Computational Visual Media
https://doi.org/10.1007/s41095-022-0273-9 Vol. 9, No. 1, March 2023, 141–154

Research Article

Joint specular highlight detection and removal in single images via Unet-Transformer

Zhongqi Wu[1,2], **Jianwei Guo**[1,2] (✉), **Chuanqing Zhuang**[2], **Jun Xiao**[2] (✉), **Dong-Ming Yan**[1,2], **and Xiaopeng Zhang**[1,2]

Abstract Specular highlight detection and removal is a fundamental problem in computer vision and image processing. In this paper, we present an efficient end-to-end deep learning model for automatically detecting and removing specular highlights in a single image. In particular, an encoder–decoder network is utilized to detect specular highlights, and then a novel Unet-Transformer network performs highlight removal; we append transformer modules instead of feature maps in the Unet architecture. We also introduce a highlight detection module as a mask to guide the removal task. Thus, these two networks can be jointly trained in an effective manner. Thanks to the hierarchical and global properties of the transformer mechanism, our framework is able to establish relationships between continuous self-attention layers, making it possible to directly model the mapping between the diffuse area and the specular highlight area, and reduce indeterminacy within areas containing strong specular highlight reflection. Experiments on public benchmark and real-world images demonstrate that our approach outperforms state-of-the-art methods for both highlight detection and removal tasks.

Keywords specular highlight detection; specular highlight removal; Unet-Transformer

1 Introduction

Specular highlights are commonly observed in images. However, they can interfere with solutions for many computer vision and image processing tasks, including image segmentation [1–3], photometric stereo [4], binocular stereo [5], and text detection [6]. Hence, effective detection and removal of specular highlights can be beneficial in various real-world applications.

In recent decades, many approaches have been proposed to address the challenging problems of specular highlight detection and removal. Most existing detection methods are based on various forms of thresholding operations [7, 8]. These methods are based on the strict premise that the light is white, and the brightest pixels form specular highlights. As for specular highlight removal, traditional methods can be roughly classified into three categories [9], using color [10, 11], polarization information [5], or illumination estimation [12–14]. Most methods make simple assumptions concerning specific scenes or specific materials, so are difficult to apply to the complex situations in real-world images. We have observed that scenes with specular highlights in the real-world have two common characteristics: firstly, specular highlights are usually small and sparsely distributed, and secondly, the colors of the highlights are similar to the color of the light source. However, the brightest areas in a real image may not be highlights, but caused by overexposure or excessive reflections between objects (see the bottom row of Fig. 1). As a result, existing traditional methods cannot accurately locate specular highlights, nor can they eliminate the semantic ambiguity between white (or close to white) materials and highlights in complex real scenes, especially

1 National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China. E-mail: Z. Wu, wuzhongqi2019@ia.ac.cn; J. Guo, jianwei.guo@nlpr.ia.ac.cn (✉), D.-M. Yan, yandongming@gmail.com; X. Zhang, xiaopeng.zhang@ia.ac.cn.

2 The School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China. E-mail: C. Zhuang, zhuangchuanqing19@mails.ucas.ac.cn; J. Xiao, xiaojun@ucas.ac.cn (✉).

Manuscript received: 2021-11-10; accepted: 2022-02-03

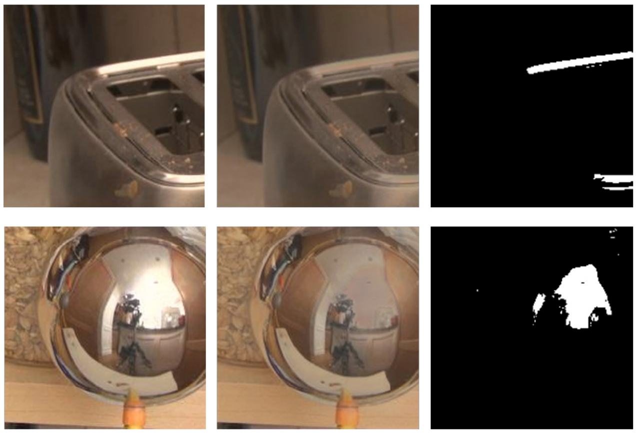

Fig. 1 Specular highlight removal and detection results from our method. Left: input images with specular highlights. Middle: after highlight removal. Right: detected highlights.

when the image simultaneously contains refraction, reflection, and transmission, particularly involving metal and glass. Some recent approaches have started to utilize deep neural networks [15–17] to remove specular highlights from a single image. However, most are trained on synthetic data in a supervised manner, and their detection and removal abilities usually deteriorate significantly for real-world images. Furthermore, existing methods usually regard highlight detection and removal as two separate tasks, and do not use detection results to guide specular highlight removal.

In this paper, we propose a new deep neural model which jointly detects and removes specular highlights in a single image. To accomplish this, we utilize the popular Unet to detect specular highlight areas, and use this information to guide the specular highlight removal network. Inspired by the great success of transformers in recent computer vision tasks [18], which construct hierarchical feature expressions by dividing images or feature maps into smaller windows, we integrate an efficient Swin transformer into our highlight removal network. Doing so is also supported by our previous observations about specular highlights, since the Swin transformer [19] works well to capture global features and establish relationships between continuous self-attention layers. This enables interaction and connection between windows of the previous layer, which greatly improves the expressive ability of the model. As a result of introducing the transformer module, our specular highlight removal network is capable of leveraging high-level contextual clues to reduce indeterminacy within areas containing strong specular highlight reflections. Furthermore, we also use highlight detection results as a mask to guide the removal task, which can reduce color defects. To sum up, the key contributions of this work are:

- a joint specular highlight detection and removal network that works well for single real-world images, and
- an efficient Unet-Transformer module for specular highlight removal, where detection results are used as a guidance mask to reduce the effects of chromatic aberration.

2 Related work

In this section, we briefly review previous specular highlight detection and removal methods.

