Prior-guided Hierarchical Harmonization Network for Efficient Image Dehazing

Xiongfei Su[1,2], Siyuan Li[1,2], Yuning Cui[3], Miao Cao[1,2], Yulun Zhang[4], Zheng Chen[4], Zongliang Wu[1,2], Zedong Wang[2], Yuanlong Zhang[5], Xin Yuan[2*]

[1]Zhejiang University, Hangzhou, China, e-mail: xsuac@zju.edu.cn
[2]Westlake University, Hangzhou, China
[3]Technical University of Munich, Munich, Germany
[4]Shanghai Jiao Tong University, Shanghai, China
[5]Tsinghua University, Beijing, China

Abstract

Image dehazing is a crucial task that involves the enhancement of degraded images to recover their sharpness and textures. While vision Transformers have exhibited impressive results in diverse dehazing tasks, their quadratic complexity and lack of dehazing priors pose significant drawbacks for real-world applications. In this paper, guided by triple priors, Bright Channel Prior (BCP), Dark Channel Prior (DCP), and Histogram Equalization (HE), we propose a *P*rior-guided Hierarchical *H*armonization Network (PGH^{2}Net) for image dehazing. PGH^{2}Net is built upon the UNet-like architecture with an efficient encoder and decoder, consisting of two module types: (1) Prior aggregation module that injects B/DCP and selects diverse contexts with gating attention. (2) Feature harmonization modules that subtract low-frequency components from spatial and channel aspects and learn more informative feature distributions to equalize the feature maps. Inspired by observing the lower sparsity of B/DCP and the histogram equalization, we harmonize the deep features using a histogram equation-guided module and further leverage B/DCP to guide spatial attention through a sandwich module as the bottleneck. Comprehensive experiments demonstrate that our model efficiently attains the highest level of performance among existing methods across four different datasets for image dehazing.

Introduction

Image dehazing aims to recover clear images from hazy ones (Zheng et al. 2023; Cui et al. 2025). It is crucial in fields like surveillance, autonomous driving, and remote sensing. Estimating clean backgrounds, textures, and colors from a single hazy image is complex and ill-posed. Solutions fall into three categories: conventional, deep learning, and hybrid methods. (i) **Conventional methods** (Zhang et al. 2017; He, Sun, and Tang 2010) rely on physical model assumptions and manual feature engineering, often failing in real-world situations due to the problem's ill-posed nature. These priors are only effective in specific scenarios, as handcrafted features are too simple for complex phenomena like haze and have difficulty selecting optimal transforms and tuning parameters. (ii) **Deep learning methods** use CNN-based approaches (Bai et al. 2022; Cui et al. 2024), including encoder-decoder structures, dilated convolution, and attention mechanisms, achieving impressive restoration performance. These methods employ deep learning as a black box for image restoration but rely on 2D images, with computational time increasing rapidly with larger image sizes. Recently, Transformer models have shown significant improvements in image restoration (Chen et al. 2021; Song et al. 2023), but these often increase model complexity, training costs, and convergence issues due to numerous parameters. (iii) **Hybrid methods** (Zheng et al. 2023; Mo 2022) reduce dependence on training data by combining inherent priors with the representation ability of deep neural networks. (Cai, Zuo, and Zhang 2020; Zheng et al. 2023; Dai et al. 2022) use physics priors in the feature space to enhance interpretability aligned with the hazing process. However, these priors are limited to shallow layers, which lose rich information in deep layers. Thus, universal priors and hierarchical mechanisms are necessary to advance hybrid methods.

In this paper, we empirically reveal the guidance mechanism of Bright Channel Prior (BCP) and Dark Channel Prior (DCP) in the hierarchical feature domain and the distribution matching mechanism of Histogram Equalization (HE). Specifically, we first analyze the deep feature maps

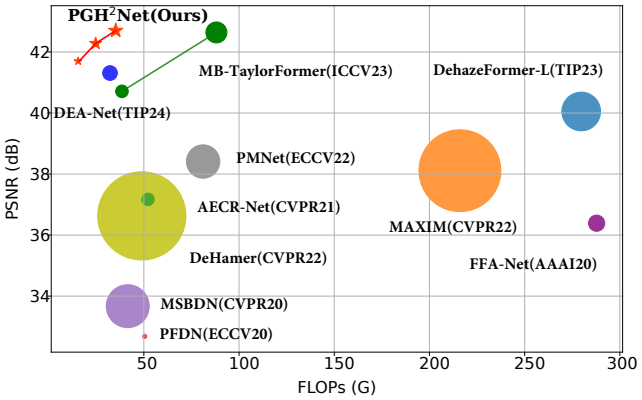

Figure 1: Reconstruction quality (PSNR) and computational complexity (FLOPs) on the SOTS-Indoor (Li et al. 2018) dataset. The size of the dots indicates the model size.

*Corresponding author.

Figure 2: Visualization of the relationship between the spatial haze degradation(a)(d) and BCP(b) and DCP(c) in the deep feature domain. The error maps are differential values with reference, indicating haze distribution, shown in red boxes.

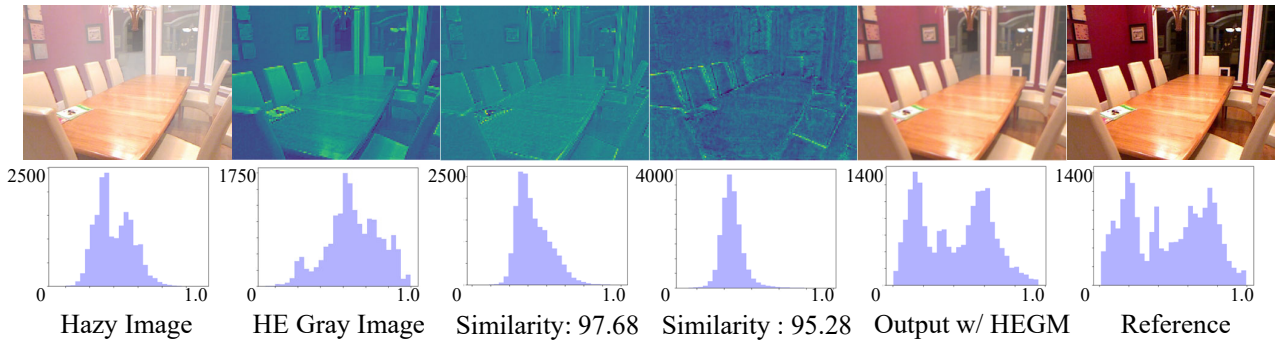

Figure 3: Visualization of the relationship between the value distribution and feature channels. The second row shows histograms assigned to the first row of each image/feature. Similarities are calculated by Cosine similarity. The horizontal/x-axis is the normalized value from 0 to 1, and vertical/y-axis is the number of the value distribution of images/feature.

