Depth-Aware Blind Image Decomposition for Real-World Adverse Weather Recovery

Chao Wang[1], Zhedong Zheng[2], Ruijie Quan[3], and Yi Yang[3]

[1] ReLER Lab, AAII, University of Technology Sydney, Australia
[2] FST and ICI, University of Macau, China
[3] ReLER Lab, CCAI, Zhejiang University, China
chao.wang-11@student.uts.edu.au, zhedongzheng@um.edu.mo,
{quanruijie,yangyics}@zju.edu.cn

Abstract. In this paper, we delve into Blind Image Decomposition (BID) tailored for real-world scenarios, aiming to uniformly recover images from diverse, unknown weather combinations and intensities. Our investigation uncovers one inherent gap between the controlled lab settings and the complex real-world environments. In particular, existing BID methods and datasets usually overlook the physical property that adverse weather varies with scene depth rather than a uniform depth, thus constraining their efficiency on real-world photos. To address this limitation, we design an end-to-end Depth-aware Blind Network, namely **DeBNet**, to explicitly learn the depth-aware transmissivity maps, and further predict the depth-guided noise residual to jointly produce the restored output. Moreover, we employ neural architecture search to adaptively find optimal architectures within our specified search space, considering significant shape and structure differences between multiple degradations. To verify the effectiveness, we further introduce two new BID datasets, namely BID-CityScapes and BID-GTAV, which simulate depth-aware degradations on real-world and synthetic outdoor images, respectively. Extensive experiments on both existing and proposed benchmarks show the superiority of our method over state-of-the-art approaches.

Keywords: Image decomposition · Scene depth · Weather recovery

1 Introduction

Image restoration against adverse weather remains a classical yet challenging task for many real-world applications, *e.g.*, auto-driving car, with increasing demands on safety and robustness. Traditional methods usually focus on specific pre-defined weather conditions, such as image deraining [10, 22, 48, 55], dehazing [17, 24, 45, 66], desnowing [3, 4, 35], low-light enhancement [14, 19, 29, 37], adherent raindrop removal [41, 44, 62], *etc.* In an attempt to increase model scalability, researchers further propose all-in-one networks [2, 32, 58, 63, 64] to universally handle different degradations with multiple model ensembles. However, this line of methods requires extra individual training for each specific

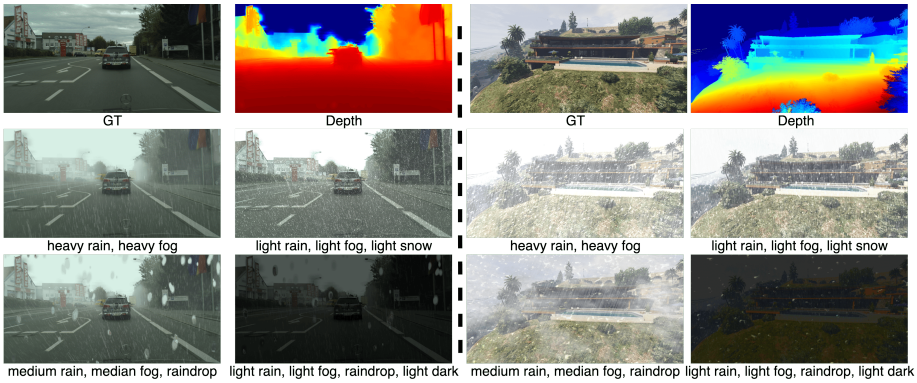

Fig. 1: Example images in two proposed depth-aware datasets (**left:** BID-CityScapes *(real-world)*, **right:** BID-GTAV *(synthesized)*). Different from existing datasets, we explicitly involve depth into the simulation process. For instance, rain exists in different formats such as rain streak and raindrop. Rain streak usually appears in the remote area, while raindrops are closer to the camera. Similarly, we also consider snow, haze and dark illumination, which often co-occur during rain. When testing, BID setting demands the model to handle images with an arbitrary combination of multiple adverse weathers, which is challenging yet more overarching.

weather task. Considering efficiency, some researchers resort to a unified architecture [5,26,31,50,57]. Such an approach can handle a single type of corruption at one time, but still suffers from multiple adverse weather combinations. In nature, adverse weather conditions tend to occur simultaneously in a random combination, *e.g.*, rain usually comes with fog and dim lighting. Therefore, in this work, we study Blind Image Decomposition (BID) task [15,51], which takes a step closer to real-world practice. BID considers the corrupted images as an arbitrary combination of degradation layers and separates the superimposed image into constituent underlying images in a **blind** setting, *i.e.*, both the source components involved in mixing as well as the mixing mechanism are unknown.

Existing methods and datasets for BID [15, 51] have achieved competitive performance in the public academic datasets, but usually ignored an inherent property that adverse weather varies with scene depth instead of a uniform depth, limiting the scalability to real-world cases. For instance, objects closer to the camera are mainly affected by the rain streaks with more light reflected into the camera, while objects far away are affected more heavily by the fog and the low-light condition. There remain two problems: (1) How to leverage the depth? Despite a few works [18, 47] have discussed degradation modeling combined with depth information, they are still limited to specific weather conditions and fail to work under the BID setting. (2) The scarcity of adverse weather data with depth. The lack of depth-aware BID datasets comprising different weather combinations, intensities and their corresponding ground truth also impedes decomposition algorithm development, which is closer to real-world applications.

To address these two limitations, we propose a new depth-aware network, dubbed **DeBNet** and two large-scale datasets, *i.e.*, BID-CityScapes and BID-

GTAV. (1) In particular, DeBNet simultaneously restores arbitrary hybrid adverse weather conditions in a generic framework. DeBNet is an end-to-end network encapsulating the underlying physics principles behind the formation of depth-aware BID. Specifically, DeBNet takes the corrupted image as inputs, predicts a depth map, and then produces a clean image as the output. Considering the features of different weathers usually do not share the same characteristics, we further resort to a Neural Architecture Search (NAS) method to achieve the optimal accuracy-efficiency trade-offs. Specifically, we design a specific BID search space that consists of several effective fundamental restoration operations, such as multi-scale convolutions [49] and self-attention modules [65]. (2) Meanwhile, in order to enable depth-aware BID training, we first construct a BID dataset in real-world autonomous driving scenes based on the CityScape [6] dataset, dubbed **BID-CityScapes**, including five types of adverse weather mixed under BID setting, as well as their corresponding depth maps and clean images. Moreover, to enrich the diversity and enable DeBNet to generalize well on real-world images with random viewpoints, we further generate a high-resolution synthetic dataset using a commercial video game Grand Theft Auto (GTAV) [46], called **BID-GTAV**. Examples of the two datasets are illustrated in Fig. 1. Unlike existing methods that are limited to type-specific or depth-independent degradation modeling, we analyze the physical properties of weather degradations and formulate a depth-aware BID setting including five different weather conditions: rain streak, fog, raindrop, snow and low-light degradation. Our contributions are as follows:

