# Blind Image Decomposition

Junlin Han<sup>1,2</sup> Weihao Li<sup>1</sup> Pengfei Fang<sup>1,2</sup> Chunyi Sun<sup>2</sup> Jie Hong<sup>1,2</sup> Mohammad Ali Armin<sup>1</sup> Lars Petersson<sup>1</sup> Hongdong Li<sup>2</sup>

Data61-CSIRO<sup>1</sup> Australian National University<sup>2</sup>

Abstract. We propose and study a novel task named Blind Image Decomposition (BID), which requires separating a superimposed image into constituent underlying images in a blind setting, that is, both the source components involved in mixing as well as the mixing mechanism are unknown. For example, rain may consist of multiple components, such as rain streaks, raindrops, snow, and haze. Rainy images can be treated as an arbitrary combination of these components, some of them or all of them. How to decompose superimposed images, like rainy images, into distinct source components is a crucial step toward real-world vision systems. To facilitate research on this new task, we construct multiple benchmark datasets, including mixed image decomposition across multiple domains, real-scenario deraining, and joint shadow/reflection/watermark removal. Moreover, we propose a simple yet general Blind Image Decomposition Network (BIDeN) to serve as a strong baseline for future work. Experimental results demonstrate the tenability of our benchmarks and the effectiveness of BIDeN. Codes and datasets are available at GitHub.

Keywords: Image Decomposition, Low-level Vision, Rain Removal

### 1 Introduction

Various computer vision and computer graphics tasks [23,98,40,44,26,37,28,2,15,1] can be viewed as image decomposition, which aims to separate a superimposed image into distinct components/layers with only a single observation. For example, foreground-background segmentation [16,69,50,3,52] aims at decomposing a holistic image into foreground objects and background stuff. Image dehazing [31,5,47] can be treated as decomposing a hazy image into a haze-free image and a haze map (medium transmission map, atmosphere light). Shadow removal [44,13,18,38,8,38] decomposes a shadow image into a shadow-free image and a shadow mask. Other tasks like transparency separation [91,17,46,56], watermark removal [53,7], image deraining [65,82,86,95,51,79], texture separation [26], underwater image restoration [22,29], image desnowing [55,68], stereo mixture decomposition [92], 3D intrinsic mixture image decomposition [2], fence removal [14,80,54], flare removal [78,4] are covered in image decomposition.

Vanilla image decomposition tasks come with a fixed and known number of source components, and the number is most often set to two [23,98,40,44,26,86,43].



Fig. 1. Example of raining cases. Rain exists in different formats such as rain streak and raindrop. Snow and haze often co-occur during raining. BID setting treats rainy images as an arbitrary combination of these components. Deraining under the BID setting becomes more challenging yet more overarching

Such a setting does capture some basic real-world cases. However, real-world scenarios are more complex. Consider autonomous driving on rainy days, where the visual perception quality is degraded by different forms of precipitation and the co-occurring components, shown in Figure 1. Some natural questions emerge: Can a vision system assume precipitations to be of a specific form? Should a vision system assume raindrops always exist or not? Shall a vision system assume the haze or snow comes along with rain or not? These questions are particularly important based on their relevance in real-world applications. The answer to these questions should be NO. A comprehensive vision system is supposed to be robust with the ability to handle many possible circumstances [61,72,48]. Yet, with the previous setting in deraining [89,93,65,82,49,51,88,67,57,94,90,87] there remains a gap toward sophisticated real-world scenarios.

This paper aims at addressing the aforementioned gap, as a step toward robust real-world vision systems. We propose a task that: (1) does not fix the number of source components, (2) considers the presence and varying intensities of source components, and (3) amalgamates every source component as potential combinations. To disambiguate with previous tasks, we refer to our proposed task as *Blind Image Decomposition (BID)*. This name is inspired by the Blind Source Separation (BSS) task in the field of signal processing.

The task format is straightforward. We no longer set the number of source components to a fixed value. Instead, we set a maximum number of potential source components, where each component can be arbitrarily part of the mix. Let 'A', 'B', 'C', 'D', 'E' denote five source components and 'a', 'b', 'c', 'd', 'e' denote images from the corresponding source components. The mixed image can be either 'a', 'd', 'ab', 'bc', 'abd', 'ade', 'acde', 'abcde' ..., with up to 31 possible combinations in total. Given any of the 31 possible combinations as input, a BID method is required to predict and reconstruct the individual source components involved in mixing.

As different components can be arbitrarily involved in the mixing, no existing datasets support such a training protocol. Thus, we construct three benchmark datasets, they are: (I) mixed image decomposition across multiple domains, (II) real-scenario deraining, and (III) joint shadow/reflection/watermark removal.

To perform multiple BID tasks, we design a simple yet flexible model, dubbed BIDeN (Blind Image Decomposition Network). BIDeN is a generic model that supports diverse BID tasks with distinct objectives. BIDeN is based on the framework of GANs (Generative Adversarial Networks) [25], and we explore some critical design choices of BIDeN to present a general model. Designed for a more challenging BID setting, BIDeN still outperforms the current state-of-theart image decomposition model [98,23] and shows competitive results compared to models designed for specific tasks. Lastly, a comprehensive ablation study is conducted to analyze the design choices of BIDeN.

# 2 Related Work

**Image Decomposition.** This task is a general task covering numerous computer vision and computer graphics tasks. Double-DIP [23] couples multiple DIPs [74] to decompose images into their basic components in an unsupervised manner. Deep Generative Priors [37] employs a likelihood-based generative model as a prior, performing the image decomposition task. Deep Adversarial Decomposition (DAD) [98] proposed a unified framework for image decomposition by employing three discriminators. A crossroad L1 loss is introduced to support pixel-wise supervision when domain information is unknown. Different from the conventional image decomposition task that usually aims to solve a particular degradation, in the BID setting, we treat degraded images as an arbitrary combination of individual components, and aim to solve sophisticated compound degradation in a unified framework.

Blind Source Separation, BSS [35,9,20,21,34], also known as the "cocktail party problem", is an important research topic in the field of signal processing. Blind refers to a setting where sources and mixing mechanism are both unknown. The task requires performing the separation of a mixture signal into the constituent underlying source signals with limited prior information. A representative algorithm is the Independent Component Analysis (ICA) [35,45,63]. The BID task shares common properties with BSS, where the blind settings are similar but not identical. The setting of BID assumes an unknown number of source components involved in the mixing and unknown mixing mechanisms. With such a setting, the number of source components involved in mixing is also unknown. However, the domain information of each source component is considered known as the goal of BID is to advance real-world vision systems. The setting of including known domain information can be better applied to computer vision tasks. For instance, the goal of the shadow removal task is to separate a shadow image into a shadow-free image and a shadow mask, where the domain information is clear.