($128 \times 64 \times 64$) of a basic UNet and split the feature maps into bright channel and dark channel. The error map in Fig. 2(d) is the difference between input and reference images, indicating haze distribution. So red and blue are the thick and thin hazy regions, respectively. The distribution of DCP is basically consistent with the error map, which is sufficient to serve as coarse guidance for hazy regions. Fig. 2(c) distinguishes the cleaner window and hazy wall (red box) in the top row, as well as the cleaner table and hazy back room (red circle) in the bottom row. For BCP, it highlights the major high-frequency information, which is shown in Fig. 2(b). However, limitations of B/DCP exist in Fig. 2(e): The spatial guidance (Ruikun Zhang 2024; Yao et al. 2023b) with B/DCP can remove the major haze, but it is necessary to harmonize with the HE prior, verified in Fig. 2(f) and Tab. 2,3.

To further explore the priors in hierarchical levels, we notice that distribution (such as histogram) with only one dimension vector is easy to transport in deep layers without the limitation of hierarchical sizes. In Fig. 3, we first generate the histogram of the hazy image and processed image by HE. It is apparent that hazy images own more voxel values close to white, causing a peak in the histogram results. HE flattens the histogram with a remapping algorithm. Subsequently, we calculate the distribution similarity between each channel feature and HE. We show the distribution of two channels in Fig. 3 with similarities and find that similar distributions with HE indicate cleaner and sharper corresponding channel features.

Based on the above observation and analysis, we design a novel hierarchical pure convolutional architecture, *dubbed PGH^{2}Net*, for image dehazing tasks. From a fresh perspective, we solve the ill-posed problem by jointly introducing channel and distribution priors into deep layers of the network to guide the restoration from hierarchical levels:

- We reveal the spatial guidance mechanism of B/DCP in the hierarchical feature domain and propose a design philosophy of **aggregating priors**. The Prior Aggregator injects B/DCP and selects diverse contexts via gating attention, while the Sandwich Module as bottleneck injects B/DCP with complementary spatial attention.
- We reveal the distribution guidance from HE prior as another principle, **harmonizing feature distributions**. Spatial and Channel Harmonization Modules enrich and equalize features by adaptively removing low-frequency components, while the HE Guidance Module as bottleneck provides channel-wise weighting harmonization.
- Extensive experiments demonstrate that the proposed PGH^{2}Net performs favorably against previous state-of-the-art algorithms. Meanwhile, PGH^{2}Net significantly reduces the computational complexity and achieves a sweet point in the performance-parameters trade-off.

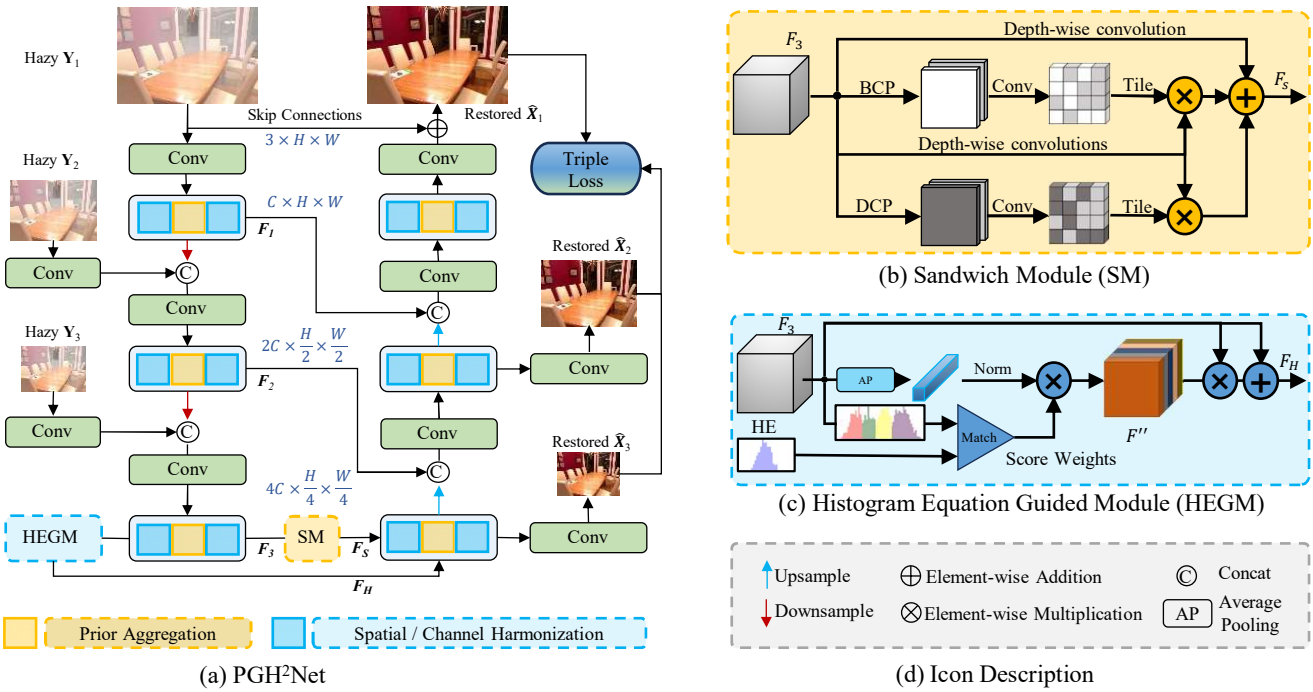

Figure 4: PGH²Net architecture. (a) The encoders and decoders with a stack of Prior Aggregation and Spatial/Channel Harmonization modules learn hierarchical features with diverse distributions. Then, the bottleneck with the (b) Sandwich Module (SM) and (c) Histogram Equation Guided Module (HEGM) transports equalized deep features to the decoders.

Related Work

Image Dehazing

Researchers have begun using deep neural networks for image dehazing (Liu et al. 2019b; Song et al. 2023; Bai et al. 2022; Zheng et al. 2023; Cui et al. 2023c,a). CNN architectures significantly outperform traditional physical-based methods (Abuolaim and Brown 2020; Chen et al. 2022). The encoder-decoder paradigm, used for hierarchical representations (Mao et al. 2021), has been enhanced with modules like dynamic filters (Lee et al. 2021), dilated convolution (Zou et al. 2021), shortcut connections (Cho et al. 2021; Cui et al. 2023b,d), and attention modules (Qin et al. 2020). Vision Transformers have shown impressive results in image dehazing but suffer from quadratic complexity, leading researchers to restrict operation regions (Liang et al. 2021) or switch operation dimensions (Zamir et al. 2022). Hybrid methods have also been introduced, combining statistics-based priors with deep learning, such as DCA-CycleGAN (Mo 2022), which integrates dark channel prior.

