Single-Image Specular Highlight Removal via Real-World Dataset Construction

Zhongqi Wu, Chuanqing Zhuang, Jian Shi, Jianwei Guo, Jun Xiao, Xiaopeng Zhang, Dong-Ming Yan

***Abstract*—Specular reflections pose great challenges on various multimedia and computer vision tasks, *e.g.*, image segmentation, detection and matching. In this paper, we build a large-scale Paired Specular-Diffuse (PSD) image dataset, where the images are carefully captured by using real-world objects and the ground-truth specular-free diffuse images are provided. To the best of our knowledge, this is the first real-world benchmark dataset for specular highlight removal task, which is useful for evaluating and encouraging new deep learning-based approaches. Given this dataset, we present a novel Generative Adversarial Network (GAN) for specular highlight removal from a single image by introducing the detection of specular reflection information as a guidance. Our network also makes full use of the attention mechanism and is able to directly model the mapping relation between the diffuse area and the specular highlight area without any explicit estimation of the illumination. Experimental results demonstrate that the proposed network is more effective to remove specular reflection components with the guidance of specular highlight detection than recent state-of-the-art methods.**

***Index Terms*—Specular highlight removal, PSD-Dataset, Deep learning.**

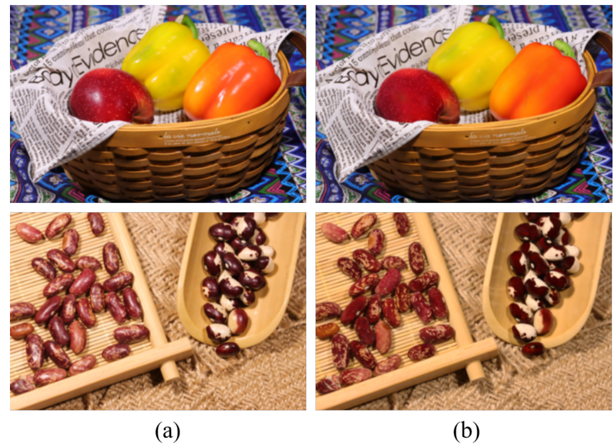


Fig. 1. Specular highlight removal results on real-world images. (a) Input specular highlight images, (b) removal results by our proposed method.

I. Introduction

SPECULAR highlight, as the reflection of the light source on shiny surfaces when illuminated, often creates undesired discontinuities in the object diffuse part and reduces the image contrast in a local window. Therefore, removing specular highlight in color images plays an important role to facilitate many multimedia and computer vision tasks, such as image segmentation [3], [52], [37], intrinsic image decomposition [48], [7], object detection [20], illumination estimation [65], [19], [12] and text detection [59], [40], [11], etc.

An effective specular highlight removal technique is widely required. Traditional model-based methods can be classified into two categories: multiple-image and single-image approaches. A common strategy for multiple-image highlight removal is to use the viewpoint dependence to find matching specular and diffuse pixels from several images [25], [33], [34]. While achieving good results, these methods are time consuming. Single-image specular highlight removal is more challenging, where some prior knowledge (*e.g.*, the Dichromatic Reflection Model [44]) is often used. However, their performances highly depend on the quality of the estimated geometry, illumination, reflectance and material properties. Essentially, existing traditional algorithms cannot semantically disambiguate highlights and pure white from complex real-world scenarios.

Recent deep-learning-based approaches [18], [32], [36] rely heavily on training data to learn a robust model. However, so far there is no public real-world dataset aiming for learning based specularity removal (an artificially rendered synthetic dataset is presented in [32]). In particular, when the training data is insufficient, the color distortion, the highlight residual or other problems are often present in the final results. The lack of a large real-world dataset also hinders the development of new specular highlight removal techniques. In this work, we built a large-scale dataset for the first time to promote deep specular highlight removal for real-world images. Our dataset includes 13,380 images, which are captured on a wide variety of scenes and materials, each with corresponding ground-truth diffuse images. It contains many daily shiny materials, such as plastic, organic material, leather, and wood, on which specular highlight often appears. Based on our dataset, we have also proposed a novel deep-learning-based specular highlight removal method. Different from existing approaches, we consider the highlight detection task and present an iterative two-branch network, which can detect and remove the specular highlight from single image at the same time. To sum up, the contributions of this work are presented include:

Z. Wu, J. Shi, J. Guo, X. Zhang, D.-M. Yan are with NLPR, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China. (Zhongqi Wu and Chuanqing Zhuang are joint first authors. Jianwei Guo and Jun Xiao are the corresponding authors. E-mail: jianwei.guo@nlpr.ia.ac.cn, xiaojun@ucas.ac.cn)

C. Zhuang and J. Xiao are with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.

- We present a first large-scale Specular-Diffuse benchmark dataset with real-world images for specular highlight removal task, in which each specular highlight image is paired with the ground-truth specular-free diffuse image.
- We propose a new two-branch neural network that captures the channel-wise global context information by using a distribution-based channel attention module. We also introduce the mask of the detected highlight area as guidance to achieve state-of-the-art performance.

II. Related work

A. Model-based methods

Multiple-image approaches. A number of studies attempted to separate reflection components by using multiple input images. Based on the dichromatic reflection model [44], Sato *et al*. [43] separated the specular and diffuse reflection components at each pixel from a sequence of color images. They captured multiple images under a moving light source. Lin *et al*. [34] presented a method based on the neutral interface reflection model for separating two reflection components with two photometric images. Lee *et al*. [30] presented a model for specular highlight region detection. They used multiple color images from different viewing directions. Lin *et al*. [33] presented a method based on color analysis that simultaneously estimates the separation of specular reflections. They speckled out pixels in the specular area as outliers and matched the remaining diffuse portions on other views. Guo *et al*. [22] proposed a method to separate these two layers from multiple images, which exploits the correlation of the transmitted layer across multiple images. Although multiple-image approaches can achieve better results, this method is less practical, because it requires multiple images and increases algorithm complexity.

Single-image approaches. Additional priors are required to solve the single-image specular highlight removal problem in traditional color segmentation methods [44], [29], [6]. Shen *et al*. [46] separated the specular highlight reflections in a color image based on the error analysis of chromaticity and the appropriate selection of body color for each pixel. Shen and Cai [45] further extended this work to improve the robustness of the algorithm. For natural images, specular highlight can be effectively removed based on the dichromatic reflection model [47], [61], [62]. Yang *et al*. [61], [62] proposed to separate diffuse and specular reflection components in the HSI color space, which is suitable for real-time applications. Shen and Zheng [47] considered color space to analyse the distribution of the diffuse and specular components and used this information for separation. Tan *et al*. [39] presented an interactive method by introducing specular highlight removal as an inpainting process. Akashi and Okatani [1] presented a modified version of sparse *non-negative matrix factorization* (NMF) without spatial prior.

Besides, the illumination estimation methods can coarsely remove highlights [9], [14], [42]. There are two approaches for estimating illumination color, one is to analyze the surface color based on the color constant of the a prior model [15], [23], [27] and the other approach is to estimate illumination color from specular reflections [24], [50]. Tan *et al*. [49], [51] separated specular illumination using the concept of inverse intensity space. Xia *et al*. [58] formulated specular highlight removal problem as an energy minimization, which can simultaneously estimate diffuse and specular highlight images. However, these methods tend to be vulnerable to complex chromatic aberrations. Several recent methods attempt to utilize intrinsic image decomposition [48], [7], [2] to handle specular highlight removal. Although existing specular highlight methods have achieved remarkable progress, they fail to produce satisfactory results for real-world images with complex ambient light and different scene content. For more details, please refer to the recent survey [4].