**Fig. 2.** The architecture of the Blind Image Decomposition Network (BIDeN). We show an example, where N = 4, L = 2,  $x = \{a, b, c, d\}$ , and  $I = \{1, 3\}$ . a, c are selected then passed to the mixing function f, and outputs the mixed input image z, which is f(a, c) here. The generator consists of an encoder E with three branches and multiple heads H.  $\bigotimes$  denotes the concatenation operation. Depth and receptive field of each branch are different to capture multiple scales of features. Each specified head points to the corresponding source component, and the number of heads varies with the maximum number of source components N. All reconstructed images (a', c') and their corresponding real images (a, c) are sent to an unconditional discriminator. The discriminator also predicts the source components involved in the mixing of the input image z. The outputs from other heads (b', d') do not contribute to the optimization

**Generative Adversarial Networks.** GANs [25] include two key components, a generator, and a discriminator, where the generator is trying to generate realistic samples while the discriminator is trying to identify real samples and generated samples. The adversarial training mechanism helps the output from the generator match the distribution of real data. GANs are especially successful in image generation tasks [85,41] and image-to-image translation tasks [36,30,97]. GANs are also a common tool for image decomposition tasks, where GANs have been successfully employed in image deraining [86,65], transparency separation [56,91], and image dehazing [47].

### 3 Blind Image Decomposition Formulation

Given a set of N ( $N \ge 2$ ) source components, *i.e.*, image domains, denoted by  $\mathcal{X} = \{X_m\}_{m=1}^N$ . Each source component  $X_m$  contains some images  $x_m, x_m \in X_m$ . L ( $1 \le L \le N$ ) source components are randomly selected from  $\mathcal{X}$ . Let  $I = \{I_j\}_{j=1}^L$  indicate the index set for the selected source components, where  $I_j \in \{1, ..., N\}$ . Hence, the selected source components are denoted by  $\{X_{I_j}\}_{j=1}^L$ . Each selected source component  $X_{I_j}$  contains some images  $x_{I_j}$ . With a predetermined mixing function f, the mixed image z is given by  $z = f(\{x_{I_j}\}_{j=1}^L)$ . The mixed image z can be identical to a single image when L = 1. The BID task requires the BID method to find a function g to separate z, as  $g(z) = \{x'_{I_j}\}_{j=1}^L$ . Each reconstructed image  $x'_{I_j}$  is close to its corresponding image  $x_{I_j}$ . That is, given a mixed image as input, the task requires the BID method to: (1) predict the source components involved in mixing, (2) reconstruct images preserving the fidelity of the corresponding images involved in mixing.

For the example shown in Figure 2, where N = 4,  $\mathcal{X} = \{A, B, C, D\}$ ,  $x = \{a, b, c, d\}$ , L can be 1, 2, 3, 4, z can be a, b, c, d, f(a, b), f(a, c), f(a, d), f(b, c), f(b, d)f(c, d), f(a, b, c), f(a, b, d), f(a, c, d), f(b, c, d), f(a, b, c, d). Let L = 2 and  $I = \{1, 3\}$ . Given z = f(a, c) as the input, knowing there are four different source components A, B, C, D without any other information, the task requires the method to find a function g, so that g(f(a, c)) = a', c', where  $a' \to a$  and  $c' \to c$ . Also, the method should correctly predict the source components involved in the mixing, that is, predicting  $I = \{1, 3\}$ .

The BID task is challenging for the following reasons: (1) When N increases, the number of possible z increases rapidly. The BID setting forces the method to deal with  $2^N - 1$  possible combinations. For instance, when N increases to 8, there are 255 variants of z. (2) The task requires the method to predict the source components involved in the mixing. Source components are difficult to be predicted when N is large and L varies a lot. (3) The mixing mechanism, or the mixing function f, is unknown to the method. The mixing function f varies with different source components and can be non-linear/complex in specific circumstances, such as rendering raindrop images, adding shadows or reflections to images. (4) As L increases, each source component contributes a decreasing amount of information to z, making the task highly ill-posed.

### 4 Blind Image Decomposition Network

To perform diverse BID tasks, a unified framework is required. Inspired by the success of image-to-image translation models [36,30,97], we design our Blind Image Decomposition Network (BIDeN) as follows. Figure 2 presents an overall architecture of BIDeN when the maximum number of source components is four.

The generator G consists of two parts: a multi-scale encoder E and multiple heads H. We design a multi-scale encoder containing three branches to capture multiple scales of features. This design is beneficial to the reconstruction of source components. We concatenate different scales of features and send them to multiple heads, where the number of heads is identical to the maximum number of source components. Each head is specific to reconstructing a particular kind of source component. Such multiple-head domain-specific autoencoders have been adopted particularly in inverse rendering task [28,77].

The discriminator D consists of two branches and most weights are shared. The reconstructed images and corresponding real images are sent to the discriminator branch  $D_S$  (Separation) individually. The function of  $D_S$  is similar to a typical discriminator, *i.e.*, classifying whether the input to  $D_S$  is generated or real to direct the generator generating realistic images. The discriminator branch  $D_P$  (Prediction) predicts the source components involved in the mixed image z, with a confidence threshold of zero. A successful prediction unveils the correct index set of the selected source components I.

Taking the example of Figure 2, we name four heads as  $H_A$ ,  $H_B$ ,  $H_C$ , and  $H_D$ . For an input f(a, c),  $H_A$  and  $H_C$  aim to reconstruct a', c', so that  $a' \to a$  and  $c' \to c$  while  $H_B$ ,  $H_D$  are free to output anything or are simply turned

6 J. Han et al.

off. We employ adversarial loss [25], perceptual loss [39], L1/L2 loss, and binary cross-entropy loss. The details of the objective function are expressed below.

### 4.1 Objective

We employ the adversarial loss [25] to encourage the generator G to output well separated and realistic images, regardless of the source components. For the function  $g(z) = \{x'_{I_i}\}_{j=1}^L$ , the GAN loss is expressed as:

$$\mathcal{L}_{\text{GAN}}(G, D_S) = \mathbb{E}_x \left[ \log D_S(x) \right] \\ + \mathbb{E}_z \left[ \log \left( 1 - D_S(G(z)) \right] \right], \tag{1}$$

where G behaves as g. It tries to separate the input mixed image z and reconstruct separated outputs  $x'_{I_j}$ , while  $D_S$  attempts to distinguish between G(z)and real samples  $x_{I_j}$ . Note x inside equations 1-5 denotes  $x_{I_j}$ . We employ the LSGAN [58] loss and the Markovian discriminator [36].