2.1 Specular highlight detection

Early studies proposed various methods for specular highlight detection task based on a color constancy model [20–22]. Zhang et al. [23] formulated specular highlight detection as non-negative matrix factorization (NMF) [24] based on the assumption that the number of specular highlights is small. Li et al. [8] proposed an adaptive robust principal component analysis (Adaptive-RPCA) method to robustly detect specular highlights in endoscope image sequences. Recently, Fu et al. [25] presented a large-scale dataset for specular highlight detection in real-world images. Based on the dataset, they also proposed a deep neural network leveraging multi-scale context-contrasted features to accurately detect specular highlights. However, this dataset does not contain corresponding diffuse images, so it cannot be used for learning specular highlight removal.

2.2 Specular highlight removal

2.2.1 Using color

Tan et al. [26] proposed to use the spatial distribution of color to overcome specular reflection separation ambiguity in real images. Shen et al. [11] separated highlight reflections in a color image by selecting an appropriate body color for each pixel by error analysis of chromaticity. Shen and Cai [27] extended this work to improve the robustness of the algorithm. For real images, specular reflections can be effectively removed [28–30] using a dichromatic reflection model [10]. Akashi and Okatani [31] formulated the separation of reflections as a sparse non-negative matrix factorization problem without spatial priors. Likewise, Guo et al. [32] imposed the non-negativity constraint on the weighting matrix and enhanced robustness. Fu et al. [33] designed an optimization framework for simultaneously estimating diffuse and specular highlight images from a single image. They recovered the diffuse components of specular highlight regions by encouraging sparseness of the encoding coefficients. However, such methods cannot semantically distinguish whether a bright area in the image is a highlight or a white object: when light-colored areas and highlights coexist, these algorithms may suffer from large errors.

2.2.2 Using polarization

Unlike color-based methods, to avoid the color distortion caused by illumination, polarization-based methods take advantage of polarimetric information. Nayar et al. [34] presented a method that separates the diffuse and specular components of brightness in single images, using color and polarization information at the same time to obtain constraints on the reflection component of each pixel. By using two or more images of surface reflection, Umeyama and Godin [35] proposed a stable separation algorithm for diffuse and specular reflections based on independent component analysis. Further, Wang et al. [36] presented an efficient specular removal method based on polarization images through global energy minimization. Wen et al. [37] recently introduced a new polarization guided model to generate a polarization chromaticity image. They conducted specular reflection separation by optimizing a global energy function. However, these polarization-based methods rely on strictly controlled light sources and are only suitable in certain specific scenarios. In real scenes, some specular highlight components are still retained in the diffuse reflection images.

2.2.3 Using illumination estimation

Illumination-estimation-based methods can coarsely remove highlights [12, 13, 38]; they either focus on estimating the distribution direction of the light source or estimating the illumination color. There are two methods for estimating illumination color. One is to estimate illumination color based on the specular reflections [14, 39], and the other is to analyze the surface color based on the color constant of the prior model [40–42]. Lin et al. [43] presented an interactive method by introducing specular highlight removal as an inpainting process. Tan et al. [44, 45] separated specular illumination using the concept of inverse intensity space. However, these methods are often susceptible to complex lighting and chromaticity issues.

2.2.4 Using deep-learning

Deep-learning-based methods have been widely used for removing specular highlights in single image; handcrafted priors are replaced by data-driven learning [15–17, 46–48]. Shi et al. [15] presented an encoder–decoder convolutional neural network (CNN) to handle the non-Lambertian object intrinsic decomposition problem. Funke et al. [46] proposed a CycleGAN based network for specular highlight removal from a single endoscopic image. To train this network, they constructed a dataset by extracting small image patches with specular highlights and patches without highlights from the endoscopic video. Lin et al. [16] presented a fully CNN, trained on a synthetic dataset. The network can work out the intricate relationships between an image and its diffuse parts. Muhammad et al. [17] presented Spec-Net and Spec-CGAN, aimed at removing high intensity specularities from low chromaticity facial images. Yi et al. [49] presented an unsupervised method for specular reflection layer separation using multi-view images. Wu et al. [47, 48] proposed a novel generative adversarial network (GAN) for specular highlight removal based on polarization theory. Fu et al. [50] developed a multi-task network for joint highlight detection and removal, based on a large-scale dataset with manual annotation. However, while numerous researchers have considered specularity removal, current methods still leave specular highlight residuals and chromatic aberrations in real-world scenes.

3 Methodology

3.1 Overview

In this paper, we propose an end-to-end network structure to jointly detect and remove the specular highlights from real-world images. Given a single image with specular highlights as input, our goal is to detect the locations of the highlights and restore diffuse reflection in such highlight areas. Our network architecture comprises two branches and is summarized in Fig. 2. The highlight detection branch uses a pure Unet to output a set of masks locating the highlights. Then the second branch uses a novel Unet-Transformer-based highlight removal module to obtain the corresponding diffuse reflection image (i.e., without specular highlights). We exploit the assumption that the locations of highlights provide a strong prior for highlight removal, so we use the highlight detection result to guide the removal task.

Details of our network architecture and loss functions follow, then implementation details, including training dataset and settings.

3.2 Specular highlight detection

The specular highlight detection task can be regarded as a binary classification task, which outputs 0 in non-highlight areas and 1 in highlights. To this end, our proposed specular highlight detection network is based on an encoder–decoder framework [51]. As shown in Fig. 2, the network takes the input image with specular highlights I_{s} and outputs a mask M' indicating specular highlight regions. We adopt a fully convolutional architecture consisting of four downsampling layers (the encoder) and corresponding upsampling layers (the decoder). The purpose of the encoder is to extract feature maps from the highlight image. Specifically, each downsampling layer consists of two 3×3 convolutions, each followed by a ReLU, and a 2×2 max-pooling operation. The decoder is used to output the pixel classification result. It consists of three parts: upsampling of a feature map followed by a 2×2 convolution in which the number of feature channels is halved, concatenation with the corresponding cropped feature map from downsampling, and two 3×3 convolutions each of which is followed by a ReLU. Cropping is necessary due to the loss of border pixels in every convolution.