- **What is the remaining gap between lab and real-world images against adverse weather?** We identify one overlooked practical problem between the depth property and the corrupted observation, and introduce a new Depth-aware Blind Network, named **DeBNet**, to explore the inherent depth prior of multiple adverse weathers. The key idea underpinning the network design follows the reverse process to disentangle the noise with the aid of depth prediction. In particular, we leverage the neural architecture search to adaptively find the optimal architecture for processing depth and visual features in an end-to-end manner.
- To verify the effectiveness of the proposed method, we introduce two new BID datasets, called BID-CityScapes and BID-GTAV. To our knowledge, the datasets are the first two depth-aware BID datasets including different types of degradations captured across various scenes and viewpoints. Extensive experiments on two proposed BID benchmarks substantiate that our learned model achieves competitive PSNR and SSIM scores, and is scalable to other existing image restoration benchmarks, including SPAdata [56], DeRaindrop [41], SOTS [25], Snow100K [36], SICE [1] and RainDS [43].

2 Related Work

Blind Image Decomposition. Aiming at the adverse weather removal task, several restoration works [9, 33, 41, 60] have achieved satisfactory results on var-

ious degradations. However, these methods still require individual training for each type of degradation, impeding the real-world all-in-one implementation. Based on blind source separation problem [13, 23], Han *et al*. first propose the "Blind Image Decomposition" (BID) [15], regarding rain and other real-world weather corruptions as superimposing and separable to a clean image. Wang *et al*. further introduce a two-stage BID learning paradigm [51] to utilize a pre-trained masked autoencoder (MAE) [16] for efficient fine-tuning on the restoration network. However, these BID methods remain heavily dependent on tedious multi-scale multi-head reconstruction or time-consuming pertaining, and neglect the inherent depth-related physical properties of weather degradations. In comparison, our end-to-end method explicitly leverages the depth prior for a better understanding of the corrupted image against adverse weathers.

Neural Architecture Search (NAS). Neural Architecture Search [21, 34, 40, 67] automates the designing of neural network architectures for optimal performance while minimizing human hours and efforts. For instance, Li *et al*. first applies NAS to all-in-one weather removal task [31], while Gou *et al*. further propose an efficient search space [12] with three task-flexible modules. Quan *et al*. introduce a two-stage searching and training strategy [43] for the joint removal of raindrops and rain streaks. Yet these methods all tend to search the combinations of each individual operation for specific degradations, which are not suitable for the BID problem. Our work is closely related to [7, 31, 34] to further formulate the searching task into an end-to-end optimization problem. Different from previous methods, we adopt a unified multi-branch search space to explore the weightings of different operations, considering the inherent patterns of varying feature interference among multiple weathers.

3 Method

3.1 Problem Formulation

Existing methods are usually type-specified and ignore the scene depth during weather formulation. In this paper, we first propose a uniform BID imaging model inherited with depth annotations. Without loss of generality, the degraded image I_{noisy} can be formulated as a composition of a clean background I_{clean} and multiple weather degradation layers, including raindrop (RD), snowflakes (SN), lighting (L), rain streak (RS), and fog (A). The formulation can be written as:

$$I_{noisy} = \overbrace{\underbrace{L}_{Lighting} \odot [\underbrace{I_{clean}(1-RS-A)}_{Transparency} + \underbrace{RS + A_0 A}_{Scene\ Occlusion}]}^{Depth-variant} \odot \overbrace{\underbrace{(1 - M_{rd} \odot M_{sn})}_{Non-occluded\ Area} + \underbrace{RD + SN}_{Lens\ Occlusion}}^{Depth-invariant}. \tag{1}$$

where $\odot$ denotes the element-wise multiplication. From left to right, $L \in [0, 1]$ represents the illumination intensity. $RS \in [0, 1]$, $A \in [0, 1]$ are single-channel soft masks for rain streak and fog with a larger value indicating a higher weather intensity (lower transparency). A_0 is the atmospheric light intensity [47]. $M_{rd} \in \{0, 1\}$ and $M_{sn} \in \{0, 1\}$ denote the masks of raindrop and snow area, which

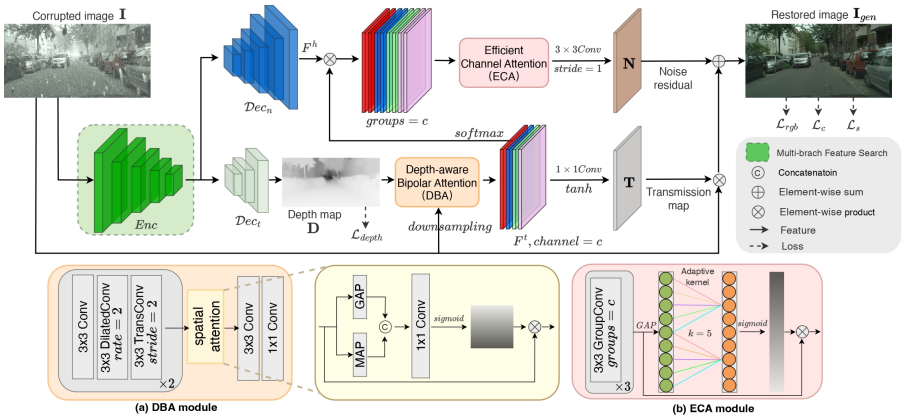

Fig. 2: The architecture of our **D**epth-aware **B**lind image decomposition **N**etwork (**DeBNet**). Given the corrupted image I_{noisy} with adverse weather, we extract the visual features via a searchable multi-branch encoder Enc (See Sec. 3.3 and Fig. 3). Then two decoder branches (Sec. 3.2) are deployed to progressively upsample the feature map with skip-connected features from the encoder. In particular, the transmittance decoder branch Dec_t is to predict the scene depth D. The noise branch Dec_n combines the learned attention weights with reconstructed high-resolution features. As shown in (a), we apply Depth-aware Bipolar Attention (DBA) module to compare the depth and the input image as the transmissivity map T. Besides, we also adopt an Efficient Channel Attention (ECA) module to fuse the depth and visual features to produce the depth-guided noise residual N as shown in (b). Finally, we fuse T, N, and I_{noisy} for joint restoration following Eq. (3).

are usually transparent to some extent. RD is the visual occlusion brought by adherent raindrops on lens, representing the blurred imagery reflected by the environment. Similarly, SN represents snowflakes on lens, which totally occlude the scene. It is worth noting that the scene area is usually depth-variant, while the occlusion on lens is depth-invariant. Therefore, for the first term, we should involve the depth prior, and not consider depth for the occlusion on lens. In particular, for **rain streak & fog:** we follow the formulation [18] to simulate the observation changes according to the depth; for **low-light**, we follow existing low-light image enhancement works [8, 28] to regard inverted low-lighted inputs as haze images. We also introduce depth information into low-light modeling, which aligns with human intuition that the farther the distance, the heavier the fog, and the weaker the illumination. for **raindrop & snow:** we follow the degradation in [36, 41], to split occlusion as two different types, *i.e.*, scene and lens. More details on the degradation modeling are given in the **suppl**.

