A Multi-Task Network for Joint Specular Highlight Detection and Removal

Gang Fu[1], Qing Zhang[2], Lei Zhu[3], Ping Li[4], and Chunxia Xiao[1,*]
[1]School of Computer Science, Wuhan University, China
[2]School of Computer Science and Engineering, Sun Yat-sen University, China
[3]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
[4]Department of Computing, The Hong Kong Polytechnic University, China

Abstract

Specular highlight detection and removal are fundamental and challenging tasks. Although recent methods have achieved promising results on the two tasks by training on synthetic training data in a supervised manner, they are typically solely designed for highlight detection or removal, and their performance usually deteriorates significantly on real-world images. In this paper, we present a novel network that aims to detect and remove highlights from natural images. To remove the domain gap between synthetic training samples and real test images, and support the investigation of learning-based approaches, we first introduce a dataset with about 16K real images, each of which has the corresponding ground truths of highlight detection and removal. Using the presented dataset, we develop a multi-task network for joint highlight detection and removal, based on a new specular highlight image formation model. Experiments on the benchmark datasets and our new dataset show that our approach clearly outperforms state-of-the-art methods for both highlight detection and removal.

1. Introduction

Specular highlight, as a common physical phenomenon in the real world, often presents as bright spot on shiny material surfaces when illuminated. Highlight detection and removal have long been fundamental problems in computer vision. The reason is twofold. First, detecting where the highlight is allows us to infer the light direction, scene geometry [19] and camera location. Second, removing the effect from highlight can help improve the performance of many vision tasks, such as object detection [14], intrinsic image decomposition [2], and tracking [9]. Note that, for simplicity, we refer *highlight* to *specular highlight* in this paper, except stated definitely.

Early works detect highlight by treating the brightest pixels in an image as highlights [41, 18], which has low accuracy since it may mistake pixels with high intensities as highlight. As for highlight removal, traditional methods are often based on optimization [13, 20, 7], clustering [29] and filtering [37, 38] *etc*, and may fail to handle large-scale highlight removal due to lack of image semantics. More recent methods on highlight detection and removal are mostly deep learning-based. Although learning-based highlight detection and removal methods have achieved remarkable progress [28, 39, 6], they basically have two limitations. First, these methods are typically trained on synthetic data or a very small number of real data, so that may

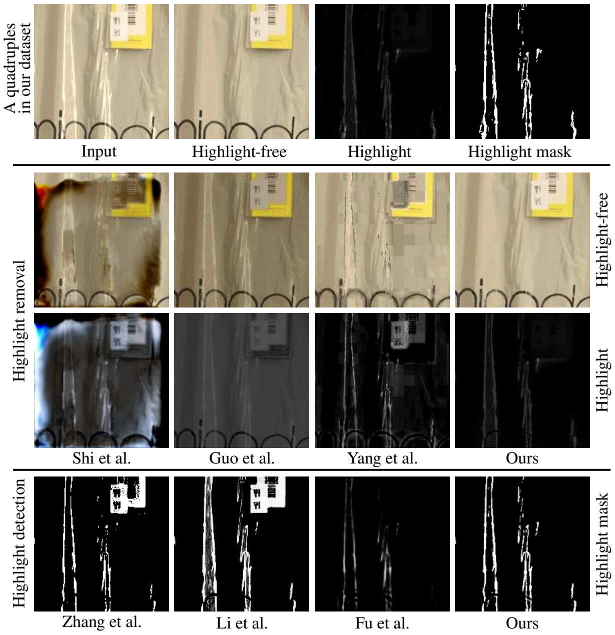


Figure 1. Visual comparison of highlight detection and removal on an example image from our dataset. We compare our method with the state-of-the-art removal methods including Shi *et al.* [28], Guo *et al.* [10], and Yang *et al.* [37], and with the state-of-the-art detection methods including Zhang *et al.* [41], Li *et al.* [18], and Fu *et al.* [6].

[*] Corresponding author.

not work well for real images due to the domain gap between training and test images. Second, they are either designed for highlight detection or removal, and are not able to be employed for joint highlight detection and removal.

To enable effective training and comprehensive evaluation for highlight detection and removal, we in this paper introduce a large-scale dataset for highlight detection and removal. It contains about 16K real images that cover a wide range of scenes, subjects, and lighting conditions. Each image in the dataset has the corresponding highlight detection, removal, and highlight intensity images. Based on the presented dataset and a new region-aware highlight image formation model, we develop a deep learning framework for joint highlight detection and removal. Particularly, a multi-task network with multiple Dilated Spatial Contextual Feature Aggregation (DSCFA) modules is designed to harvest contextual features of different scales, for accurately detecting and removing highlights of varying sizes.

To sum up, our contributions are as follows:

- We present the first large-scale highlight detection and removal dataset with about 16K real images.
- We develop a multi-task network for jointly detecting and removing highlights from natural images.
- Experimental results on benchmark datasets and our new dataset show that our network performs favorably against the previous methods for both highlight detection and removal.

2. Related work

Highlight detection. Most of existing detection methods [18, 22] are typically based on different forms of a thresholding scheme. Based on an assumption that only a small portion of a scene contains highlights, Zhang *et al.* [41] formulated highlight detection as Non-negative Matrix Factorization (NMF) [41] problem. Although these methods are efficient and easy to implement, they often incorrectly detect pixels with white appearance or high intensities as highlights. To overcome this issue, Fu *et al.* [6] recently proposed a deep learning-based network leveraging context contrast feature for detection.

Single-image highlight removal. Tan *et al.* [30] proposed a method based on chromaticity analysis without requiring any geometrical information. Yang *et al.* [37] utilized low-pass filter to propagate information from the diffuse pixels to the specular pixels. The method proposed by Kim *et al.* [13] is based on an observation that the dark channel usually provides an approximate highlight-free image. Liu *et al.* [20] presented a two-step saturation-preserving method, where it first produces an over-saturation highlight-free image, and then corrects the saturation. Akashi *et al.* [1] formulated the highlight removal problem as a sparse non-negative matrix factorization model. Guo *et al.* [10] presented a sparse and low-rank reflection model for highlight removal. Shen *et al.* [27] used the idea of intensity ratio to estimate the specular fractions of the image. Li *et al.* [16, 17] presented a method for removing specular highlight in facial images. Recently, Shi *et al.* [28] proposed a deep learning-based method to handle the non-Lambertian object-level intrinsic decomposition problem. Yi *et al.* [39] designed a unified framework for joint intrinsic image decomposition and highlight separation. However, they fail to handle images with complex illuminations and textures. In comparison, we present a unified framework for jointly learning highlight detection and removal, which can effectively remove highlights while preserving saturation.