Triple Priors

He *et al.* (He, Sun, and Tang 2010) introduced the dark channel prior, positing that its values in an unobstructed image tend to be close to zero. (Pan et al. 2018) verifies its effectiveness in deblurring tasks. However, this approach works well for most outdoor hazy images but struggles with hazy images featuring bright areas, especially in the sky. (Yan et al. 2017) leverages B/DCP and incorporates a prior using both bright and dark information. (Cai, Zuo, and Zhang 2020) uses B/DCP in a multi-branch network layer to extract feature information, increasing computational complexity.

Histogram adjustment is another widely used prior, helpful in industry, such as *Photoshop*. The histogram of a hazy image typically peaks around a specific value, with few voxels close to zero, while the histogram of a clear image is more evenly distributed from 0 to 255. (Chi et al. 2020) learns the ground truth histogram distribution with a full connection network, but it lacks spatial guidance. We aim to guide our network's attention on the channel dimension with a histogram by a one-dimensional vector.

Proposed Method

In this section, we first present the overall architecture of PGH²Net, shown in Fig. 4. Following this, we describe the core components of PGH²Net: spatial harmonization module, prior aggregation module, channel harmonization module, sandwich module, and histogram equation guide module (HEGM). Finally, the training loss function is defined.

Overall Architecture

As shown in Fig. 4, the proposed PGH²Net uses triple levels of architecture to efficiently learn hierarchical representations. Both the encoder and decoder networks comprise three scales. Specifically, given a hazy image of dimensions $3 \times H \times W$, a convolutional layer with a kernel size of 3×3 is applied to extract shallow features. These shallow features, which have dimensions $C \times H \times W$, then pass through

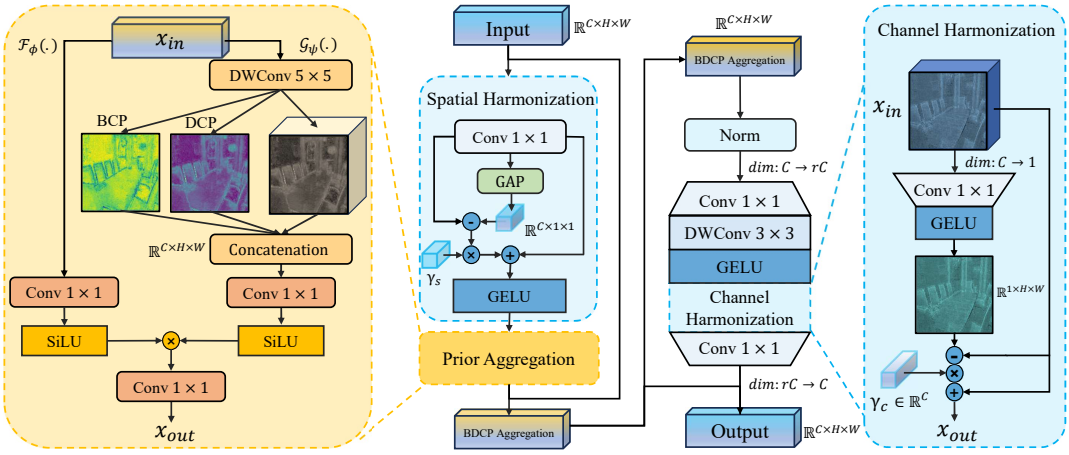

Figure 5: Structure of the encoder and decoder blocks: Spatial Harmonization Module $\mathrm{SH}(\cdot)$, Prior Aggregation Module $\mathrm{PA}(\cdot)$, and Channel Harmonization Module $\mathrm{CH}(\cdot)$ are cascaded. $\mathrm{SH}(\cdot)$ and $\mathrm{PA}(\cdot)$ combine to aggregate spatial information.

a three-scale symmetric encoder-decoder structure. Transformation yields enhanced features with comprehensive image information through n stages. Each stage includes prior aggregation, spatial harmonization, and channel harmonization, illustrated in Fig. 5. To integrate our proposed sandwich module, we mount it in the deepest layer since the BCP and DCP information is easier to learn in the shallow feature extractor but harder to learn as layers increase. Therefore, the HEGMs are utilized among different layers since the HE remains constant in different sizes.

Spatial Aggregation Block

We propose the spatial aggregation block (SA) to learn the Harmonized representations of B/DCP by a pure convolutional design, as shown in Fig. 5 (left part), which consists of two cascaded components.

Spatial Harmonization. According to our proposed harmonization prior, we extract diverse features with both *static* and *adaptive* locality perceptions in the SH module. Since convolutions are inherently high-pass filters (Park and Kim 2022; Wang et al. 2022a), there are two complementary counterparts, fine-grained local texture and complex global shape, which are instantiated by $\mathrm{Conv}_{1\times1}(\cdot)$ and $\mathrm{GAP}(\cdot)$, respectively. To counter the network's inherent interaction bias strengths (Li et al. 2023), we design $\mathrm{SH}(\cdot)$ to adaptively exclude the trivial (overlooked) interactions, defined as:

$$\mathbf{Y} = \mathrm{Conv}_{1\times1}(\mathbf{X}), \tag{1}$$

$$\mathbf{Z} = \mathrm{GELU}\Big(\mathbf{Y} + \boldsymbol{\gamma}_s \otimes \big(\mathbf{Y} - \mathrm{GAP}(\mathbf{Y})\big)\Big), \tag{2}$$

where $\gamma_s \in \mathbb{R}^{C\times1}$ denotes a scaling factor initialized as zeros. Reweighting the complementary interaction component $\mathbf{Y} - \mathrm{GAP}(\mathbf{Y})$, $\mathrm{SH}(\cdot)$ also increases spatial feature diversities (Park and Kim 2022; Wang et al. 2022a).

Prior Aggregation. Since the difference between clean and degraded images of B/DCP in Fig. 2, the associated priors and sparse constraints aid in restoring clear images. Then, we ensemble the B/DCP and local edge features in the context branch and adaptively select the informative channels by the gating aggregation in the PA module. The establishment of this prior principle is predicated upon empirical observation (He, Sun, and Tang 2010; Yan et al. 2017; Yao et al. 2024), revealing that within the vast majority of patches in natural scenes, there consistently exists a feature tensor in which the highest and lowest intensity values of voxels tend to exhibit a pronounced prominence. In this paper, the B/DCP of the feature domain is defined by

$$\mathrm{D}(\mathbf{F})(\boldsymbol{x}) = \min_{\boldsymbol{y}\in\Omega(\boldsymbol{x})}\left(\min_{c\in\{0\ldots N\}}\mathbf{F}^c(\boldsymbol{y})\right), \tag{3}$$

$$\mathrm{B}(\mathbf{F})(\boldsymbol{x}) = \max_{\boldsymbol{y}\in\Omega(\boldsymbol{x})}\left(\max_{c\in\{0\ldots N\}}\mathbf{F}^c(\boldsymbol{y})\right), \tag{4}$$

where $\boldsymbol{y}$ is the patch of location $\boldsymbol{x}$ in the feature map, $\mathbf{F}^c$ is a channel of $\mathbf{F}$, and $\Omega(\boldsymbol{x})$ denotes a local patch centered at $\boldsymbol{x}$. N denotes the number of channels.