The reconstructed images  $x'_{I_j}$  should be separated, as well as to be near the corresponding  $x_{I_j}$  in a distance sense. Hence, we employ perceptual loss (VGG loss) [39] and L1/L2 loss. They are formalized as:

$$\mathcal{L}_{\text{VGG}}(G) = \mathbb{E}_{x,z}[\sum_{l} \lambda_{l} \left[ \left\| \Phi_{l}(x) - \Phi_{l}(G(z)) \right) \right\|_{1} \right], \tag{2}$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,z} \left[ \| x - G(z) \|_1 \right],$$
(3)

$$\mathcal{L}_{L2}(G) = \mathbb{E}_{x,z} \left[ \|x - G(z)\|_2^2 \right],$$
(4)

where  $\Phi$  is a trained VGG19 [71] network,  $\Phi_l$  denotes a specific layer, and  $\lambda_l$  denotes the weights for the *l*-th layer. The choice of layers and weights is identical to pix2pixHD [76]. We use L2 loss for masks and L1 loss for other source components. For simplification, we denote L1/L2 loss as  $\mathcal{L}_{\rm L}$ .

For the source prediction task, we find that the discriminator performs better than the generator. The goal of the discriminator is to classify between reconstructed samples and real samples. It naturally learns an embedding. Such an embedding is beneficial even when the input is a mixed image z. The discriminator D is capable of performing an additional source prediction task. Thus we design a source prediction branch  $D_P$ . The binary cross-entropy loss is employed for the source prediction task:

$$\mathcal{L}_{BCE}(D_P) = \mathbb{E}_z \left[ -\sum_{m=1}^N [GT(z)_m \log(D_P(z)_m) + (1 - GT(z)_m) \log(1 - D_P(z)_m)] \right],$$
(5)

where N denotes the maximum number of source components, GT denotes the binary label of the source components involved in the mixing of input image z.  $D_P$  is the source prediction branch of the discriminator.

7

Our final objective function is:

$$\mathcal{L}(G, D_S, D_P) = \lambda_{\text{GAN}} \mathcal{L}_{\text{GAN}} (G, D_S) + \lambda_{\text{VGG}} \mathcal{L}_{\text{VGG}} (G) + \lambda_{\text{L}} \mathcal{L}_{\text{L}} (G) + \lambda_{\text{BCE}} \mathcal{L}_{\text{BCE}} (D_P).$$
(6)

We set  $\lambda_{\text{GAN}}, \lambda_{\text{VGG}}, \lambda_{\text{L}}, \lambda_{\text{BCE}}$  to be 1, 10, 30, 1 respectively. This setting is a generic setting that is applied to all tasks.

### 4.2 Training details

Throughout all experiments, we use the Adam optimizer [42] with  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$  for both G, D. BIDeN is trained for 200 epochs with a learning rate of 0.0003. The learning rate starts to decay linearly after half of the total epochs. We use a batch size of 1 and instance normalization [73]. All training images are loaded as  $286 \times 286$  then cropped to  $256 \times 256$  patches. Horizontal flip is randomly applied. At test time, we load test images in a  $256 \times 256$  resolution. More details on the training settings, the architecture, the number of parameters, and training speed are provided in the supplementary material. The training details of all baselines are also provided there.

### 5 Blind Image Decomposition Tasks

We construct benchmark datasets from different views to support practical usages of BID. For each benchmark dataset, we involve multiple source components that may occur together. To explore the generality of BIDeN and the tenability of constructed datasets, we test BIDeN on three new challenging datasets. BIDeN is trained under the BID setting, which is more difficult than the conventional image decomposition setting. During training, mixed images are randomly synthesized. At test time, the input mixed images are fixed. As BID is a novel task not previously investigated, no existing baselines are available for comparison. For different tasks, we choose different evaluation strategies and baselines.

Throughout all tasks, BIDeN is trained under the BID setting, that is, BIDeN is facing more challenging requirements than other baselines. Also, BIDeN is a generic model designed to perform all kinds of BID tasks. These two constraints limit the performance of BIDeN. We compare BIDeN to other baselines designed for specific tasks, where BIDeN is still able to show very competitive results on all tasks. All qualitative results are *randomly picked*. Additional task settings, discussion on the order of mixing, dataset construction details, results, including the detailed case results of BIDeN, are provided in the supplementary material.

### 5.1 Task I: Mixed image decomposition across multiple domains

**Dataset.** This dataset contains eight different image domains, *i.e.*, source components. Each domain has approximately 2500 to 3000 images in the training set, and the test set contains 300 images for each domain. Image domains are



Fig. 3. Qualitative results of Task I (Mixed image decomposition across multiple domains). We train BIDeN 7 times, setting different maximum numbers of source components (2-8). Double-DIP fails to separate the mixed input. DAD shows blurry, non-clean results while the results shown by BIDeN are well-separated and visually satisfying. When BIDeN is trained on a greater maximum number of source components, the quality of the results drops progressively as expected

designed to be big and inclusive to cover multiple categories, like animal, fruit, vehicle, instead of being comparatively small domains such as horse, cat, car. The eight domains are Fruit (2653), Animal (2653), Flower (2950), Furniture (2582), Yosemite (2855), Vehicle (2670), Vegetable (2595), and CityScape (2975). CityScape and Flower are selected from the CityScape [11] dataset and the VGG flower [62] dataset. The remaining six image domains are mainly gathered from Flicker using the corresponding keyword, except for Yosemite, which also combines the Summer2Winter dataset from CycleGAN [97]. The order of the eight domains is randomly shuffled. The mixing mechanism is linear mix.

**Table 1.** Quantitative results on Task I (Mixed image decomposition across multiple domains). The testing condition is identical, using Fruit (A) + Animal (B) mixture as inputs. N in BIDeN (N) denotes the maximum number of source components. Double-DIP [23] performs poorly. Under a more challenging BID setting, BIDeN (2,3,4) still outperforms DAD [98] overall, suggesting the superiority of BIDeN. Please refer to Appendices for detailed case results

Method	Fruit (A PSNR↑ SSIM <sup>*</sup>	\)    FID↓	Ar PSNR↑	imal (  SSIM1	$ \begin{array}{c} B \\ \downarrow F I D \downarrow \end{array} $	Acc (AB) $\uparrow$	Acc (All)	$\uparrow$ Model size
Double-DIP [23] DAD [98]	$\begin{array}{c cccccc} 13.14 & 0.49 \\ 17.59 & 0.72 \end{array}$	$\begin{vmatrix} 257.80 \\ 137.66 \end{vmatrix}$	$13.11 \\ 17.52$	$\left  \begin{array}{c} 0.39 \\ 0.62 \end{array} \right $	$\begin{array}{c} 221.76 \\ 126.32 \end{array}$	0.996	-	669.0 MB
BIDeN (2) BIDeN (3) BIDeN (4) BIDeN (5) BIDeN (6) BIDeN (7) BIDeN (8)	$\begin{array}{c c c c c c c c c c c c c c c c c c c $	$\begin{array}{ c c c c c c c c c c c c c c c c c c c$	$\begin{array}{c} 19.89 \\ 18.75 \\ 18.03 \\ 17.27 \\ 16.57 \\ 16.54 \\ 15.79 \end{array}$	$ \begin{smallmatrix} 0.69 \\ 0.61 \\ 0.58 \\ 0.54 \\ 0.51 \\ 0.49 \\ 0.45 \\ \end{smallmatrix} $	$\begin{array}{c} 69.35\\ 88.23\\ 97.16\\ 114.40\\ 118.00\\ 126.66\\ 191.29\\ \end{array}$	$\begin{array}{c} 1.0\\ 0.836\\ 0.716\\ 0.676\\ 0.646\\ 0.413\\ 0.383\end{array}$	$\begin{array}{c} 0.957 \\ 0.807 \\ 0.733 \\ 0.603 \\ 0.483 \\ 0.310 \\ 0.278 \end{array}$	144.9 MB 147.1 MB 149.3 MB 151.5 MB 153.7 MB 155.9 MB 158.1 MB

