Deep Adversarial Decomposition: A Unified Framework for Separating Superimposed Images

Zhengxia Zou[1]*, Sen Lei[2], Tianyang Shi[3], Zhenwei Shi[2], Jieping Ye[1,4]
[1]University of Michigan, Ann Arbor, [2]Beihang University, [3]NetEase Fuxi AI Lab, [4]Didi Chuxing
{zzhengxi, jpye}@umich.edu, {senlei, shizhenwei}@buaa.edu.cn, shitianyang@corp.netease.com

Abstract

Separating individual image layers from a single mixed image has long been an important but challenging task. We propose a unified framework named "deep adversarial decomposition" for single superimposed image separation. Our method deals with both linear and non-linear mixtures under an adversarial training paradigm. Considering the layer separating ambiguity that given a single mixed input, there could be an infinite number of possible solutions, we introduce a "Separation-Critic" - a discriminative network which is trained to identify whether the output layers are well-separated and thus further improves the layer separation. We also introduce a "crossroad l_1" loss function, which computes the distance between the unordered outputs and their references in a crossover manner so that the training can be well-instructed with pixel-wise supervision. Experimental results suggest that our method significantly outperforms other popular image separation frameworks. Without specific tuning, our method achieves the state of the art results on multiple computer vision tasks, including the image deraining, photo reflection removal, and image shadow removal.

1. Introduction

In the computer vision field, many tasks can be considered as image layer mixture/separation problems. For example, when we take a picture on rainy days, the image obtained can be viewed as a mixture of two layers: a rain-streak layer and a clean background layer. When we look through a transparent glass, we see a mixture of the scene beyond the glass and the scene reflected by the glass.

Separating superimposed images with single observation has long been an important but challenging task. On one hand, it forms the foundation of a large group of real-world applications, including transparency separation, shadow removal, deraining, etc. On the other hand, it is naturally a massively ill-posed problem, where the difficulty lies not only in the absence of the mixture function but also in the lack of constraints on the output space. In recent literature, most of the above-mentioned tasks are investigated individually despite the strong correlation between them [8, 23, 48, 54, 59, 60]. In this paper, we propose a new framework for single superimposed image separation which deals with all the above tasks under a unified framework.

Learning *a priori* for image decomposition. Given a single superimposed image, as we have no extra constraint on the output space, there could be an infinite number of possible decomposition. Previous works often integrate hard-crafted priors to apply additional constraints to their separation outputs. For example, in recent literature [17, 67], researchers introduced the "gradient exclusiveness" [67] and the "internal self-similarity" [17], where the former one emphasizes the independence of the layers to be separated in their gradient domain, and the latter one assumes that the distribution of small patches within each separate layer should be "simpler" (more uniform) than in the original mixed one. However, these hand-crafted priors may introduce unexpected bias and thus fails in complex mixture conditions. In this paper, we investigate an interesting question: can a good prior be learned from data? To answer this question, we re-examine the prior under a totally different point of view by taking advantage of the recent success of generative adversarial networks [19, 28, 47]. We introduce a "Separation-Critic" - a discriminative network D_C, which is trained to identify whether the output layers are well-separated. The layer separation can be thus gradually enforced by fooling the Critic under an adversarial training paradigm.

Crossroad loss function. In addition to the Critic D_C, we also introduce a layer separator G and train it to minimize the distance between the separated outputs and the ground truth references. However, a standard l_1 or l_2 loss does not apply to our task since the G may predict unordered outputs. We, therefore, introduce a "crossroad l_1" loss which computes the distance between the outputs and their references in a crossover fashion. In this way, the training can be well-instructed by pixel-wise supervision.

*Corresponding author: Zhengxia Zou (zzhengxi@umich.edu)

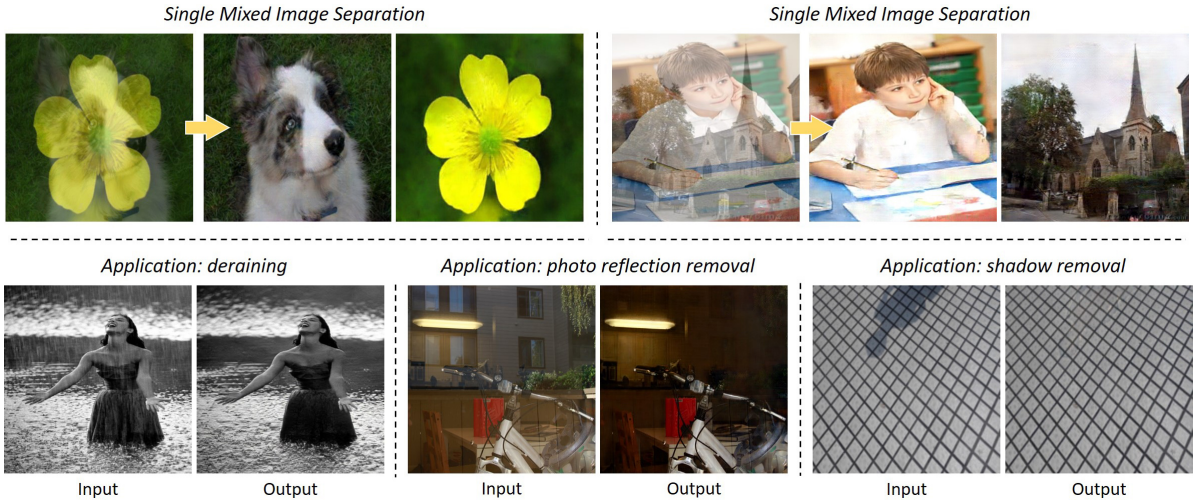

Figure 1: We propose a unified framework for single mixed image separation under an adversarial training paradigm. Our method can be applied to a variety of real-world tasks, including image deraining, photo reflection removal, image shadow removal, etc.