3.3 Specular highlight removal

Existing specular highlight removal methods do not achieve satisfactory results for transparent objects and complex scenes (e.g., in which transmission and reflection exist at the same time), which usually require global contextual reasoning. To solve this problem, our key idea is to exploit the self-attention mechanism [52] of transformers to enhance the connection between a specular highlight area and the surrounding area. Our method appears to be the first application of transformers to specular highlight removal.

As Fig. 2 shows, we first perform a patch partition operation to preprocess the input images I_{s} and M'. This partition changes the image from $H \times W \times 3$ to a tensor of size $H/4 \times W/4 \times 48$. Then in stage 1 (see Fig. 2), a linear embedding layer is used to convert an image with dimensions $(H/4, W/4, 48)$ into a feature vector with dimensions $(H/4, W/4, C)$. In stages 2–4, the patch merging layer divides the

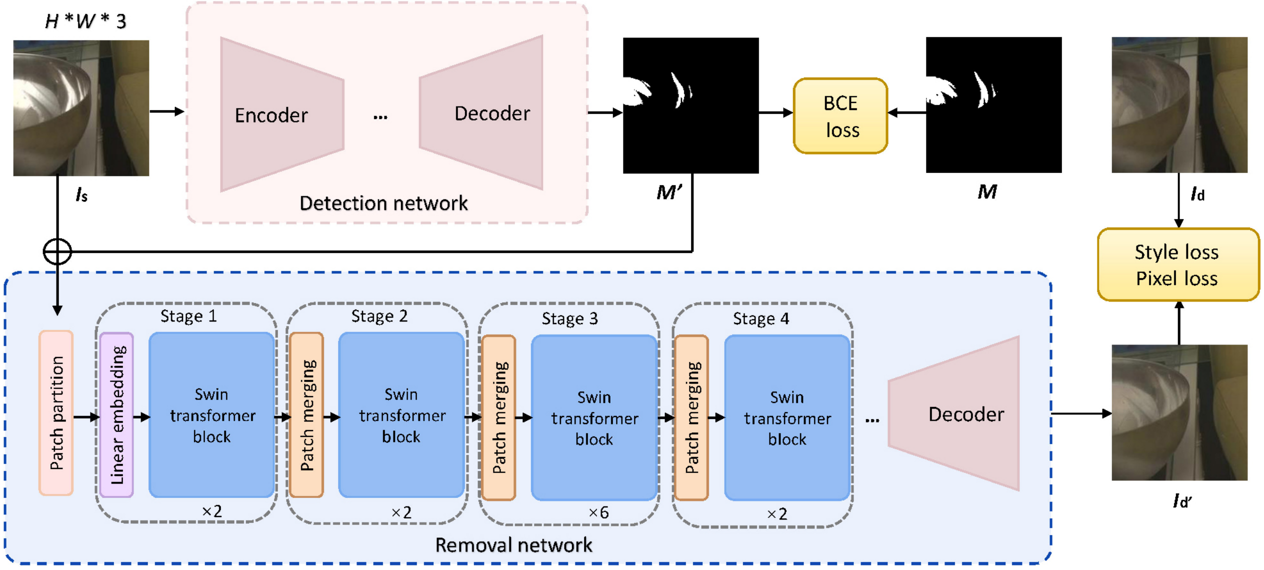

Fig. 2 Proposed specular highlight detection and removal framework. Given an input image I_{s} with specular highlights, an encoder–decoder detection network outputs a mask M' indicating highlight areas. Using the detection results as guidance, a Unet-Transformer provides a specular-free diffuse image $I_{\mathrm{d}'}$.

input patches into 4 parts and concatenates them. This processing down-samples the feature resolution by a factor of 2. Since the concatenation operation increases the feature dimension by a factor of 4, a linear layer is applied to the concatenated features to unify the feature dimension to twice the original dimension.

Here we adopt a powerful transformer architecture, a Swin transformer [19], recently used in image classification and segmentation, to generate the hierarchical feature maps in linear time. The Swin transformer is constructed using shifted windows, as shown in Fig. 3; two consecutive Swin transformer blocks are presented. Each Swin transformer block is composed of a LayerNorm (LN) layer, a multi-head self attention module, a residual connection, and a 2-layer multilayer perceptron (MLP) with a non-linear Gaussian error linear unit (GELU). The window-based multi-head self attention (W-MSA) module and the shifted window-based multi-head self attention (SW-MSA) module are applied in the two successive transformer blocks. The transformer uses linear projections to compute a set of queries Q, keys K, and values V, and takes a weighted sum of value vectors according to a similarity distribution between query and key vectors:

$$Q = BHWC_1, \quad K = BHWC_2, \quad V = BHWC_3 \tag{1}$$

$$\text{Attention}(Q, K, V) = \text{Softmax}(QK^{\mathrm{T}}/\sqrt{d} + B)V \tag{2}$$

where Q is a matrix of n_{q} query vectors, B is the learnable relative positional encoding, K and V both contain n_{k} keys and values, all with the same dimensionality, and d is a scaling factor.

In the implementation of the encoder, the C-dimensional tokenized inputs with resolution $H/4 \times W/4$ are fed into the Swin transformer block to perform representation learning. Note that in this process the feature dimension and resolution remain unchanged. Then the patch merging layer reduces the number of tokens (using 2× downsampling) and increases the feature dimension by a factor of 2. This procedure is repeated three times in the encoder. Finally, the decoder, consisting of four upsampling layers, is used to output the generated diffuse image $I_{\mathrm{d}'}$.