B. Deep-learning-based methods

Recently, there is an emerging interest in applying deep learning for single image specular highlight removal such that the handcrafted priors can be replaced by data-driven learning [18], [32], [36], [57]. Funke *et al*. [18] presented a GAN-base method for automatic specular highlight removal from a single endoscopic image. To train this network, small image patches with specular highlights and patches without highlights are extracted from endoscopic videos. Lin *et al*. [32] presented a novel learning approach, in the form of a fully convolutional neural network (CNN), which automatically and consistently removes specular highlights from a single image by generating its diffuse component. They also rendered a synthetic dataset to help the network generalize well. Later, Muhammad *et al*. [36] presented the Spec-Net and Spec-CGAN for removing high intensity specularity from low chromaticity facial images. Wu *et al*. [57] presented a new data-driven approach for automatic specular highlight removal from a single image. However, these methods rely heavily on the training data to learn a robust model. Due to the lack of a general real dataset, the performance of these methods on real images is far beyond satisfactory. So a challenging problem which arises in this domain is to build a large scale real-world dataset.

Specular highlight removal is also highly related to reflection removal and intrinsic image decomposition. Wan *et al*. [54], [55] presented a novel deep learning based framework to effectively remove reflection using the first captured single-image reflection removal dataset [53]. Zhang *et al*. [66] created a dataset of real-world images with reflection and corresponding ground-truth transmission layers. Then, they proposed to use a deep neural network with perceptual losses for single image reflection separation.

C. Dataset for specularity removal and detection

Shi *et al*. [48] developed a new rendering-based object-centric intrinsics dataset with specular reflection based on ShapeNet [10]. They picked 31,072 models from several common categories: car, chair, bus, sofa, airplane, etc. Yi *et al*. [63] constructed a multi-view dataset, which consists of 228 products with 10–520 photos for each product. In total, the dataset consists of 9,472 images. Beigpour *et al*. [7] created a real dataset with precise ground-truth for intrinsic image research. However, the number of specular-diffuse images are too small to support network training. Lin *et al*. [32] built 20,000 rendering training dataset using Blender and the Cycles engine. In the process of making the dataset, they considered the influence of colored lights, objects texture, white objects and environment maps. However, they did not make their synthetic data public accessible. Fu *et al*. [16] presented a large-scale dataset for specular highlight detection of real-world images. This dataset includes 4,310 images featuring a wide variety of scenes and materials, each with a labeled ground truth highlight mask. However, this dataset does not have a corresponding diffuse image, so it cannot be used for specular highlight removal.

Atkinson *et al*. [5] studied the underlying physics of polarization by reflection, based on the Fresnel equations. Then [31], [35], [67] mentioned the basic principles related to polarized light, and used polarized and unpolarized images to solve the problems of reflection removal or depth estimation. Nayar *et al*. [38] proposed to separate reflection components from color images by placing a polarization filter in front of the imaging sensor. Inspired by these works, we also used the polarizer to capture the diffuse images according to the Fresnel equations.

III. Dataset

In this section, we first introduce the mechanism of obtaining the specular-free images, then describe the setup and capturing details of our dataset.

A. Theoretical background

As the Fresnel reflection [8] indicates, when the incident light is linearly polarized along any direction, the reflected light and refracted light are still linearly polarized light (we provide the detailed formula derivation in the supplemental materials). In the real world, the common lighting source is natural light which is a unpolarized light. Fortunately, under controlled experimental conditions, we can convert the lighting source into a polarized light by adding a line polarizer in front of the source. In such case, specular reflection at a smooth surface interface is still linearly polarized. In contrast, the diffuse reflection tends to be more or less unpolarized due to the random nature of the diffuse scattering process [38].

Next, we describe how to remove the linearly polarized specular reflection. From the Malus's law[1], we know that when a perfect polarizer is placed in a polarized beam of light, the irradiance, I_ϕ of the light that passes through is given by

$$I_\phi = I_0 cos^2 \phi, \tag{1}$$

where I_0 is the initial intensity and ϕ is the angle between the polarization direction of reflected light and the axis of the polarizer. Therefore, by using a polarizer in front of the camera with a special angle (*i.e*., $\pi/2$ rotation) with respect to the polarization direction of the linearly polarized specular reflection, the specular reflection component is entirely blocked out to produce an image with just the diffuse reflection component.

[1]https://en.wikipedia.org/wiki/Polarizer

TABLE I
Detailed statistics of our PSD dataset.

Statistics		Object' classification	
Image resolution	6960*4640	Fruits and vegetables	163
Total dataset	13380	Toys	71
Training set	9481	Packages	405
Testing set	2526	Flowers	44
Validation set	1373	Office supplies	48
Total scenes	2210	Daily items	56

Similarly, this effect can be achieved by fixing the Circular Polarizing Filter (CPL) in front of the camera and rotating the polarizing film in front of the light source, so that the angle between the specular reflection and the polarizer in front of the lens is $\pi/2$.

B. Paired Specular-Diffuse Image Dataset

Although some synthetic datasets are presented [48], [32], creating a large scale real-world dataset is still missing for the task of specularity removal. To build such a dataset, we establish a studio with controlled lighting for photography. We place three light sources that can adjust the color temperature and brightness. A rotatable device with a linear polarizer is fabricated and placed in front of each light source. In addition, a rotating circular polarizer (CPL) is placed in front of a Canon EOS 90D camera. By rotating the two polarizers separately, when the angle ϕ between the polarization direction of reflected light and the axis of the CPL is π/2, the specular-free diffuse image of the object can be obtained.

We collect a total of 13,380 images captured on 2,210 different scenes. We use different objects and backgrounds to build our dataset. Table I reports the detailed statistics of the dataset. There is no overlap of objects and tablecloths between the training and testing set. We also carefully control the experimental conditions such as the number of lights, lighting intensity, and object size in each scene to ensure they have similar distributions across the training and the testing set. We separate the PSD dataset into training set 9,481 images, testing set 2,526 images, and validation set 1,373 images. A detailed statistical analysis on our proposed dataset is summarized below:

Diversity of objects. Our dataset contains a rich variety of objects, including 163 fruits and vegetables, 71 toys, 105 packages, 44 flowers, 48 office supplies, and 56 daily items. Fig. 2 shows some examples of objects used in our dataset.

Diversity of ambient light in the scene. The change of ambient light source will have an effect on the surface highlight of the object. To simulate the real environments, we randomly adjust the number, color temperature, brightness, and positions of the lights.

Diversity of highlights. The number and morphological distribution of highlights pose challenges to the specularity removal algorithms. Our dataset contains a variety of scenes with different degrees of difficulties. As shown in Fig. 2, the number of objects in a simple scene is small, while a complex scene

Fig. 2. Examples of specular highlight images in our dataset.

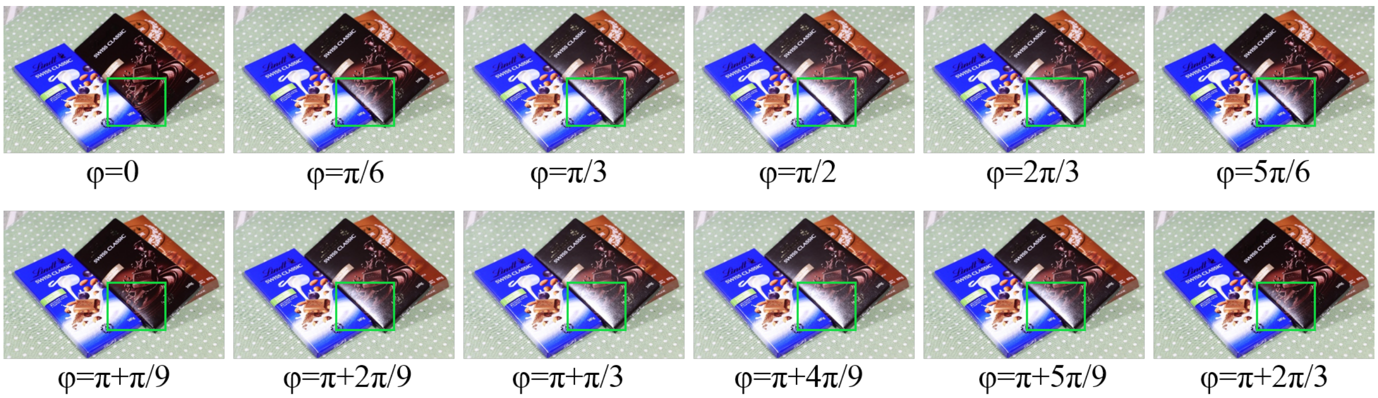


Fig. 3. A group of specular highlight images with 12 fixed polarization angles.

contains more objects, more light sources and different shape forms of the highlights.