Multi-image highlight removal. Some methods are proposed to remove highlight from multiple images. By assuming that surface geometry is known, Wei *et al.* [35] leveraged the principal component analysis to separate highlights and estimate the position of light source. Feris *et al.* [4] removed highlight by solving a Poisson equation in the gradient domain. Guo *et al.* [11] proposed RPCA, which focuses on removing specular reflection from superimposed multiple images. Despite methods in this category produce promising highlight removal results, the requirement of multiple images limits their applicability.

3. Dataset Preparation

3.1. Background

Existing works [37] often use the cross-polarization technique to capture images without highlights in a rigorous laboratory environment. Table 1 reports the number of existing datasets made by this technique. These image pairs are very few and are object-level without background for quality assurance. In fact, it is difficult to control the cross-polarization process in general settings to produce high-quality highlight-free images for everyday objects. Also, it is very difficult and impractical (taking a lot of manpower) to precisely annotate many highlight regions often with dotted and very thin shapes in a real image. In contrast, we propose a semi-automatic method to construct a large-scale real dataset with about 16K Specular Highlight Image Quadruples (SHIQ), by leveraging multi-illumination sequences to produce ground truths.

3.2. Building Dataset

We first discuss how we obtain an initial set of multi-illumination sequences, then describe the pipeline of the task performed on these images and chosen high-quality region pairs. Figure 2 shows the overall workflow of our data creation. In the following, we will describe the steps in detail.

Stage 1: Collecting multi-illumination images. We col-

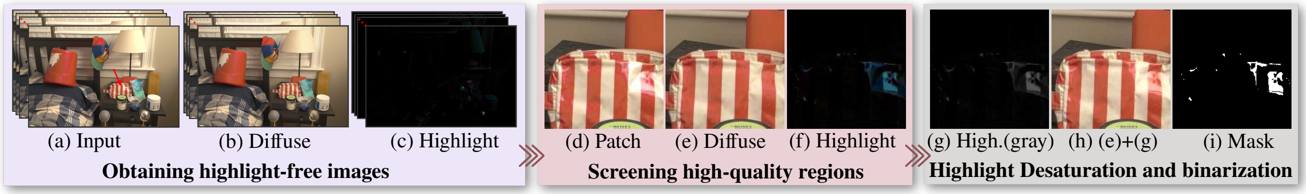

(a) Input
(b) Diffuse
(c) Highlight
Obtaining highlight-free images
(d) Patch
(e) Diffuse
(f) Highlight
Screening high-quality regions
(g) High.(gray)
(h) (e)+(g)
(i) Mask
Highlight Desaturation and binarization

Figure 2. Pipeline of our data generation. We provide a typical example in each stage. Please see Section 3.2 for more details.

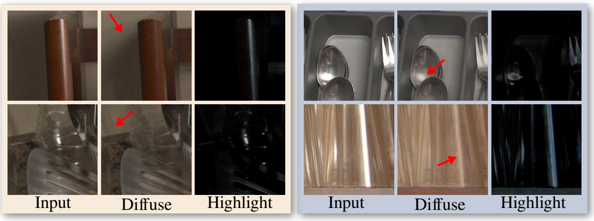


Figure 3. *Bad* example regions. **Left**: environment lighting distortion; **Right**: highlight residual. These pairs need to be removed.

lect multi-illumination image sequences from the MIW dataset [23], which consists of 1016 scenes, each photographed under 25 predetermined lighting directions for a total of 25,000 high-resolution images. The MIW dataset contains many everyday shiny materials on which characteristically appear highlights.

Stage 2: Obtaining highlight-free images. We choose the state-of-the-art RPCA method proposed by Guo *et al.* [11] to generate highlight-free images in the MIW dataset.

Stage 3: Screening high-quality regions. As RPCA may fail to produce satisfactory highlight removal results for sequences with complex illumination variations, we thus screen high-quality input/output image pairs from the results produced by RPCA. The detailed procedure is as follows.

We recruited 50 subjects to manually select the high-quality highlight removal results. To avoid distractions from other image regions, we propose to collect high-quality input/output image pairs for the highlight regions rather than the entire image. To this end, we first randomly cropped each image pair into overlapping image patches of size $k \times k$ with a step size of l, where k and l are set as 200 and 50, respectively. In total, we produced 3,197,250 image pairs. Next, each subject was shown with 63K image pairs (*i.e.*, initial input highlight and corresponding highlight-free images), and was asked to select the high-quality image pairs. Finally, we obtain about 16K image pairs.

Stage 4: Highlight desaturation and binarization. For simplicity, following previous highlight removal works [13, 37], we assume that the specular chromaticity is uniform (*i.e.*, highlight is colorless). So we simply desaturate the specular layers derived from the previous stage to remove colors in it. Then, we combine the results with the highlight-free layers to produce the final images considered as inputs in our dataset. In addition, we perform the binarization on the highlight layers to produce their masks for highlight detection task. Here, elements in the masks are binary values, where “1” indicates highlight regions and “0” indicates non-highlight regions. Finally, (d), (h), (f) and (i) in Figure 2 (*i.e.*, input and its corresponding highlight-free, highlight as well as highlight mask images) are an example quadruples in our dataset. Our dataset can be used simultaneously for highlight detection and removal tasks.