Considering the multiplication and addition operation in Eq. (1), we can simplify it as follows:

$$I_{noisy} = W I_{clean} + b, \tag{2}$$

where W and b denote the multiplicative and additive factors, respectively. Both factors are partially related to the scene depth. Our depth-aware BID task is a

reverse prediction task, and thus can be formulated as:

$$I_{gen} = TI_{noisy} + N, \tag{3}$$

where the **Transmittance** map $T = W^{-1}$ and the **Noise** map $N = -W^{-1}b$. In general, T denotes the positional and intensity information (*e.g.*, the density of rain and fog), while N tends to characterize the intrinsic information of the degradation itself (*e.g.*, blurry or obstruction effects of the raindrops). As shown in Fig. 2, this paper leverages the deep neural network to predict the value of T and N. Different from previous methods that directly learn an additive and non-convex degradation residual [18,43], our method explicitly separates the learning of transmittance and noise factors, which is more conducive for feature learning under multiple overlapping weather conditions.

3.2 Depth-aware BID Network

Depth-aware Transmittance Map. As shown in the bottom branch of Fig. 2, the transmittance decoder Dec_t predicts the scene depth with direct supervision. Following [18], we calculate the depth reconstruction loss, which is the $\mathcal{L}_2$ distance between the predicted depth values D_{gen} and ground truth depth D_{gt}:

$$\mathcal{L}_{depth} = \|D_{gen} - D_{gt}\|_2 , \tag{4}$$

where D_{gen} and D_{gt} are normalized into $[0, 1]$ with size of a quarter of the input image. To further utilize the learned depth, we propose a Depth-aware Bipolar Attention (DBA) module to produce the transmittance map, *i.e.*, T in Eq. (3). Considering that different weather conditions lead to particular pixel enhancement (*e.g.*, fog, rain) or attenuation (*e.g.*, low light), we adopt a continuous bipolar mask between $[-1, 1]$ for the transmittance map T, indicating both contextual information and coarse weather types underlying the images. As shown in Fig. 2 (a), the proposed DBA module takes both downsampled input image and depth map as inputs, containing two cascaded convolution blocks followed by an attention block [59] to emphasize the spatial information. Finally, we apply tanh function to normalize features and generate the transmittance map T.

Depth-guided Noise Residual. According to Eq. (3), the additive noise map N is also related to the scene depth D, thus we further introduce the noise decoder branch Dec_n to generate the final high-resolution feature map F^h, and then combine it with the learned transmittance features F^t from transmittance branch to produce the noise map N.

Similar to [18], we consider the output of the last convolutional block in the DBA module as a set of un-normalized attention weights, representing the complete transmittance features. In general, each weight $F_i^t (i = 1, 2, ..., c)$ corresponds to a certain type of weather degradation after softmax. We divide the features F^h into c groups as $F_i^h (i = 1, 2, ..., c)$, where c is the channel number of our transmittance features F^t ($c = 32$ in this work). Then, we re-weight each submap F_i^h through an element-wise multiplication with F_i^t to produce the final noise map. After the multiplication, we further perform group convolutions [61]

in c groups to individually refine the features of different types of degradation. Since the channel number of a weather condition varies under different combinations, we further adopt an efficient channel attention mechanism (ECA) [54] with the adaptive kernel size k as $|\frac{\log_2(C)+1}{2}|_{odd}$, where C is the channel number of the input features. $|t|_{odd}$ indicates the nearest odd number of t. As shown in Fig. 2 (b), ECA can adaptively learn a local cross-channel interaction, thus improving the robustness against different hybrid weather conditions. Finally, we merge all the features from different groups using a 1×1 convolution to produce the noise residual N, with which we combine the transmittance map T to produce the restored clean image I_{gen} following Eq. (3). The reconstruction loss on the restored image can be written as:

$$\mathcal{L}_{rgb} = \|I_{gen} - I_{gt}\|_2 , \tag{5}$$

where I_{gen} is the restored image and I_{gt} denotes the ground truth.

Conditional Adversarial Training. To mimic the distributions of the clean image I_{gt}, we first introduce a discriminator Dis to distinguish whether the input image is real or generated. Our source adversarial loss $\mathcal{L}_s$ is written as:

$$\mathcal{L}_s = \mathbb{E}\left[\log\left(1 - Dis\left(I_{gen}\right)\right)\right] + \mathbb{E}\left[\log Dis\left(I_{gt}\right)\right], \tag{6}$$

where $Dis(I)$ predicts the possibility that the image I is uncorrupted. For this discriminator, we hope $Dis(I_{gt}) = 1$ and $Dis(I_{gen}) = 0$, so we maximize the $\mathcal{L}_s$. Meanwhile, we also deploy a degradation classifier Cls to classify the weather combinations from the restored image. The multi-class conditional loss $\mathcal{L}_c$ can be defined as:

$$\mathcal{L}_c = -\sum_{i=0}^{N-1} \log\left(1 - Cls(I_{gen})_i\right), \tag{7}$$

where N denotes the number of degenerate types, $Cls(I)_i$ denote the possibility that the image I suffer from degeneration type i. In this paper, we select five common weather types, *e.g.*, rainstreak, raindrop, snow, fog and low-light condition. It is worth noting that we have multiple combined weather during training. Therefore, we apply 5-dimensional multi-class vectors to indicate the combined weather type. The weather type is not fixed but randomly generated during training. The overall loss function $\mathcal{L}_{gen}$ for DeBNet is formulated as:

$$\mathcal{L}_{gen} = \mathcal{L}_{depth} + \mathcal{L}_{rgb} + \lambda_s \mathcal{L}_s + \mathcal{L}_c, \tag{8}$$

here we empirically set $\lambda_s = 0.1$ to balance the relative weight of $\mathcal{L}_s$.