Diversity of polarization angles. The dataset can be divided into two different polarization conditions. One consists of 1010 groups of images photographed with fixed polarization angles, and the other one consists of 1200 pairs of images photographed with random polarization angles. For the former, each group contains 12 images photographed with 12 fixed polarization angles. As shown in Fig. 3, the first row is the polarization angle for the interval [0, π], and the second row is the polarization angle for the interval [π, 2π]. For the latter, each pair of images is photographed with a random polarization angles ranging from [0, 2π].

There will be image misalignment in the captured image pairs. We deal with this problem in three aspects. First, during the image capturing, we only rotate the polarizer in front of the light source and keep the camera still. Then we filter out image pairs with visible offset. Finally, in our neural network, we use the perceptual loss, i.e., VGG loss, which has a certain degree of translation invariance in the feature extraction process.

IV. Specularity Removal Network

In this section, we propose an end-to-end neural network structure to remove the specular highlights in the given single image. Our network follows the GAN framework with a generator G and a discriminator D (see Fig. 4). One of the key contributions is a novel distribution-based channel attention method, and the other is that we introduce the highlights detection result as a guidance for highlights removal. Specifically, given an image I with specular highlight, the generator G produces an output image I' without highlight as well as a probability map P of the detected highlight area, then the discriminator D determines whether I' is a real specular-free diffuse image.

A. The Generator

We follow the encoder-decoder framework to build our generator, which is an iterative two-branch network. As shown in Fig. 4, given an input image I, the generator extracts feature maps with an encoder-decoder framework. Then the coarse detection block D_c estimates the coarse highlight probability map P_c, and the coarse removal block R_c produces the coarse diffuse image I'_c with the guidance of P_c. In some cases, the highlight locates in large area with strong intensity and can not be fully removed in a single step, so we use the refined detection block D_r and the refined removal block R_r to iteratively refine the coarse results.

More specifically, in order to remove the specular highlight in larger area, we use the dilation layer to expand the receptive field. It consists of dilated convolutions and residual connections. Since we detect the specular highlights as the guidance to remove that, we introduce the gated convolution [64] into our network with the guidance of highlight mask. Besides, we

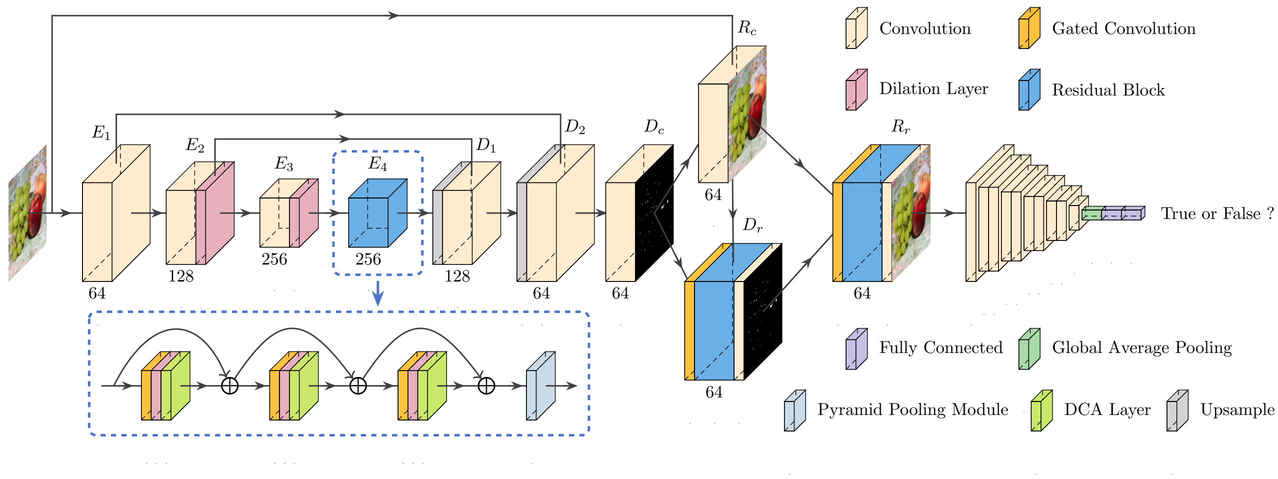


Fig. 4. **The architecture of the proposed network.** Every convolution layer is followed by an ELU activation function, except the output layer in highlight detection blocks where we use sigmoid activation function instead. Each convolution block has three convolution layers, and the DCA layer means our distribution-based channel attention method. The number of output channels of D_c and D_r is 1 and the number of output channels of R_c and R_r is 3.

also improve the existing channel attention module to better utilize the global context information.

Gated convolution. The process of specular highlight removal is similar to image inpainting, since the original color of the object is completely covered in areas of strong specular highlight. Yu *et al.* [64] explain why vanilla convolution leads to visual artifacts, and proposes an element-wise attention-based method, called *gated convolution*. Given the input feature maps F_i, the output feature maps F_o can be expressed as

$$F_o = \sigma(W_g \otimes F_i) \odot \varphi(W_f \otimes F_i) \tag{2}$$

where σ is a sigmoid function and φ can be any activation function. The symbols $\otimes$ and $\odot$ are convolution operator and element-wise product operator, while W_g and W_f are convolution kernels. Different from vanilla convolution, gated convolution assigns different weights to each feature element, which avoids the pollution of features from highlight areas.

Distribution-based channel attention. We propose a distribution-based channel attention method to introduce the global contextual information across channels to feature maps in the network. Denoting the feature maps with the number of channels C as F, we calculate the mean values, standard deviations and maximum values in each channel of F to get three descriptor vectors:

$$mean = [mean_1, ..., mean_C], \tag{3}$$

$$std = [std_1, ..., std_C], \tag{4}$$

$$max = [max_1, ..., max_C]. \tag{5}$$

Then we use a MLP with a sigmoid activation function $\sigma(\cdot)$ to predict the attention scores $S \in \mathbb{R}^C$ for each channel of F, which is represented as:

$$S = \sigma(MLP(mean \oplus std \oplus max)), \tag{6}$$

where $\oplus$ indicates the concatenation operator. The new feature map is calculated as $F' = F \cdot S$. Different from the descriptor given by simply applying global average pooling to each feature map as in [56], our distribution-based descriptor vectors preserve more information about the distribution mode of feature maps, which is important to process images in different scenarios.

B. The Discriminator

Our discriminator includes ten convolution layers and two FC layers. The number of channels in first convolution layer is 64 and doubles every two convolution layers, and the stride of convolution is 1 for every odd layer and 2 for every even layer. Each convolution layer except the first one is followed by a group normalization layer and a leaky ReLU activation function, and the first FC layer is followed by a leaky ReLU activation function.

C. Training Loss

Given the input image I with ground truth diffuse image I_0 and highlight segmentation image T, the generator outputs two highlight detection probability maps P_c, P_r and two diffuse images I'_c, I'_r, while the output of the discriminator is defined as $D(\cdot)$. Our loss function consists of four parts. In the following, I' represents diffuse image, P represents probability map.

Adversarial loss. The relativistic average SGAN loss is proposed in [26] to get a more realistic visual effect, which can be expressed as:

$$\begin{aligned} L_{RaSGAN}(I', I_0) = 0.5 \cdot (&BCE(\sigma(D(I') - D(I_0)), y') \\ &+ BCE(\sigma(D(I') - D(I_0)), y)), \end{aligned} \tag{7}$$

where (y', y) are set as $(1, 0)$ for generator and $(0, 1)$ for discriminator, respectively, and BCE measures the binary cross entropy.

