916
M_e	✓	✓	✓	2.73	**40.23**	**0.9920**

4.3 Model Complexity Analyses

We further evaluate the model complexity (*e.g.,* network parameters, FLOPs, and inference time) of deep learning-based dehazing approaches over the last three years. The average running time is evaluated on the SOTS-indoor dataset, FLOPs are obtained based on 256×256 resolution image patch, and each result is acquired by repeating the experiment five times to guarantee a fair comparison. From Figure 6, we can easily conclude that our MITNet obtains the best performance with the fewest number of parameters, the fewest FLOPs, and the second-fast inference time. Thus, our MITNet achieves an excellent model complexity versus performance trade-off. FSDGN [54] utilizes dual-domain information while having a noticeable performance gap with ours, which shows it is necessary for adequate feature interaction and redundant feature removal.

4.4 Ablation Studies

Two-stage Architecture. Here, we design several variants of networks to validate the significance and effectiveness of the two-stage architecture. All ablation studies are conducted without using the proposed ATIM and MIC. M_1 denotes that the Fourier prior is not used in both stages, that is, the overall process is performed on the

spatial domain. M_2 denotes that the RABs used in the first stage are only performed in the spatial domain. M_3 denotes that RPBs used in the second stage are only performed in the spatial domain. M_4 denotes exchanging the order of the first stage and the second stage. M_5 denotes that the network architecture is consistent with our MITNet, but uses Y_1 as the input of the second stage. M_6 denotes the proposed MITNet without using the ATIM and MIC.

As shown in Table 3, M_4 presents the worse performance. In this case, the inputs of these two stages contain amplitude spectrum information, which means the whole process is trained on the degraded hazy images. Both M_2 and M_3 exceed the M_1, showing the effectiveness of spatial-frequency dual domain information integration. M_6 achieves 0.21 dB PSNR gain compared to M_6, demonstrating that keeping the $\mathcal{P}(X_{hazy})$ invariant is vital for the training of the second stage. Simultaneously, this phenomenon reveals that the first stage also utilizes the phase spectrum information to some extent, which provides insight guidance for removing redundant information using mutual information constraint.
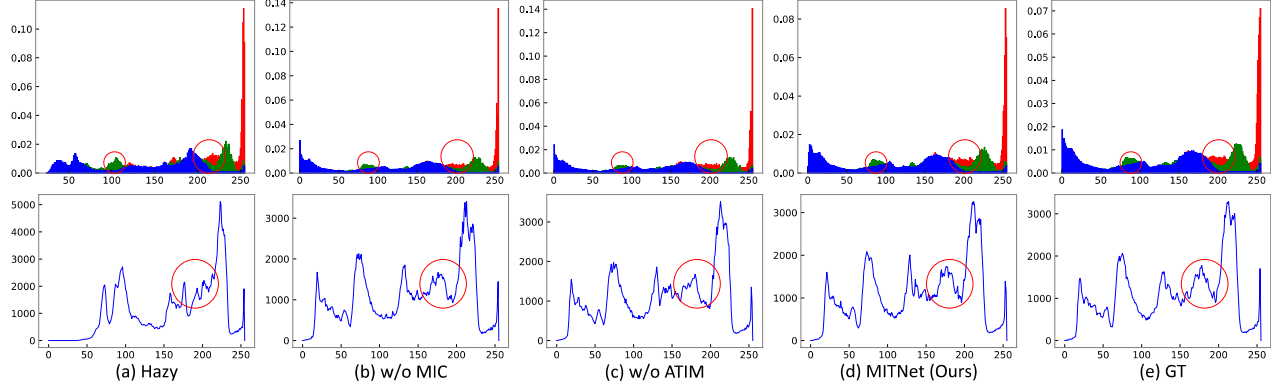
Figure 7: Visualization of the histograms of various images. (a) Input hazy images, (b) output of MITNet without mutual information constraint (w/o MIC), (c) output of MITNet without adaptive triple interaction module (w/o ATIM), (d) output of our MITNet, and (e) ground truth images.

Adaptive Triple Interaction Module. To verify the effectiveness of the proposed ATIM, we ablate the use of components "Triple Interaction" and ADFB. As shown in Table 4, one can see from the top three rows that the proposed triple interaction manner and ADFB significantly outperform the baseline with slightly increasing parameters. Further, when these two modules are employed, M_d achieves 1.77 dB performance gains in terms of PSNR, which suggests utilizing the interactive cross-domain, cross-scale, and cross-stage features to generate content-adaptive dynamic filters and applying them to subsequent decoder features is beneficial for improving representation capability.

Mutual Information Constraint. We also verify the proposed MIC scheme and list the results in the last two rows of Table 4. Because the MIC is only employed during the training phase, it has no effect on inference speed or model parameters. It is clearly seen that Model (e), the final MITNet, gets the only result over 40dB, meaning that using MIC forces complementary feature learning.