3.3 Multi-branch Architecture Search

Multi-branch feature connection is a potential solution for the BID task [53], since image restoration is a classical position-sensitive vision problem requiring precise spatial information and rich semantics. To adaptively restore images with various weather combinations, we adopt a multi-branch search method [7]

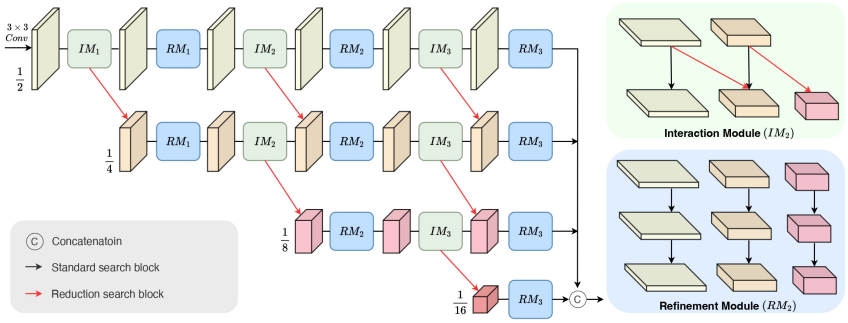

Fig. 3: Illustration of the multi-branch feature search, which contains three-stage of alternated Interaction Module IM and Refinement Module RM. In this work, we deploy the multi-branch learning after a standard 3×3 convolution layer to decrease the resolution by half. Each black line represents a standard search block, and the red line denotes a reduction search block for downsampling. Taking the second stage IM_2, RM_2 as an example, IM_2 generates an extra branch from the previous lowest-resolution branch by a reduction searching block (red line), while RM_2 further refine the features using cascaded standard search blocks (black line) within the same resolution.

to explore the optimal architecture for various weather combinations. Different from simply utilizing Neural Architecture Search (NAS) to select features from different encoders [31, 43], we deeply involve NAS to extract depth and visual features from different convolution kernels and the transformer branch in an efficient way, considering different types of weather correspond to varying sizes of receptive field. In this way, we introduce a channel/token-wise fine-grained search strategy embedded in a multi-branch high-resolution feature space.

Search Space. Our multi-branch search space is embedded in the encoder Enc. Specifically, the network consists of two modules: the refinement module RM and the interaction module IM, as shown in Fig. 3. Each module is composed of several search blocks operating at different resolutions. We alternate between using two modules to construct a multi-branch search space. The interaction module $IM_i(i = 1, 2, 3, ..., n)$ achieves high-to-low resolution feature transformation to maintain more semantic details, while the refinement module $RM_i(i = 1, 2, 3, ..., n)$ obtains larger receptive fields and multi-scale features by stacking searching blocks in each branch. Finally, multi-branch features are resized and concatenated together, connected to the double-branch decoder.

Search Block. Unlike previous NAS-based restoration methods [31,43] that are designed for specific tasks, we aim to customize the network for various weather combinations. Our searching block contains two paths: a MixConv [49] path for multi-scale feature extraction, and a lightweight Transformer [7] to provide more global contexts in a residual manner. Instead of selecting different sizes or stacking order of the single-scale convolution kernels, Mix-Conv divides all channels into groups and applies multi-scale convolution kernels to each group in a depth-wise complementary way, which can extract features with different receptive field sizes. The number of convolutional channels and the number of tokens in the Transformer are searchable parameters as channels in depth-wise convo-

lutions are independent in the searching block [7, 38], any convolution channels or transformer queries can be easily removed without affecting the other search blocks. More details on the search block are given in the **suppl**.

Search Algorithm. Following Darts [34], we use the importance factors which are learned jointly with the network weights of each search block. We also adopt a resource-aware $\mathcal{L}_1$ penalty [7] to push the importance factors of high computational costs to zero. More details on the resource-aware penalty can be found in the **suppl**. Combing with this resource-aware penalty term $\mathcal{L}_{l_1}$ with an attenuation coefficient λ set as 0.01, the overall training loss is:

$$\mathcal{L}_{total} = \mathcal{L}_{gen} + \lambda \mathcal{L}_{\ell 1}. \tag{9}$$

4 Datasets

BID-CityScapes. There are several large-scale synthetic datasets [15, 18, 30] available for training restoring networks. However, none of them considers depth effects under the BID setting, thus impeding the performance for real-world images. To enable the depth supervision for DeBNet, we introduce a new depth-aware BID dataset named BID-CityScapes, using images from CityScapes [6] dataset as background. We first generate synthetic degradation layers based on the provided camera parameters and scene depth. We apply the rendering strategy in [18, 47] to smooth the degradation layers, and then generate the final corrupted images according to Eq. (1). Altogether, our BID-CityScapes dataset has 6,000 training image pairs and 800 pairs for testing. Each pair contains a clean background image, a depth map from the original dataset, and 14 types of different weather combinations (see Fig. 1 left), including different types and intensities. Although BID-CityScapes dataset has effectively expanded the practicality for BID tasks, its imaging perspective is restricted to autonomous driving scenes, which limits the generalization on real-world photos with various imaging scenes and viewpoints. More details and examples are given in the **suppl**.

BID-GTAV. We further propose the **BID-GTAV** dataset with more diverse scenes as a supplement (see Fig. 1 right). The dataset is rendered from Grand Theft Auto V (GTAV) [46], an open-world game with large-scale city models. We capture the images and corresponding ground-truth depth maps from the game with two plugins, Script Hook V and Script Hook V.NET [20]. We endeavor to leverage the rich virtual worlds created for major video games to simulate real-world scenarios with high-level fidelity and various viewpoints. We extract 8,000 pairs of images for training and 1,000 pairs for testing. The extracted pairs contain various weather conditions using graphics debugging mods, as well as their clear counterparts and precise depth maps.