Unlike previous work that simply combined DWConv with self-attention to model local and global interactions (Zhang, Hu, and Wang 2022; Pan, Cai, and Zhuang 2022; Si et al. 2022; Rao et al. 2022; Huang et al. 2023; Guo et al. 2025; Yao et al. 2023a; Guan et al. 2023), we employ three different branches in parallel to capture DCP, BCP and identical interactions: Given the input feature $\mathbf{X} \in \mathbb{R}^{C\times HW}$, $\mathrm{DW}_{5\times5,d=1}$ is first applied for extracting features; then, the output is factorized into $\mathrm{B}(\mathbf{X})$, $\mathrm{D}(\mathbf{X})$ and an identical mapping $\mathbf{X}$; finally, the outputs are concatenated to form B/DCP contexts, $\mathbf{Y}_C = \mathrm{Concat}(B(\mathbf{X}), D(\mathbf{X}), \mathbf{X})$.

After injecting B/DCP into feature maps, we utilize the gating aggregation to adaptively fuse priors and contextual features (*e.g.,* edges). Taking the output from $\mathrm{SH}(\cdot)$ as the input, the output of the Prior Aggregation is written as:

$$\mathbf{Z} = \underbrace{\mathrm{SiLU}\big(\mathrm{Conv}_{1\times1}(\mathbf{X})\big)}_{\mathcal{F}_\phi} \otimes \underbrace{\mathrm{SiLU}\big(\mathrm{Conv}_{1\times1}(\mathbf{Y}_C)\big)}_{\mathcal{G}_\psi}. \tag{5}$$

It produces informative representations with similar parameters and FLOPs as $\mathrm{DW}_{7\times7}$ in ConvNeXt, which is beyond

the reach of existing methods.

Channel Harmonization Block

Due to channel redundancy (Woo et al. 2018; Cao et al. 2019; Tan and Le 2019; Wang et al. 2020), vanilla MLP needs numerous parameters (r default to 4 or 8) for optimal performance. To overcome this, most methods insert a channel enhancement module, *e.g.,* SE module (Hu, Shen, and Sun 2018). We propose a lightweight channel harmonization module $\mathrm{CH}(\cdot)$ to reallocate channel-wise features in high-dimensional spaces, further developing it into a channel harmonization (CH) block. As shown in Fig. 5,

$$\begin{aligned}\mathbf{Y} &= \mathrm{GELU}\Big(\mathrm{DW}_{3\times3}\big(\mathrm{Conv}_{1\times1}(\mathrm{Norm}(\mathbf{X}))\big)\Big),\\ \mathbf{Z} &= \mathrm{Conv}_{1\times1}\big(\mathrm{CH}(\mathbf{Y})\big) + \mathbf{X}.\end{aligned} \tag{6}$$

Concretely, $\mathrm{CH}(\cdot)$ is implemented by a channel-reducing projection $W_r : \mathbb{R}^{C\times HW} \rightarrow \mathbb{R}^{1\times HW}$ and GELU to gather and reallocate channel-wise information:

$$\mathrm{CH}(\mathbf{Y}) = \mathbf{Y} + \boldsymbol{\gamma}_c \otimes \big(\mathbf{Y} - \mathrm{GELU}(\mathbf{Y}\mathbf{W}_r)\big), \tag{7}$$

where γ_c is the channel-wise scaling factor initialized as zeros. It harmonizes the channel-wise feature with the complementary interactions $(\mathbf{Y} - \mathrm{GELU}(\mathbf{Y}\mathbf{W}_r))$.

Sandwich Module

The higher sparsity of dark and bright channels in a sharp image compared to a degraded image, along with sparse constraints from these priors, aids in the restoration of clear images. The proposed sandwich module can regularize the spatial attention space. Specifically, there are two branches in the sandwich module, and the main branch squeezes feature map $\mathbf{F}$ from the encoder along the channel dimension with B/DCP, shown in Fig. 4. In addition, the average pooling $\mathrm{GAP}(\cdot)$ is used for complementary degradation locations. Subsequently, a concatenation operation is used to merge these three feature maps $\mathrm{B}(\mathbf{F})$, $\mathrm{D}(\mathbf{F})$ as Eq. (3), and $\mathrm{GAP}(\mathbf{F})$, facilitating the integration of representations from the degraded image and the B/DCP as follows:

$$\mathbf{F}_B = \mathrm{Conv}([\mathrm{B}(\mathbf{F}), \mathrm{GAP}(\mathbf{F})]), \tag{8}$$

$$\mathbf{F}_D = \mathrm{Conv}([\mathrm{D}(\mathbf{F}), \mathrm{GAP}(\mathbf{F})]), \tag{9}$$

As each channel exhibits distinct degradation patterns, we proceed to create channel-specific representations by applying channel-separated transformations to the input feature $\mathbf{F}$ using depth-wise convolutions, followed by modulation as:

$$\mathbf{F}_s = \mathrm{DW}(\mathbf{F}) \otimes (\mathbf{F}_B + \mathbf{F}_D) + \mathrm{DW}(\mathbf{F}), \tag{10}$$

To incorporate constraints based on dark and bright channel priors into a network, we also utilize a l_1-regularization term to enforce sparsity during training.

Histogram Equation Guided Module

As another parallel branch to B/DCP in the bottleneck of the network, we introduce an innovative approach for single-image dehazing utilizing HEGM. Unlike other attention mechanism that operates on the 2D feature map, our model takes 1D histogram equation distribution data, specifically histograms in the image domain, as guided input. This phenomenon is visually demonstrated in Fig. 3. Additionally, all guided features in the intermediate layers are 1D, simplifying our model and improving the ease of training compared to other CNN-based methods.