Pixel loss. In order to restrain color and texture distortion, we use the pixel loss introduced in [13]:

$$L_{Pixel}(I', I_0) = \alpha \cdot \|I' - I_0\|_2^2 + \beta \cdot (\|\nabla_x I' - \nabla_x I_0\|_1 + \|\nabla_y I' - \nabla_y I_0\|_1), \tag{8}$$

where we set $\alpha = 0.2$ and $\beta = 0.4$.

Feature loss. To improve the similarity between I' and I_0 and reduce the blur of I', we use the feature loss defined in [66], [56] with a VGG-19 network pretrained on ImageNet [41], which can be formulated as:

$$L_{VGG}(I', I_0) = \sum_l \lambda_l \cdot \|\phi_l(I') - \phi_l(I_0)\|_1, \tag{9}$$

where ϕ_l defines the output feature of l-th layer in VGG-19 network, and $\{\lambda_l\}$ are weighting factors. We use the layers 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', 'conv5_2' in VGG-19 network and set $\{\lambda_l\}$ as $\{1.0/2.6, 1.0/4.8, 1.0/3.7, 1.0/5.6, 1.0/0.15\}$.

Focal loss. We use the focal loss [21] to train the network to detect the specular highlight area, which performs well on tasks with foreground-background class imbalance encountered. The focal loss is defined as:

$$L_{Focal}(P_i, T_i) = \begin{cases} -\alpha(1-P_i)^\gamma \log P_i & T_i = 1 \\ -(1-\alpha)P_i^\gamma \log(1-P_i) & T_i = 0, \end{cases} \tag{10}$$

where i is the element index in P and T, and we set $\alpha = 0.25$ and $\gamma = 2$.

Finally, our loss function is defined as:

$$\begin{aligned} L = {} & \omega_1 L_{Pixel}(I'_c, I_0) + \omega_2 L_{Pixel}(I'_r, I_0) \\ & + \omega_3 L_{VGG}(I'_r, I_0) + \omega_4 L_{Focal}(P_c, T) \\ & + \omega_5 L_{Focal}(P_r, T) + \omega_6 L_{RaSGAN}(I'_r, I_0). \end{aligned} \tag{11}$$

In all our experiments, we set $\omega_1 = 1.0$, $\omega_2 = 0.5$, $\omega_3 = 0.01$, $\omega_4 = 1.0$, $\omega_5 = 1.0$ and $\omega_6 = 0.01$.

V. Experimental Results

In this section, we first demonstrate the effectiveness of the proposed framework and compare to state-of-the-art approaches with qualitative and quantitative evaluations. Then we conduct various ablation studies to verify the validity of the specularity removal network and our new dataset, respectively.

A. Training Details

Our network is implemented in PyTorch on an NVIDIA Tesla V100 graphics card. We compress the captured PSD dataset, and each input image into the network is resized to 512×768. We train the network on our training set for 50 epochs with Adam optimizer [28]. The initial learning rate is set to 10^{-4} and reduced with attenuation coefficient of 0.8 every 5 epochs until 10^{-5}. In the first 10 epochs, the generator is trained without adversarial loss. In our implementations, we also augment our dataset by randomly mirror-flipping the images and adding noise. Besides, in order to train the highlight detection parts of the network, we get the highlight area based on the gray value difference between I and I_0 with a threshold t instead of labeling every pixel manually. The threshold is calculated as $t = 0.7\max(I - I_0)$.

TABLE II
Quantitative comparison on our dataset. The best result of each measurement is marked in **bold** font.

Scenes	Methods	MSE/1e$^{-2}\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
Chocolate	Ours	***0.10***	***0.9876***	***28.4635***
	Multi-class GAN [32]	0.16	0.9728	26.3812
	Spec-CGAN[18]	0.14	0.9745	26.7707
	Shen *et al.* [45]	0.30	0.9758	23.8261
	Yamamoto *et al.* [60]	15.45	0.3599	6.9627
Balls	Ours	***0.03***	***0.9946***	***32.7073***
	Multi-class GAN [32]	0.08	0.9833	28.9527
	Spec-CGAN[18]	0.08	0.9771	27.8457
	Shen *et al.* [45]	0.13	0.9797	26.9133
	Yamamoto *et al.* [60]	5.89	0.6946	10.9937
Toys	Ours	***0.15***	***0.9849***	***26.5272***
	Multi-class GAN [32]	0.16	0.9801	26.1796
	Spec-CGAN[18]	0.31	0.9559	22.8340
	Shen *et al.* [45]	0.40	0.9731	22.4474
	Yamamoto *et al.* [60]	1.17	0.9441	17.9891
Beans	Ours	***0.07***	***0.9850***	***29.6032***
	Multi-class GAN [32]	0.23	0.9636	24.4730
	Spec-CGAN[18]	0.19	0.9467	24.3184
	Shen *et al.* [45]	0.70	0.9095	20.2650
	Yamamoto *et al.* [60]	0.85	0.9028	19.4611
Fruits	Ours	***0.10***	***0.9730***	***27.9719***
	Multi-class GAN [32]	0.29	0.9290	22.8184
	Spec-CGAN[18]	0.26	0.9104	22.9580
	Shen *et al.* [45]	1.28	0.8121	17.2048
	Yamamoto *et al.* [60]	12.14	0.5488	7.8656
All test set	Ours	***0.14***	***0.9916***	***30.4694***
	Multi-class GAN [32]	0.50	0.8550	23.5240
	Spec-CGAN[18]	0.36	0.9172	25.7082
	Shen *et al.* [45]	1.07	0.8826	20.6226
	Yamamoto *et al.* [60]	8.46	0.6264	11.8587