Figure 7 shows the statistical distribution with or without using the proposed modules. The first row is the histogram of the corresponding ablated model's output, representing the illumination and color distribution. The histogram of the corresponding grayscale images is shown in the second row, displaying the variation of texture, light, and contrast. Our MITNet obviously produces a more comparable distribution to the clean photographs. However, the results of "w/o ATIM" and "w/o MIC" (zoom in on the red circle) have apparent discrepancies. We further show the difference of feature maps in two stages before and after employing MIC and ATIM in Figure 8. We can get some intuitive clues from the visualization. (1) It is apparent that integrating MIC enhances the complementary features learning and reduces the redundant features, thereby producing different textures. (2) The ATIM utilizes the dual domain and spatial-channel contextual information to guide the generation of subsequent features. The resulting feature maps focus on informative structural contents and fine texture details. These experiments indicate that both modules are the key ingredients of our method.
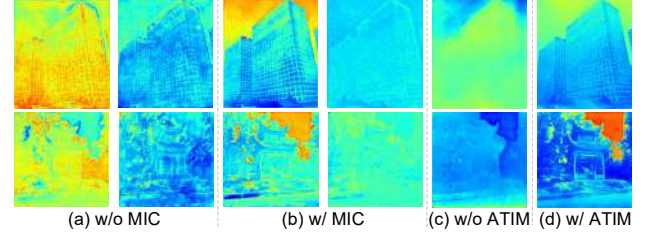

Figure 8: The features difference between two stages in our model without and with integrating the MIC and ATIM. The first and second columns in (a) and (b) denote features extracted from the first and second stages, respectively.

5 CONCLUSION

In this paper, a two-stage network named MITNet is proposed for image dehazing to address three issues. For the limited investigation of frequency domain information, we propose a two-stage design to construct the network, which progressively restores the haze-free images based on Fourier's amplitude and phase spectrum priors. For the inadequate information interaction, we propose an adaptive triple interaction module to fully integrate cross-domain, cross-scale, and cross-stage features, enhancing model representational capability. For the information redundancy, we utilize the mutual information constraint to learn complementary information and alleviate feature redundancy from two stages. Extensive experiments confirm the validity and generality of our MITNet.

REFERENCES

[1] Codruta O Ancuti, Cosmin Ancuti, Mateu Sbert, and Radu Timofte. 2019. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *IEEE International Conference on Image Processing*. 1014–1018.

[2] Codruta O Ancuti, Cosmin Ancuti, and Radu Timofte. 2020. NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 444–445.

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

[4] Dana Berman, Shai Avidan, et al. 2016. Non-local image dehazing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1674–1682.

[5] E Oran Brigham and RE Morrow. 1967. The fast Fourier transform. *IEEE spectrum* 4, 12 (1967), 63–70.

[6] Bolun Cai, Xiangmin Xu, Kui Jia, Chunmei Qing, and Dacheng Tao. 2016. De-hazenet: An end-to-end system for single image haze removal. *IEEE Transactions on Image Processing* 25, 11 (2016), 5187–5198.

[7] Chen Chen, Minh N Do, and Jue Wang. 2016. Robust image and video dehazing with visual artifact suppression via gradient residual minimization. In *Proceedings of the European Conference on Computer Vision*. 576–591.

[8] Liangyu Chen, Xin Lu, Jie Zhang, Xiaojie Chu, and Chengpeng Chen. 2021. Hinet: Half instance normalization network for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 182–192.

[9] Xiang Chen, Zhentao Fan, Pengpeng Li, Longgang Dai, Caihua Kong, Zhuoran Zheng, Yufeng Huang, and Yufeng Li. 2022. Unpaired deep image dehazing using contrastive disentanglement learning. In *European Conference on Computer Vision*. 632–648.

[10] Xiang Chen, Hao Li, Mingqiang Li, and Jinshan Pan. 2023. Learning A Sparse Transformer Network for Effective Image Deraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5896–5905.

[11] Xiang Chen, Jinshan Pan, Jiyang Lu, Zhentao Fan, and Hao Li. 2023. Hybrid cnn-transformer feature fusion for single image deraining. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 378–386.

[12] Xiaofeng Cong, Jie Gui, Kai-Chao Miao, Jun Zhang, Bing Wang, and Peng Chen. 2020. Discrete haze level dehazing network. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1828–1836.

[13] Yuning Cui, Yi Tao, Zhenshan Bing, Wenqi Ren, Xinwei Gao, Xiaochun Cao, Kai Huang, and Alois Knoll. 2023. Selective Frequency Network for Image Restoration. In *Proceedings of the International Conference on Learning Representations*.