**Experiments and results.** We compare BIDeN to Double-DIP [23] and DAD [98]. DAD is trained on the first two domains (Fruit, Animal) with mixed input only. For BIDeN, we train it 7 times under the BID setting, varying from 2 domains to 8 domains. At test time, we evaluate the separation results on Fruit + Animal mixture with PSNR, SSIM [96], and FID [32].

Table 1 and Figure 3 present the results of Task I. In terms of PSNR/SSIM, BIDeN outperforms Double-DIP by a large margin and outperforms DAD when N is less than 5. For FID, BIDeN shows preferable results than DAD even when N = 7, showing the superiority of BIDeN.

#### 5.2Task II: Real-scenario deraining

We design two sub-tasks, real-scenario deraining in driving (Task II.A) and realscenario deraining in general (Task II.B).

### Task II.A: Real scenario deraining in driving

Dataset. Based on the CityScape [11] dataset, we construct our real-scenario deraining in driving dataset. We use the test set from the original CityScape dataset as our training set (2975), and the validation set from the original CityScape dataset as our test set (500). The test set for all source components contains a fixed number of 500 images. We use three different masks, including rain streak (1620), raindrop (3500), and snow (3500). These masks cover different intensities. For haze, we use the corresponding transmission maps (2975 x 3) with three different intensities acquired from Foggy CityScape [70]. The masks for rain streak are acquired from Rain100L and Rain100H [82] while the masks for snow are selected from Snow100K [55]. For raindrop masks, we model the droplet shape and property using the metaball model [6]. The locations, numbers, and sizes of raindrops are randomly sampled. Paired refraction maps are generated using refractive model [10,64]. The mixing mechanism for this dataset is based on physical imaging models [82,55,31,10,64].



Fig. 4. CityScape, masks (Rain Streak, Snow, Raindrop), and transmission map (Haze) generated by BIDeN for case (6), rain streak + snow + moderate haze + raindrop. All generated images are perceptually faithful and visually close to the ground truth (GT)

**Experiments and results.** We train BIDeN, MPRNet [84], Restormer [83], and RCDNet [75] under the BID setting. For all baselines, we do not require the prediction of the source component and the generation of masks. Thus, BIDeN is still trained with a more challenging requirement. We report the results for 6 10 J. Han et al.

**Table 2.** Results of BIDeN on Task II.A (Real-scenario deraining in driving). We employ PSNR and SSIM metrics for both CityScape images, masks, and transmission maps. We report the results for 6 test cases as presented in Figure 1, the 6 cases are (1): rain streak, (2): rain streak + snow, (3): rain streak + light haze, (4): rain streak + heavy haze, (5): rain streak + moderate haze + raindrop, (6) rain streak + snow + moderate haze + raindrop. Note that only haze is divided into light/moderate/heavy intensities. Both training set and test set of Rain Streak, Snow, and Raindrop already consist of different intensities

Method		$_{ m SSIM\uparrow}$	Rain S PSNR↑	$Streak   SSIM\uparrow $	$_{\mathrm{PSNR}\uparrow}^{\mathrm{Sn}}$	$^{\mathrm{ow}}$	Ha  PSNR↑	ze	Rain PSNR↑	drop  SSIM↑	$\mathrm{Acc}\uparrow$
BIDeN (1) BIDeN (2) BIDeN (3) BIDeN (4) BIDeN (5) BIDeN (6)	30.89 29.34 28.62 26.77 27.11 26.44	$\begin{array}{c} 0.932 \\ 0.899 \\ 0.919 \\ 0.898 \\ 0.898 \\ 0.870 \end{array}$	32.13 29.24 31.48 30.57 30.54 28.31	$\begin{array}{c} 0.924 \\ 0.846 \\ 0.914 \\ 0.897 \\ 0.898 \\ 0.823 \end{array}$	25.77 	0.692	- 30.77 33.73 30.52 29.83	$ \begin{array}{c c} -\\ 0.960\\ 0.957\\ 0.952\\ 0.948 \end{array} $	- - 20.20 21.47	- - - 0.908 0.893	$\begin{array}{c} 0.998 \\ 0.996 \\ 0.994 \\ 0.998 \\ 0.994 \\ 0.998 \\ 0.998 \end{array}$

cases, as the examples presented in Figure 1. Note that only haze is divided into light/moderate/heavy intensities. Both training set and test set of rain streak, snow, and raindrop already contain different intensities. We report the results in SSIM and PSNR for CityScape images, masks, and transmission maps. For



**Fig. 5.** Results of Task II.A (Real-scenario deraining in driving). Row 1-6 presents 6 cases as presented in Table 2. Baselines performs well at case (1) and they effectively removes rain streak but is not strong at removing other components. BIDeN is more robust at the removal of all components. BIDeN generates artifact-free, visually pleasing results while all baselines leaves some components that are not completely removed, especially when hazy intensity is moderate or heavy, as shown in case (4), (5), and (6)

all 6 cases, we report the detailed results of BIDeN in Table 2. BIDeN shows excellent quantitative results on accuracy. For the PSNR/SSIM metrics on all source components, BIDeN performs well except for the raindrop masks. An

**Table 3.** Comparison on task II.A (Real-scenario deraining in driving) between MPR-Net [84], RCDNet [75], and BIDeN. MPRNet and Rostormer shows superior results for case (1) and case (2). In contrast, BIDeN is better at other cases. For the details of 6 test cases, please refer to Table 2 and Figure 1