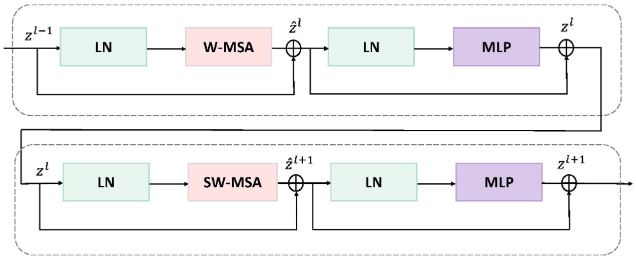

Fig. 3 A Swin transformer block [19]. LN: LayerNorm layer. W-MSA: window-based multi-head self-attention module. SW-MSA: shifted window-based multi-head self-attention module.

However, in some cases, the specular highlight may be located in a large area with strong intensity, and direct removal will cause chromatic aberration in non-highlight areas. In order to reduce chromatic aberration, we use the specular highlight detection mask to guide specular highlight removal.

3.4 Loss functions

To jointly train the network, we integrate the two modules above in a unified network architecture. Network training is supervised by an efficient loss function with three components: specular detection loss, context loss, and style loss.

3.4.1 Specular detection loss

Cross-entropy loss is commonly used for solving edge detection and semantic segmentation problems. Similarly, we use this loss for the specular highlight detection task, formulating it as L_{BCE}:

$$L_{\mathrm{BCE}} = -\sum_i [M_i \log(M_i') + (1 - M_i)\log(1 - M_i')] \tag{3}$$

where i indexes each pixel, M_i is an element of the ground-truth highlight mask M, and M_i' is the predicted probability of the pixel belonging to a specular highlight area.

3.4.2 Pixel loss

Following Ref. [53], we use *pixel loss* to reduce the intensity and texture difference between the generated diffuse image $I_{\mathrm{d}'}$ and the ground-truth image I_{d}:

$$\begin{aligned} L_{\text{pixel}} = \alpha \|I_{\mathrm{d}} - I_{\mathrm{d}'}\|_2^2 + \beta(\|\nabla_x I_{\mathrm{d}} - \nabla_x I_{\mathrm{d}'}\|_1 + \\ \|\nabla_y I_{\mathrm{d}} - \nabla_y I_{\mathrm{d}'}\|_1) \end{aligned} \tag{4}$$

We set $\alpha = 0.2$ and $\beta = 0.4$ in all of our experiments.

3.4.3 Style loss

Style loss is usually used in image style transfer tasks [54]. We use this loss to add constraints on the pixel and feature space:

$$L_{\text{style}} = \sigma \left\|\psi(I_{\mathrm{d}'}) - \psi(I_{\mathrm{d}})\right\|_1 \tag{5}$$

where $\sigma = 120$, $\psi(\cdot) = \phi(\cdot)\phi(\cdot)^{\mathrm{T}}$ is the Gram matrix [54], where ϕ are feature maps of pre-trained

VGG-16 [55]. The selected layer indices in VGG-16 for style loss are 0, 5, 10, 19, 28.

3.4.4 Total loss

To summarize, our total loss function is defined as

$$L_{\mathrm{G}} = \omega_1 L_{\mathrm{BCE}} + \omega_2 L_{\mathrm{pixel}} + \omega_3 L_{\mathrm{style}} \tag{6}$$

where we set $\omega_1 = 1.0$, $\omega_2 = 1.0$, and $\omega_3 = 0.08$ in our experiments.

3.5 Implementation details

3.5.1 Datasets

We trained our method on the SHIQ dataset [50], which provides specular highlight images, corresponding diffuse images, and highlight mask images. These specular highlight images were collected from the MIW dataset [56], which contains many hard shiny materials (e.g., metal, plastics, glass). Note that instead of capturing real diffuse images, the corresponding diffuse images were generated by the RPCA method [57]. However, the results generated by RPCA still contain specular residuals, so Fu et al. [50] only cropped high-quality local images (with paired specular image and specular-free diffuse image) to build the SHIQ dataset. We used 9825 groups for training and 999 groups for testing. The resolution of each image was 200 × 200.

3.5.2 Training settings

Our network was implemented in PyTorch on an NVIDIA GeForce GTX1080Ti graphics card. We trained the network on our training set for 60 epochs using the Adam optimizer [58]. The initial learning rate was set to 10^{-4} and reduced using an attenuation coefficient of 0.8 every 5 epochs until reaching 10^{-5}. We also augmented the SHIQ dataset by randomly mirror-flipping images and adding noise.

4 Experimental results

In this section, we start with several experiments through visually inspecting our results on the dataset to demonstrate the effectiveness of our proposed neural network. Then we compare our detection and removal results with current state-of-the-art approaches with qualitative and quantitative evaluations.

4.1 Detection and removal results

Figure 4 shows our specular highlight detection and removal results on some representative images selected from the SHIQ dataset. These examples include objects with different materials such as transparent plastic, glass, and metal. Specular highlight removal is particularly challenging for transparent objects as there are both reflection and transmission components, making it difficult restore specular highlight areas. As the figure illustrates, our method can accurately locate the specular highlights, and can also effectively remove them from such objects. The fourth row shows that our method can still give satisfactory results for large highlight areas, and areas including reflections.

4.2 Comparisons

4.2.1 Highlight detection

We first compare highlight detection results from our method with those from previous methods, including two traditional methods (NMF [8] and ATA [23]), and two state-of-the-art deep-learning-based methods (SHDN [25] and JSHDR [50]). For quantitative evaluation, we adopt two commonly used metrics, namely, detection accuracy and balance error rate (BER). Their definitions are

$$\mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{7}$$

$$\mathrm{BER} = \frac{1}{2}\left(\frac{\mathrm{FP}}{\mathrm{TN}+\mathrm{FP}} + \frac{\mathrm{FN}}{\mathrm{TN}+\mathrm{TP}}\right) \tag{8}$$

where TP is the number of true positives, TN true negatives, FP false positives, and FN false negatives.