Confronting nonlinear mixture and degradation. In some real-world image separation tasks, the mixture of images is usually beyond linear. For example, the formation of a reflection image may depend not only on the relative position of the camera to the image plane but also on lighting conditions [59]. Besides, the degradation (e.g., overexposure and noise) may further increase the difficulty of the separation. In these conditions, one may need injecting "imagination" into the algorithm to recover the hidden structure of degraded data. Inspired by the recent image translation methods [28], we further introduce two Markovian discriminators (PatchGAN) to improve the perceptual quality of the outputs.

Experimental results show that our method significantly outperforms other popular image separation frameworks [17, 35]. We apply our method to a variety of computer vision tasks. Without specifically tuning, we achieve the state of the art (sota) results on nine datasets of three different tasks, including image deraining, photo reflection removal, and image shadow removal. To our best knowledge, this is the first unified framework for solving these problems as most previous solutions on these tasks are separately investigated and designed.

2. Related Work

Superimposed image separation. In signal processing, a similar topic to our paper is Blind Source Separation (BSS) [5, 14–16, 27], which aims to separate source signals from a set of mixed ones. The research of this topic can be traced back to the 1990s [26], where the Independent Component Analysis (ICA) [27] was a representative of the methods at the time. The key to estimating the ICA model is the Central Limit Theorem, i.e., the distribution of a sum of two images tends toward a Gaussian distribution, under certain conditions. Some statistics-based criteria thus have been introduced to measure the independence and non-Gaussianity of the images, e.g., Kurtosis and negative entropy.

The main difference between the BSS and our task is that the former one typically requires multiple mixed inputs [14–16, 27] or additional user interactions [35], while the latter one does not. We focus on the latter case since multiple mixed inputs or user interactions are not always available in practice. Recently, Gandelsman *et al*. proposed a deep learning-based method called Double-DIP [17] that can separate superimposed images with single observation under certain conditions. However, their method can only handle the input with regular mixed patterns.

Related application. Many real-world tasks can be viewed as special cases of superimposed image separation:

1) Single Image Deraining. A rainy image can be simply viewed as a superposition of a clean background image and rain streaks. Some early deraining methods were designed based on low-rank constraints [4, 41, 42, 71] and sparse coding methods [20, 29, 43, 57], where the rain-streaks are considered as high frequency noise. However, these methods usually lead to over-smoothed results. Recent deraining methods typically formulate the deraining as a deep learning based "image-to-image" translation process which is trained with pixel-wise regression loss [12, 13, 23, 39, 55, 62, 64].

2) Reflection-removal. Early reflection removal methods often require additional input images [35] and hand-crafted priors on estimating the reflection layer. Such priors include

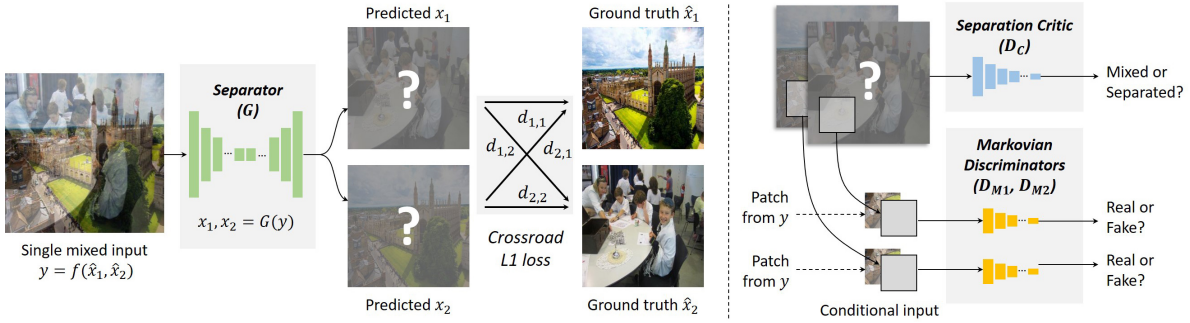

Figure 2: An overview of our method. Our method consists of a separator G and several discriminators. The G, which is trained under a "crossroad l_1" loss, aims to decompose a single mixed input into two individual images. To identify whether the separation is good or not, we introduce an "Critic" D_C, which is trained together with G under an adversarial training paradigm. We further use two Markovian Discriminators (D_{M1}, D_{M2}) to improve the perceptual quality of the outputs.

the smoothness prior [40, 51], gradient sparsity constraint [2, 36, 37], ghost cues [51], etc. In Recent methods, the priors are usually explored by data-driven methods [9, 44, 53, 60] and adversarial training/synthesis [59, 67], which better handle more complex reflections.

3) Shadow-removal. To remove shadows, some early works designed physical models based on illumination invariant assumption [10, 11]. Later, more methods based on hand-crafted features were proposed [1, 21, 25, 32, 52, 66, 69]. Similar to deraining and reflection removal, recent researches on shadow removal also suggest using deep learning or adversarial training techniques, which brings additional improvements especially on complex lightening conditions [8, 24, 30, 33, 46, 50, 54, 68].

Generative Adversarial Networks (GAN). GAN has received a great deal of attention in recent years and has achieved impressive results in a variety of computer vision tasks, e.g., image generation [7, 47], image style transfer [28, 70], image super-resolution [34], etc. A typical GAN [19] consists of two neural networks: a generator G and a discriminator D. The key to the GAN's success is the idea of adversarial training where the G and D will contest with each other in a minimax two-player game and forces the generated data to be, in principle, indistinguishable from real ones. More recently, GAN has also been applied to some image separation tasks to improve perceptual quality of the recovered images, including image deraining [38, 65], image reflection removal [44, 59, 67] and image de-shadowing [8, 24, 33, 54].

3. Methodology

We frame the training of our model as a pixel-wise regression process with the help of adversarial losses. Our method consists of an image separator G, a Separation Critic D_C and two Markovian discriminators D_{M1} and D_{M2}. Fig. 2 shows an overview of our method.