5 Experiment

5.1 Implementation Details

Our model is trained in an end-to-end manner following [31,34]. The training set is split into a search training part Set_1 (70%) and a search validation part Set_2

Table 1: Quantitative results on our synthetic BID-CityScapes and BID-GTAV datasets. We evaluate the performance in Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) under 5 BID cases, which are (1) heavy rain + heavy fog, (2) light rain + light fog + snow, (3) medium rain + medium fog + raindrop, (4) light rain + light fog + light dark, (5) light rain + light fog + light dark + raindrop. The best performance under each case is marked in **bold**.

case	BID-CityScapes								BID-GTAV							
	TransWeather		BIDeN		CPNet		DeBNet		TransWeather		BIDeN		CPNet		DeBNet	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
(1)	23.07	0.845	25.70	0.869	26.97	0.902	**29.51**	**0.916**	18.45	0.623	19.55	0.699	22.30	0.725	**23.88**	**0.751**
(2)	24.33	0.877	26.05	0.880	27.18	0.904	**29.72**	**0.920**	18.88	0.631	19.80	0.720	22.57	0.731	**24.31**	**0.763**
(3)	24.20	0.865	25.61	0.867	27.05	0.901	**29.54**	**0.917**	18.42	0.622	19.61	0.700	22.31	0.727	**24.05**	**0.757**
(4)	22.39	0.801	25.11	0.861	25.87	0.869	**28.35**	**0.909**	18.33	0.619	18.78	0.643	20.90	0.713	**23.02**	**0.744**
(5)	22.10	0.792	24.07	0.849	25.44	0.857	**28.18**	**0.907**	18.01	0.609	18.51	0.631	20.67	0.701	**22.71**	**0.737**

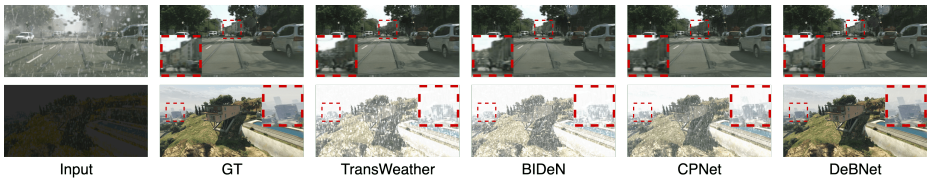

Fig. 4: Qualitative comparisons on BID-CityScapes and BID-GTAV under several mixed cases. DeBNet produces clean and precise restored images without visible artifacts or color shifts, especially for depth-related scenarios. Please zoom in to see details.

(30%). We simultaneously optimize the architecture parameters on Set_1 and the network parameters on Set_2. Each search block contains two operating paths. For the MixConv path, we set the expansion rate as 4. In the reduction block, we set the stride of the depthwise convolution as 2. For the Transformer path, we set $s = 8$. The number of attention heads is 1, and the hidden dimension of the attended subspaces is set as 64. We resize the features into half size of the former resolution during inverse projection in Transformer to achieve downsampling. We progressively remove the search units with less importance [34] and re-calibrate the running statistics of BN layers after every 5 epochs. After the architecture search, DeBNet can fully explore global and local information across different weather degradations and restore clean images. Besides, since the network training and architecture search are conducted in a unified end-to-end manner, the resulting network can be used directly without fine-tuning. More implementation details are given in the **suppl**. Code will be released in `https://github.com/Oli-iver/Depth-BID`.

5.2 Synthetic BID Analysis

Instead of directly combining existing single-degradation datasets [31], we further propose two depth-aware BID datasets, BID-CityScapes and BID-GTAV, across various scenes and viewpoints by incorporating depth information during simulation. We conduct extensive experiments on the proposed two datasets to evaluate the performance of the proposed DeBNet and the existing state-of-the-art BID methods BIDeN [15] and CPNet [51], as well as an all-in-one method

Table 2: Quantitative results for benchmarking the proposed DeBNet and the state-of-the-art methods on real-world deraining SPAdata [56] and dehazing SOTS-outdoor [25] test sets. The original pre-trained weights of all these models are directly used for evaluation. We have also trained these methods on our BID-CityScapes (red) and BID-GTAV (blue) datasets. Results of the two-stage refined DeBNet$^+$ are marked in green. The best performance is marked in **bold**.

Method	SPAdata [56]						SOTS-outdoor [25]					
	RCDNet [52]			AirNet [27]			DeBNet			FFA-Net [42]			AirNet [27]			DeBNet		
PSNR	34.08	36.72	34.11	34.05	35.32	34.40	36.69	34.78	**36.77**	33.07	**34.60**	31.25	33.09	34.51	31.10	33.50	31.57	33.82
SSIM	0.953	0.969	0.958	0.948	0.963	0.960	0.965	0.954	**0.973**	0.980	0.982	0.975	0.979	**0.983**	0.968	0.986	0.982	0.989

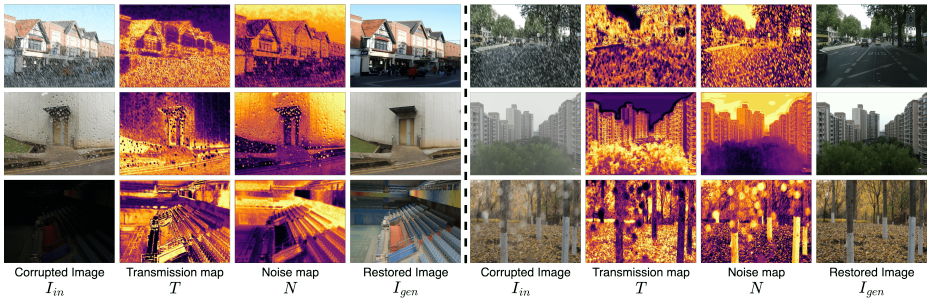

Fig. 5: Visualization of the learned Transmittance map T and Noise map N on different weather scenarios. T and N are re-normalized into $[0, 1]$ for better visualization.

TransWeather [50]. We select five common weather combinations existing in the real world for evaluation as: (1) heavy rain + heavy fog, (2) light rain + light fog + snow, (3) medium rain + medium fog + raindrop, (4) light rain + light fog + light dark, (5) light rain + light fog + light dark + raindrop. All the methods are trained and evaluated under the same settings for fair comparisons. The quantitative results on the two proposed synthetic datasets are reported in Table 1. It can be found that our DeBNet delivers state-of-the-art performance under the BID setting on both datasets. Notably, our method performs favorably against other counterparts under the scenes where, *e.g.*, DeBNet exceeds CPNet by 2.54dB on PSNR under heavy rain and fog combinations. This phenomenon can be attributed to the explicit utilization of scene depth in DeBNet, which encourages the model to leverage spatial information while extracting features. We also show some qualitative comparisons in Fig. 4. Due to the page limitation, more analyses and comparisons on synthetic datasets are given in the **suppl**.

5.3 Real-world BID Analysis

To validate the effectiveness of our approach, we further conducted experiments on real-world images. As existing real-world datasets are limited with single-type degradation and lack the ground-truth scene depth, we directly evaluated the generalization ability of models trained respectively on BID-CityScapes and BID-GTAV datasets with real-world test sets [25, 56]. To further validate the effectiveness of our BID-GTAV dataset constructed for multi-view natural image restoration, we further prepared a two-stage learned model (DeBNet$^+$) which

is first trained on the BID-CityScapes dataset, and then fine-tune the searched model on the BID-GTAV dataset. Table 2 shows the quantitative results of different restoration tasks. It can be found that though the performance of DeBNet is constrained by the BID training setting, our method remains competitive (DeBNet$_a$) or even better performance (DeBNet$^+$) against other task-specific counterparts. This experiment also indicates the quality of our proposed datasets, as the depth-aware modeling method effectively improves the generalization ability on real-world images. More results on other real-world restoration tasks can be found in the **suppl**.