A higher value of accuracy and a lower value of BER indicate better detection results. Table 1 reports quantitative comparison results on the SHIQ testing data. As we can see, JSHDR [50] and our method achieve the best results, while our method is slightly better in terms of accuracy.

4.2.2 Highlight removal

We also compare our approach to various highlight removal competitors, including two traditional approaches (Shen et al. [11], Yamamoto et al. [59]), and

Table 1 Quantitative comparison of our method with state-of-the-art highlight detection methods. The best results are given in bold

Method	Acc↑	BER↓
NMF	0.70	18.8
ATA	0.71	24.4
SHDN	0.91	6.18
JSHDR	0.93	**5.92**
Ours	**0.97**	**5.92**

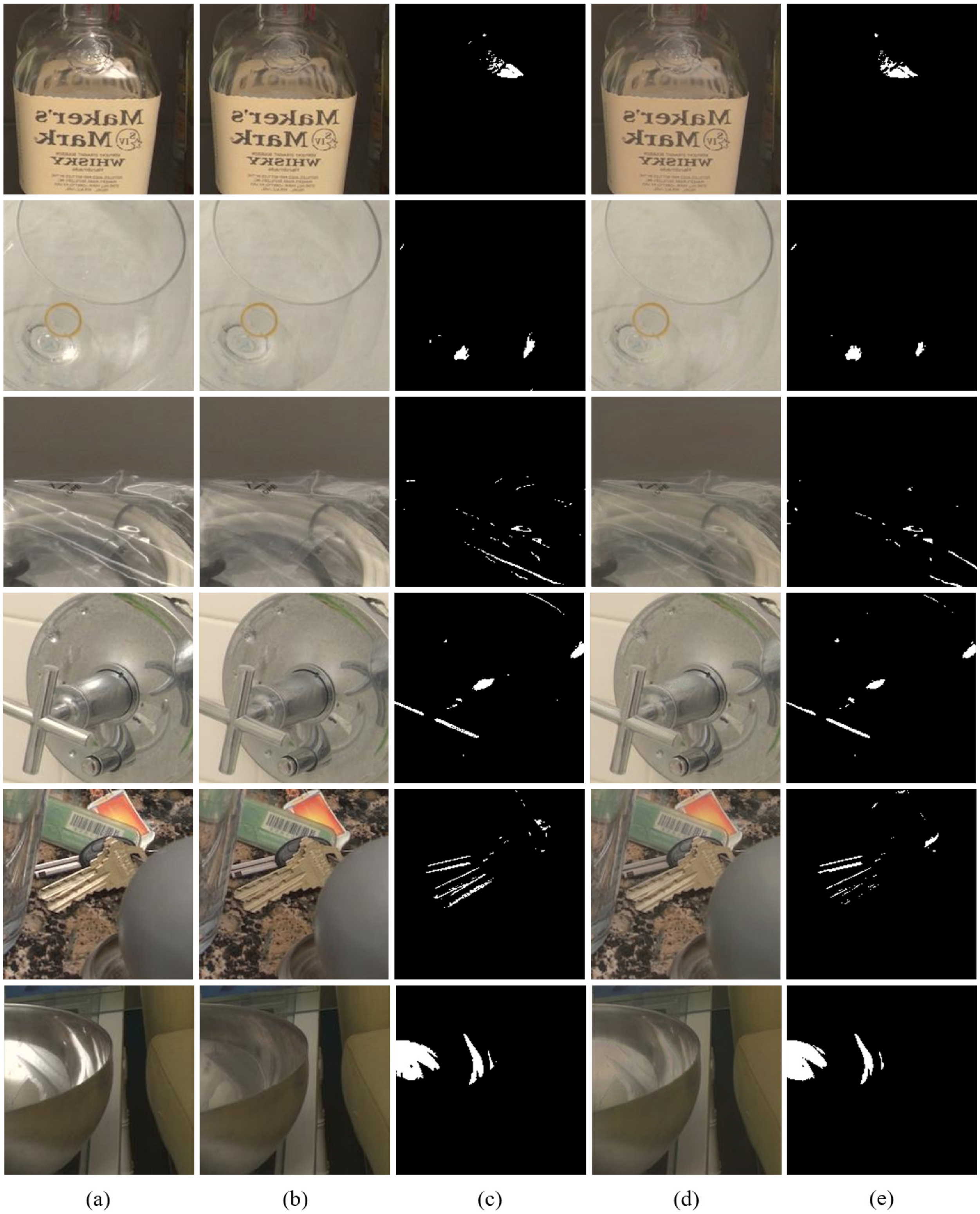

Fig. 4 Specular highlight detection and removal results using our neural network: (a) input images with specular reflections, (b) ground-truth diffuse images, (c) ground-truth masks of the specular highlights, (d) our removal results, and (e) our detection results.

three state-of-the-art deep-learning-based approaches (Multi-class GAN [16], Spec-CGAN [46], and JSHDR [50]). For a fair comparison, we re-trained Multi-class GAN and Spec-CGAN on the SHIQ dataset. Figure 5 shows results using the SHIQ testing data for evaluation. We observe that Shen and Cai [27] and