Case	Inµ PSNR↑	$ _{\rm SSIM\uparrow} $	$^{\rm MPR}_{\rm PSNR\uparrow}$	${ m Net} \\  { m SSIM}\uparrow $	$\begin{array}{c} \operatorname{Restc} \\ \operatorname{PSNR}\uparrow \end{array}$	$\operatorname{rmer}_{ \mathrm{SSIM}\uparrow }$	$\operatorname{RCD}_{\operatorname{PSNR}\uparrow}$	$ $ SSIM $\uparrow$	BIE PSNR↑	eN  SSIM↑
$(1) \\ (2) \\ (3) \\ (4) \\ (5) \\ (6)$	$\begin{array}{c} 25.69 \\ 18.64 \\ 17.45 \\ 11.12 \\ 14.05 \\ 12.38 \end{array}$	$ \begin{vmatrix} 0.786 \\ 0.564 \\ 0.712 \\ 0.571 \\ 0.616 \\ 0.461 \end{vmatrix} $	$\begin{array}{c} 33.39 \\ 30.52 \\ 23.98 \\ 18.54 \\ 21.18 \\ 20.76 \end{array}$	$\begin{array}{c} 0.945 \\ 0.909 \\ 0.900 \\ 0.829 \\ 0.846 \\ 0.812 \end{array}$	$\begin{array}{c} 34.29\\ 30.60\\ 23.74\\ 20.33\\ 22.17\\ 21.24 \end{array}$	$\begin{array}{c} 0.951 \\ 0.917 \\ 0.905 \\ 0.853 \\ 0.859 \\ 0.821 \end{array}$	32.38 28.45 27.14 19.67 24.23 22.93	$\begin{array}{c} 0.937 \\ 0.892 \\ 0.911 \\ 0.865 \\ 0.889 \\ 0.846 \end{array}$	30.89 29.34 28.62 26.77 27.11 26.44	$\begin{array}{c} 0.932 \\ 0.899 \\ 0.919 \\ 0.898 \\ 0.898 \\ 0.870 \end{array}$

example of all components generated by BIDeN is shown in Figure 4. Table 3 and Figure 5 presents the comparison between BIDeN and baselines. For better visualization, we resize the resolution of visual examples to match the original CityScape resolution.

### Task II.B: Real scenario deraining in general

**Dataset.** The training set contains 3661 natural images as rain-free images, where 861 images are adopted from the training set of [65], 1800 images are borrowed from the training set of Rain1800 [82], and the rest 1000 images are selected from the training set of Snow100K [55]. We adopt identical rain streak, snow, and raindrop masks from Task II.A. The test set contains real images only, including rain streak (185, from [49]), raindrop (249, from [65]) and snow (1329, from [55]) images.

 Table 4. Quantitative results of Task II.B (Real-scenario deraining in general). \*

 indicates the BID setting. For all methods and testing cases, models trained under the

 BID setting performs better than the conventional image decomposition setting

Method	Rain NIQE↓ B	Streak RISQUI	$\mathbb{E}_{\downarrow} _{\mathrm{NIQE}\downarrow \mathbb{E}}^{\mathrm{Rain}}$	ndrop BRISQUE	${\rm E}_{\downarrow} _{\rm NIQE}_{\downarrow {\rm B}}$	ow RISQUE↓
Input	4.86	27.84	5.61	24.85	4.74	22.68
${}^{\mathrm{MPRNet}}_{\mathrm{MPRNet}}$	$\begin{array}{c c} 4.14 \\ 4.34 \end{array}$	$\begin{array}{c} 28.72 \\ 28.00 \end{array}$	$\left \begin{array}{c}4.94\\4.81\end{array}\right $	$\begin{array}{c} 29.42\\ 25.86 \end{array}$	$\left \begin{array}{c}4.60\\4.24\end{array}\right $	$\substack{25.93\\24.55}$
$_{ m BIDeN}^{ m BIDeN}$	$\begin{array}{c c} 4.71 \\ 4.31 \end{array} \right $	$25.39 \\ 26.55$	$5.39 \\ 4.71$	$22.94 \\ 21.22$	$\left \begin{array}{c} 4.97\\ 4.28 \end{array}\right $	$22.64 \\ 22.40$

**Experiments and results.** We aim to validate whether models trained on the synthetic dataset can generalize to real testing samples and whether the BID training setting generalizes better than the conventional image decomposition setting. We train BIDeN and MPRNet under the BID setting as well as the conventional image decomposition setting with identical training recipes, that is, the training recipes used in Task II.A. For the conventional image decomposition setting, models are trained on a large-scale synthetic dataset [84] (13712 pairs). We report the quantitative results with no-reference metrics, NIQE [60] and BRISQUE [59].



Fig. 6. Qualitative results of Task II.B (Real-scenario deraining in general). \* indicates the BID setting. Rows 1-2, 3-4, and 5-6 presents the results of real rain streak, raindrop, and snow images, respectively. For all cases, models trained under the BID setting are more robust in restoring real-world images

All trained models are evaluated with real-world images. Results are presented in Table 4 and Figure 6. With only 26.6% training data and iterations, both BIDeN and MPRNet models trained under the BID setting are more robust in restoring real images, showing the generality and robustness of the BID setting in real-world scenarios.

### 5.3 Task III: Joint shadow/reflection/watermark removal



**Fig. 7.** Images and masks (Shadow, Watermark) produced by BIDeN for three cases, (a), (c), and (abc). All generated masks are faithful to the ground truth (GT)

**Dataset.** This task is designed to jointly perform multiple tasks with uncertainty in one go. Once a model is trained on this dataset under the BID setting, the model is capable of performing multiple tasks. We construct two versions for this

**Table 5.** Results of Task III (Joint shadow/reflection/watermark removal). We employ RMSE $\downarrow$  to measure shadow region, non-shadow region, and all region. For BIDeN, we report the performance of all cases. a,b,c denotes shadow, reflection, and watermark, respectively. BIDeN (ab) is the result of BIDeN tested on shadow + reflection inputs. Results for all baselines are reported by [12,19]. The generality of BIDeN and the challenging BID training setting limit the performance of BIDeN