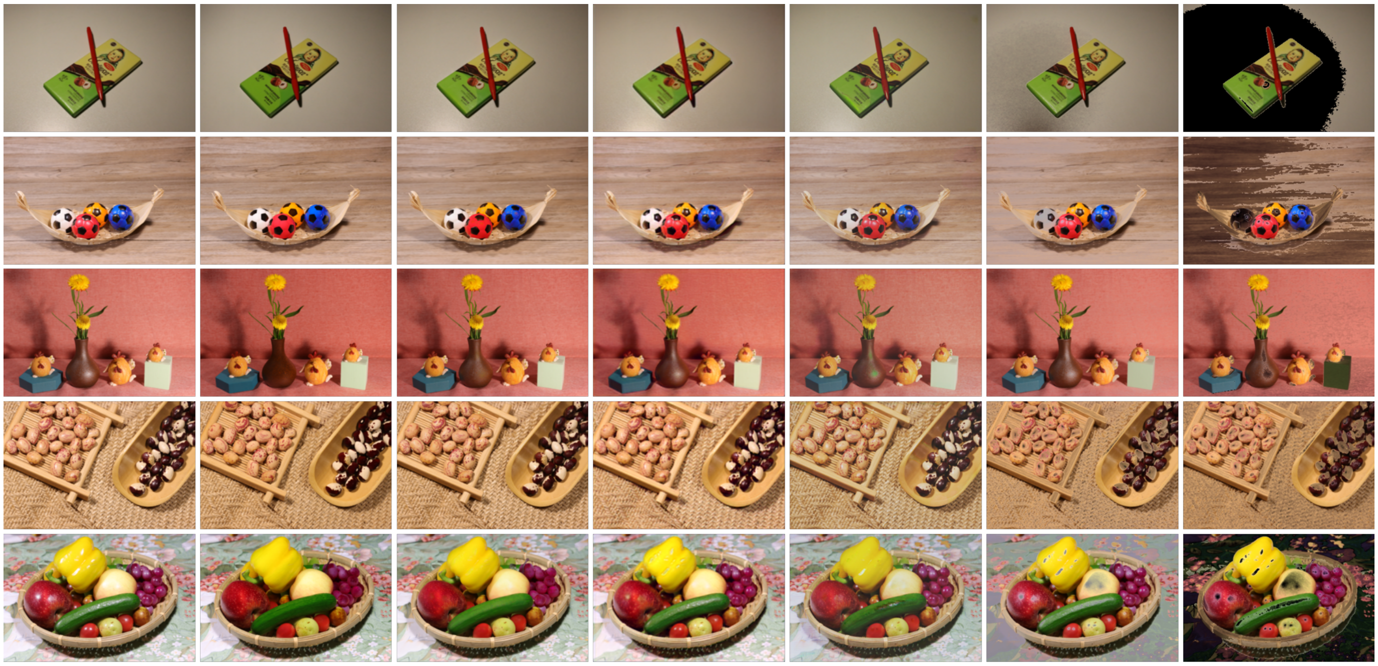
B. Comparisons

We now compare our approach against various highlight removal competitors, including five traditional approaches ([46], [45], [60], [1], [17]) and two state-of-the-art learning-based approaches (*i.e.*, Spec-CGAN[18] and Multi-class GAN [32]). For a fair comparison, we re-train Multi-class GAN and Spec-CGAN on our training dataset. Due to the page limit, we only show some selected results, and more exhaustive comparisons are provided in the supplemental materials.

Comparison on our proposed dataset. Fig. 5 visually compares specular highlight removal results on our proposed dataset, where the ground truth results are provided in Fig. 5 (b). Note that none of the objects in our testing set appear in the training set. As can be seen, previous methods induce visual artifacts including enhanced textures/structures and color distortion. In contrast, our specular-free results are more similar to the ground truths, *i.e.*, we remove the specular highlights more cleanly with more realistic texture filling in highlight area.

For quantitative comparison, we adopt three commonly used metrics including *mean-squared error* (MSE), *structural similarity index* (SSIM), and *peak signal to noise ratio* (PSNR). The numerical statistics are reported in Table II. As we can see, among all the competing methods, our network achieves the best performance, which demonstrates the ability of our approach to remove specular highlights.

Comparison on publicly available datasets. In Fig. 6, we use the real images introduced in [17] for evaluation. We



(a) (b) (c) (d) (e) (f) (g)

Fig. 5. Visual comparison on the testing set of our dataset. From top to bottom, the scenes we selected are *chocolate, balls, toys, beans, and fruits*. (a) Input, (b) ground-truth, (c) our results, (d)-(g) are the results of Multi-class GAN [32], Spec-CGAN [18], Shen *et al*. [45], Yamamoto *et al*. [60], respectively.

(a) (b) (c) (d) (e) (f) (g)

Fig. 6. Visual comparison on the testing data in [17]. (a) Input, (b) our results, (c)-(g) are the results of Multi-class GAN [32], Spec-CGAN [18], Fu *et al*. [17], Shen *et al*. [45], Yamamoto *et al*. [60], respectively.

observe that Yamamoto *et al*. [60] induces color distortion on the surface of lighting objects, resulting in more black areas (see the third row), while Fu *et al*. [17] and Shen *et al*. [45] result in local chromatic aberrations, especially on the toy face of the models (see second and third rows). Besides, Multi-class GAN [32] and Spec-CGAN [18] have obvious specular highlight residuals. In comparison, our network removed most of the highlights and produced no black shadows and chromatic aberrations.

Comparison on natural images. In order to prove the validity and fairness of our approach in natural images in the wild, we capture several pictures using mobile phones and take some web images. As shown in Fig. 7, traditional methods (Shen *et al*. [45] and Yamamoto *et al*. [60]) generate distinct black color in the white and specular highlight regions. Although deep-learning-based methods (Multi-class GAN [32] and Spec-CGAN [18]) are likely to have the ability to exclude non-specular regions, they usually fail to locate tiny highlight details (see pomegranate in third row), and there are still specular remnants in uneven areas (see the crab in fourth row). In comparison, our method can effectively detect and remove the specular highlights. This attests to a better generalization

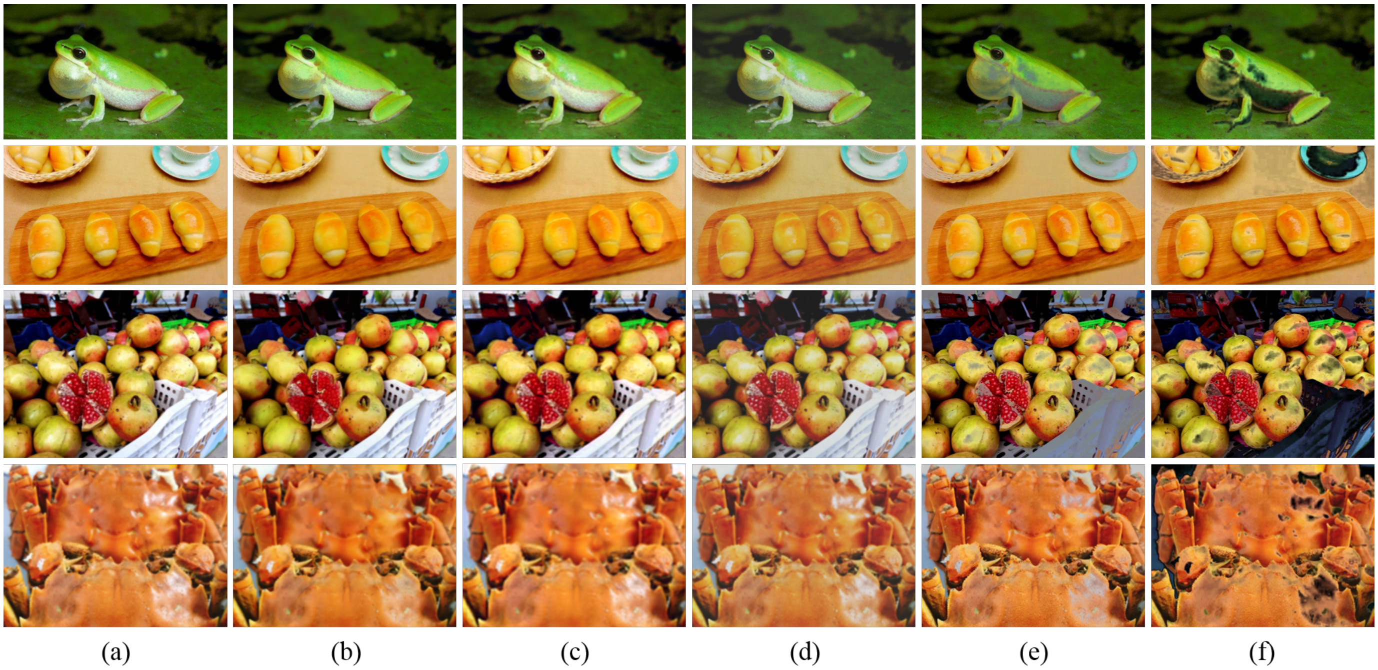


Fig. 7. Visual comparison on natural images in the wild. (a) Input, (b) our results, (c)-(f) are the results of Multi-class GAN [32], Spec-CGAN [18], Shen *et al*. [45], Yamamoto *et al*. [60], respectively.