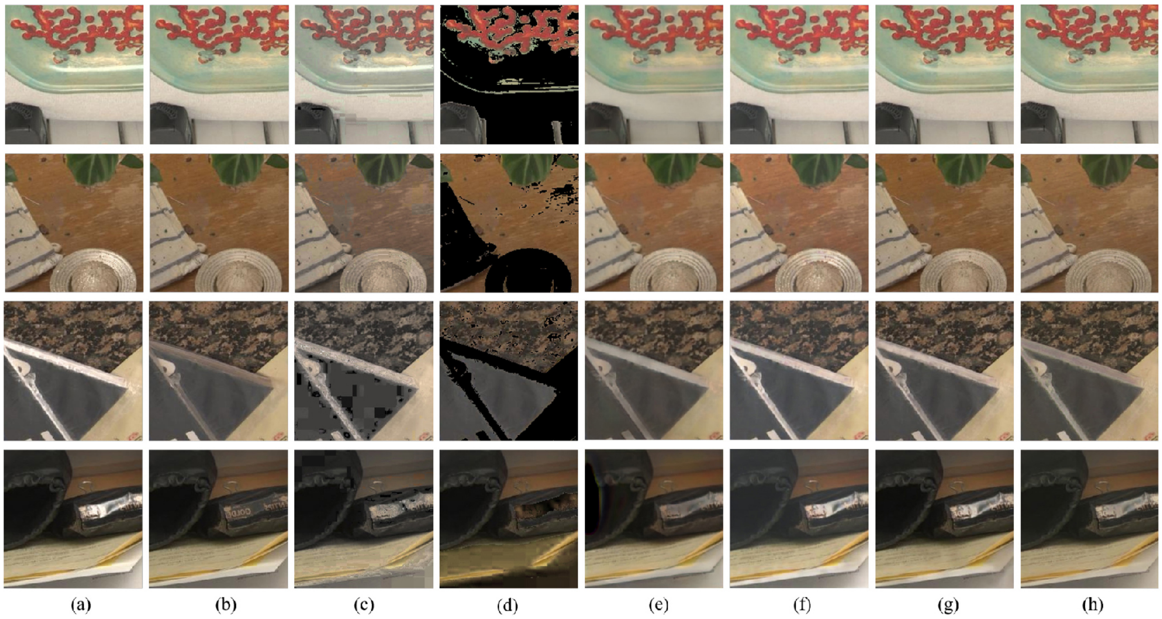

Fig. 5 Visual comparison of highlight removal methods on the SHIQ dataset: (a) input image, (b) ground-truth specular-free diffuse image, and results from (c) Shen et al. [11], (d) Yamamoto et al. [59], (e) Multi-class GAN [16], (f) Spec-CGAN [46], (g) JSHDR [50], and (h) our method.

JSHDR [50] have local specular highlight residuals, especially on the Garniture and Wrapper scenes (in first and fourth rows). Yamamoto et al. [59] induced color distortion on the surfaces of light color objects, resulting in dark areas. Multi-class GAN [16] and Spec-CGAN [46] leave obvious specular highlight residuals. In comparison, our network removes most of the highlights and produces no dark shadows and chromatic aberrations. For quantitative comparison, we adopt three commonly used metrics: mean-squared error (MSE), structural similarity index (SSIM), and peak-signal-to-noise ratio (PSNR). Table 2 reports these values for different methods for Fig. 5. Our network has better scores than all compared methods.

Finally, in order to compare the capabilities of approaches using natural images, we captured a real-world specular testing dataset using a cellphone. As Fig. 6 shows, traditional methods like Shen et al. [11] either fail to effectively remove specular highlights (see first row), or produce chromatic aberrations (see rows 2, 3, 5). Yamamoto et al. [59] generated distinct dark areas in light and specular highlight regions. Multi-class GAN [16] can remove some specular highlights from the images, but chromatic aberration appears (see rows 2, 4). Spec-CGAN [46] has obvious specular highlight residuals (see rows 3, 4, 5). JSHDR [50] achieves good results overall, but it still generates distinct dark patches (see rows 2, 3) and has obvious specular highlight residuals (see

Table 2 Quantitative comparison of highlight removal methods on the SHIQ dataset

Scene	Method	MSE/$10^{-2}\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
Garniture	Shen et al.	5.93	0.3707	12.27
	Yamamoto et al.	27.12	0.0679	5.67
	Multi-class GAN	0.45	0.9373	23.51
	Spec-CGAN	0.14	0.9597	28.38
	JSHDR	0.08	0.9738	31.00
	Ours	**0.04**	**0.9812**	**34.47**
Metal	Shen et al.	0.21	0.9046	26.79
	Yamamoto et al.	10.87	0.5450	9.63
	Multi-class GAN	0.23	0.9666	26.33
	Spec-CGAN	0.21	0.9712	26.85
	JSHDR	0.07	**0.9923**	**34.30**
	Ours	**0.05**	0.9894	32.43
Plastic	Shen et al.	4.87	0.2728	13.13
	Yamamoto et al.	6.89	0.1828	11.62
	Multi-class GAN	0.36	0.8627	24.38
	Spec-CGAN	1.24	0.8869	19.07
	JSHDR	0.51	0.9282	22.93
	Ours	**0.20**	**0.9374**	**26.99**
Wrapper	Shen et al.	10.75	0.2903	24.61
	Yamamoto et al.	6.56	0.3824	9.69
	Multi-class GAN	0.31	0.8746	23.75
	Spec-CGAN	1.04	0.8439	19.83
	JSHDR	0.35	0.9515	24.54
	Ours	**0.28**	**0.9598**	**25.51**