4. Highlight Image Formation

The dichromatic reflection model [26] has been commonly used in the highlight removal field. This model formulates a color image (denoted as $\mathbf{I}$) as a linear combination of diffuse component (*i.e.*, highlight-free layer, denoted as $\mathbf{D}$) and specular component (*i.e.*, highlight layer, denoted as $\mathbf{S}$):

$$\mathbf{I} = \mathbf{D} + \mathbf{S}\,. \tag{1}$$

Based on this model, highlight removal could be considered as a two-signal separation problem. This means that given the observation $\mathbf{I}$, how to separate the highlight-free layer $\mathbf{D}$ from the highlight layer $\mathbf{S}$ by relying on their unique characteristics. Existing highlight removal methods based on Equation (1) have two issues. First, the existing methods based on Equation (1) without distinguishing the highlight and non-highlight regions would suffer from the saturation ambiguity problem [13]. Namely, the less saturated surface colors are incorrectly treated as highlights to be removed, leading to color distortion in the non-highlight regions, for example, a white material surface (see Figure 1). Second, highlights in real-world scenes usually have a wide range of intensity values, and have different spatial distributions. So, nearly all traditional optimization-based methods [13, 20] leverage a smoothness prior and do not effectively model $\mathbf{S}$ to produce satisfactory results. In essence, the main reason for the above issues lies in the inherent ambiguity of modeling $\mathbf{D}$ and $\mathbf{S}$.

To address the above issues, we present a generalized

Table 1. Comparison of the proposed dataset with existing publicly available laboratory datasets. As can be seen, our dataset has more images (several orders of magnitude larger than exiting highlight datasets). Also, it has a wider range of highlight intensity. Thus, it is more challenging to remove highlights (which are more strong) in our dataset. Here, CP: cross-polarization, MI: multi-illumination, and HI: highlight intensity.

Dataset	Tan [30]	Shen [27]	Yang [37]	Yi [39]	**Ours**
Making method	CP	CP	CP	CP	MI
Number (tuples)	2	4	4	9	16K
Background	✗	✗	✗	✗	✓
Highlight mask	✗	✗	✗	✗	✓
Mean (HI)	0.020	0.012	0.018	0.023	0.138
Std (HI)	0.029	0.027	0.028	0.021	0.182

highlight image formation model, expressed as

$$\mathbf{I} = \mathbf{D} + \mathbf{M} \otimes \mathbf{S}\,, \tag{2}$$

where $\otimes$ denotes the element-wise multiplication, and $\mathbf{M}$ denotes the highlight mask to indicate the locations of individually visible highlights. The above highlight image formation model has two desirable advantages for learning-based highlight removal methods: (1) it provides additional position information for the network to learn about highlight regions; (2) it allows a new highlight removal framework to first detect highlights, and subsequently to operate differently on the highlight and non-highlight regions, benefiting for producing saturation-preserving results with natural-looking appearances.

5. Joint Highlight Detection and Removal

We design a multi-task network for Joint Specular Highlight Detection and Removal (JSHDR), based on the inverse problem in Equation (2). To accurately detect and remove highlights of varying sizes, we further propose a Dilated Spatial Contextual Feature Aggregation (DSCFA) module.

5.1. Multi-Task Network for Joint Highlight Detection and Removal

According to Equation (2), $\mathbf{D}$, $\mathbf{S}$, and $\mathbf{M}$ are inherently correlated, and thus computing $\mathbf{S}$ and $\mathbf{M}$ benefits for the estimation of $\mathbf{D}$. Motivated by this observation, we develop a multi-task convolutional neural network with DSCFA modules to jointly predicting $\mathbf{D}$, $\mathbf{S}$, and $\mathbf{M}$ in an end-to-end manner. Figure 5 shows the schematic illustration of the developed network. As an encoder-decoder framework, our network first passes an input image into a series of DSCFA modules (see Section 5.2) of an encoder and decoder framework to extract highlight features $\mathbf{F}$. Then, $\mathbf{M}$, $\mathbf{S}$ and $\mathbf{D}$ are predicted in a sequential order from $\mathbf{F}$:

(1) $\mathbf{M}$ is estimated by using a convolutional block with "Conv(3×3) → Conv(3×3) → Conv(3×3)" on $\mathbf{F}$,

(2) $\mathbf{S}$ is estimated by applying another convolutional block with three 3×3 convolutions on the concatenation of $[\mathbf{F}, \mathbf{M}]$,

(3) $\mathbf{D}$ is estimated by feeding the concatenation of $[\mathbf{F}, \mathbf{M}, \mathbf{S}, \mathbf{I}-\mathbf{MS}]$ into a convolutional block consisting of three 3×3 convolutions.

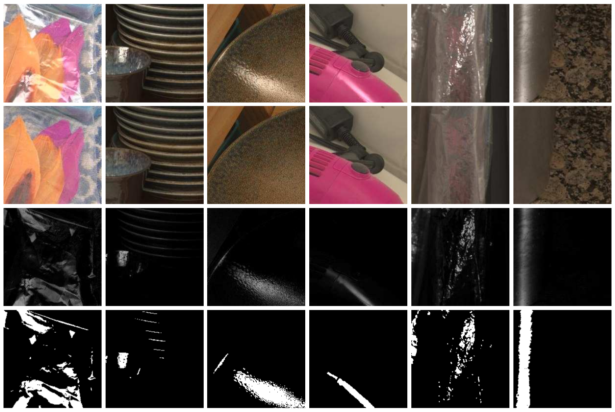


Figure 4. An illustration of several highlight (**1st row**), highlight-free (**2nd row**), highlight intensity (**3rd row**) and highlight mask (**4th row**) image quadruples in the proposed dataset.

5.2. Dilated Spatial Contextual Feature Aggregation Module

Figure 6 shows the schematic illustration of the proposed DSCFA module, which extracts and aggregates multi-scale dilated spatial contextual information simultaneously for detecting and removing highlights at varied region sizes by developing a series of DSCFA blocks.