3.1. Crossroad l_1 Loss Function

Suppose $\hat{x}_1$ and $\hat{x}_2$ represent two individual images and $y = f(\hat{x}_1, \hat{x}_2)$ represents their mixture. We assume the operation $f(\cdot)$ is unknown and could be either a linear or a non-linear mixing function. Given a mixed input y, our separator aims to predict two individual outputs x_1 and x_2:

$$x_1,\ x_2 = G(y), \tag{1}$$

that recover the two original images $\hat{x}_1$ and $\hat{x}_2$.

We train the separator G to minimize the distance between its outputs $(x_1,\ x_2)$ and their ground truth $(\hat{x}_1,\ \hat{x}_2)$. Note that since we can not specify the order of the two outputs for a typical image decomposition problem (especially when the $\hat{x}_1$ and $\hat{x}_2$ are from the same image domain), the standard pixel-wise l_1 or l_2 loss functions do not apply to our task. The solution to this problem is to introduce new loss functions that can deal with unordered outputs. We therefore propose a new loss function called "crossroad l_1" loss for our task. The main idea behind is to crossly compute the distance by exchanging the order of the outputs and then take their minimum value as the final response:

$$\begin{aligned} &l_{cross}((x_1, x_2),\ (\hat{x}_1, \hat{x}_2)) \\ &= \min\{d_{1,1} + d_{2,2},\ \ d_{1,2} + d_{2,1}\} \end{aligned} \tag{2}$$

where $d_{i,j} = \|x_i - \hat{x}_j\|_1, i, j \in \{1, 2\}$. We use the standard l_1 function rather than l_2 in $d_{i,j}$ since it encourages less blurring effect. The G can be therefore trained to minimize the loss $\mathcal{L}_{cross}$ on an entire dataset:

$$\mathcal{L}_{cross}(G) = E_{\hat{x}_i \sim p_i(\hat{x}_i)}\{l_{cross}((x_1, x_2),\ (\hat{x}_1, \hat{x}_2))\}, \tag{3}$$

where $p_i(\hat{x}_i)$ represents the distribution of the image data, and $i \in \{1, 2\}$.

3.2. Separation Critic

Considering the layer ambiguity, instead of applying any hand-crafted [67] or statistics-based constraints [27] to

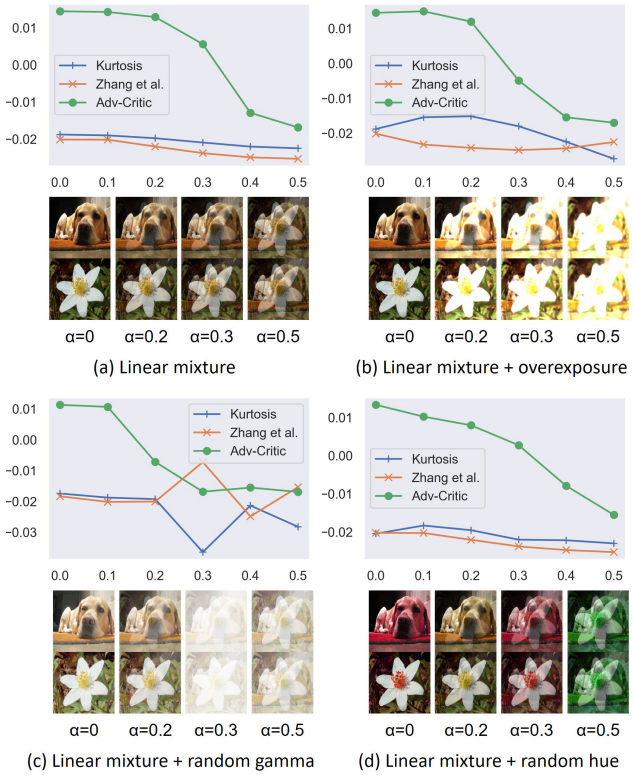

Figure 3: We compare different decomposition priors for image separation, including the "exclusion loss" [17, 67], "Kurtosis" [27], and the proposed "Separation-Critic". For either of the three metrics, a lower score indicates a heavier mixture. In sub-figure (a), we plot the response of the three metrics given a set of mixed inputs that are synthesized based on Eq. (5). Clearly, if a metric is good enough, the response should be monotonically decreasing as α increases. We also test on additional nonlinear corruptions, including (b) overexposure, (c) random gamma correction, and (d) random hue transform. The proposed Critic shows better robustness in all conditions.

our output space, we learn a decomposition prior through an adversarial training process. We therefore introduce a "Separation-Critic" D_C which is trained to distinguish between the outputs (x_1, x_2) and a pair of clean images $(\hat{x}_1, \hat{x}_2)$. We express its objective function as follows:

$$\begin{aligned}\mathcal{L}_{critic}(G, D_C) = & E_{\hat{x}_i \sim p_i(\hat{x}_i)}\{\log D(\hat{x}_1, \hat{x}_2)\} \\ & + E_{x_i \sim p_i(x_i)}\{\log(1 - D(x_1, x_2))\} \\ & + E_{\hat{x}_i \sim p_i(\hat{x}_i)}\{\log(1 - D(\text{mix}(\hat{x}_1, \hat{x}_2)))\}.\end{aligned} \tag{4}$$

Note that when training the D_C with fake samples, in addition to the decomposed output (x_1, x_2), we also synthesize a set of "fake" images by mixing two clean images with random linear weights $(x'_1, x'_2) = \text{mix}(\hat{x}_1, \hat{x}_2)$ to enhance its discriminative ability on mixed images:

$$x'_1 = \alpha\hat{x}_1 + (1-\alpha)\hat{x}_2, \quad x'_2 = (1-\alpha)\hat{x}_1 + \alpha\hat{x}_2. \tag{5}$$

At the input end of the D_C, we simply concatenate two images together in the channel dimension for modeling their joint probability distribution. The adversarial training of G and D_C is essentially a minimax optimization process, where G tries to minimize this objective while D_C tries to maximize it: $G^\star = \arg\min_G \max_{D_C} \mathcal{L}_{critic}(G, D_C)$.