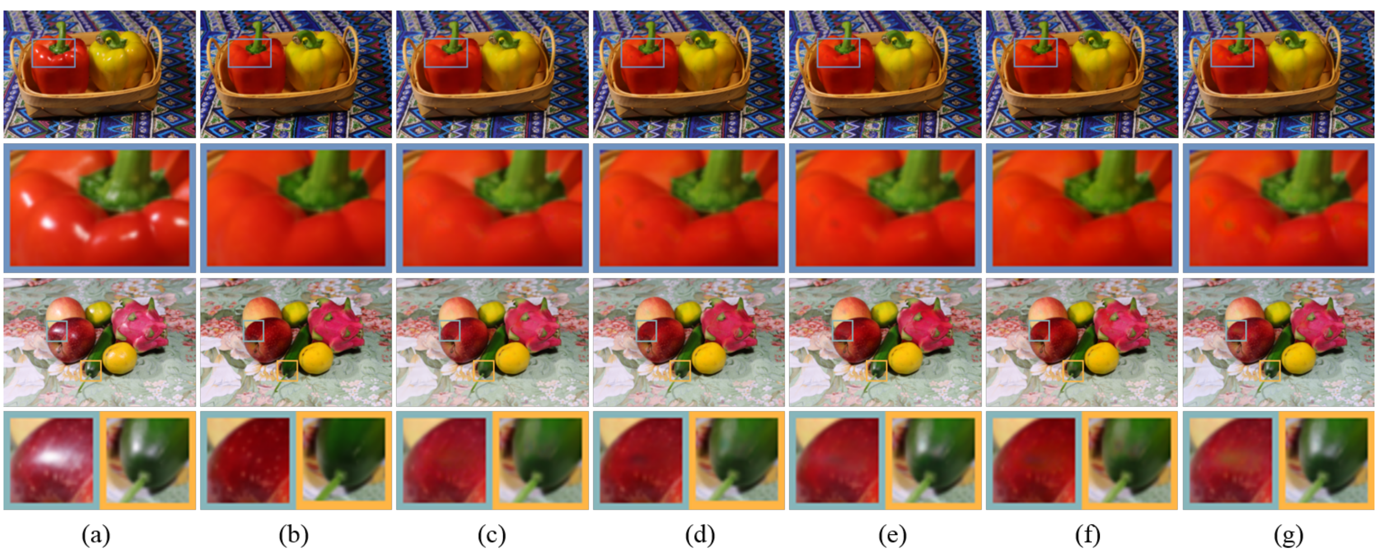


Fig. 8. Ablation studies of the proposed network. (a) Input; (b) ground-truth; (c) our results; (d) ours without highlight detection branch; (e) results of using directly iterative refinement; (f) ours without distribution-based channel attention; (g) ours without gated convolution.

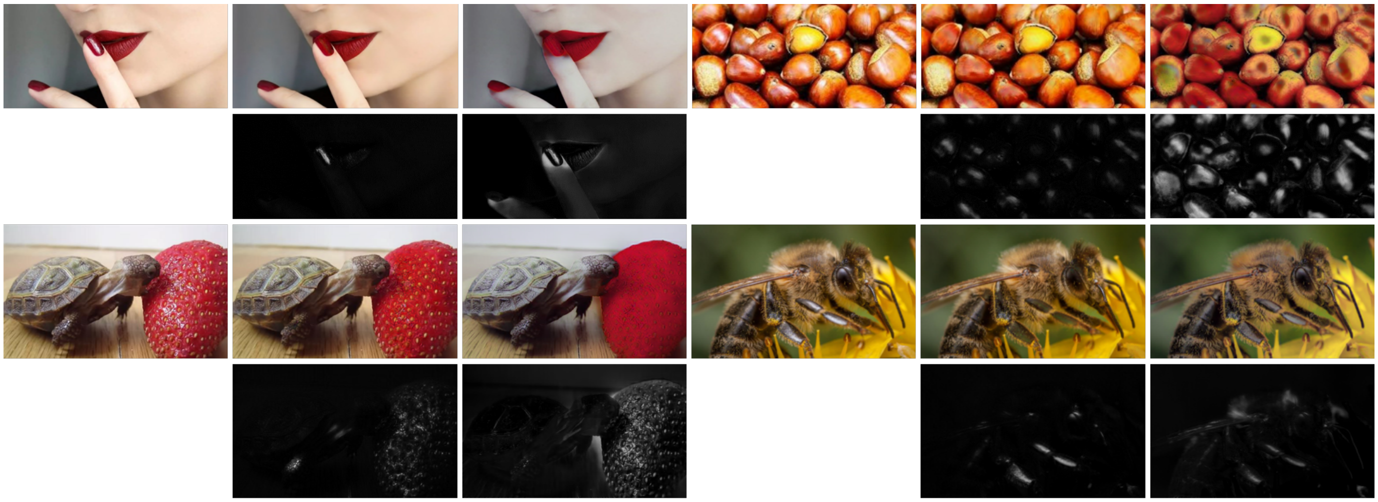
from our network.

C. Ablation Studies

Network architecture. We first conduct experiments to evaluate the influence of different components of our designed network.

- *Highlight detection branch.* We remove the highlight detection branch to validate the effectiveness of our multi-task design, which is shown in Fig. 8 (d).
- *Directly iterative refinement.* The refinement structure in our current network works with additional inputs such as I. Instead, another design is to iteratively feed the coarse result I'_c to the whole network as a new input. We implement such directly iterative refinement method and show the result in Fig. 8 (e).
- *Detail structures.* We test the effectiveness of our distribution-based channel attention method and the gated convolution, as shown in Fig. 8 (f) and (g).

Fig. 8 demonstrates that the complete implementation of our network performs better with more efficient removal of highlights and fewer color distortion than other variants. We also compute the quantitative results to reflect the contribution



Input Ours Shi *et al.* Input Ours Shi *et al.*

Fig. 9. Comparison of our PSD dataset with a synthetic dataset [48]. Separated specular reflection are shown in the second and fourth rows.

of each component of the proposed network. Table III reports the average quantitative results of this ablation study for the three local windows in Fig. 8. As observed, our full network (Fig. 8 (c)) obtains the best results.

TABLE III
COMPARISON OF DIFFERENT NETWORK SETTINGS. EACH NETWORK STRUCTURE IS USED IN THE SUB-FIGURE AS SHOWN IN FIGURE 8.

	Fig. 8 (c)	Fig. 8 (d)	Fig. 8 (e)	Fig. 8 (f)	Fig. 8 (g)
MSE	**0.0029**	0.0036	0.0064	0.0032	0.0074
SSIM	**0.8813**	0.8670	0.8227	0.8387	0.8073
PSNR	**28.6**	26.7	24.4	26.7	23.9

Weights of loss functions. Our loss function consists of textural loss, perceptual loss, adversarial loss and segmentation loss. The textural loss restricts the color and texture of the generated image to be consistent with ground truth. The perceptual loss helps the network to obtain the clearer images. The adversarial loss makes network generate more realistic results, while the segmented loss is used to guide the network to identify the highlight area. The weights of these four terms are adjusted according to our experiments, and the weights of the internal components of the perceptual and textural loss are set according to existing works [13], [66], [56]. Table IV shows the quantitative results on the data used in Fig. 8. It shows that the complete loss function leads to better results. As for the trade-off parameters ω_i, larger parameters for Pixel loss can maintain the consistent texture and color compared with the input images, while the VGG loss and GAN loss with small weights make the results have better visual effects. Moreover, the Focal loss hardly affects the results of highlight removal. As shown in Fig. 10 (d) and (e), if the GAN loss or the VGG loss is too large, the highlight area is filled with dark color. This is because the GAN loss and the VGG loss are based on the high-level feature expression of the images, and they ignore the basic color and textures in the local area.

Dataset. Next, we train our network on other datasets to verify the validity of our proposed dataset. The number of specular-diffuse images in the rendering dataset of [7] is too small to support network training. Lin *et al.* [32] did not make their synthetic data public accessible. Therefore, we compare our dataset with the synthetic dataset of [48]. We train the exact same network on these two datasets, and test them using natural images, as shown in the Fig. 9. It can be seen that the deep model trained on our dataset obtains cleaner results with less color distortion and no black shadows.

(a) Input (b) Ground truth (c) Ours (d) GAN loss 1.0 (e) VGG loss 1.0

Fig. 10. Ablation study of different parameters ω_i.

TABLE IV
ABLATION STUDY OF LOSS FUNCTIONS.