Take into account a feature tensor $\mathbf{F}$, where n_i represents the number of voxel value i. The probability of encountering a voxel with level i in the image is the ratio of occurrences

$$p_{\mathbf{F}}(i) = \frac{n_i}{n} \quad 0 \le i < L, \tag{11}$$

where L is typically 256 and n is the total number of voxels. Cumulative distribution function (CDF) is defined as

$$cdf_{\mathbf{F}}(i) = \sum_{j=0}^{i} p_{\mathbf{F}}(j), \tag{12}$$

which is also the feature's accumulated normalized histogram. The general histogram equalization formula is:

$$\mathbf{H}_{\bar{\mathbf{F}}}(i) = \mathrm{round}\left(\frac{cdf_{\mathbf{F}}(i) - cdf_{\min}}{1 - cdf_{\min}} \times (L-1)\right) \tag{13}$$

where $\mathbf{H}_{\bar{\mathbf{F}}}(i)$ is the remapped voxel value in the new equalized image, $cdf_{\min}$ is the minimum non-zero value of the cumulative distribution function, and L is the number of levels being used. Then we obtain a new image $\bar{\mathbf{F}}$ with histogram equalization. New probability $p_{\bar{\mathbf{F}}}$ is calculated by Eq. 11.

As illustrated in Fig. 4(c), the feature voxel $\mathbf{F}$ is split along the channel, and the histogram of each channel is calculated and normalized to probability p_c individually. Match feature voxel probability to $p_{\bar{\mathbf{F}}}$ so that the histogram of each channel generates a corresponding score with Cosine similarity:

$$\mathbf{F}'' = \mathrm{Norm}(\mathrm{Sim}(p_{\mathbf{F}}, p_{\bar{\mathbf{F}}})) \times \mathrm{Norm}\left(\mathrm{GAP}(\mathbf{F})\right), \tag{14}$$

where Sim is Cosine similarity. After broadcasting, the score weights attention map is used to element-wise multiply with input $\mathbf{F}$, and residual addition, which is formulated as:

$$\mathbf{F}_H = \mathbf{F} \otimes \mathbf{F}'' + \mathbf{F}. \tag{15}$$

Triple Learning Objective

To facilitate the sparsity of B/DCP, we adopt ℓ_1 loss in the spatial, frequency, and structural domains:

$$\mathcal{L}_{\text{spatial}} = \sum_{i=1}^{3} \frac{1}{P_i} \left\| \hat{\mathbf{X}}_i - \mathbf{X}_i \right\|_1, \tag{16}$$

$$\mathcal{L}_{\text{frequency}} = \sum_{i=1}^{3} \frac{1}{P_i} \left\| \mathrm{F}\left(\hat{\mathbf{X}}_i\right) - \mathrm{F}\left(\mathbf{X}_i\right) \right\|_1, \tag{17}$$

$$\mathcal{L}_{\text{ssim}} = \sum_{i=1}^{3} \frac{1}{P_i} \left\| \mathrm{SSIM}\left(\hat{\mathbf{X}}_i\right) - \mathrm{SSIM}\left(\mathbf{X}_i\right) \right\|_1, \tag{18}$$

where i is the index of multiple outputs, as illustrated in Fig. 4(a); $\hat{\mathbf{X}}$ and $\mathbf{X}$ denote the predicted image and ground truth, respectively; P_i represents the total elements of the image for normalization, and F represents the fast Fourier

Method	Venue	SOTS-Indoor PSNR ↑	SOTS-Indoor SSIM ↑	SOTS-Outdoor PSNR ↑	SOTS-Outdoor SSIM ↑	Dense-Haze PSNR ↑	Dense-Haze SSIM ↑	O-HAZE PSNR ↑	O-HAZE SSIM ↑	Params (M)	FLOPs (G)
DehazeNet (Cai et al. 2016)	TIP'16	19.82	0.821	24.75	0.927	13.84	0.43	17.57	0.77	0.009	**0.581**
AOD-Net (Li et al. 2017)	ICCV'17	20.51	0.816	24.14	0.920	13.14	0.41	15.03	0.54	**0.002**	0.115
GridDehaze (Liu et al. 2019a)	CVPR'19	32.16	0.984	30.86	0.982	-	-	-	-	0.956	21.49
MSBDN (Dong et al. 2020)	CVPR'20	33.67	0.985	33.48	0.982	15.37	0.49	24.36	0.75	31.35	41.54
FFA-Net (Qin et al. 2020)	AAAI'20	36.39	0.989	33.57	0.984	14.39	0.45	22.12	0.77	4.456	287.8
AECR-Net (Wu et al. 2021)	CVPR'21	37.17	0.990	-	-	15.80	0.47	-	-	2.611	52.20
DeHamer (Guo et al. 2022)	CVPR'22	36.63	0.988	35.18	0.986	16.62	0.56	-	-	132.45	48.93
PMNet(Ye et al. 2022)	ECCV'22	38.41	0.990	34.74	0.985	16.79	0.51	24.64	0.83	18.90	81.13
MAXIM-2S (Tu et al. 2022)	CVPR'22	38.11	0.991	34.19	0.985	-	-	-	-	14.10	216.0
DehazeFormer(Song et al. 2023)	TIP'23	38.46	0.994	34.29	0.983	16.29	0.51	-	-	14.10	216.0
TaylorFormer (Qiu et al. 2023)	ICCV'23	40.71	0.992	37.42	0.989	16.66	0.56	25.05	0.788	2.68	38.50
LH-Net (Yuan et al. 2023)	MM'23	37.04	0.989	36.05	0.986	18.87	0.561	-	-	35.64	-
MITNet (Shen et al. 2023)	MM'23	40.23	0.992	35.18	0.988	16.97	0.606	-	-	2.83	16.25
DEA (Chen, He, and Lu 2024)	TIP'24	41.31	0.995	36.59	0.989	-	-	-	-	3.65	32.23
PGH^{2}Net	**Ours**	**41.70**	**0.996**	**37.52**	**0.989**	**17.02**	**0.61**	**25.47**	**0.88**	1.76	16.05

Table 1: Image dehazing results on both synthetic dataset (Li et al. 2018) and real-world datasets (Ancuti et al. 2019, 2018).

transform. Structural Similarity Index Measure (SSIM) to measure the local structural similarity between images or patches, which is formulated as the combination of three metrics: luminance, contrast, and structure. The final loss function is given by the three terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{spatial}} + \lambda_1 \mathcal{L}_{\text{frequency}} + \lambda_2 \mathcal{L}_{\text{ssim}}, \qquad (19)$$

where loss weight λ requires fine-tuning in practice and we used $\lambda_1 = 0.5$ and $\lambda_2 = 1$ as the final setting.

Experimental Results

In this section, we evaluate our proposed PGH^{2}Net in four data sets for image dehazing tasks, including indoor synthetic data, outdoor synthetic data, and two real data.

Datasets and Evaluation Metrics

Evaluation. We calculate the Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity index (SSIM) (Wang et al. 2004) between the predicted results and ground-truth images for all datasets. Floating point operations (FLOPs) are measured on the patch of size $3 \times 256 \times 256$.