In Fig. 3, we give four examples to illustrate the effectiveness of our Critic. We compare a well-trained D_C with two popular metrics for image separation, 1) the exclusion loss [17, 67], and 2) the Kurtosis [27], where the former one enforces separation of two images on the image gradient domain, and the latter one is widely used in BSS for measuring the independence (non-Gaussianity) of the recovered signals: $\text{Kurtosis}(u) = E\{u^4\} - 3(E\{y^2\})^2$. For either of the three metrics, a lower score indicates a higher degree of mixture[1]. We mix two images by using Eq. (5) with different α. Clearly, if a metric is good enough, the curve should be monotonically decreasing as α increases. Fig. 3 (a) shows the response the above three metrics. We further add some additional nonlinear corruptions on the mixed images, including (b) random overexposure, (c) random gamma correction, and (d) random hue transform. We can see our Critic shows better robustness, especially for nonlinear degradation.

3.3. Improving Perceptual Quality

To improve the perceptual quality of the decomposed images, we further introduce another two conditional discriminators D_{M1} and D_{M2} to enhance high-frequency details. We follow Isola *et al.* [28] and build D_{M1} and D_{M2} as two local perception networks - that only penalize structure at the scale of patches (a.k.a the Markovian discriminator or "PatchGAN"). The D_{M1} and D_{M2} try to classify if each $N \times N$ patch in an image is a clean image (real) or a decomposed one (fake). This type of architecture can be equivalently implemented by building a fully convolutional network with $N \times N$ perceptive fields, which is more computationally efficient since the responses of all patches can be obtained by taking only one time of forward propagation. We express the objective of D_{M1} and D_{M2} as follows:

$$\begin{aligned}\mathcal{L}_{Mi}(G, D_{Mi}) = & E_{(\hat{x}_i, y) \sim p_i(\hat{x}_i, y)}\{\log D(\hat{x}_i|y)\} \\ & + E_{(x_i, y) \sim p_i(x_i, y)}\{\log(1 - D(x_i|y))\},\end{aligned} \tag{6}$$

where $i = 1, 2$. Our final objective function is defined as follows:

$$\begin{aligned}\mathcal{L}(G, D_C, D_{Mi}) = & \mathcal{L}_{cross}(G) + \beta_C \mathcal{L}_{critic}(G, D_C) \\ & + \beta_M \sum_{i=1,2} \mathcal{L}_{Mi}(G, D_{Mi})\end{aligned} \tag{7}$$

[1]To ensure the monotonic consistency of the three metrics, we plot their negative values when computing the exclusion loss and Kurtosis

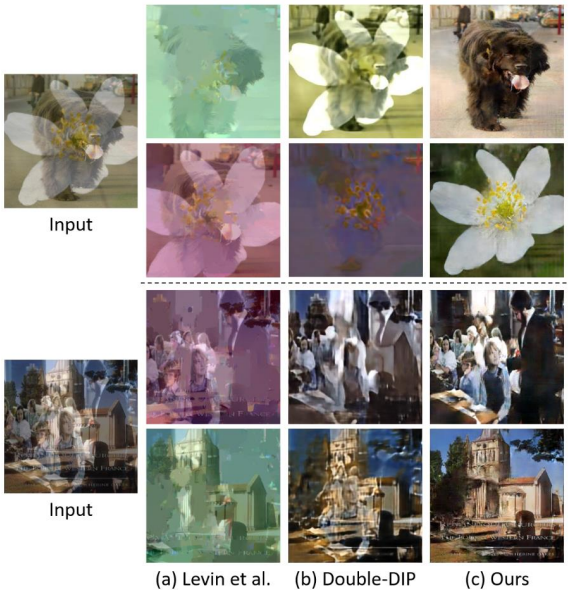

Figure 4: A comparison between our method and two image separation methods: Double-DIP [17] and a user-assisted framework proposed by Levin *et al*. [35]. Check out our supplementary material for more separation results.

where $\beta_C > 0$ and $\beta_M > 0$ control the balance between the different components of the objective. We aim to solve:

$$G^\star = \arg\min_G \max_{D_C, D_{Mi}} \mathcal{L}(G, D_C, D_{Mi}), \;\; i = 1, 2. \quad (8)$$

The networks G, D_C and D_{Mi} thus can be alternatively updated in an end-to-end training process.

3.4. Implementation Details

We follow the configuration of the "UNet" [49] when designing the architecture of our separator G. We build our D_C, D_{M1} and D_{M2} as three standard FCNs with 4, 3, and 3 convolutional layers. The perceptive field of D_{M1} and D_{M1} is set to $N = 30$. We resize the input of D_C to a relatively small size, e.g., 64×64, to capture the semantics of the whole image instead of adding more layers. We do not use batch-normalization in G as it may introduce unexpected artifacts. As our default settings, we train our model for 200 epochs by using the Adam optimizer with batch_size $= 2$ and learning_rate $= 0.0001$. We set $\beta_C = \beta_M = 0$ for the first 10 epochs and set $\beta_C = \beta_M = 0.001$ for the rest epochs. For more implementation details, please refer to our supplementary material.

4. Experimental Analysis

We evaluate our methods on four tasks: 1) superimposed image separation, 2) image deraining, 3) image reflection removal, and 4) image shadow removal.