	Data	MSE	SSIM	PSNR	Data	MSE	SSIM	PSNR
Full loss		**0.0010**	**0.969**	29.5		**0.0013**	**0.960**	**29.2**
Only Pixel loss		0.0011	0.953	29.5		**0.0013**	0.937	29.0
Without GAN loss	Pepper	0.0011	0.966	**29.6**	Fruits	0.0026	0.948	25.8
Without Pixel loss		**0.0013**	0.935	28.7		0.0028	0.947	25.6
Without VGG loss		0.0012	0.960	29.0		0.0026	0.952	25.9

D. Limitations

We successfully applied our method for removing specular highlights from a variety of images. However, our neural network may fail to remove large specular-highlight areas, as shown in Fig. 11, where the large areas with bright specularity are quite challenging since most pixels are over-exposed and the textural information is lost. In addition, our PSD-Dataset only contains Lambertian objects, whose diffused reflection on surface is isotropic. The limitation is that for non-Lambertian objects, e.g., metal, our method cannot completely remove the specular highlights.

Fig. 11. Test on images with various area of specularity. Top row: input specular highlight images. Bottom row: our removal results. While our method performs well on small- and middle-size specular regions, it is hard for large area with bright specularity.

VI. Conclusion and Future Work

In this work, we construct a large-scale Paired Specular-Diffuse (PSD) image dataset consisting of 13,380 real-world images. Based on this new dataset, we propose an attention-based Generative Adversarial Network for removing specular highlight. Our method outperforms existing approaches on the proposed dataset as well as many challenging real-world images. We believe that our dataset will inspire more advanced methods to tackle the tasks of specular highlight removal or detection. In future work, we will conduct in-depth research on bright specularity, large-area specularity and the specularity on metal materials and build a ourdoor scencs dataset. Moreover, we will manually mark out the specular highlight areas to improve the detection results of the specular highlights, thereby improving the performance of specular highlights removal.

Acknowledgments

We thank anonymous reviewers for their valuable comments. We also thank Prof. Xuedong Bai, Dr. Jianlin Wang, Bohan Yu and Chaojie Ma (Institute of Physics, Chinese Academy of Sciences) for the support of constructing the dataset capturing equipment and the discussion of the polarization theory formula. This work is partially funded by the National Natural Science Foundation of China (62172416, 62172415, U2003109, 62002358, 61972459), the National Key R&D Program of China (2019YFB2204104), Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDA23090304), the Key Research Program of Frontier Sciences CAS (QYZDY-SSW-SYS004), the Youth Innovation Promotion Association of the Chinese Academy of Sciences (Y201935), the Fundamental Research Funds for the Central Universities, and the Alibaba Group through Alibaba Innovative Research Program.

References

[1] Yasuhiro Akashi and Takayuki Okatani. Separation of reflection components by sparse non-negative matrix factorization. In *Asian Conference on Computer Vision*, pages 611–625. Springer, 2014. 2, 6

[2] Anna Alperovich and Bastian Goldluecke. A variational model for intrinsic light field decomposition. In *Asian Conference on Computer Vision*, pages 66–82. Springer, 2016. 2

[3] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2010. 1

[4] Alessandro Artusi, Francesco Banterle, and Dmitry Chetverikov. A survey of specularity removal methods. 30(8):2208–2230, 2011. 2

[5] Gary A Atkinson and Edwin R Hancock. Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing*, 15(6):1653–1664, 2006. 3

[6] Ruzena Bajcsy, Sang Wook Lee, and Aleš Leonardis. Detection of diffuse and specular interface reflections and inter-reflections by color image segmentation. *Int. Journal of Computer Vision*, 17(3):241–272, 1996. 2

[7] Shida Beigpour, Andreas Kolb, and Sven Kunz. A comprehensive multi-illuminant dataset for benchmarking of the intrinsic image algorithms. In *IEEE International Conference on Computer Vision (ICCV)*, pages 172–180, 2015. 1, 2, 9

[8] Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. Elsevier, 2013. 3

[9] David H Brainard and William T Freeman. Bayesian color constancy. *JOSA A*, 14(7):1393–1411, 1997. 2

[10] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[11] Pengwen Dai, Hua Zhang, and Xiaochun Cao. Deep multi-scale context aware feature aggregation for curved scene text detection. *IEEE Trans. Multimedia*, 22(8):1969–1984, 2019. 1

[12] Bo Du, Mengfei Zhang, Lefei Zhang, Ruimin Hu, and Dacheng Tao. Pltd: Patch-based low-rank tensor decomposition for hyperspectral images. *IEEE Trans. Multimedia*, 19(1):67–79, 2016. 1

[13] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3238–3247, 2017. 6, 9

[14] Graham D. Finlayson, Steven D. Hordley, and Paul M. Hubel. Color by correlation: A simple, unifying framework for color constancy. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(11):1209–1221, 2001. 2

[15] David A Forsyth. A novel algorithm for color constancy. *Int. Journal of Computer Vision*, 5(1):5–35, 1990. 2

[16] Gang Fu, Qing Zhang, Qifeng Lin, Lei Zhu, and Chunxia Xiao. Learning to detect specular highlights from real-world images. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1873–1881, 2020. 3

[17] Gang Fu, Qing Zhang, Chengfang Song, Qifeng Lin, and Chunxia Xiao. Specular highlight removal for real-world images. 38(7):253–263, 2019. 6, 7

[18] Isabel Funke, Sebastian Bodenstedt, Carina Riediger, Jürgen Weitz, and Stefanie Speidel. Generative adversarial networks for specular highlight removal in endoscopic images. In *Medical Imaging: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 1057604, 2018. 1, 2, 6, 7, 8

[19] Yuanyuan Gao, Hai-Miao Hu, Bo Li, and Qiang Guo. Naturalness preserved nonuniform illumination estimation for image enhancement based on retinex. *IEEE Trans. Multimedia*, 20(2):335–344, 2017. 1

[20] Theo Gevers and Harro Stokman. Classifying color edges in video into shadow-geometry, highlight, or material transitions. *IEEE Trans. Multimedia*, 5(2):237–243, 2003. 1

[21] Priyal Goyal and He Kaiming. Focal loss for dense object detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 39:2999–3007, 2018. 6

[22] Xiaojie Guo, Xiaochun Cao, and Yi Ma. Robust separation of reflection from multiple images. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2187–2194, 2014. 2

[23] Thorsten Hansen, Maria Olkkonen, Sebastian Walter, and Karl R Gegenfurtner. Memory modulates color appearance. *Nature neuroscience*, 9(11):1367–1368, 2006. 2

[24] Yoshie Imai, Yu Kato, Hideki Kadoi, Takahiko Horiuchi, and Shoji Tominaga. Estimation of multiple illuminants based on specular highlight detection. In *International workshop on computational color imaging*, pages 85–98. Springer, 2011. 2

[25] Jan Jachnik, Richard A Newcombe, and Andrew J Davison. Real-time surface light-field capture for augmentation of planar specular surfaces. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 91–97. IEEE, 2012. 1

[26] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. In *International Conference on Learning Representations*, 2019. 5

[27] Hamid Reza Vaezi Joze and Mark S Drew. Exemplar-based color

constancy and multiple illumination. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 36(5):860–873, 2013. 2

[28] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 6

[29] Gudrun J Klinker, Steven A Shafer, and Takeo Kanade. The measurement of highlights in color images. *Int. Journal of Computer Vision*, 2(1):7–32, 1988. 2

[30] Sang Wook Lee and Ruzena Bajcsy. Detection of specularity using color and multiple views. In *European Conference on Computer Vision (ECCV)*, pages 99–114. Springer, 1992. 2

[31] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1750–1758, 2020. 3

[32] John Lin, Mohamed El Amine Seddik, Mohamed Tamaazousti, Youssef Tamaazousti, and Adrien Bartoli. Deep multi-class adversarial specularity removal. In *Scandinavian Conference on Image Analysis*, pages 3–15. Springer, 2019. 1, 2, 3, 6, 7, 8, 9