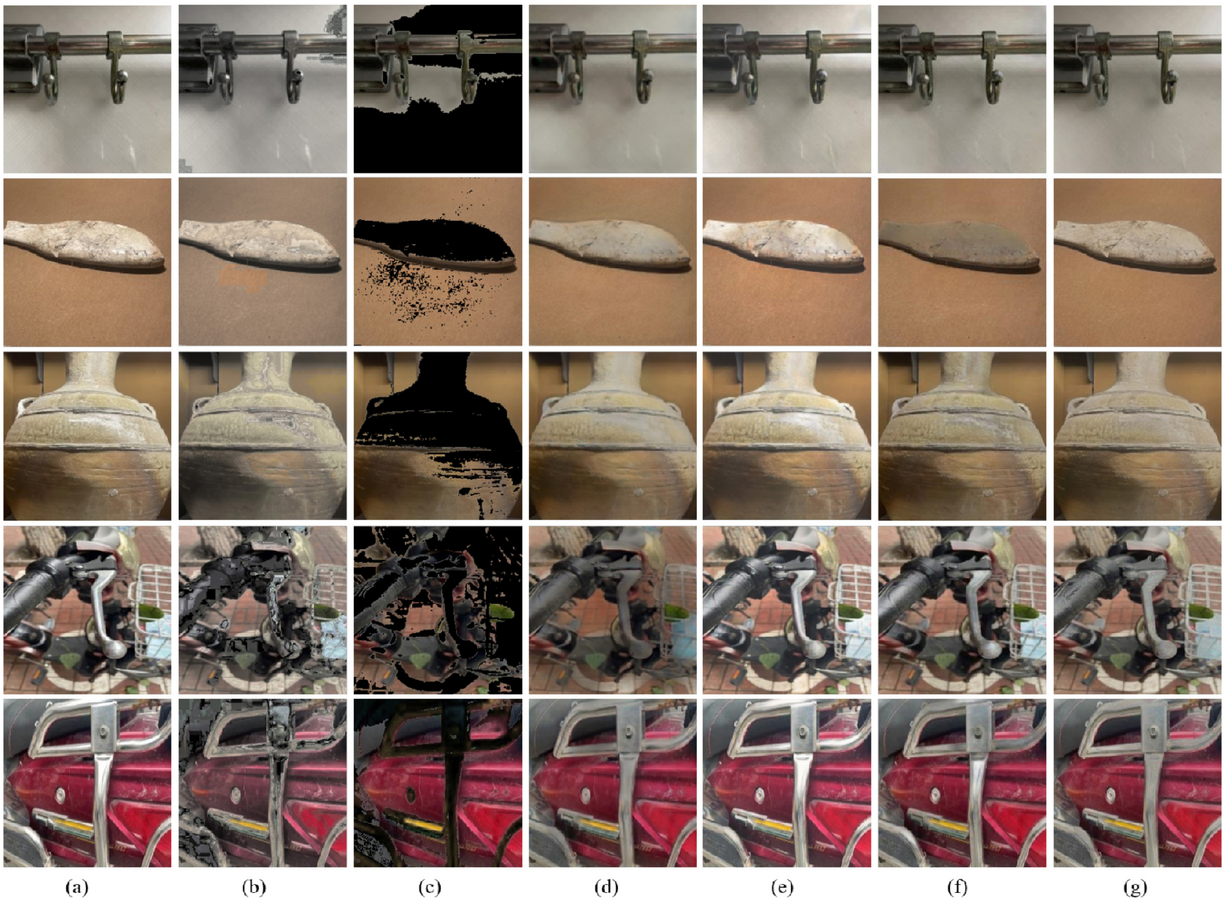

Fig. 6 Visual comparison on natural images in the wild: (a) input, and results from (b) Shen et al. [11], (c) Yamamoto et al. [59], (d) Multi-class GAN [16], (e) Spec-CGAN [46], (f) JSHDR [50], and (g) our method.

rows 4, 5). In contrast, our network can effectively remove specular highlights without generating dark shadows or chromatic aberrations. Our neural network generalises well.

4.3 Ablation studies

4.3.1 Highlight detection

It should be noted that, when the highlight detection and removal networks are jointly trained, the detection results are slightly inferior to when only using the highlight detection network. This is because the highlight detection probability is used as a weight when the diffuse reflection image is finally generated; the two have a connection relationship via the gradient which affects the detection result. To demonstrate this, we conducted experiments to evaluate the influence of the specular highlight removal network and Swin transformer on the specular detection network. Table 3 reports quantitative results of this ablation study. As observed, when the removal network or the Swin transformer is removed from our network, the balance error rate metric is slightly worse than for SHDN [25] and JSHDR [50], while our full method is slightly better than the other methods.

Table 3 Quantitative comparison of detection network settings. The best result in each case is shown in bold

Method	Accuracy ↑	BER ↓
Ours without removal	0.92	6.98
Ours without Swin transformers	0.93	6.31
Our full method	**0.97**	**5.92**

4.3.2 Highlight removal

To verify the effectiveness of our network architecture and loss functions, we compared our network with ablated versions. Visual examples from the ablation studies are shown in Fig. 7 and corresponding quantitative results are given in Table 4. As we can see, our full model achieves the best performance. As shown in Fig. 7(c), Fig. 7(d), and Fig. 7(g), the network without Swin transformers produces obvious chromatic aberration. The network without

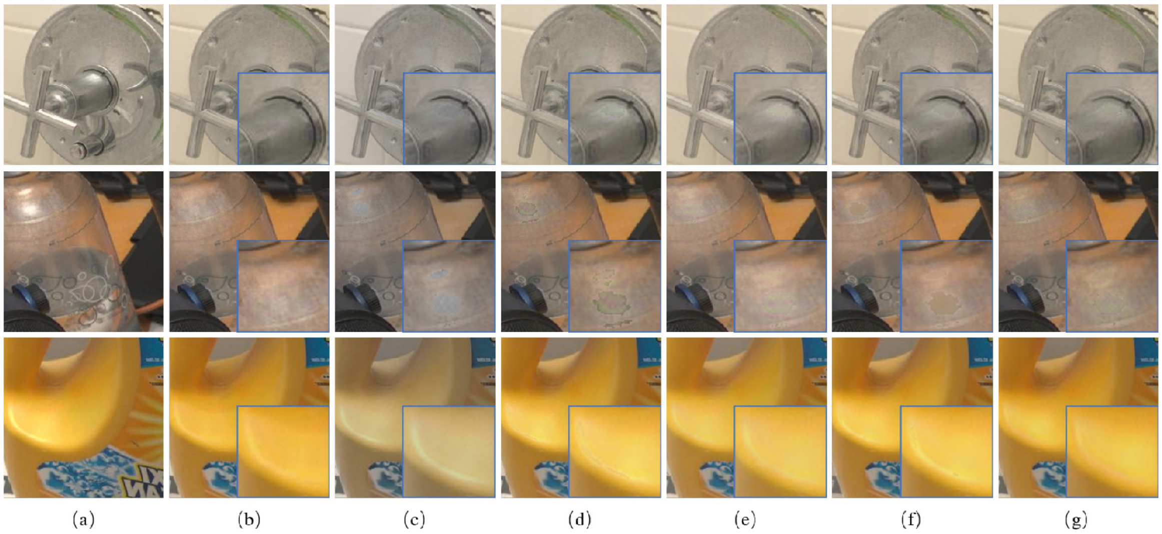

Fig. 7 Visual examples from ablation studies for the proposed network: (a) input, (b) ground-truth, and results from (c) our method without Swin transformer, (d) our method without highlight mask, (e) our method without pixel loss, (f) our method without style loss, and (g) our overall method.