DSCFA block. Contextual information has been demonstrated to be useful for highlight detection [6] and low-level image processing [3]. Our DSCFA block learns dilated spatial features (DSF) to extract and aggregate dilated contextual features from four directions. To achieve this, we first replace the common convolution of the spatial CNN module [24] with dilated convolutions to enlarge the receptive field for more contextual information, and then obtain four features (*i.e.*, DSCNN_L, DSCNN_R, DSCNN_D, DSCNN_U) along with four directions. Figure 7 shows a dilated spatial module, which learns a DSCNN_D, DSCNN_U, DSCNN_L, and DSCNN_R from input features. Spatial aggregation from four directions adopt slice-by-slice convolutions within feature maps from downward, upward, rightward, and leftward directions, thus enabling rich message passing between pixels across rows and columns in a layer. Then, given an input feature map, we first apply a 3×3 convolution and a ReLU Layer to produce a new feature map H. Based on H, we apply two branches to learn contextual features. The first branch is the whole procedure of Figure 7 while the second branch is to replace the order of feature learning at four directions, where DSCNN_L, DSCNN_R,

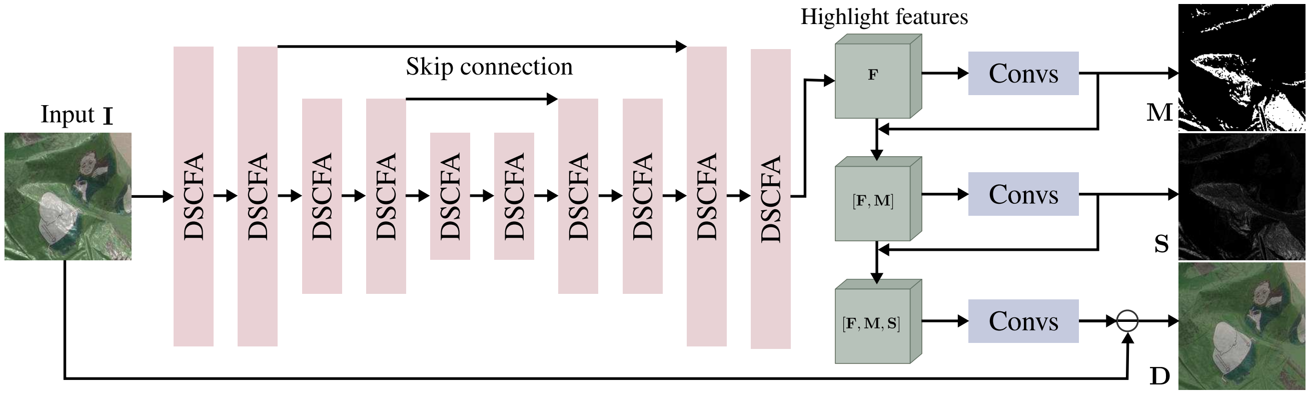


Figure 5. The pipeline of our joint highlight detection and removal network. Our network applies an encoder-decoder structure with DSCFA modules to extract highlight features $\mathbf{F}$ from the input highlight image $\mathbf{I}$. Based on $\mathbf{F}$, $\mathbf{M}$, $\mathbf{S}$ and $\mathbf{D}$ are then subsequently predicted to perform the joint highlight detection, estimation and removal.

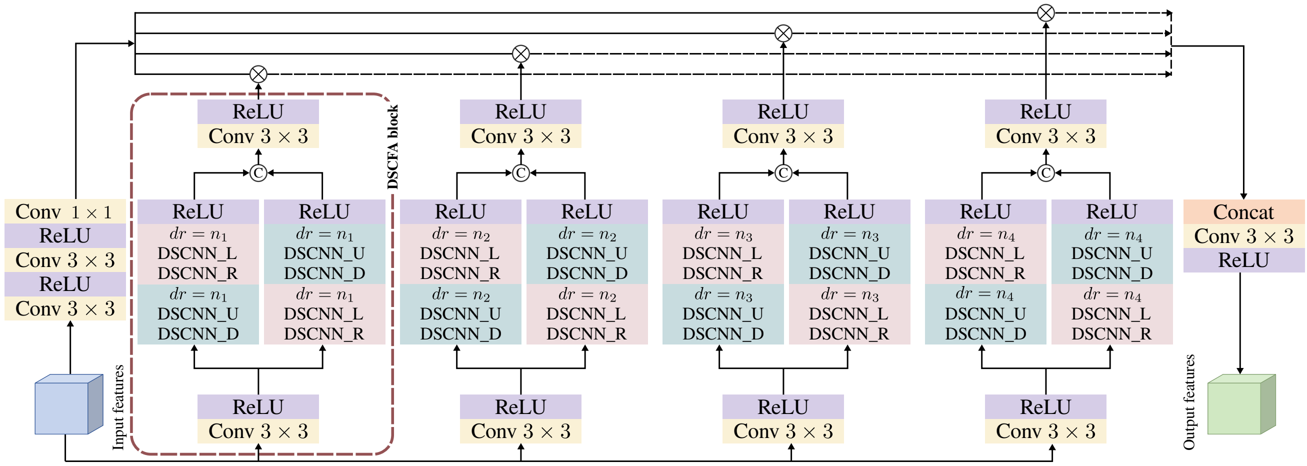


Figure 6. The schematic illustration of the proposed DSCFA module. The input features are pass through four parallel DSCFA blocks, and the output features of four DSCFA blocks are performed weighted fusion to produce multi-scale dilated spatial contextual features. In each DSCFA block (dark red dashed box), input features are fed to two parallel dilated spatial convolutions with opposite convolution orders to obtain abundant contextual infromation with different highlight characteristics.