[14] Hang Dong, Jinshan Pan, Lei Xiang, Zhe Hu, Xinyi Zhang, Fei Wang, and Ming-Hsuan Yang. 2020. Multi-scale boosted dehazing network with dense feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2157–2167.

[15] Raanan Fattal. 2014. Dehazing using color-lines. *ACM Transactions on Graphics* 34, 1 (2014), 1–14.

[16] Jie Gui, Xiaofeng Cong, Yuan Cao, Wenqi Ren, Jun Zhang, Jing Zhang, Jiuxin Cao, and Dacheng Tao. 2023. A comprehensive survey and taxonomy on single image dehazing based on deep learning. *Comput. Surveys* 55, 13s (2023), 1–37.

[17] Chun-Le Guo, Qixin Yan, Saeed Anwar, Runmin Cong, Wenqi Ren, and Chongyi Li. 2022. Image dehazing transformer with transmission-aware 3D position embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5812–5820.

[18] Kaiming He, Jian Sun, and Xiaoou Tang. 2010. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 12 (2010), 2341–2353.

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670* (2018).

[20] Ming Hong, Jianzhuang Liu, Cuihua Li, and Yanyun Qu. 2022. Uncertainty-driven dehazing network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 906–913.

[21] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861* (2017).

[22] Yifan Jiang, Bartlomiej Wronski, Ben Mildenhall, Jonathan T Barron, Zhangyang Wang, and Tianfan Xue. 2022. Fast and High Quality Image Denoising via Malleable Convolution. In *Proceedings of the European Conference on Computer Vision*. 429–446.

[23] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical review E* 69, 6 (2004), 066138.

[24] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. 2017. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4770–4778.

[25] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. 2018. Benchmarking single-image dehazing and beyond. *IEEE Transactions on Image Processing* 28, 1 (2018), 492–505.

[26] Zechao Li, Hao Tang, Zhimao Peng, Guo-Jun Qi, and Jinhui Tang. 2023. Knowledge-guided semantic transfer network for few-shot image recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[27] Xudong Lin, Lin Ma, Wei Liu, and Shih-Fu Chang. 2020. Context-gated convolution. In *Proceedings of the European Conference on Computer Vision*. 701–718.

[28] Xiaohong Liu, Yongrui Ma, Zhihao Shi, and Jun Chen. 2019. Griddehazenet: Attention-based multi-scale network for image dehazing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7314–7323.

[29] Yu Liu, Guanlong Zhao, Boyuan Gong, Yang Li, Ritu Raj, Niraj Goel, Satya Kesav, Sandeep Gottimukkala, Zhangyang Wang, Wenqi Ren, et al. 2018. Improved techniques for learning to dehaze and beyond: A collective study. *arXiv preprint arXiv:1807.00202* (2018).

[30] Ye Liu, Lei Zhu, Shunda Pei, Huazhu Fu, Jing Qin, Qing Zhang, Liang Wan, and Wei Feng. 2021. From synthetic to real: Image dehazing collaborating with unlabeled real data. In *Proceedings of the 29th ACM International Conference on Multimedia*. 50–58.

[31] Srinivasa G. Narasimhan and Shree K. Nayar. 2003. Contrast restoration of weather degraded images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25, 6 (2003), 713–724.

[32] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[33] Xu Qin, Zhilin Wang, Yuanchao Bai, Xiaodong Xie, and Huizhu Jia. 2020. FFA-Net: Feature fusion attention network for single image dehazing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11908–11915.

[34] Yanyun Qu, Yizi Chen, Jingying Huang, and Yuan Xie. 2019. Enhanced pix2pix dehazing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8160–8168.

[35] Wenqi Ren, Lin Ma, Jiawei Zhang, Jinshan Pan, Xiaochun Cao, Wei Liu, and Ming-Hsuan Yang. 2018. Gated fusion network for single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3253–3261.

[36] Wenqi Ren, Jingang Zhang, Xiangyu Xu, Lin Ma, Xiaochun Cao, Gaofeng Meng, and Wei Liu. 2018. Deep video dehazing with semantic segmentation. *IEEE Transactions on Image Processing* 28, 4 (2018), 1895–1908.

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 234–241.

[38] Christos Sakaridis, Dengxin Dai, Simon Hecker, and Luc Van Gool. 2018. Model adaptation with synthetic and real data for semantic dense foggy scene understanding. In *Proceedings of the European Conference on Computer Vision*. 687–704.

[39] Aditya Sanghi. 2020. Info3d: Representation learning on 3d objects using mutual information maximization and contrastive learning. In *Proceedings of the European Conference on Computer Vision*. 626–642.