	Dogs+Flwrs.	LSUN
Double-DIP [17] (CVPR'19)	14.70 / 0.661	13.83 / 0.590
Levin *et al*. [35] (TPAMI'07)	10.54 / 0.444	10.46 / 0.366
Our method (tr. on ImageNet)	<u>23.32</u> / <u>0.803</u>	<u>21.63</u> / <u>0.773</u>
Our method (w/ default tr. set)	**25.51 / 0.849**	**26.32 / 0.883**

Table 1: Comparisons (PSNR/SSIM) of different methods on mixed image separation: 1) Stanford-Dogs [31] + VGG-Flowers [45], 2) LSUN Classroom + LSUN Church [63]. To test on the cross-domain generalization ability of our method, we also train on ImageNet and test on the above datasets. Higher scores indicate better.

4.1. Separating Superimposed Images

We evaluate our method on two groups of well-known datasets: 1) Stanford-Dogs [31] + VGG-Flowers [45], 2) LSUN Classroom + LSUN Church [63]. During training phase, we randomly select two images ($\hat{x}_1$, $\hat{x}_2$) from one group of the datasets and then linearly mix them as $y = \alpha\hat{x}_1 + (1-\alpha)\hat{x}_2$ with a random linear mixing factor α from the range of $[0.4, 0.6]$. During testing, we set the mixing factor as a constant $\alpha = 0.5$. All images are resized to 256x256 pixels. We follow the datasets' original train/test split when performing training and evaluation. For the LSUN dataset, due to its large number of images, we only train our method for 20 epochs.

We compare our method with other two popular methods for single mixed image separation: the Double-DIP (CVPR'19) [17] and Levin's method (TPAMI'07) [35], where the former one is an unsupervised deep learning based method, and the latter one is designed based on image statistics and requires additional user-interactions. Fig. 4 shows two typical results of the above three methods. Table 1 shows their quantitative evaluations[2]. We use the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) index [58] as two basic evaluation metrics. The accuracy is crossly computed between the outputs and their references. We record the best of the two scores as the final accuracy. Our method significantly outperforms the other two methods in terms of both visual quality and quantitative scores. As Double-DIP and Levin's methods do not require training data of specific domain, we also test on the cross-domain generalization ability of our method by training on ImageNet-1M (∼1.28M images, randomly mixed on 1K classes, with 6 training epochs) [6] and test on the above datasets. Note that this time we do not train specifically on their own training sets and our method again shows superiority over the other two methods.

[2]Since Levin's method requires heavy user interactions and the Double-DIP is extremely slow (∼40 minutes / image on a GTX-1080Ti GPU), we only evaluate these two methods on 10 images that are randomly selected from each of our test set.

Figure 5: A comparison of image separation results with the standard l_1 loss and the proposed crossroad l_1 loss.

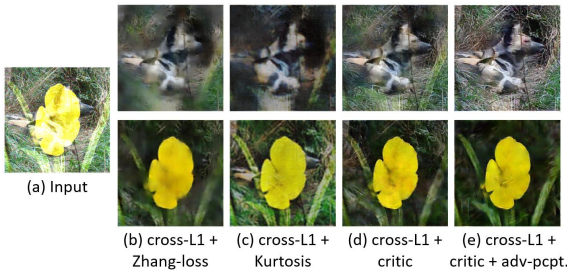

Figure 6: A comparison of the results based on different image separation priors: Zhang's exclusion loss [67], Kurtosis [27], and the Separation Critic (ours).

Ablation	PSNR	SSIM	PI
Ours (with standard l_1 loss)	16.26	0.6768	23.34
Ours (with cross-l_1 loss)	**22.74**	**0.7782**	**21.92**

Table 2: A comparison of the image separation results with standard l_1 loss and the proposed crossroad l_1 loss on Stanford-Dogs dataset [31]. PSNR/SSIM: higher scores indicate better. PI: lower scores indicate better.

Ablation	PSNR	SSIM	PI
cross-l_1 only	17.34	0.6693	22.41
cross-l_1 + Zhang's loss [67]	17.77	0.6840	23.09
cross-l_1 + kurtosis [27]	18.02	<u>0.7141</u>	22.28
cross-l_1 + Sep-Critic	**18.29**	**0.7231**	<u>22.16</u>
cross-l_1 + Sep-Critic + adv-pcpt.	<u>18.27</u>	0.6938	**21.97**

Table 3: Evaluation of decomposition results with different priors on Stanford-Dogs [31] + VGG-Flowers [45] datasets (with overexposure and noise). PSNR/SSIM: higher scores indicate better. PI: lower scores indicate better.

4.2. Controlled Experiments

Analysis on the crossroad l_1 loss. To evaluate the importance of our crossroad l_1 loss, we replace it with a standard l_1 while keeping other settings unchanged. We train the above two models on the Stanford-Dogs [31]. Table 2 shows the evaluation results of the two models and Fig. 5 shows a group of visual comparisons. We can see our method clearly separates the two images while the standard l_1 fails to do that and encourages "averaged" outputs.

Analysis on the adversarial losses. We compare our method with different decomposition priors on Stanford-Dogs [31] + VGG-Flowers [45] datasets. In addition to the linear mixing inputs, we also apply overexposure and noises to increase the separation difficulties. To better evaluate the perceptual quality, we introduce another metric called Perception Index (PI) [3]. The PI was originally introduced as a no-reference image quality assessment method based on the low-level image statistics and is recently widely used for evaluating super-resolution results [3, 56]. In Fig. 6 and Table 3, we compare our adversarial losses with the exclusion loss [17, 67] and the Kurtosis [27]. As we can see, the integration of our adversarial losses yields noticeable improvements in the output quality. We found the exclusion loss encourages blurred output and it is hard to balance it with other losses. We also found the Kurtosis may introduce a slight color-shift on its outputs.