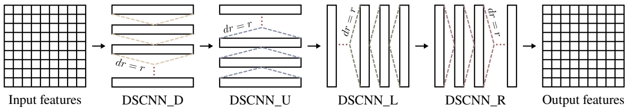
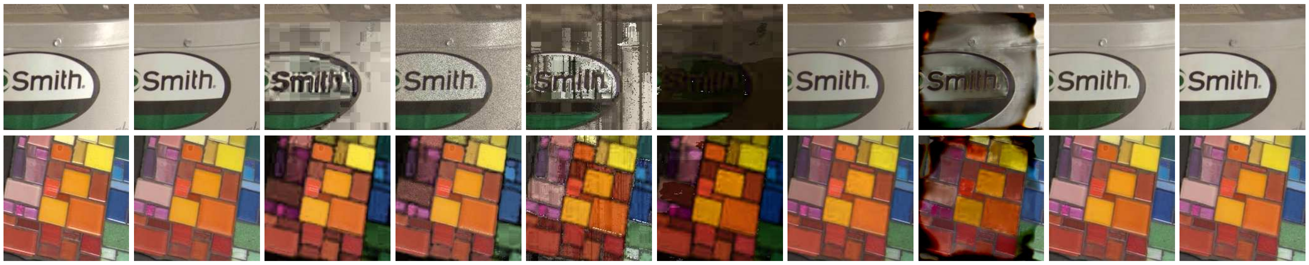


Figure 7. Illustrations of a DSCFA block. The DSCNN module with suffix ‘D’, ‘U’, ‘R’, and ‘L’ indicates DSCNN with downward, upward, rightward, and leftward directions respectively.

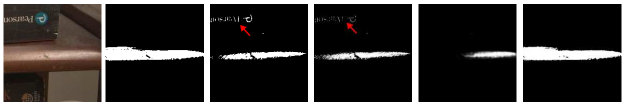
DSCNN_D, and DSCNN_U are obtained from H. After that, we concatenate features from two branches and apply a 3×3 convolution and a ReLU layer to obtain the output features of our DSCFA block; as shown in Figure 6.

DSCFA module. Highlight regions in an image often have a wide range of region sizes. Figure 4 shows an example, where highlights of the input image cover from a dotted region to a very long thin shape across a whole object. Note that the DSCFA block with a given dilation rate n tends to extract contextual information from a receptive field of a fixed size. Hence, it suffers from two issues. First, the receptive fields are maybe larger to detect small highlight regions, thereby incurring a false positive result due to much incurred noise. Second, the receptive fields are sometimes small for inferring a large highlight region. As a result, the target highlight regions are partially detected and cannot be removed completely from the input highlight image due to insufficient contextual information to eliminate it. To overcome this issue, we devise a multi-scale contextual module, namely DSCFA module, to harvest contextual information from multiple receptive fields at varied scales. In detail, as shown in Figure 6, we feed the input features into four parallel DSCFA blocks to obtain four DS features and then use a convolutional block with “Conv(3×3) → ReLU → Conv(3×3) → ReLU → Conv(1×1)” to learn an attention map with four channels to weight features from four



(a) Input (b) G.T. (c) Yang [37] (d) Akashi [1] (e) Yama. [36] (f) Tan [30] (g) Guo [10] (h) Shi [28] (i) Yi [39] (j) Ours

Figure 8. Visual comparison of our method against state-of-the-art highlight removal methods on the testing set of the proposed dataset.



(a) Input (b) G.T. (c) NMF [41] (d) ATA [18] (e) SHDN [6] (f) Ours

Figure 9. Visual comparison of our method against recent state-of-the-art highlight detection methods on our dataset.

DSCFA blocks to obtain an output feature (F_{out}) of our DSCFA module:

$$F_{out} = conv(cat(W_1 \times DSF_1, W_2 \times DSF_2, W_3 \times DSF_3, W_4 \times DSF_4)) \ , \tag{3}$$

where W_1, W_2, W_3, and W_4 denote the four channels of the learned attention map. DSF_1, DSF_2, DSF_3, and DSF_4 represent the obtained features from four DSCFA blocks of our DSCFA module. $conv$ denotes a 3×3 convolution and a ReLU layer, while cat is a feature concatenation operation. In the DSCFA module, the dilation rates in four parallel DSCFA blocks are typically set to 1, 2, 3, and 4, respectively.

5.3. Network Training

The total training loss $\mathcal{L}$ of our network consists of three prediction losses of highlight detection, highlight intensity estimation, and highlight removal. The definition of $\mathcal{L}$ is given by:

$$\mathcal{L} = \lambda_s \Phi_{L2}(\mathbf{S}, \tilde{\mathbf{S}}) + \lambda_d \Phi_{L2}(\mathbf{D}, \tilde{\mathbf{D}}) + \lambda_m \Phi_{BCE}(\mathbf{M}, \tilde{\mathbf{M}}) \ , \tag{4}$$

where λ_s, λ_d and λ_m are the weighting parameters, we empirically set them as $\lambda_s = \lambda_d = \lambda_m = 1$ for all experiments. $\mathbf{M}$ and $\tilde{\mathbf{M}}$ are the predicted highlight detection map and the associate ground truth. $\mathbf{S}$ and $\tilde{\mathbf{S}}$ are the predicted highlight intensity map and the associate ground truth. $\mathbf{D}$ and $\tilde{\mathbf{D}}$ are the predicted highlight removal result and the associate ground truth.

5.4. Implementation details

We randomly split our dataset into 12K quadruples for training and 4K quadruples testing. We have implemented JSHDR in PyTorch on a PC equipped with NVIDIA GeForce GTX 2080Ti. The Adam optimizer [15] is used to train our network. We use 100 epochs to train our network with batch size of 8, and the whole training process requires about 3 days. The initial learning rate is 2×10^{-5} and we then divide the learning rate by 10 after 20 epochs. We also use Huawei MindSpore platform to partly verify our JHSDR. What's more, we adopt the highlight attenuation and boosting editing [25] as a data augmentation to produce more training images with diverse highlights, enabling our network to better address weak and strong highlights.