[40] Hao Shen, Zhong-Qiu Zhao, Wenrui Liao, Weidong Tian, and De-Shuang Huang. 2022. Joint operation and attention block search for lightweight image restoration. *Pattern Recognition* 132 (2022), 108909.

[41] Hao Shen, Zhong-Qiu Zhao, and Wandi Zhang. 2023. Adaptive Dynamic Filtering Network for Image Denoising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2227–2235.

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

[43] Robby T Tan. 2008. Visibility in bad weather from a single image. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[44] Hao Tang, Zechao Li, Zhimao Peng, and Jinhui Tang. 2020. Blockmix: meta regularization and self-calibrated inference for metric-based meta-learning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 610–618.

[45] Hao Tang, Chengcheng Yuan, Zechao Li, and Jinhui Tang. 2022. Learning attention-guided pyramidal features for few-shot fine-grained recognition. *Pattern Recognition* 130 (2022), 108792.

[46] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision*. 776–794.

[47] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. 2022. Maxim: Multi-axis mlp for image processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5769–5780.

[48] Bo Wang, Zhao Zhang, Jicong Fan, Mingbo Zhao, Choujun Zhan, and Mingliang Xu. 2022. FineFormer: Fine-Grained Adaptive Object Transformer for Image Captioning. In *2022 IEEE International Conference on Data Mining*. IEEE, 508–517.

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

[50] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. 2021. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*

Pattern Recognition. 10551–10560.

[51] Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, and Chongyi Li. 2023. RIDCP: Revitalizing Real Image Dehazing via High-Quality Codebook Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 22282–22291.

[52] Shuanglin Yan, Hao Tang, Liyan Zhang, and Jinhui Tang. 2022. Image-specific information suppression and implicit local alignment for text-based person search. *arXiv preprint arXiv:2208.14365* (2022).

[53] Tian Ye, Mingchao Jiang, Yunchen Zhang, Liang Chen, Erkang Chen, Pen Chen, and Zhiyong Lu. 2021. Perceiving and modeling density is all you need for image dehazing. *Proceedings of the European Conference on Computer Vision.*

[54] Hu Yu, Naishan Zheng, Man Zhou, Jie Huang, Zeyu Xiao, and Feng Zhao. 2022. Frequency and spatial dual guidance for image dehazing. In *Proceedings of the European Conference on Computer Vision.* 181–198.

[55] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2021. Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 14821–14831.

[56] Zican Zha, Hao Tang, Yunlian Sun, and Jinhui Tang. 2023. Boosting few-shot fine-grained recognition with background suppression and foreground alignment. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).

[57] Jing Zhang, Yang Cao, Shuai Fang, Yu Kang, and Chang Wen Chen. 2017. Fast haze removal for nighttime image using maximum reflectance prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 7418–7426.

[58] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. 2021. RGB-D saliency detection via cascaded mutual information minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 4338–4347.

[59] Yulun Zhang, Donglai Wei, Can Qin, Huan Wang, Hanspeter Pfister, and Yun Fu. 2021. Context reasoning attention network for image super-resolution. In

Proceedings of the IEEE/CVF International Conference on Computer Vision. 4278–4287.

[60] Zhao Zhang, Huan Zheng, Richang Hong, Mingliang Xu, Shuicheng Yan, and Meng Wang. 2022. Deep color consistent network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1899–1908.

[61] Suiyi Zhao, Zhao Zhang, Richang Hong, Mingliang Xu, Yi Yang, and Meng Wang. 2022. FCL-GAN: A lightweight and real-time baseline for unsupervised blind image deblurring. In *Proceedings of the 30th ACM International Conference on Multimedia.* 6220–6229.

[62] Suiyi Zhao, Zhao Zhang, Richang Hong, Mingliang Xu, Yi Yang, and Meng Wang. 2022. FCL-GAN: A lightweight and real-time baseline for unsupervised blind image deblurring. In *Proceedings of the 30th ACM International Conference on Multimedia.* 6220–6229.

[63] Huan Zheng, Zhao Zhang, Yang Wang, Zheng Zhang, Mingliang Xu, Yi Yang, and Meng Wang. 2021. GCM-Net: Towards effective global context modeling for image inpainting. In *Proceedings of the 29th ACM International Conference on Multimedia.* 2586–2594.

[64] Huan Zheng, Zhao Zhang, Haijun Zhang, Yi Yang, Shuicheng Yan, and Meng Wang. 2022. Deep multi-resolution mutual learning for image inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia.* 6359–6367.

[65] Man Zhou, Keyu Yan, Jie Huang, Zihe Yang, Xueyang Fu, and Feng Zhao. 2022. Mutual information-driven pan-sharpening. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1798–1808.

[66] Qingsong Zhu, Jiaming Mai, and Ling Shao. 2014. Single Image Dehazing Using Color Attenuation Prior. In *British Machine Vision Conference.*

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