Method	Ver Shadow	sion one (V1 Non-Shadow	), ISTI v  All	Acc	Vei Shadow	rsion two (V2  Non-Shadow	2), SRI   All	)   Acc
Bilateral [81] Regions [27] Interactive [24] DSC [33] DHAN [12] Auto-Exposure [19] CANet [8]	$19.82 \\ 18.95 \\ 14.98 \\ 9.48 \\ 8.14 \\ 7.77 \\ 8.86$	$14.83 \\ 7.46 \\ 7.29 \\ 6.14 \\ 6.04 \\ 5.56 \\ 6.07$	$\begin{array}{c} 15.63 \\ 9.30 \\ 8.53 \\ 6.67 \\ 6.37 \\ 5.92 \\ 6.15 \end{array}$		$\begin{array}{c c} 23.43 \\ 29.89 \\ 19.58 \\ 10.89 \\ 8.94 \\ 8.56 \\ 7.82 \end{array}$	$\begin{array}{c c} 22.26 \\ 6.47 \\ 4.92 \\ 4.99 \\ 4.80 \\ 5.75 \\ 5.88 \end{array}$	$\begin{array}{c} 22.57 \\ 12.60 \\ 8.73 \\ 6.23 \\ 5.67 \\ 6.51 \\ 5.98 \end{array}$	
BIDeN (a) BIDeN (ab) BIDeN (ac) BIDeN (abc)	$11.55 \\ 12.96 \\ 11.89 \\ 13.20$	$10.24 \\ 10.77 \\ 10.23 \\ 10.76$	$  \begin{array}{c} 10.45 \\ 11.12 \\ 10.50 \\ 11.16 \\ \end{array}  $	$\begin{array}{c} 0.359 \\ 0.661 \\ 0.694 \\ 0.929 \end{array}$	$\begin{array}{c c} 12.06 \\ 14.10 \\ 13.29 \\ 15.28 \end{array}$	$ \begin{array}{c c} 7.47 \\ 8.16 \\ 8.08 \\ 8.85 \end{array} $	$\begin{vmatrix} 8.73 \\ 9.79 \\ 9.51 \\ 10.62 \end{vmatrix}$	$\begin{array}{c} 0.919 \\ 0.911 \\ 0.943 \\ 0.936 \end{array}$
BIDeN (b) BIDeN (c) BIDeN (bc)	- -	-	$  \begin{array}{c} 10.85 \\ 10.20 \\ 10.77 \\ \end{array}  $	$\begin{array}{c} 0.559 \\ 0.461 \\ 0.727 \end{array}$		- - -	8.01 7.92 8.71	$\begin{array}{c} 0.891 \\ 0.914 \\ 0.879 \end{array}$

dataset, Version one (V1) is based on ISTD [76], and Version two (V2) is based on SRD [66,12]. We use paired shadow masks, shadow-free images, and shadow images from ISTD/SRD. ISTD consists of 1330 training images and 540 test images while SRD contains 2680 training images and 408 testing images. The algorithm for adding reflection to images is acquired from [91], we select 3120 images from the reflection subset [91] as the reflection layer. The watermark generation algorithm as well as the paired watermark masks, RGB watermark images are acquired from LVM [53], we select 3000 paired watermark images and masks from the training set of LVW [53].

Following the data split of ISTD and SRD, both V1 and V2 share 2580 reflection layer images and 2460 watermark images/masks. V1 contains 1330 paired shadow-free images/shadow masks and the test set contains 540 images for every source component. V2 includes 2680 paired shadow-free images/shadow masks and the test set contains 408 images for every source component. Note that we do not require the reconstruction of reflection layer images.

**Experiments and results.** We mainly compare the shadow removal results to multiple shadow removal baselines, including [81,27,24,33,12,19,8]. We train BIDeN under the BID setting. The trained BIDeN is capable of dealing with all combinations between shadow/reflection/watermark removal tasks. At test time, we report the results for all cases. We employ the root mean square error (RMSE) in LAB color space, following [12,19].

The quantitative results for Task III are reported in Table 5. Constrained by the generality of BIDeN and the challenging BID training setting, BIDeN does not show superior quantitative results compared to other baselines designed for the shadow removal task only. Please refer to the supplementary material for qualitative results.

### 14 J. Han et al.

# 6 Ablation Study and Analysis

We perform ablation experiments to analyze the effectiveness of each component inside BIDeN. Evaluation is performed on Task I (Mixed image decomposition across multiple domains). We set the maximum number of source components to be 4 throughout all ablation experiments. The results are shown in Table 6.

Ablation	$\begin{array}{c} {\rm Fruit} \ ({\rm A}) \\ {\rm PSNR} \uparrow \end{array}$	Animal (B) PSNR↑	$ Acc (AB) \uparrow$	Acc (All) $\uparrow$
I III IV V VI VII VIII	$\begin{array}{c} 17.26 \\ 17.95 \\ 16.67 \\ 15.56 \\ 18.04 \\ 18.19 \\ 18.13 \\ 15.68 \end{array}$	$17.05 \\ 17.41 \\ 16.34 \\ 13.65 \\ 17.98 \\ 17.98 \\ 17.98 \\ 15.64$	$ \begin{vmatrix} 0.730 \\ 0.566 \\ 0.706 \\ 0.733 \\ 0.0 \\ 0.634 \\ 0.520 \\ 0.716 \end{vmatrix} $	$\begin{array}{c} 0.732 \\ 0.616 \\ 0.722 \\ 0.753 \\ 0.06 \\ 0.698 \\ 0.609 \\ 0.683 \end{array}$
BIDeN	18.19	18.03	0.716	0.733

**Table 6.** Ablation study on the design choices of BIDeN. (I) Single-scale encoder. (II) No adversarial loss. (III) No perceptual loss. (IV) No L1/L2 loss. (V) No binary cross-entropy loss. (VI) Source prediction branch inside the generator. (VII) No weights sharing between two branches of discriminator. (VIII) Zeroed loss

Multi-scale encoder (I). We present the results of using a single-scale encoder to replace the multi-scale encoder. BIDeN yields better performance when the multi-scale encoder is employed, which validates the effectiveness of our design. Choice of losses (II, III, IV, V). BIDeN consists of four different losses, we show that removing either one of the losses leads to a performance drop.

**Source prediction branch (VI, VII)**. We move the prediction branch  $D_P$  to the generator. This change degrades the performance, showing that the source prediction task is better to be performed by the discriminator. We report the results for a variant where  $D_P$  does not share weights with the separation branch  $D_S$ . The performance of this variant is worse than vanilla BIDeN, indicating that the embedding learned by  $D_S$  is beneficial to  $D_P$ .

**Zeroed loss (VIII)**. Taking the example of Figure 2, four heads are  $H_A$ ,  $H_B$ ,  $H_C$ , and  $H_D$ . By default, BIDeN ignores the outputs from  $H_B$  and  $H_D$ . Here, we encourage the outputs from  $H_B$  and  $H_D$  to be zero pixels. Such a zeroed loss forces the generator to perform the source prediction task implicitly, however, the results after applying zeroed loss are not comparable to default BIDeN.

# 7 Conclusion

We believe BID is a novel computer vision task advancing real-world vision systems. We form a solid foundation for the future study and we invite the community to further explore its potential, including discovering interesting areas of application, developing novel methods, extending the BID setting, and constructing benchmark datasets. We expect more application areas related to image decomposition, especially in image deraining, to apply the BID setting.