6. Experiments

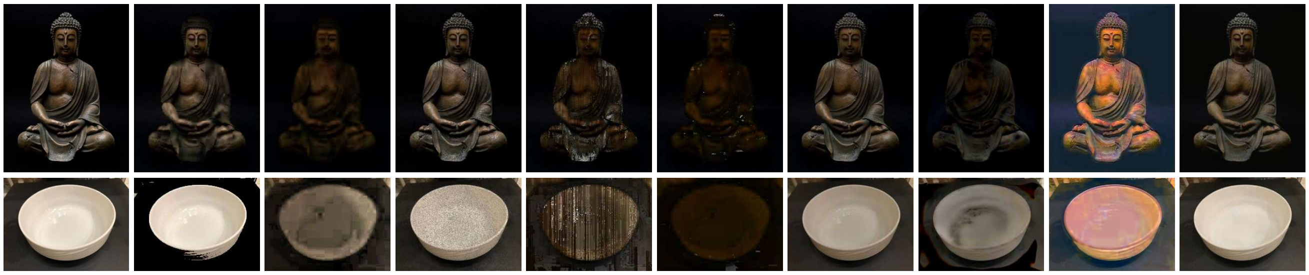
6.1. Datasets

Since our dataset SHIQ has the associate ground truths of the highlight detection and highlight removal, we use SHIQ to evaluate highlight detection and removal, respectively. Apart from our SHIQ, we also include a common SRW [6] to evaluate different highlight detection methods (available at https://github.com/fu123456/SHDNet). Regarding highlight removal, we first create a dataset (namely CLH) by collecting testing images from existing highlight removal works [30, 27, 37, 39], and introduce an existing synthesized LIME dataset [21]. Overall, we use SRW and our SHIQ as two datasets for highlight detection evaluation, while CLH, LIME, and our SHIQ are considered as three datasets for highlight removal evaluation.

6.2. Compared methods and Metrics

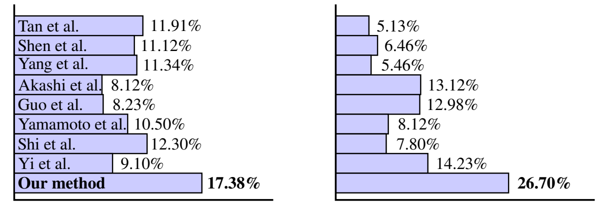
Highlight detection. We compare our method with the state-of-the-art highlight detection methods including two traditional methods of NMF [41] and ATA [18], and a deep learning-based method of SHDN [6]. To evaluate the highlight detection performance quantitatively, we follow two commonly used metric terms [12] including the accuracy and the balance error rate (BER). Higher value of accuracy and lower value of BER indicate better detection results.

Highlight removal. We compare our method against traditional methods [27, 37, 30, 1, 28, 10, 36] and recent CNN-based methods [28, 39]. PSNR and SSIM [34] are adopted to quantitatively compare different highlight removal methods. In general, larger PSNR and SSIM scores indicate better removal results.



(a) Input (b) Shen [27] (c) Yang [37] (d) Akashi [1] (e) Yama. [36] (f) Tan [30] (g) Guo [10] (h) Shi [28] (i) Yi [39] (j) Ours

Figure 10. Visual comparison of our method against state-of-the-art highlight removal methods on real-world images from the Internet.



(a) Q1:Is specularity removed effectively? (b) Q2:Is environment lighting not distored?

Figure 11. Rating percentage distribution in the user study.

6.3. Comparison with SOTA Highlight Detectors

Quantitative comparison. Table 2 reports the accuracy and BER scores of different highlight detectors on SRW [6] and our collected dataset SHIQ. Apparently, our method achieves the best quantitative results on the accuracy and BER metrics, demonstrating that our network can more accurately detect highlight regions.

Visual comparison. Figure 9 visually compares highlight detection maps produced by our network and state-of-the-art methods. Specifically, the traditional methods [18, 41] often wrongly detect white text texture as highlight, since they fail to semantically distinguish highlight regions from white material surfaces. In addition, the method proposed in [6] fails to locate weak highlight regions. By contrast, our method can more accurately locate spatially-varying highlight regions, and our results are more consistent with the ground truths.

Table 2. Quantitative comparison of our method with state-of-the-art highlight removal methods on SHIQ, CLH, and LIME datasets. The best results are marked in **bold**.

Dataset	SHIQ		CLH		LIME [21]	
Metric	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Tan [30]	11.04	0.40	19.20	0.59	13.21	0.52
Shen [27]	13.90	0.42	19.43	0.60	14.08	0.51
Yang [37]	14.31	0.50	20.10	0.64	17.64	0.58
Akashi [1]	14.01	0.52	17.39	0.54	16.13	0.55
Guo [10]	17.18	0.58	19.57	0.61	18.03	0.60
Yamamoto [36]	19.54	0.63	20.45	0.64	19.89	0.63
Shi [28]	18.21	0.61	18.65	0.58	24.21	0.76
Yi [39]	21.32	0.72	21.86	0.73	26.77	0.79
Ours	**34.13**	**0.86**	**35.69**	**0.88**	**37.01**	**0.91**

Table 3. Quantitative comparison of our method state-of-the-art detection methods. The best results are marked in **bold**.

Dataset	SHIQ		SRW [6]	
Metric	accuracy↑	BER↓	accuracy↑	BER↓
Li [18]	0.70	18.8	0.72	20.2
Zhang [41]	0.71	24.4	0.64	24.8
Fu [6]	0.91	6.18	0.90	6.21
Ours	**0.93**	**5.92**	**0.92**	**6.04**

6.4. Comparison with SOTA Highlight Removal Methods

Quantitative comparison. Table 2 reports PSNR and SSIM values of different methods on three datasets (*i.e.*, our SHIQ, CLH, and LIME). It shows that our network has larger PSNR and SSIM scores than all the compared methods.

Visual comparison. In Figure 8, we show the highlight removal results of different methods on our dataset. From the results, we can see that traditional methods based on optimizations and color analysis either fail to effectively remove highlight, or produce color/shading distortion. Regarding two CNN-based methods, Shi *et al.* [28] suffers from the over-despecular problem and thus is not able to preserve shading information, thereby resulting in unnatural-looking results, Yi *et al.* [39] tends to leave parts of highlight in the highlight removal results. Fortunately, our method can produce high-quality results, which are considerably similar to the ground truths.

User study. To further evaluate the performance of our network on real-world highlight images, we collect 200 images and conduct a user study to evaluate the results of different methods. Specifically, we download 200 test images from *Pinterest* by searching with keywords like *jade*, *sculpture* and *mask*. Then, we test all methods on 200 collected images to produce their results of estimating underlying highlight-free counterparts, and recruit 20 students to rate different results, which are listed for rating in a random

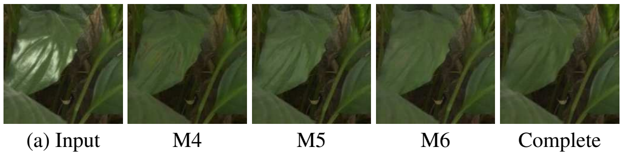


Figure 12. Ablation study for our DSCFA module.