Image Dehazing Datasets. We train and evaluate our models on synthetic and real-world datasets for image dehazing. Following (Wang et al. 2022b, 2024), we train separate models on the RESIDE-Indoor and RESIDE-Outdoor datasets (Li et al. 2018), and evaluate the resulting models on the corresponding test sets of RESIDE, *i.e.,* SOTS-Indoor and SOTS-Outdoor, respectively. In addition, we adopt two real-world datasets, *i.e.,* Dense-Haze (Ancuti et al. 2019) and O-HAZE (Ancuti et al. 2018), to verify the robustness of our model in more challenging real-world scenarios.

Implementation Details

The models are trained using Adam (Kingma and Ba 2014) with initial learning rate as $8e^{-4}$, which is gradually reduced to $1e^{-6}$ with cosine annealing (Loshchilov and Hutter 2016). For data augmentation, we adopt random horizontal flips with a probability of 0.5. Models are trained on 32 samples of size 256×256 for each iteration.

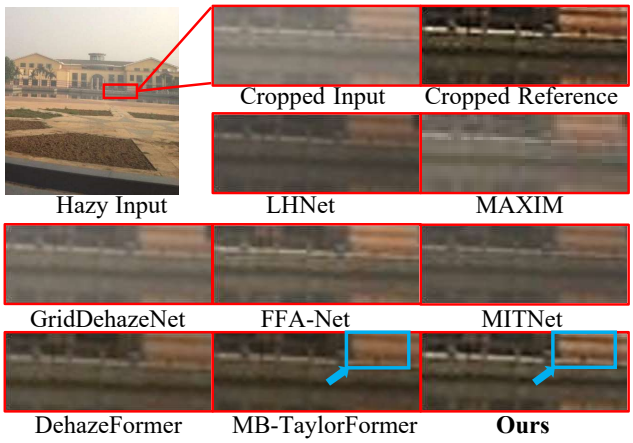

Figure 6: Image dehazing comparisons on the SOTS-Outdoor (Li et al. 2018) test sets. The red box is zoomed in by 6× for visualization. The cyan arrows point to the superior performance of our method.

Figure 7: Comparisons on the Indoors (Li et al. 2018) data.

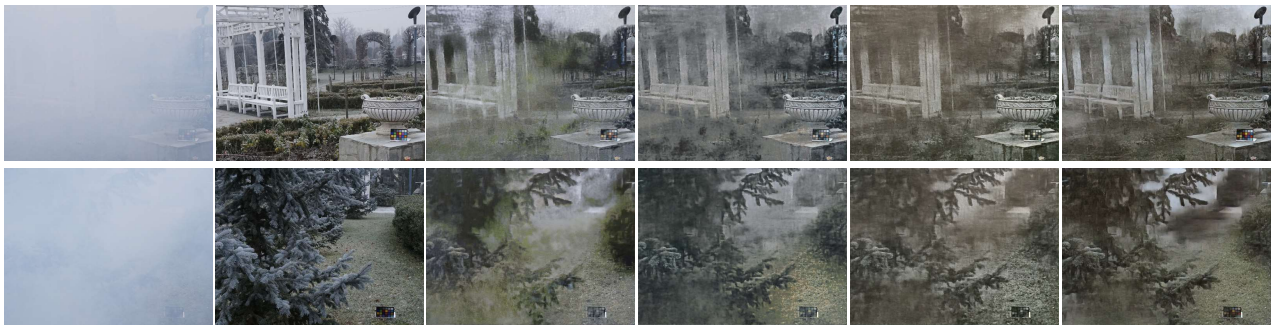

Figure 8: Comparisons on the Dense-Haze (Ancuti et al. 2019).

Image Dehazing Results

Quantitative Comparisons. We report the quantitative performance of image dehazing approaches on both synthetic (Li et al. 2018) and real-world (Ancuti et al. 2019, 2018) datasets in Tab. 1. Overall, our method receives higher performance on all datasets than other state-of-the-art algorithms. Specifically, on the daytime synthetic dataset SOTS-Indoor (Li et al. 2018), our method outperforms MB-TaylorFormer (Qiu et al. 2023) by 0.99 dB PSNR with only 42% parameters and 67% FLOPs. Furthermore, our model yields a significant performance gain of 3.23 dB in terms of PSNR over Transformer model DeHamer (Guo et al. 2022) on SOTS-Outdoor (Li et al. 2018) with fewer parameters.

Visual Comparisons in synthetic dataset. The daytime visual results produced by several dehazing methods are illustrated in Fig. 6, 7. Our method is more effective in removing haze blurs in both indoor and outdoor scenes than other algorithms, such as blurs on the doors in the top two images of Fig. 6. The proposed PGH^{2}Net can retrieve not only the sharp shapes of the objects but also the colorful, fine textures and details. Simultaneously, PGH^{2}Net avoids noise and artefacts in the background that appear in other methods.

Visual Comparisons in a real-world dataset. In Fig. 8, our method is well generalized to the more challenging real-world scenarios following USCFormer (Wang et al. 2023) and obtains the best performance. Our method recovers details of the grasses and trees. The color plate at the lower right corner indicates color correction ability of our method.

Ablation Studies

We conduct ablation studies to demonstrate the effectiveness of our modules by training the model on RESIDE-Indoor (Li et al. 2018) and testing on SOTS-Indoor (Li et al. 2018). The model is trained with an initial learning rate of $8e^{-4}$ and a batch size of 32, ending in epoch 100.

Effects of the encoder and decoder module. We first ablate the spatial aggregation module and the channel aggregation module CH$(\cdot)$ in Tab. 2. Spatial modules include SH$(\cdot)$ and PA$(\cdot)$, containing the gating branch. We found that all proposed modules yield improvements with favorable costs.

Effects of bottleneck attention module. As shown in Tab. 3a, the baseline receives 34.69 dB PSNR. sandwich module (Tab. 3b) and HEGM (Tab. 3c) yield accuracy gains of 1.40 and 0.63 dB over the baseline, respectively.

In addition, the visual results of our sandwich module SM$(\cdot)$ are illustrated in Fig. 2. The sandwich module helps the model focus more on the severe degradation regions, *e.g.,* metal fence. HEGM further highlights the accurate voxel value distribution (see Fig. 3).

Modules	PSNR↑	SSIM↑	Params. (M)	FLOPs (G)
Baseline	35.64	0.985	1.34	15.96
+Gating branch	36.27	0.987	1.68	15.98
+PA$(\cdot)$	37.14	0.987	1.72	16.00
+SH$(\cdot)$	37.48	0.990	1.75	16.01
+CH$(\cdot)$	**38.00**	**0.992**	1.76	16.05

Table 2: Ablation of our modules on the SOTS-Indoor (Li et al. 2018) dataset. The baseline uses the non-linear projection and DW$_{5\times5}$ as SH$(\cdot)$ and the MLP as CH$(\cdot)$.