5.4 Discussion and Ablation Study

Map visualization. To validate the effectiveness of our proposed depth-aware BID modeling approach, we visualized the learned transmittance map T and noise map N under real-world scenarios. As shown in Fig. 5, both T and N are highly spatially related to the scene depth. **(1).** In highly depth-related scenes like low-light and fog, the transmittance map T tends to capture low-frequency patterns while the noise map N focuses on high-frequency details, *e.g.*, small points. **(2).** For raindrops and snow scenes, it is challenging to distinguish transmittance and noise maps from a frequency perspective, as the chaotic distribution of such noise constitutes low-frequency interference.

Depth branch. The design of the depth branch is motivated by two points: (1) The off-the-shelf depth models are not robust to complicated weather, due to various occlusions and illumination. Therefore, we decide to re-train the depth estimation task from scratch. (2) Combining with the depth estimation is complementary to our main task, *i.e.*, image decomposition, since the low-level feature of depth tasks focuses on the distance and object edges. As shown in Fig. 6, we further select real images from KITTI dataset [11] under autonomous driving scenarios, and randomly generate outputs using the modeling method as Eq. (3) for the scene depth D visualization. It can be found that our method can not only achieve ideal BID restoration but also predict reasonable scene depths.

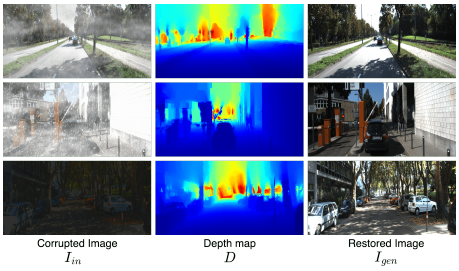

Fig. 6: Predicted depths on real images.

Comparisons with diffusion-based method. We also provide comparisons with another diffusion-based method WeatherDiff [39] in Table 3a. We retrained the official model from scratch on our proposed dataset and tested it on two real-world datasets. It can be observed that, benefiting from the depth guidance and NAS targeted at various degradation combinations, our method DeBNet clearly outperforms the compared method, particularly in real-world datasets and complex scenarios.

Datasets. To further validate the effectiveness, we conduct experiments on three real-world datasets [1, 36, 41]. As shown from Table 3b, we could observe two

Table 3: More experiment results on different settings. The best performance is marked in **bold**. case-1: heavy rain + heavy fog, case-2: light rain + light fog + snow, case-5: light rain + light fog + light dark + raindrop. (a) Comparison with the diffusion-based method in terms of PSNR and SSIM. (b) More experiment results on different training datasets in terms of PSNR and SSIM. *ft* DeBNet means the model is pre-trained on target dataset / previous BIDeN / BID-CityScapes and fine-tuned on the training set of each evaluated dataset. (c) Ablation on DBA and ECA modules with BID-CityScapes dataset in terms of PSNR and SSIM. (d) Ablation on hyperparameters with BID-CityScapes dataset in terms of PSNR and number of parameters.

(a)

Methods	BID-CityScapes		SPAdata [56]	SOTS-outdoor [25]
	case-1	case-5	real-world rain	real-world haze
TransWeather [50]	23.07/0.845	22.10/0.792	33.75/0.942	29.89/0.959
WeatherDiff [39]	28.11/0.910	25.07/0.887	35.32/0.955	33.17/0.979
DeBNet	**29.51**/**0.916**	**28.18**/**0.907**	**36.69**/**0.965**	**33.50**/**0.986**

(b)

Variants	Pre-training Dataset	Deraindrop [41]	Snow100k-M [36]	SICE [1]
DeBNet_1	BIDeN dataset [15]	29.75/0.931	31.15/0.933	20.03/0.719
DeBNet_2	BID-CityScapes	32.79/0.942	33.80/0.948	21.79/0.730
ft DeBNet	from scratch	33.05/0.942	**33.99**/**0.953**	22.10/**0.738**
ft DeBNet_1	BIDeN dataset	32.89/0.941	33.75/0.942	22.03/0.735
ft DeBNet_2	BID-CityScapes	**33.11**/**0.945**	33.96/**0.953**	**22.12**/**0.738**

(c)

Ablation	w/o DBA	w/o ECA	DeBNet (full)
case-1	28.87 / 0.906	29.23 / 0.911	**29.51** / **0.916**
case-5	27.50 / 0.868	27.95 / 0.882	**28.18** / **0.907**

(d)

Variants		case-2	case-5
NAS $\mathcal{L}_{t_1}$	Uformer [58]	25.64 (50.90M)	22.05 (50.90M)
	$\lambda = 0.05$	26.82 (**30.55M**)	24.91 (**35.23M**)
	$\lambda = 0.005$	29.73 (44.73M)	**28.19** (47.53M)
	w/o NAS	**29.79** (53.15M)	27.88 (53.15M)
GAN $\mathcal{L}_s$	$\lambda_s = 0.05$	29.70 (39.08M)	**28.20** (42.50M)
	$\lambda_s = 0.5$	29.68 (38.95M)	28.05 (43.11M)
DeBNet ($\lambda = 0.01, \lambda_s = 0.1$)		29.72 (39.15M)	28.18 (42.37M)

points. **(1).** Since existing BID dataset [15] does not consider the impact of depth on weather imaging, there is a notable performance loss with DeBNet_1. In contrast, DeBNet_2 trained on our dataset generalizes well to real-world images, indicating our dataset's efficacy in prompting depth-related feature learning which is vital for real-world applications. **(2).** Pre-training on our dataset and fine-tuning on downstream datasets also yields better performance and shows scalability to different real-world scenarios. For instance, *ft* DeBNet_2 pretrained on our dataset surpassed the performance of models pretrained on existing datasets or from scratch, which is also evidenced in Table 2.

DBA and ECA modules. In our modeling with Eq. (1) and Eq. (3), the transmittance map primarily characterizes the location and density information of adverse weather conditions, which are intimately associated with depth. Thus we design a bipolar transmittance mechanism in DBA to fully exploit the learned depth features. Moreover, DBA further embeds diverse weather characteristics into the reconstructed image features through the ECA module. To understand the contributions of these two attention modules, we respectively replace the DBA and ECA modules into 1×1 convolution to build two variants. The quantitative results in Table 3c show that the absence of these two attention modules in linking depth information with image restoration leads to a significant decline in performance, particularly in complex composite scenes. In conclusion, high-quality BID requires spatial attention (DBA) to capture common features (*e.g.*, scene depth), as well as channel attention (ECA) to enhance distinctive features of different degradations.