# References

- Alayrac, J.B., Carreira, J., Zisserman, A.: The visual centrifuge: Model-free layered video representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2457–2466 (2019)
- Alhaija, H.A., Mustikovela, S.K., Thies, J., Jampani, V., Nießner, M., Geiger, A., Rother, C.: Intrinsic autoencoders for joint deferred neural rendering and intrinsic image decomposition. In: 2020 International Conference on 3D Vision (3DV). pp. 1176–1185. IEEE (2020)
- Alpert, S., Galun, M., Brandt, A., Basri, R.: Image segmentation by probabilistic bottom-up aggregation and cue integration. IEEE transactions on pattern analysis and machine intelligence 34(2), 315–327 (2011)
- Asha, C., Bhat, S.K., Nayak, D., Bhat, C.: Auto removal of bright spot from images captured against flashing light source. In: 2019 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER). pp. 1–6. IEEE (2019)
- 5. Berman, D., Avidan, S., et al.: Non-local image dehazing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1674–1682 (2016)
- Blinn, J.F.: A generalization of algebraic surface drawing. ACM transactions on graphics (TOG) 1(3), 235–256 (1982)
- Chen, X., Wang, W., Bender, C., Ding, Y., Jia, R., Li, B., Song, D.: Refit: a unified watermark removal framework for deep learning systems with limited data. In: Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. pp. 321–335 (2021)
- Chen, Z., Long, C., Zhang, L., Xiao, C.: Canet: A context-aware network for shadow removal. In: ICCV. pp. 4743–4752 (2021)
- 9. Cichocki, A., Amari, S.i.: Adaptive blind signal and image processing: learning algorithms and applications. John Wiley & Sons (2002)
- Cohen, J., Olano, M., Manocha, D.: Appearance-preserving simplification. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques. pp. 115–122 (1998)
- 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)
- Cun, X., Pun, C.M., Shi, C.: Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting gan. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10680–10687 (2020)
- Ding, B., Long, C., Zhang, L., Xiao, C.: Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10213–10222 (2019)
- Du, C., Kang, B., Xu, Z., Dai, J., Nguyen, T.: Accurate and efficient video defencing using convolutional neural networks and temporal information. In: 2018 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2018)
- Fadili, M.J., Starck, J.L., Bobin, J., Moudden, Y.: Image decomposition and separation using sparse representations: An overview. Proceedings of the IEEE 98(6), 983–994 (2009)
- Faktor, A., Irani, M.: Co-segmentation by composition. In: Proceedings of the IEEE international conference on computer vision. pp. 1297–1304 (2013)

- 16 J. Han et al.
- Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: A generic deep architecture for single image reflection removal and image smoothing. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 3238–3247 (2017)
- Finlayson, G.D., Drew, M.S., Lu, C.: Entropy minimization for shadow removal. International Journal of Computer Vision 85(1), 35–57 (2009)
- Fu, L., Zhou, C., Guo, Q., Juefei-Xu, F., Yu, H., Feng, W., Liu, Y., Wang, S.: Autoexposure fusion for single-image shadow removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10571–10580 (2021)
- Gai, K., Shi, Z., Zhang, C.: Blindly separating mixtures of multiple layers with spatial shifts. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
- Gai, K., Shi, Z., Zhang, C.: Blind separation of superimposed images with unknown motions. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1881–1888. IEEE (2009)
- Galdran, A., Pardo, D., Picón, A., Alvarez-Gila, A.: Automatic red-channel underwater image restoration. Journal of Visual Communication and Image Representation 26, 132–145 (2015)
- Gandelsman, Y., Shocher, A., Irani, M.: "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11026–11035 (2019)
- 24. Gong, H., Cosker, D.: Interactive shadow removal and ground truth for variable scene categories. In: BMVC. pp. 1–11. Citeseer (2014)
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems (2014)
- Gu, S., Meng, D., Zuo, W., Zhang, L.: Joint convolutional analysis and synthesis sparse representation for single image layer separation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1708–1716 (2017)
- Guo, R., Dai, Q., Hoiem, D.: Paired regions for shadow detection and removal. IEEE transactions on pattern analysis and machine intelligence **35**(12), 2956–2967 (2012)
- Halperin, T., Ephrat, A., Hoshen, Y.: Neural separation of observed and unobserved distributions. In: International Conference on Machine Learning. pp. 2566– 2575. PMLR (2019)
- Han, J., Shoeiby, M., Malthus, T., Botha, E., Anstee, J., Anwar, S., Wei, R., Armin, M.A., Li, H., Petersson, L.: Underwater image restoration via contrastive learning and a real-world dataset. arXiv preprint arXiv:2106.10718 (2021)
- Han, J., Shoeiby, M., Petersson, L., Armin, M.A.: Dual contrastive learning for unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2021)
- He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. IEEE transactions on pattern analysis and machine intelligence **33**(12), 2341–2353 (2010)
- 32. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems (2017)
- 33. Hu, X., Fu, C.W., Zhu, L., Qin, J., Heng, P.A.: Direction-aware spatial context features for shadow detection and removal. IEEE transactions on pattern analysis and machine intelligence 42(11), 2795–2808 (2019)
- Hyvärinen, A., Oja, E.: A fast fixed-point algorithm for independent component analysis. Neural computation 9(7), 1483–1492 (1997)

- Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural networks 13(4-5), 411–430 (2000)
- Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
- 37. Jayaram, V., Thickstun, J.: Source separation with deep generative priors. In: International Conference on Machine Learning. pp. 4724–4735. PMLR (2020)
- Jin, Y., Sharma, A., Tan, R.T.: Dc-shadownet: Single-image hard and soft shadow removal using unsupervised domain-classifier guided network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5027–5036 (2021)
- Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. pp. 694–711. Springer (2016)
- 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)
- Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
- 42. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) (2014)
- Kong, Q., Xu, Y., Wang, W., Jackson, P.J., Plumbley, M.D.: Single-channel signal separation and deconvolution with generative adversarial networks. arXiv preprint arXiv:1906.07552 (2019)
- Le, H., Samaras, D.: Shadow removal via shadow image decomposition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8578–8587 (2019)
- Lee, T.W., Lewicki, M.S., Sejnowski, T.J.: Ica mixture models for unsupervised classification of non-gaussian classes and automatic context switching in blind signal separation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(10), 1078–1089 (2000)
- 46. Li, C., Yang, Y., He, K., Lin, S., Hopcroft, J.E.: Single image reflection removal through cascaded refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3565–3574 (2020)
- 47. Li, R., Pan, J., Li, Z., Tang, J.: Single image dehazing via conditional generative adversarial network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8202–8211 (2018)
- 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)
- Li, S., Araujo, I.B., Ren, W., Wang, Z., Tokuda, E.K., Junior, R.H., Cesar-Junior, R., Zhang, J., Guo, X., Cao, X.: Single image deraining: A comprehensive benchmark analysis. In: CVPR. pp. 3838–3847 (2019)
- Li, W., Jafari, O.H., Rother, C.: Deep object co-segmentation. In: Asian Conference on Computer Vision. pp. 638–653. Springer (2018)
- Li, X., Wu, J., Lin, Z., Liu, H., Zha, H.: Recurrent squeeze-and-excitation context aggregation net for single image deraining. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 254–269 (2018)