	Sandwich	HEGM	PSNR↑	SSIM↑	Params. (M)	FLOPs (G)
(a)			34.69	0.985	1.74	15.98
(b)	✓		36.09	0.989	1.76	16.04
(c)		✓	35.32	0.985	1.74	16.00
(d)	✓	✓	**38.00**	**0.992**	1.76	16.05

Table 3: Ablation studies for different bottleneck attention of PGH^{2}Net on the SOTS-Indoor (Li et al. 2018).

Conclusion

In this paper, we present a triple priors guided network for image dehazing, dubbed PGH^{2}Net, which is effective and computationally efficient. To our knowledge, this is the first work to reveal the relationship between B/DCP and spatial guidance in hierarchical feature representation. We are the first to utilize HE matching similarity to harmonize the channel-wise features, which is effective and low-cost. By collaborating with triple priors, PGH^{2}Net can leverage their individual strengths and provide complementary information in a harmonious manner, shown in Fig. 2(d-f). Our future work will explore the framework in other image tasks.

Acknowledgements

This work was supported by the National Key R&D Program of China (grant number 2024YFF0505603, 2024YFF0505600), the National Natural Science Foundation of China (grant number 62271414), Zhejiang Provincial Outstanding Youth Science Foundation (grant number LR23F010001), Zhejiang "Pioneer" and "Leading Goose"R&D Program(grant number 2024SDXHDX0006, 2024C03182), the Key Project of Westlake Institute for Optoelectronics (grant number 2023GD007), the 2023 International Sci-tech Cooperation Projects under the purview of the "Innovation Yongjiang 2035" Key R&D Program (grant number 2024Z126) and the Zhejiang Province Postdoctoral Research Excellence Funding Program (grant number ZJ2024086).

References

Abuolaim, A.; and Brown, M. S. 2020. Defocus deblurring using dual-pixel data. In *ECCV*.

Ancuti, C. O.; Ancuti, C.; Sbert, M.; and Timofte, R. 2019. Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images. In *ICIP*.

Ancuti, C. O.; Ancuti, C.; Timofte, R.; and De Vleeschouwer, C. 2018. O-HAZE: A Dehazing Benchmark With Real Hazy and Haze-Free Outdoor Images. In *CVPRW*.

Bai, H.; Pan, J.; Xiang, X.; and Tang, J. 2022. Self-guided image dehazing using progressive feature fusion. *TIP*.

Cai, B.; Xu, X.; Jia, K.; Qing, C.; and Tao, D. 2016. Dehazenet: An end-to-end system for single image haze removal. *TIP*.

Cai, J.; Zuo, W.; and Zhang, L. 2020. Dark and Bright Channel Prior Embedded Network for Dynamic Scene Deblurring. *TIP*.

Cao, Y.; Xu, J.; Lin, S.; Wei, F.; and Hu, H. 2019. GCNet: Non-Local Networks Meet Squeeze-Excitation Networks and Beyond. In *ICCVW*.

Chen, H.; Wang, Y.; Guo, T.; Xu, C.; Deng, Y.; Liu, Z.; Ma, S.; Xu, C.; Xu, C.; and Gao, W. 2021. Pre-Trained Image Processing Transformer. In *CVPR*.

Chen, L.; Chu, X.; Zhang, X.; and Sun, J. 2022. Simple baselines for image restoration. In *ECCV*.

Chen, Z.; He, Z.; and Lu, Z.-M. 2024. DEA-Net: Single Image Dehazing Based on Detail-Enhanced Convolution and Content-Guided Attention. *TIP*.

Chi, J.; Li, M.; Meng, Z.; Fan, Y.; Zeng, X.; and Jing, M. 2020. Single Image Dehazing using a Novel Histogram Tranformation Network. In *ISCAS*.

Cho, S.-J.; Ji, S.-W.; Hong, J.-P.; Jung, S.-W.; and Ko, S.-J. 2021. Rethinking Coarse-To-Fine Approach in Single Image Deblurring. In *ICCV*.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2023a. Focal network for image restoration. In *ICCV*.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2023b. Image restoration via frequency selection. *PAMI*.

Cui, Y.; Ren, W.; Cao, X.; and Knoll, A. 2024. Revitalizing convolutional network for image restoration. *PAMI*.

Cui, Y.; Tao, Y.; Bing, Z.; Ren, W.; Gao, X.; Cao, X.; Huang, K.; and Knoll, A. 2023c. Selective frequency network for image restoration. In *ICLR*.

Cui, Y.; Tao, Y.; Ren, W.; and Knoll, A. 2023d. Dual-domain attention for image deblurring. In *AAAI*.

Cui, Y.; Wang, Q.; Li, C.; Ren, W.; and Knoll, A. 2025. EENet: An effective and efficient network for single image dehazing. *PR*.

Dai, T.; Feng, Y.; Chen, B.; Lu, J.; and Xia, S.-T. 2022. Deep image prior based defense against adversarial examples. *Pattern Recognition*, 122: 108249.

Dong, H.; Pan, J.; Xiang, L.; Hu, Z.; Zhang, X.; Wang, F.; and Yang, M.-H. 2020. Multi-Scale Boosted Dehazing Network With Dense Feature Fusion. In *CVPR*.

Guan, Y.; Xu, R.; Yao, M.; Wang, L.; and Xiong, Z. 2023. Mutual-guided dynamic network for image fusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1779–1788.

Guo, C.-L.; Yan, Q.; Anwar, S.; Cong, R.; Ren, W.; and Li, C. 2022. Image Dehazing Transformer with Transmission-Aware 3D Position Embedding. In *CVPR*.

Guo, H.; Li, J.; Dai, T.; Ouyang, Z.; Ren, X.; and Xia, S.-T. 2025. Mambair: A simple baseline for image restoration with state-space model. In *ECCV*, 222–241. Springer.

He, K.; Sun, J.; and Tang, X. 2010. Single image haze removal using dark channel prior. *PAMI*.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *CVPR*.

Huang, Y.; Chen, B.; Qin, S.; Li, J.; Wang, Y.; Dai, T.; and Xia, S.-T. 2023. Learned distributed image compression with multi-scale patch matching in feature domain. In *AAAI*, 4322–4329.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lee, J.; Son, H.; Rim, J.; Cho, S.; and Lee, S. 2021. Iterative filter adaptive network for single image defocus deblurring. In *CVPR*.