[33] Stephen Lin, Yuanzhen Li, Sing Bing Kang, Xin Tong, and Heung-Yeung Shum. Diffuse-specular separation and depth recovery from image sequences. In *European Conference on Computer Vision (ECCV)*, pages 210–224. Springer, 2002. 1, 2

[34] Stephen Lin and Heung-Yeung Shum. Separation of diffuse and specular reflection in color images. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages I–I. IEEE, 2001. 1, 2

[35] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolarized and polarized images. *Advances in neural information processing systems*, 32:14559–14569, 2019. 3

[36] Siraj Muhammad, Matthew N Dailey, Muhammad Farooq, Muhammad F Majeed, and Mongkol Ekpanyapong. Spec-net and spec-cgan: Deep learning models for specularity removal from faces. *Image and Vision Computing*, 93:103823, 2020. 1, 2

[37] Bernd Münzer, Klaus Schoeffmann, and Laszlo Böszörmenyi. Content-based processing and analysis of endoscopic images and videos: A survey. *Multimedia Tools and Applications*, 77(1):1323–1362, 2018. 1

[38] Shree K Nayar, Xi-Sheng Fang, and Terrance Boult. Separation of reflection components using color and polarization. *Int. Journal of Computer Vision*, 21(3):163–186, 1997. 3

[39] Long Quan, Heung-Yeung Shum, et al. Highlight removal by illumination-constrained inpainting. In *IEEE International Conference on Computer Vision (ICCV)*, pages 164–169. IEEE, 2003. 2

[40] Xiaohang Ren, Yi Zhou, Jianhua He, Kai Chen, Xiaokang Yang, and Jun Sun. A convolutional neural network-based chinese text detection algorithm via text structure modeling. *IEEE Trans. Multimedia*, 19(3):506–518, 2016. 1

[41] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *Int. Journal of Computer Vision*, 115(3):211–252, 2015. 6

[42] Guillermo Sapiro. Color and illuminant voting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(11):1210–1215, 1999. 2

[43] Yoichi Sato and Katsushi Ikeuchi. Temporal-color space analysis of reflection. *JOSA A*, 11(11):2990–3002, 1994. 2

[44] Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985. 1, 2

[45] Hui-Liang Shen and Qing-Yuan Cai. Simple and efficient method for specularity removal in an image. *Applied optics*, 48(14):2711–2719, 2009. 2, 6, 7, 8

[46] Hui-Liang Shen, Hong-Gang Zhang, Si-Jie Shao, and John H Xin. Chromaticity-based separation of reflection components in a single image. *Pattern Recognition*, 41(8):2461–2469, 2008. 2, 6

[47] Hui-Liang Shen and Zhi-Huan Zheng. Real-time highlight removal using intensity ratio. *Applied optics*, 52(19):4483–4493, 2013. 2

[48] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-lambertian object intrinsics across shapenet categories. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 1685–1694, 2017. 1, 2, 3, 9

[49] Robby T Tan and Katsushi Ikeuchi. Separating reflection components of textured surfaces using a single image. In *Digitally Archiving Cultural Objects*, pages 353–384. Springer, 2008. 2

[50] Robby T Tan, Ko Nishino, and Katsushi Ikeuchi. Color constancy through inverse-intensity chromaticity space. *JOSA A*, 21(3):321–334, 2004. 2

[51] TT Tan, Ko Nishino, and Katsushi Ikeuchi. Illumination chromaticity estimation using inverse-intensity chromaticity space. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2003. 2

[52] Michael W Tao, Jong-Chyi Su, Ting-Chun Wang, Jitendra Malik, and Ravi Ramamoorthi. Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 38(6):1155–1169, 2015. 1

[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. 2

[54] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2018. 2

[55] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Corrn: Cooperative reflection removal network. *IEEE transactions on pattern analysis and machine intelligence*, 42(12):2969–2982, 2019. 2

[56] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 8178–8187, 2019. 5, 6, 9

[57] Zhongqi Wu, Chuanqing Zhuang, Jian Shi, Jun Xiao, and Jianwei Guo. Deep specular highlight removal for single real-world image. In *SIGGRAPH Asia 2020 Posters*, pages 1–2. 2020. 2

[58] Wenyao Xia, Elvis CS Chen, Stephen E Pautler, and Terry M Peters. A global optimization method for specular highlight removal from a single image. *IEEE Access*, 7:125976–125990, 2019. 2

[59] Minglong Xue, Palaiahnakote Shivakumara, Chao Zhang, Yao Xiao, Tong Lu, Umapada Pal, Daniel Lopresti, and Zhibo Yang. Arbitrarily-oriented text detection in low light natural scene images. *IEEE Trans. Multimedia*, 2020. 1

[60] Takahisa Yamamoto, Toshihiro Kitajima, and Ryota Kawauchi. Efficient improvement method for separation of reflection components based on an energy function. In *IEEE international conference on image processing (ICIP)*, pages 4222–4226. IEEE, 2017. 6, 7, 8

[61] Jianwei Yang, Lixing Liu, and Stan Li. Separating specular and diffuse reflection components in the hsi color space. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 891–898, 2013. 2

[62] Qingxiong Yang, Jinhui Tang, and Narendra Ahuja. Efficient and robust specular highlight removal. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 37(6):1304–1311, 2014. 2

[63] Renjiao Yi, Ping Tan, and Stephen Lin. Leveraging multi-view image sets for unsupervised intrinsic image decomposition and highlight separation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12685–12692, 2020. 2

[64] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4471–4480, 2019. 4, 5

[65] Ling Zhang, Qingan Yan, Zheng Liu, Hua Zou, and Chunxia Xiao. Illumination decomposition for photograph with multiple light sources. *IEEE Trans. Image Process.*, 26(9):4114–4127, 2017. 1

[66] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 4786–4794, 2018. 2, 6, 9

[67] Dizhong Zhu and William AP Smith. Depth from a polarisation+ rgb stereo pair. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7586–7595, 2019. 3

Zhongqi Wu received her master's degree in the School of Artificial Intelligence of University of Chinese Academy of Sciences in 2019. She is currently working toward her Ph.D degeree at the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. Her research interests include image processing and computer vision.

Chuanqing Zhuang is working toward the master's degree in University of Chinese Academy of Sciences, School of Artificial Intelligence. He received his bachelor's degree of engineering from Tsinghua University in 2019. His research interests include computer vision and image processing.

Jian Shi received the Ph.D. degree from the University of Chinese Academy of Sciences in 2017. He is an Assistant Professor with the Institute of Automation, Chinese Academy of Sciences. He works on computer vision and graphics problems, including intrinsic images and illumination estimation.

Jianwei Guo is an associate professor in National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). He received his Ph.D. degree in computer science from CASIA in 2016, and bachelor degree from Shandong University in 2011. His research interests include computer vision, computer graphics and image processing,.

Jun Xiao is a professor in University of Chinese Academy of Sciences, Beijing. He obtained his Ph.D. degree in communication and information system from the Graduate University of the Chinese Academy of Sciences, Beijing, in 2008. His research interests include computer graphics, computer vision, image processing and 3D reconstruction.

Xiaopeng Zhang received the PhD degree in computer science from Institute of Software, Chinese Academic of Sciences in 1999. He is a professor in National Laboratory of Pattern Recognition at Institute of Automation, Chinese Academy of Sciences. He received the National Scientific and Technological Progress Prize (second class) in 2004 and the Chinese Award of Excellent Patents in 2012. His main research interests include image processing, computer graphics and computer vision.

Dong-Ming Yan is a professor in National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences(CAS). He received his Ph.D. degree in computer science from Hong Kong University in 2010, and his master and bachelor degrees in computer science and technology from Tsinghua University in 2005 and 2002, respectively. His research interests include image processing, geometric processing, and visualization.