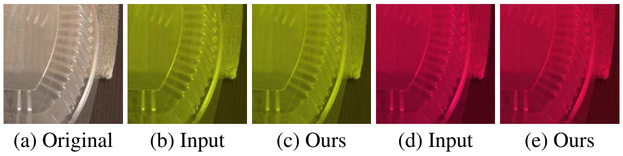


Figure 13. Example failure cases with colored lighting. (b) and (d) are two composited color versions of (a).

Table 4. Component analysis of the proposed method on our dataset for highlight removal. The best results are marked in **bold**.

Methods	**Ours**	M_1	M_2	M_3	M_4	M_5	M_6	M_7
PSNR↑	**34.13**	32.09	31.81	26.19	31.53	33.67	34.02	18.34
SSIM↑	**0.86**	0.82	0.80	0.76	0.80	0.81	0.83	0.61

order to avoid subjective bias. We use the PoV metric [6] to evaluate the users' preference, and a larger PoV value indicates a better highlight removal result. Figure 11 summarizes the average rating results on all 200 images for two questions of the user study, showing our method achieves the largest PoV values among all methods. It indicates that our method has the highlight removal results of our method are more preferred by human subjects. Moreover, Figure 10 visually shows the results of different methods for two downloaded images, demonstrating that our method has a superior performance of highlight removal.

6.5. Discussions

Ablation study. To validate the effectiveness of the major components of our network, we accordingly modify our network to construct seven baselines:

- M_1: JSHDR *w/o* highlight detection.
- M_2: JSHDR *w/o* highlight detection and estimation.
- M_3: JSHDR with only highlight estimation.
- M_4: use only one DSCFA block in the DSCFA module of our JSHDR.
- M_5: use two DSCFA blocks in the DSCFA module of our JSHDR.
- M_6: use three DSCFA blocks in the DSCFA module of our JSHDR.
- M_7: replace DSCFA modules of our JSHDR with simple convolution operations.

Table 4 lists the PSNR and SSIM results of our network and seven baselines. From the results, we have the following observations: (1) Our network has a superior PSNR and SSIM performance over M_1, M_2, and M_3, showing that additional highlight detection and estimation in the multi-task learning of our network help our network to better recover the highlight-free results. (2) The higher PSNR and SSIM scores of our method than M_4 and M_5, and M_6 demonstrate that progressively adding the number of DSCFA blocks improves the highlight removal performance, and our network with four DSCFA blocks has the best performance. Figure 12 shows visual results of our network, M_4, M_5, and M_6, showing that our network has the best performance of highlight removal. (3) Our network has a better PSNR and SSIM performance over M_7, showing that the effectiveness of DSCFA over a basic convolution. In the nutshell, it indicates that with the help of the designed multi-task scheme and the DSCFA module, our method can effectively remove highlight while preserving shading/saturation very well, thereby leading to natural-looking results.

Limitations. Our method has two limitations. First, our method is less able to effectively remove highlights in colored lighting. Figure 13 presents two examples where our method, as well as previous methods, all fail to produce visually satisfactory results. Second, our method and even state-of-the-art inpainting methods are not able to recover text textures since there are no meaningful and reliable contextual cues to help restore them.

7. Conclusion

In this paper, we have presented a large-scale real dataset with about 16K image quadruples that covers a diversity of real-world highlight scenes. Besides, based on the classic dichromatic reflection model, a new region-dependent highlight image formation model is proposed for highlight detection, which provides useful information for highlight removal. Based on this model, we proposed a multi-task convolution network for joint highlight detection and removal. Extensive experiments illustrate that in comparison to previous methods, our method can effectively handle spatially-varying highlights, while preserving shading well.

In the future, we will incorporate an illumination color estimation module into our network and extent our method to handling colored-illumination scenes. Moreover, we will take our method as a pre-processing step for intrinsic image decomposition [5, 8] and recoloring [40], and take our DSCFA as a common feature extraction module for deraining [32, 33] and dehazing [31].

Acknowledgments

We would like to thank Wenxi Gan and Siyuan Yang for their hard work in building the dataset. This work was partly supported by the National Key Research and Development Program of China (2017YFB1002600), the Key Technological Innovation Projects of Hubei Province (2018AAA062), NSFC (NO. 61972298), and CAAI-Huawei MindSpore Open Fund.

References

[1] Yasushi Akashi and Takayuki Okatani. Separation of reflection components by sparse non-negative matrix factorization. *Computer Vision and Image Understanding*, 100(146):77–85, 2015. 2, 6, 7

[2] Sean Bell, Kavita Bala, and Noah Snavely. Acm intrinsic images in the wild. *ACM Transactions on Graphics*, 33(4):159, 2014. 1

[3] Qifeng Chen, Jia Xu, and Vladlen Koltun. Fast image processing with fully-convolutional networks. In *ICCV*, pages 2516–2525, 2017. 4

[4] Rogerio Feris, Ramesh Raskar, Kar-Han Tan, and Matthew Turk. Specular highlights detection and reduction with multi-flash photography. *Journal of the Brazilian Computer Society*, 12(1):35–42, 2006. 2

[5] Gang Fu, Lian Duan, and Chunxia Xiao. A hybrid $l_2 - l_p$ variational model for single low-light image enhancement with bright channel prior. In *ICIP*, pages 1925–1929, 2019. 8

[6] Gang Fu, Qing Zhang, Qifeng Lin, Lei Zhu, and Chunxia Xiao. Learning to detect specular highlights from real-world images. In *ACM MM*, pages 1873–1881, 2020. 1, 2, 4, 6, 7, 8

[7] Gang Fu, Qing Zhang, Chengfang Song, Qifeng Lin, and Chunxia Xiao. Specular highlight removal for real-world images. *Computer Graphics Forum*, 38(7):253–263, 2019. 1