Li, B.; Peng, X.; Wang, Z.; Xu, J.; and Feng, D. 2017. AOD-Net: All-In-One Dehazing Network. In *ICCV*.

Li, B.; Ren, W.; Fu, D.; Tao, D.; Feng, D.; Zeng, W.; and Wang, Z. 2018. Benchmarking single-image dehazing and beyond. *TIP*.

Li, S.; Wang, Z.; Liu, Z.; Tan, C.; Lin, H.; Wu, D.; Chen, Z.; Zheng, J.; and Li, S. Z. 2023. MogaNet: Multi-order Gated Aggregation Network. In *ICLR*.

Liang, J.; Cao, J.; Sun, G.; Zhang, K.; Van Gool, L.; and Timofte, R. 2021. SwinIR: Image Restoration Using Swin Transformer. In *ICCVW*.

Liu, X.; Ma, Y.; Shi, Z.; and Chen, J. 2019a. Griddehazenet: Attention-based multi-scale network for image dehazing. In *ICCV*.

Liu, Y.; Pan, J.; Ren, J.; and Su, Z. 2019b. Learning deep priors for image dehazing. In *ICCV*.

Loshchilov, I.; and Hutter, F. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Mao, X.; Liu, Y.; Shen, W.; Li, Q.; and Wang, Y. 2021. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*.

Mo, Y. 2022. DCA-CycleGAN: Unsupervised single image dehazing using Dark Channel Attention optimized CycleGAN. *JVCIR*.

Pan, J.; Sun, D.; Pfister, H.; and Yang, M.-H. 2018. Deblurring Images via Dark Channel Prior. *PAMI*.

Pan, Z.; Cai, J.; and Zhuang, B. 2022. Fast Vision Transformers with HiLo Attention. In *NeurIPS*.

Park, N.; and Kim, S. 2022. How Do Vision Transformers Work? In *ICLR*.

Qin, X.; Wang, Z.; Bai, Y.; Xie, X.; and Jia, H. 2020. FFA-Net: Feature fusion attention network for single image dehazing. In *AAAI*.

Qiu, Y.; Zhang, K.; Wang, C.; Luo, W.; Li, H.; and Jin, Z. 2023. MB-TaylorFormer: Multi-branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing. In *ICCV*.

Rao, Y.; Zhao, W.; Tang, Y.; Zhou, J.; Lim, S. N.; and Lu, J. 2022. HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. In *NeurIPS*.

Ruikun Zhang, . L. P., Zhiyuan Yang. 2024. DehazeMamba: large multimodal model guided single image dehazing via Mamba. In *Visual Intelligence 2*.

Shen, H.; Zhao, Z.-Q.; Zhang, Y.; and Zhang, Z. 2023. Mutual Information-driven Triple Interaction Network for Efficient Image Dehazing. In *ACM MM*.

Si, C.; Yu, W.; Zhou, P.; Zhou, Y.; Wang, X.; and Yan, S. 2022. Inception Transformer. In *NeurIPS*.

Song, Y.; He, Z.; Qian, H.; and Du, X. 2023. Vision transformers for single image dehazing. *TIP*.

Tan, M.; and Le, Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*.

Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; and Li, Y. 2022. MAXIM: Multi-Axis MLP for Image Processing. In *CVPR*.

Wang, J.; Chen, Y.; Chakraborty, R.; and Yu, S. X. 2020. Orthogonal Convolutional Neural Networks. In *CVPR*.

Wang, P.; Zheng, W.; Chen, T.; and Wang, Z. 2022a. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. *ICLR*.

Wang, Y.; Xiong, J.; Yan, X.; and Wei, M. 2023. Uscformer: Unified transformer with semantically contrastive learning for image dehazing. *TITS*, 24(10): 11321–11333.

Wang, Y.; Yan, X.; Guan, D.; Wei, M.; Chen, Y.; Zhang, X.-P.; and Li, J. 2022b. Cycle-snspgan: Towards real-world image dehazing via cycle spectral normalized soft likelihood estimation patch gan. *TITS*, 23(11): 20368–20382.

Wang, Y.; Yan, X.; Wang, F. L.; Xie, H.; Yang, W.; Zhang, X.-P.; Qin, J.; and Wei, M. 2024. Ucl-dehaze: Towards real-world image dehazing via unsupervised contrastive learning. *TIP*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*.

Woo, S.; Park, J.; Lee, J.-Y.; and Kweon, I. S. 2018. CBAM: Convolutional Block Attention Module. In *ECCV*.

Wu, H.; Qu, Y.; Lin, S.; Zhou, J.; Qiao, R.; Zhang, Z.; Xie, Y.; and Ma, L. 2021. Contrastive Learning for Compact Single Image Dehazing. In *CVPR*.

Yan, Y.; Ren, W.; Guo, Y.; Wang, R.; and Cao, X. 2017. Image Deblurring via Extreme Channels Prior. In *CVPR*.

Yao, M.; He, D.; Li, X.; Li, F.; and Xiong, Z. 2023a. Towards Interactive Self-Supervised Denoising. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yao, M.; He, D.; Li, X.; Pan, Z.; and Xiong, Z. 2023b. Bidirectional Translation Between UHD-HDR and HD-SDR Videos. *IEEE Transactions on Multimedia*.

Yao, M.; Xu, R.; Guan, Y.; Huang, J.; and Xiong, Z. 2024. Neural degradation representation learning for all-in-one image restoration. *IEEE Transactions on Image Processing*.

Ye, T.; Zhang, Y.; Jiang, M.; Chen, L.; Liu, Y.; Chen, S.; and Chen, E. 2022. Perceiving and Modeling Density for Image Dehazing. In *ECCV*.

Yuan, S.; Chen, J.; Li, J.; Jiang, W.; and Guo, S. 2023. LH-Net: A Low-cost Hybrid Network for Single Image Dehazing. In *ACM MM*.

Zamir, S. W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F. S.; and Yang, M.-H. 2022. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*.

Zhang, H.; Hu, W.; and Wang, X. 2022. EdgeFormer: Improving Light-weight ConvNets by Learning from Vision Transformers. In *ECCV*.

Zhang, J.; Cao, Y.; Fang, S.; Kang, Y.; and Wen Chen, C. 2017. Fast Haze Removal for Nighttime Image Using Maximum Reflectance Prior. In *CVPR*.

Zheng, Y.; Zhan, J.; He, S.; Dong, J.; and Du, Y. 2023. Curricular contrastive regularization for physics-aware single image dehazing. In *CVPR*.

Zou, W.; Jiang, M.; Zhang, Y.; Chen, L.; Lu, Z.; and Wu, Y. 2021. Sdwnet: A straight dilated network with wavelet transformation for image deblurring. In *ICCV*.