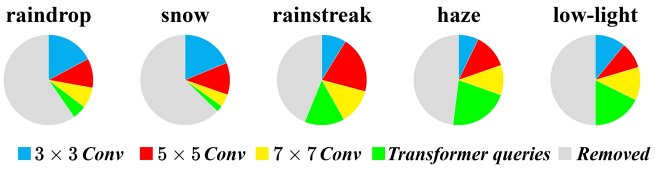

Fig. 7: Visualization of the searched architectures under different scenarios. For simplicity, we only show the averaged searching proportions in refinement modules when the feature size is scaled to 1/8 (see Fig. 3). More visualizations are given in the **suppl**.

Hyperparameters. 1. NAS $\mathcal{L}_{l_1}$. In practice, we can obtain different searched models using different λ values. As shown in Table 3d, we observe two points: (1) We usually could yield a small-size NAS model. Our model (λ=0.05) easily achieves a better PSNR than a raw end-to-end Uformer [58] with much fewer parameters. (2) For the simple scenes case-2, the proposed model (λ=0.005) could achieve a competitive performance compared with a bigger baseline model w/o NAS, while achieving a significant PSNR improvement in a challenging weather combination (case-5). Our DeBNet achieves a favorable trade-off between performance and efficiency. **2. GAN Loss $\mathcal{L}_s$.** We also conduct experiments on the hyperparameter λ_s of the source adversarial loss, it can be found that our trained model is not sensitive to the GAN loss weight λ_s.

Architecture search. Our multi-branch feature search aims to find the best structure to comprehensively handle BID in various weather combinations. To verify the role of different operations in our search space, we further use a fixed combination dataset for the search. As shown in Fig. 7, our method exhibits adaptability, finding dynamic architectures for different weather degradations. As for local weather degradations (*e.g.*, raindrop and snow), the model tends to favor smaller convolutional kernels. In comparison, for globally corrupted weather types (*e.g.*, haze and low-light), the model prefers utilizing larger-sized convolutions and attention mechanisms to achieve a broader receptive field.

6 Conclusion

In this paper, we explore the visual effects of various adverse weather conditions subject to scene depth and formulate a depth-aware BID imaging model. Considering the depth information within images, we propose a novel BID model, namely DeBNet, to restore arbitrary hybrid adverse weather conditions in a unified framework. Taking advantage of neural architecture search and the specifically designed restoring search space, we achieved an effective deraining network to remove various types of rain. To bridge the domain gap between real and synthetic images, we present two depth-aware BID datasets BID-CityScapes and BID-GTAV under real-world and synthetic scenes, respectively. Extensive experiments on our benchmarks and other real-world datasets demonstrate the effectiveness and superiority of our unified network. We hope that our dataset would take a step closer for transferring the research of blind image decomposition to real-world applications.

Acknowledgement

The paper is supported by Start-up Research Grant at the University of Macau (SRG2024-00002-FST).

References

1. Cai, J., Gu, S., Zhang, L.: Learning a deep single image contrast enhancer from multi-exposure images. IEEE Transactions on Image Processing **27**(4), 2049–2062 (2018)
2. Chen, L., Chu, X., Zhang, X., Sun, J.: Simple baselines for image restoration. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII. pp. 17–33. Springer (2022)
3. Chen, W.T., Fang, H.Y., Ding, J.J., Tsai, C.C., Kuo, S.Y.: Jstasr: Joint size and transparency-aware snow removal algorithm based on modified partial convolution and veiling effect removal. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 754–770. Springer (2020)
4. Chen, W.T., Fang, H.Y., Hsieh, C.L., Tsai, C.C., Chen, I., Ding, J.J., Kuo, S.Y., et al.: All snow removed: Single image desnowing algorithm using hierarchical dual-tree complex wavelet representation and contradict channel loss. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4196–4205 (2021)
5. Chen, W.T., Huang, Z.K., Tsai, C.C., Yang, H.H., Ding, J.J., Kuo, S.Y.: Learning multiple adverse weather removal via two-stage knowledge learning and multi-contrastive regularization: Toward a unified model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17653–17662 (2022)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
7. Ding, M., Lian, X., Yang, L., Wang, P., Jin, X., Lu, Z., Luo, P.: Hr-nas: Searching efficient high-resolution neural architectures with lightweight transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2982–2992 (2021)
8. Dong, X., Pang, Y., Wen, J.: Fast efficient algorithm for enhancement of low lighting video. In: ACM SIGGRApH 2010 posters, pp. 1–1 (2010)
9. Fan, Y., Yu, J., Mei, Y., Zhang, Y., Fu, Y., Liu, D., Huang, T.S.: Neural sparse representation for image restoration. Advances in Neural Information Processing Systems **33**, 15394–15404 (2020)
10. Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X., Paisley, J.: Removing rain from single images via a deep detail network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3855–3863 (2017)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
12. Gou, Y., Li, B., Liu, Z., Yang, S., Peng, X.: Clearer: Multi-scale neural architecture search for image restoration. Advances in Neural Information Processing Systems **33**, 17129–17140 (2020)