[8] Gang Fu, Qing Zhang, and Chunxia Xiao. Towards high-quality intrinsic images in the wild. In *ICME*, pages 175–180, 2019. 8

[9] Junyu Gao, Tianzhu Zhang, and Changsheng Xu. Graph convolutional tracking. In *CVPR*, pages 4649–4659, 2019. 1

[10] Jie Guo, Zuojian Zhou, and Limin Wang. Single image highlight removal with a sparse and low-rank reflection model. In *ECCV*, pages 268–283, 2018. 1, 2, 6, 7

[11] Xiaojie Guo, Xiaochun Cao, and Yi Ma. Robust separation of reflection from multiple images. In *CVPR*, pages 2187–2194, 2014. 2, 3

[12] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, pages 7454–7462, 2018. 6

[13] Hyeongwoo Kim, Hailin Jin, Sunil Hadap, and Inso Kweon. Specular reflection separation using dark channel prior. In *CVPR*, pages 1460–1467, 2013. 1, 2, 3

[14] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. Parallel feature pyramid network for object detection. In *ECCV*, pages 234–250, 2018. 1

[15] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, pages 1–15, 2015. 6

[16] Chen Li, Stephen Lin, Kun Zhou, and Katsushi Ikeuchi. Specular highlight removal in facial images. In *CVPR*, pages 3107–3116, 2017. 2

[17] Chen Li, Kun Zhou, and Stephen Lin. Simulating makeup through physics-based manipulation of intrinsic image layers. In *CVPR*, pages 4621–4629, 2015. 2

[18] Ranyang Li, Junjun Pan, Yaqing Si, Bin Yan, Yong Hu, and Hong Qin. Specular reflections removal for endoscopic image sequences with adaptive-rpca decomposition. *IEEE Transactions on Medical Imaging*, 39(2):328–340, 2019. 1, 2, 6, 7

[19] Stephen Lin, Yuanzhen Li, Sing Bing Kang, Xin Tong, and Heung-Yeung Shum. Diffuse-specular separation and depth recovery from image sequences. In *ECCV*, pages 210–224, 2002. 1

[20] Yuanliu Liu, Zejian Yuan, Nanning Zheng, and Yang Wu. Saturation-preserving specular reflection separation. In *CVPR*, pages 3725–3733, 2015. 1, 2, 3

[21] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Christian Richardt, and Christian Theobalt. Live intrinsic material estimation. In *CVPR*, pages 6315–6324, 2018. 6, 7

[22] Othmane Meslouhi, Mustapha Kardouchi, Hakim Allali, Taoufiq Gadi, and Yassir Benkaddour. Automatic detection and inpainting of specular reflections for colposcopic images. *Open Computer Science*, 1(3):341–354, 2011. 2

[23] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A dataset of multi-illumination images in the wild. In *ICCV*, pages 4080–4089, 2019. 3

[24] Xingang Pan, Jianping Shi, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Spatial as deep: Spatial cnn for traffic scene understanding. In *AAAI*, pages 7276–7283, 2018. 4

[25] Véronique Prinet, Michael Werman, and Dani Lischinski. Specular highlight enhancement from video sequences. In *ICIP*, pages 558–562, 2013. 6

[26] Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985. 3

[27] Hui-Liang Shen and Zhi-Huan Zheng. Real-time highlight removal using intensity ratio. *Applied optics*, 52(19):4483–4493, 2013. 2, 4, 6, 7

[28] Jian Shi, Yue Dong, Hao Su, and Stella X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, pages 1685–1694, 2017. 1, 2, 6, 7

[29] Jinli Suo, Dongsheng An, Xiangyang Ji, Haoqian Wang, and Qionghai Dai. Fast and high quality highlight removal from a single image. *IEEE Transactions on Image Processing*, 25(11):5441–5454, 2016. 1

[30] Robby T Tan and Katsushi Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):178–193, 2005. 2, 4, 6, 7

[31] Cong Wang, Wanshu Fan, Yutong Wu, and Zhixun Su. Weakly supervised single image dehazing. *Journal of Visual Communication and Image Representation*, 72:102897, 2020. 8

[32] Cong Wang, Yutong Wu, Zhixun Su, and Junyang Chen. Joint self-attention and scale-aggregation for self-calibrated deraining network. In *ACM MM*, pages 2517–2525, 2020. 8

[33] Cong Wang, Xiaoying Xing, Yutong Wu, Zhixun Su, and Junyang Chen. Dcsfn: Deep cross-scale fusion network for single image rain removal. In *ACM MM*, pages 1643–1651, 2020. 8

[34] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[35] Xing Wei, Xiaobin Xu, Jiawei Zhang, and Yihong Gong. Specular highlight reduction with known surface geometry. *Computer Vision and Image Understanding*, 168:132–144, 2018. 2

[36] Takahisa Yamamoto and Atsushi Nakazawa. General improvement method of specular component separation using high-emphasis filter and similarity function. *ITE Transactions on Media Technology and Applications*, 7(2):92–102, 2019. 6, 7

[37] Qingxiong Yang, Jinhui Tang, and Narendra Ahuja. Efficient and robust specular highlight removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1304–1311, 2015. 1, 2, 3, 4, 6, 7

[38] Qingxiong Yang, Shengnan Wang, and Narendra Ahuja. Real-time specular highlight removal using bilateral filtering. In *ECCV*, pages 87–100, 2010. 1

[39] Renjiao Yi, Ping Tan, and Stephen Lin. Leveraging multi-view image sets for unsupervised intrinsic image decomposition and highlight separation. In *AAAI*, pages 12685–12692, 2020. 1, 2, 4, 6, 7

[40] Qing Zhang, Chunxia Xiao, Hanqiu Sun, and Feng Tang. Palette-based image recoloring using color decomposition optimization. *IEEE Transactions on Image Processing*, 26(4):1952–1964, 2017. 8

[41] Wuming Zhang, Xi Zhao, Jean-Marie Morvan, and Liming Chen. Improving shadow suppression for illumination robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):611–624, 2018. 1, 2, 6, 7