	Rain100H [62]	Rain800 [65]
LP [41] (CVPR'16) ¶‡	15.05 / 0.425	20.46 / 0.730
DDN [13] (CVPR'17) ¶‡	22.26 / 0.693	21.16 / 0.732
JORDER [62] (CVPR'17) ‡	23.45 / 0.749	22.29 / 0.792
RESCAN [39] (ECCV'18) ‡	26.45 / 0.846	<u>24.09</u> / <u>0.841</u>
DID [64] (CVPR'18) †	25.00 / 0.754	- / -
DAF-Net [23] (CVPR'19) †	28.44 / 0.874	- / -
PReNet [48] (CVPR'19) ¶	<u>29.46</u> / <u>0.899</u>	- / -
Our method	**30.85 / 0.932**	**24.49 / 0.885**

Table 4: A comparison (PSNR / SSIM) of different deraining methods on two datasets: Rain100H [62] and Rain800 [65]. Results reported by: # [62], ‡ [39], † [23], ¶ [48].

	DID [64]	DDN1k [13]
LP [41] (CVPR'16) †	22.75 / 0.835	20.66 / 0.811
JORDER [62] (CVPR'17) †	24.32 / 0.862	22.26 / 0.841
DDN [13] (CVPR'17) †	27.33 / 0.898	25.63 / 0.885
JBO [71] (ICCV'17) †	23.05 / 0.852	22.45 / 0.836
DID [67] (CVPR'18) †	27.95 / 0.909	<u>26.07</u> / **0.909**
SPANet [55] (CVPR'19) #	<u>30.05</u> / <u>0.934</u>	- / -
Our method	**31.67 / 0.942**	**27.91** / <u>0.893</u>

Table 5: A comparison (PSNR / SSIM) of different deraining methods. All methods are trained on the training set of DID [64] and then tested on the test sets of DID and DDN1k [13]. Results reported by: † [64], # [55].

4.3. Application: Deraining

We conduct our deraining experiments on several datasets: Rain100H [62], Rain800 [65], and DID [64]. To better test the generalization ability of our method, we follow Zhang *et al*. [64] to train our method on DID [64], and then randomly sample 1,000 images from the dataset [13] as another testing set, denoted as DDN1k. Given a rainy input image, we use its clean background and the rain streak

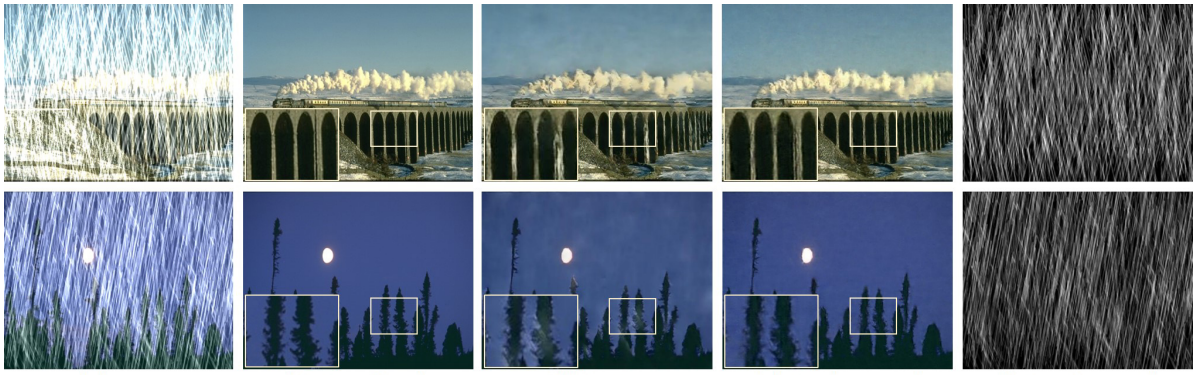

Figure 7: Deraining results of PReNet [48] (CVPR'19) and our method on the Rain100H [62] dataset. Our method encourages fewer artifacts than PReNet. Also, the rain-streak map can be well-estimated by using our method. As a comparison, the PReNet ignores this part of the output.

Reflection Removal Dataset [59]					
Method	Focused set	Defocused set	Ghosting set	Method	Dataset [60]
CEILNet [9] (ICCV'17) †	19.524 / 0.742	20.122 / 0.735	19.685 / 0.753	Li & Brown [40] (CVPR'14) ¶	16.46 / 0.745
Zhang *et al*. [67] (ICCV'18) †	17.090 / 0.712	18.108 / 0.758	17.882 / 0.738	SIRP [2] (CVPR'17) ¶	19.18 / 0.760
BDN [60] (ECCV'18) †	14.258 / 0.632	14.053 / 0.639	14.786 / 0.660	CEILNet [9] (ICCV'17) ¶	19.80 / 0.782
RmNet [59] (CVPR'19) †	21.064 / 0.770	22.896 / 0.840	21.008 / 0.780	BDN [60] (ECCV'18) ¶	23.11 / 0.835
Our method	**22.809 / 0.871**	**23.195 / 0.891**	**23.266 / 0.881**	Our method	**23.18 / 0.877**

Table 6: Reflection removal results (PSNR / SSIM) of different methods on two challenging datasets [59] and [60]. In dataset [59], the images are nonlinearly synthesized with three types of reflections: "focused", "defocused", and "ghosting". We achieve the best results in all experimental entries. Results reported by: † [59], ¶ [60].

Figure 8: Deraining results of our method on some real-world rain images [65]. 1st row: input. 2nd row: output.

	Dataset [67]
Li & Brown [40] (CVPR'14) *	18.29 / 0.750
CEILNet [9] (ICCV'17) *	19.04 / 0.762
Zhang *et al*. [67] (ICCV'18) *	21.30 / 0.821
Our method	**22.36 / 0.846**

Table 7: Reflection removal results (PSNR / SSIM) of different methods on dataset [67]. * results reported by [67].

map as our ground truth references. Since Rain800 and DID do not provide rain streak maps, we simply set the ground truth of our second output as a "zero image" when training on these two datasets. In all our following experiments, we set $\beta_C = \beta_M = 0.0001$, and set the input/output size of our separator G to 512x512 pixels.