- 18 J. Han et al.
- Lin, S., Ryabtsev, A., Sengupta, S., Curless, B.L., Seitz, S.M., Kemelmacher-Shlizerman, I.: Real-time high-resolution background matting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8762–8771 (2021)
- Liu, Y., Zhu, Z., Bai, X.: Wdnet: Watermark-decomposition network for visible watermark removal. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3685–3693 (2021)
- 54. Liu, Y.L., Lai, W.S., Yang, M.H., Chuang, Y.Y., Huang, J.B.: Learning to see through obstructions with layered decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)
- 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)
- Ma, D., Wan, R., Shi, B., Kot, A.C., Duan, L.Y.: Learning to jointly generate and separate reflections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2444–2452 (2019)
- Man, Z., Fu, X., Xiao, Z., Yang, G., Liu, A., Xiong, Z.: Unfolding taylor's approximations for image restoration. Advances in Neural Information Processing Systems 34, 18997–19009 (2021)
- Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: IEEE international conference on computer vision (ICCV). pp. 2794–2802 (2017)
- Mittal, A., Moorthy, A.K., Bovik, A.C.: No-reference image quality assessment in the spatial domain. IEEE Transactions on image processing 21(12), 4695–4708 (2012)
- Mittal, A., Soundararajan, R., Bovik, A.C.: Making a "completely blind" image quality analyzer. IEEE Signal processing letters 20(3), 209–212 (2012)
- Nayar, S.K., Narasimhan, S.G.: Vision in bad weather. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. vol. 2, pp. 820–827. IEEE (1999)
- Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 2, pp. 1447–1454. IEEE (2006)
- Oliveira, P.R., Romero, R.A.: Improvements on ica mixture models for image preprocessing and segmentation. Neurocomputing 71(10-12), 2180–2193 (2008)
- Porav, H., Bruls, T., Newman, P.: I can see clearly now: Image restoration via deraining. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7087–7093. IEEE (2019)
- Qian, R., Tan, R.T., Yang, W., Su, J., Liu, J.: Attentive generative adversarial network for raindrop removal from a single image. In: CVPR. pp. 2482–2491 (2018)
- 66. Qu, L., Tian, J., He, S., Tang, Y., Lau, R.W.: Deshadownet: A multi-context embedding deep network for shadow removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4067–4075 (2017)
- 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)
- Ren, W., Tian, J., Han, Z., Chan, A., Tang, Y.: Video desnowing and deraining based on matrix decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4210–4219 (2017)

- Rother, C., Kolmogorov, V., Blake, A.: "grabcut" interactive foreground extraction using iterated graph cuts. ACM transactions on graphics (TOG) 23(3), 309– 314 (2004)
- Sakaridis, C., Dai, D., Van Gool, L.: Semantic foggy scene understanding with synthetic data. International Journal of Computer Vision 126(9), 973–992 (2018)
- Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
- Tan, R.T.: Visibility in bad weather from a single image. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2008)
- Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016)
- Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9446–9454 (2018)
- 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)
- Wang, J., Li, X., Yang, J.: Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1788–1797 (2018)
- 77. Wang, Z., Philion, J., Fidler, S., Kautz, J.: Learning indoor inverse rendering with 3d spatially-varying lighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12538–12547 (2021)
- Wu, Y., He, Q., Xue, T., Garg, R., Chen, J., Veeraraghavan, A., Barron, J.T.: How to train neural networks for flare removal. Proceedings of the IEEE/CVF International Conference on Computer Vision (2021)
- Xiao, J., Zhou, M., Fu, X., Liu, A., Zha, Z.J.: Improving de-raining generalization via neural reorganization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4987–4996 (2021)
- Xue, T., Rubinstein, M., Liu, C., Freeman, W.T.: A computational approach for obstruction-free photography. ACM Transactions on Graphics (TOG) 34(4), 1–11 (2015)
- Yang, Q., Tan, K.H., Ahuja, N.: Shadow removal using bilateral filtering. IEEE Transactions on Image processing 21(10), 4361–4368 (2012)
- Yang, W., Tan, R.T., Feng, J., Liu, J., Guo, Z., Yan, S.: Deep joint rain detection and removal from a single image. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1685–1694 (2017). https://doi.org/10.1109/CVPR.2017.183
- 83. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. CVPR (2022)
- 84. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H., Shao, L.: Multi-stage progressive image restoration. In: CVPR (2021)
- Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-attention generative adversarial networks. In: International conference on machine learning. pp. 7354–7363. PMLR (2019)
- Zhang, H., Sindagi, V., Patel, V.M.: Image de-raining using a conditional generative adversarial network. IEEE transactions on circuits and systems for video technology **30**(11), 3943–3956 (2019)
- Zhang, K., Li, D., Luo, W., Ren, W., Liu, W.: Enhanced spatio-temporal interaction learning for video deraining: A faster and better framework. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)

- 20 J. Han et al.
- Zhang, K., Li, D., Luo, W., Ren, W., Ma, L., Li, H.: Dual attention-in-attention model for joint rain streak and raindrop removal. arXiv preprint arXiv:2103.07051 (2021)
- Zhang, K., Luo, W., Ren, W., Wang, J., Zhao, F., Ma, L., Li, H.: Beyond monocular deraining: Stereo image deraining via semantic understanding. In: European Conference on Computer Vision. pp. 71–89. Springer (2020)
- 90. Zhang, K., Luo, W., Yu, Y., Ren, W., Zhao, F., Li, C., Ma, L., Liu, W., Li, H.: Beyond monocular deraining: Parallel stereo deraining network via semantic prior. International Journal of Computer Vision pp. 1–16 (2022)
- Zhang, X., Ng, R., Chen, Q.: Single image reflection separation with perceptual losses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4786–4794 (2018)
- 92. Zhong, Y., Dai, Y., Li, H.: Stereo computation for a single mixture image. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 435– 450 (2018)
- Zhou, M., Wang, F., Wei, X., Wang, R., Wang, X.: Pid controller-inspired model design for single image de-raining. IEEE Transactions on Circuits and Systems II: Express Briefs 69(4), 2351–2355 (2021)
- Zhou, M., Wang, R.: Control theory-inspired model design for single image deraining. IEEE Transactions on Circuits and Systems II: Express Briefs 69(2), 649– 653 (2021)
- 95. Zhou, M., Xiao, J., Chang, Y., Fu, X., Liu, A., Pan, J., Zha, Z.J.: Image de-raining via continual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4907–4916 (2021)
- 96. Zhou Wang, Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
- 97. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)
- Zou, Z., Lei, S., Shi, T., Shi, Z., Ye, J.: Deep adversarial decomposition: A unified framework for separating superimposed images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12806–12816 (2020)