13. Gu, S., Meng, D., Zuo, W., Zhang, L.: Joint convolutional analysis and synthesis sparse representation for single image layer separation. In: ICCV (2017)
14. Guo, X., Li, Y., Ling, H.: Lime: Low-light image enhancement via illumination map estimation. IEEE Transactions on image processing **26**(2), 982–993 (2016)
15. Han, J., Li, W., Fang, P., Sun, C., Hong, J., Armin, M.A., Petersson, L., Li, H.: Blind image decomposition. In: Eur. Conf. Comput. Vis. (2022)
16. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
17. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE Trans. Pattern Anal. Mach. Intell. **33**(12), 2341–2353 (2010)
18. Hu, X., Fu, C.W., Zhu, L., Heng, P.A.: Depth-attentional features for single-image rain removal. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019)
19. Janner, M., Wu, J., Kulkarni, T.D., Yildirim, I., Tenenbaum, J.: Self-supervised intrinsic image decomposition. Advances in neural information processing systems **30** (2017)
20. Johnson-Roberson, M., Barto, C., Mehta, R., Sridhar, S.N., Rosaen, K., Vasudevan, R.: Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? arXiv preprint arXiv:1610.01983 (2016)
21. Jozefowicz, R., Zaremba, W., Sutskever, I.: An empirical exploration of recurrent network architectures. In: International conference on machine learning. pp. 2342–2350. PMLR (2015)
22. Kang, L.W., Lin, C.W., Fu, Y.H.: Automatic single-image-based rain streaks removal via image decomposition. IEEE transactions on image processing **21**(4), 1742–1755 (2011)
23. Levin, A., Weiss, Y.: User assisted separation of reflections from a single image using a sparsity prior. TPAMI **29**(9), 1647–1654 (2007)
24. Li, B., Peng, X., Wang, Z., Xu, J., Feng, D.: Aod-net: All-in-one dehazing network. In: Proceedings of the IEEE international conference on computer vision. pp. 4770–4778 (2017)
25. Li, B., Ren, W., Fu, D., Tao, D., Feng, D., Zeng, W., Wang, Z.: Benchmarking single-image dehazing and beyond. IEEE Transactions on Image Processing **28**(1), 492–505 (2018)
26. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17452–17462 (2022)
27. Li, B., Liu, X., Hu, P., Wu, Z., Lv, J., Peng, X.: All-in-one image restoration for unknown corruption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17452–17462 (2022)
28. Li, L., Wang, R., Wang, W., Gao, W.: A low-light image enhancement method for both denoising and contrast enlarging. In: 2015 IEEE international conference on image processing (ICIP). pp. 3730–3734. IEEE (2015)
29. Li, M., Liu, J., Yang, W., Sun, X., Guo, Z.: Structure-revealing low-light image enhancement via robust retinex model. IEEE Transactions on Image Processing **27**(6), 2828–2841 (2018)
30. Li, R., Cheong, L.F., Tan, R.T.: Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1633–1642 (2019)
31. Li, R., Tan, R.T., Cheong, L.F.: All in one bad weather removal using architectural search. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3175–3185 (2020)

32. Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R.: Swinir: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1833–1844 (2021)
33. Lin, D., WANG, X., Shen, J., Zhang, R., Liu, R., Wang, M., Xie, W., Guo, Q., Li, P.: Generative status estimation and information decoupling for image rain removal. Advances in Neural Information Processing Systems **35**, 4612–4625 (2022)
34. Liu, H., Simonyan, K., Yang, Y.: Darts: Differentiable architecture search. arXiv preprint arXiv:1806.09055 (2018)
35. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. IEEE Transactions on Image Processing **27**(6), 3064–3073 (2018)
36. Liu, Y.F., Jaw, D.W., Huang, S.C., Hwang, J.N.: Desnownet: Context-aware deep network for snow removal. IEEE Transactions on Image Processing **27**(6), 3064–3073 (2018)
37. Lore, K.G., Akintayo, A., Sarkar, S.: Llnet: A deep autoencoder approach to natural low-light image enhancement. Pattern Recognition **61**, 650–662 (2017)
38. Mei, J., Li, Y., Lian, X., Jin, X., Yang, L., Yuille, A., Yang, J.: Atomnas: Fine-grained end-to-end neural architecture search. arXiv preprint arXiv:1912.09640 (2019)
39. Özdenizci, O., Legenstein, R.: Restoring vision in adverse weather conditions with patch-based denoising diffusion models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
40. Pham, H., Guan, M., Zoph, B., Le, Q., Dean, J.: Efficient neural architecture search via parameters sharing. In: International conference on machine learning. pp. 4095–4104. PMLR (2018)
41. Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2482–2491 (2018)
42. Qin, X., Wang, Z., Bai, Y., Xie, X., Jia, H.: Ffa-net: Feature fusion attention network for single image dehazing. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11908–11915 (2020)
43. Quan, R., Yu, X., Liang, Y., Yang, Y.: Removing raindrops and rain streaks in one go. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9147–9156 (2021)
44. Quan, Y., Deng, S., Chen, Y., Ji, H.: Deep learning for seeing through window with raindrops. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2463–2471 (2019)
45. Ren, W., Ma, L., Zhang, J., Pan, J., Cao, X., Liu, W., Yang, M.H.: Gated fusion network for single image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3253–3261 (2018)
46. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 102–118. Springer (2016)
47. Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. Int. J. Comput. Vis. **126**, 973–992 (2018)
48. Shao, M.W., Li, L., Meng, D.Y., Zuo, W.M.: Uncertainty guided multi-scale attention network for raindrop removal from a single image. IEEE Transactions on Image Processing **30**, 4828–4839 (2021)
49. Tan, M., Le, Q.V.: Mixconv: Mixed depthwise convolutional kernels. arXiv preprint arXiv:1907.09595 (2019)

50. Valanarasu, J.M.J., Yasarla, R., Patel, V.M.: Transweather: Transformer-based restoration of images degraded by adverse weather conditions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2353–2363 (2022)
51. Wang, C., Zheng, Z., Quan, R., Sun, Y., Y., Y.: Context-aware pretraining for efficient blind image decomposition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
52. Wang, H., Xie, Q., Zhao, Q., Meng, D.: A model-driven deep neural network for single image rain removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3103–3112 (2020)
53. Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al.: Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence **43**(10), 3349–3364 (2020)
54. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11534–11542 (2020)
55. Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W.: Spatial attentive single-image deraining with a high quality real rain dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12270–12279 (2019)
56. Wang, T., Yang, X., Xu, K., Chen, S., Zhang, Q., Lau, R.W.: Spatial attentive single-image deraining with a high quality real rain dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12270–12279 (2019)
57. Wang, T., Zheng, Z., Sun, Y., Yan, C., Yang, Y., Chua, T.S.: Multiple-environment self-adaptive network for aerial-view geo-localization. Pattern Recognition **152**, 110363 (2024)
58. Wang, Z., Cun, X., Bao, J., Zhou, W., Liu, J., Li, H.: Uformer: A general u-shaped transformer for image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17683–17693 (2022)
59. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
60. Xiao, J., Fu, X., Wu, F., Zha, Z.J.: Stochastic window transformer for image restoration. Advances in Neural Information Processing Systems **35**, 9315–9329 (2022)
61. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
62. You, S., Tan, R.T., Kawakami, R., Mukaigawa, Y., Ikeuchi, K.: Adherent raindrop modeling, detection and removal in video. IEEE transactions on pattern analysis and machine intelligence **38**(9), 1721–1733 (2015)
63. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
64. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14821–14831 (2021)

65. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning. pp. 7354–7363. PMLR (2019)
66. Zhang, H., Patel, V.M.: Densely connected pyramid dehazing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3194–3203 (2018)
67. Zoph, B., Vasudevan, V., Shlens, J., Le, Q.V.: Learning transferable architectures for scalable image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8697–8710 (2018)