We compare with more than five sota deraining methods, including RESCAN (ECCV'18) [39], DID (CVPR'18) [64], DAF-Net (CVPR'19) [23], PReNet (CVPR'19) [48], SPANet (CVPR'19) [55], etc. Table 4 and Table 5 show the deraining results of these methods. Our method outperforms other sota methods in most entries. Fig. 7 shows two examples from the dataset Rain100H with our method and PReNet (CVPR'19) [48]. Our method encourages less artifacts. Another advantage of our method is that the rain-streak map can be also estimated. As a comparison, the PReNet ignores this part of output. Fig. 8 shows a group of our deraining results on some real-world rain images.

4.4. Application: Image Reflection Removal

We test our method on two large scale datasets for reflection removal [59, 60]. The dataset [60] consists of over 50,000 images which are synthesized by mixing their transmission layers and reflection layers (linear mixture + Gaussian blur). The dataset [59] consists of 12,000 images with three types of reflections: "focused", "defocused",

Figure 9: Results of different reflection removal methods: BDN [60] (ECCV'18), RmNet [59] (CVPR'19), and our method on a real-world reflection image from the dataset [67].

Figure 10: Reflection removal results of our method on the BDN dataset [60]. 1st row: input. 2nd row: our output.

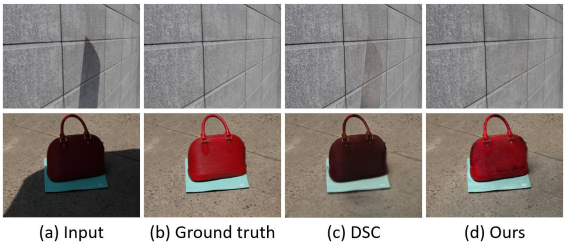

Figure 11: Results of our method and DSC [22] (TPAMI19) on two datasets: ISTD [54] (1st row), SRD [46] (2nd row).

and "ghosting", which are synthesized by using adversarial training. When we train our model, the transmission layers are used as the reference for our first output. We discard the synthesized reflection in our second output since it cannot capture ground truth reflections. We compare with several sota reflection removal methods, including the method of Zhang *et al*. (ICCV'18) [67], BDN (ECCV'18) [60], RmNet (CVPR'19) [59], etc. Table 6 shows the quantitative evaluations of these methods. Note that although RmNet [59] uses auxiliary images [53, 67] during training, we still achieve the best results in all experimental entries. We also test on a set of real-world reflection images [67]. We train our model on the synthetic training set [67] and then evaluate on its real-world testing set. Fig. 9 and Table 7 shows some comparison results.

4.5. Application: Shadow Removal

In this experiment, we test our method on two shadow removal datasets: ISTD [54] and SRD [46]. The two datasets consist of 1,870 and 3,088 shadow/shadow-free image pairs

	ISTD [54]	SRD [46]
Yang *et al*. [61] (TIP'12) *†	15.63	22.57
Guo *et al*. [21] (TPAMI'12) *†	9.300	12.60
Gong *et al*. [18] (BMVC'14) *†	8.530	8.730
DeshadowNet [46] (CVPR'17) ¶†	7.830	6.640
DSC [22] (TPAMI19) ¶†	7.100	6.210
ST-CGAN [54] (CVPR'18) *	7.470	-
ARGAN [8] (CVPR'19) ¶	6.680	-
Our method	**6.566**	**5.823**

Table 8: Shadow removal results of different methods on ISTD [54] dataset and SRD [46] dataset. We follow the evaluation metric introduced by Guo *et al*. (lower is better). Results reported by: * [54], † [22], ¶ [8].

that captured in real-world environments. We compare our methods with some sota shadow removal methods, including DSC (TPAMI19) [22], ST-CGAN (CVPR'18) [54], and ARGAN (CVPR'19) [8]. Table 8 shows the evaluation results of these methods. We do not compare ST-CGAN and ARGAN on SRD [46] because the authors did not report their accuracy on this dataset and the code has not been released yet. We follow the evaluation metric introduced by Guo *et al*. [21], where a lower score indicates a better result. Fig. 11 gives an comparison example of our method and DSC [22] on the above two datasets.

5. Conclusion

We propose a unified framework for single superimposed image separation - a group of challenging tasks in computer vision and signal processing field. Different from the previous methods that are either statistically or empirically designed, we shed light on the possibilities of the adversarial training for this task. Our method consists of a layer separator and several discriminators. We also introduce a "crossroad l_1" loss function, which minimizes the distance between the ground truth layers and the unordered outputs. Without specific tuning, our method achieves the state of the art results on multiple tasks, including image deraining, image reflection removal, and image shadow removal.

References

[1] Eli Arbel and Hagit Hel-Or. Shadow removal using intensity surfaces and texture anchor points. *IEEE transactions on pattern analysis and machine intelligence*, 33(6):1202–1216, 2010.

[2] Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Susstrunk. Single image reflection suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4498–4506, 2017.

[3] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[4] Yi Chang, Luxin Yan, and Sheng Zhong. Transformed low-rank model for line pattern noise removal. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] Andrzej Cichocki and Shun-ichi Amari. *Adaptive blind signal and image processing: learning algorithms and applications*, volume 1. John Wiley & Sons, 2002.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[8] Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. Argan: Attentive recurrent generative adversarial network for shadow detection and removal. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[9] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] Graham D Finlayson, Mark S Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85(1):35–57, 2009.

[11] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):59–68, 2005.

[12] Xueyang Fu, Jiabin Huang, Xinghao Ding, Yinghao Liao, and John Paisley. Clearing the skies: A deep network architecture for single-image rain removal. *IEEE Transactions on Image Processing*, 26(6):2944–2956, 2017.

[13] Xueyang Fu, Jiabin Huang, Delu Zeng, Yue Huang, Xinghao Ding, and John Paisley. Removing rain from single images via a deep detail network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3855–3863, 2017.

[14] Kun Gai, Zhenwei Shi, and Changshui Zhang. Blindly separating mixtures of multiple layers with spatial shifts. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[15] Kun Gai, Zhenwei Shi, and Changshui Zhang. Blind separation of superimposed images with unknown motions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1881–1888. IEEE, 2009.

[16] Kun Gai, Zhenwei Shi, and Changshui Zhang. Blind separation of superimposed moving images using image statistics. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):19–32, 2011.

[17] Yosef Gandelsman, Assaf Shocher, and Michal Irani. "double-dip": Unsupervised image decomposition via coupled deep-image-priors. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[18] Han Gong and Darren Cosker. Interactive shadow removal and ground truth for variable scene categories. In *BMVC*, 2014.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[20] Shuhang Gu, Deyu Meng, Wangmeng Zuo, and Lei Zhang. Joint convolutional analysis and synthesis sparse representation for single image layer separation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[21] Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2956–2967, 2012.

[22] X Hu, CW Fu, L Zhu, J Qin, and PA Heng. Direction-aware spatial context features for shadow detection and removal. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[23] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, and Pheng-Ann Heng. Depth-attentional features for single-image rain removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] Xiaowei Hu, Yitong Jiang, Chi-Wing Fu, and Pheng-Ann Heng. Mask-shadowgan: Learning to remove shadows from unpaired data. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[25] Xiang Huang, Gang Hua, Jack Tumblin, and Lance Williams. What characterizes a shadow boundary under the sun and sky? In *2011 International Conference on Computer Vision*, pages 898–905. IEEE, 2011.

[26] Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

[27] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[29] Li-Wei Kang, Chia-Wen Lin, and Yu-Hsiang Fu. Automatic single-image-based rain streaks removal via image decomposition. *IEEE Transactions on Image Processing*, 21(4):1742–1755, 2011.

[30] Salman H Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Automatic shadow detection and removal from a single image. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):431–446, 2015.

[31] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011.

[32] Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. Detecting ground shadows in outdoor consumer photographs. In *European conference on computer vision*, pages 322–335. Springer, 2010.

[33] Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[34] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017.

[35] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1647–1654, 2007.

[36] Anat Levin, Assaf Zomet, and Yair Weiss. Learning to perceive transparency from the statistics of natural scenes. In *Advances in Neural Information Processing Systems*, pages 1271–1278, 2003.

[37] Anat Levin, Assaf Zomet, and Yair Weiss. Separating reflections from a single image using local features. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–I. IEEE, 2004.

[38] Ruoteng Li, Loong-Fah Cheong, and Robby T. Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[39] Xia Li, Jianlong Wu, Zhouchen Lin, Hong Liu, and Hongbin Zha. Recurrent squeeze-and-excitation context aggregation net for single image deraining. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[40] Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[41] Yu Li, Robby T Tan, Xiaojie Guo, Jiangbo Lu, and Michael S Brown. Rain streak removal using layer priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2736–2744, 2016.

[42] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):171–184, 2012.

[43] Yu Luo, Yong Xu, and Hui Ji. Removing rain from a single image via discriminative sparse coding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3397–3405, 2015.

[44] Daiqian Ma, Renjie Wan, Boxin Shi, Alex C. Kot, and Ling-Yu Duan. Learning to jointly generate and separate reflections. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[45] M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1447–1454. IEEE, 2006.

[46] Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[47] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[48] Dongwei Ren, Wangmeng Zuo, Qinghua Hu, Pengfei Zhu, and Deyu Meng. Progressive image deraining networks: A better and simpler baseline. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[50] Li Shen, Teck Wee Chua, and Karianto Leman. Shadow optimization from structured deep edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2067–2074, 2015.

[51] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3193–3201, 2015.

[52] Yago Vicente, F Tomas, Minh Hoai, and Dimitris Samaras. Leave-one-out kernel optimization for shadow detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3388–3396, 2015.

[53] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3922–3930, 2017.

[54] Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[55] Tianyu Wang, Xin Yang, Ke Xu, Shaozhe Chen, Qiang Zhang, and Rynson W.H. Lau. Spatial attentive single-image deraining with a high quality real rain dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[56] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[57] Yinglong Wang, Shuaicheng Liu, Chen Chen, and Bing Zeng. A hierarchical approach for rain or snow removing in a single color image. *IEEE Transactions on Image Processing*, 26(8):3936–3950, 2017.

[58] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simon-

celli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[59] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[60] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[61] Qingxiong Yang, Kar-Han Tan, and Narendra Ahuja. Shadow removal using bilateral filtering. *IEEE Transactions on Image processing*, 21(10):4361–4368, 2012.

[62] Wenhan Yang, Robby T. Tan, Jiashi Feng, Jiaying Liu, Zongming Guo, and Shuicheng Yan. Deep joint rain detection and removal from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[63] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

[64] He Zhang and Vishal M. Patel. Density-aware single image de-raining using a multi-stream dense network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[65] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *arXiv preprint arXiv:1701.05957*, 2017.

[66] Ling Zhang, Qing Zhang, and Chunxia Xiao. Shadow remover: Image shadow removal based on illumination recovering optimization. *IEEE Transactions on Image Processing*, 24(11):4623–4636, 2015.

[67] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[68] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson W.H. Lau. Distraction-aware shadow detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[69] Jiejie Zhu, Kegan GG Samuel, Syed Z Masood, and Marshall F Tappen. Learning to recognize shadows in monochromatic natural images. In *2010 IEEE Computer Society conference on computer vision and pattern recognition*, pages 223–230. IEEE, 2010.

[70] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[71] Lei Zhu, Chi-Wing Fu, Dani Lischinski, and Pheng-Ann Heng. Joint bi-layer optimization for single-image rain streak removal. In *Proceedings of the IEEE international conference on computer vision*, pages 2526–2534, 2017.