highlight mask produces artifacts in the highlights, indicating the usefulness of the highlight mask as guidance in local highlight areas. As shown in Figs. 7(e)–7(g), results obtained without pixel loss have highlight remnants (see rows 1, 3). Results without style loss produce artifacts (see rows 1, 2) and highlight remnants (see row 3). Our full loss (see Fig. 7(g)) provides cleaner results with fewer highlight remnants and no artifacts. The quantitative results in the Table. 4 further show that our full network achieves the best results. We can also see that Swin transformers play a more important role than highlight masks for highlight removal.

Table 4 Quantitative comparison of ablation study, corresponding to Fig. 7. "−" means without

Scene	Method	MSE/10^{-2}↓	SSIM ↑	PSNR ↑
Faucet	− Swin	0.08	0.8852	30.58
	− mask	0.04	0.9798	34.19
	− pixel loss	0.09	0.9781	30.64
	− style loss	0.04	0.9815	33.65
	full method	**0.03**	**0.9822**	**34.65**
Bottle	− Swin	0.25	0.8801	25.85
	− mask	0.10	0.9859	29.86
	− pixel loss	**0.05**	0.9874	32.42
	− style loss	0.07	0.9882	31.72
	full method	**0.05**	**0.9886**	**33.08**
Yellow plastic	− Swin	1.06	0.8248	19.77
	− mask	0.05	0.9972	32.80
	− pixel loss	0.07	0.9951	31.63
	− style loss	**0.04**	0.9975	**33.92**
	full method	**0.04**	**0.9976**	33.61

4.4 Limitations

We have successfully applied our method for detecting and removing specular highlights to a variety of single images. However, our neural network in common with many state-of-the-art methods may fail to remove large specular highlight areas, as shown in Fig. 8, where the large areas lack meaningful and reliable contextual cues to help restore them. Furthermore, out network cannot handle images with text due to a lack of training data richness: when the specular highlight covers part of the text, it is challenging to

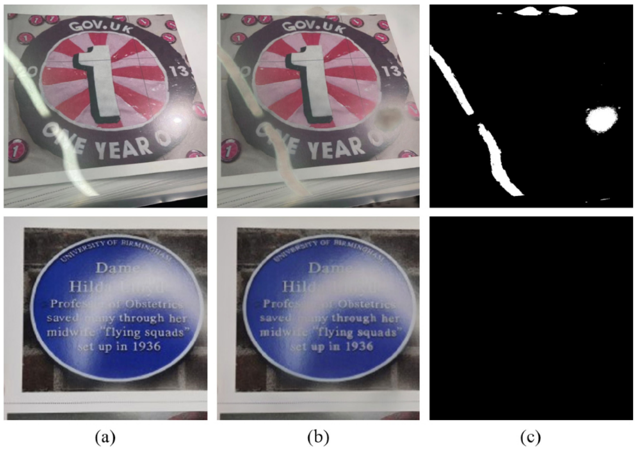

Fig. 8 Examples of failures: (a) input image, (b) our highlight removal results, and (c) our detection results.

remove the highlights. To handle such cases, a text detection branch or a text-aware loss might help [60].

5 Conclusions and future work

This work has solved the challenging problem of joint specular highlight detection and removal in a single image, using an end-to-end deep learning framework that consists of two networks: an encoder–decoder network for highlight detection, and a Unet-Transformer network for highlight removal. We also use the detection results as guidance to ensure that the highlight removal network pays more attention to the highlight areas. A variety of experiments on public benchmark datasets and many challenging real images have shown the effectiveness of our neural network. Our source code is publicly available at `https://github.com/jianweiguo/specularityRemoval`.

In future, we hope to remove specular highlights from complex scenes with rich textures. We also plan to construct a large dataset and design a more effective text-related loss to promote text-aware highlight removal. Finally, we will explore the relationship between specular highlights and object geometry, such as flat, spherical, and cylindrical highlights, which may help to accurately locate and remove specular highlights.

Acknowledgements

This work was partially funded by the National Natural Science Foundation of China (U21A20515, 62172416, 62172415, U2003109), and Youth Innovation Promotion Association of the Chinese Academy of Sciences (2022131).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Arbeláez, P.; Maire, M.; Fowlkes, C.; Malik, J. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 33, No. 5, 898–916, 2011.

[2] Tao, M. W.; Su, J. C.; Wang, T. C.; Malik, J.; Ramamoorthi, R. Depth estimation and specular removal for glossy surfaces using point and line consistency with light-field cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 38, No. 6, 1155–1169, 2016.

[3] Ramadan, H.; Lachqar, C.; Tairi, H. A survey of recent interactive image segmentation methods. *Computational Visual Media* Vol. 6, No. 4, 355–384, 2020.

[4] Khanian, M.; Boroujerdi, A. S.; Breuß, M. Photometric stereo for strong specular highlights. *Computational Visual Media* Vol. 4, No. 1, 83–102, 2018.

[5] Cui, Z. P.; Gu, J. W.; Shi, B. X.; Tan, P.; Kautz, J. Polarimetric multi-view stereo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 369–378, 2017.

[6] Xue, M. L.; Shivakumara, P.; Zhang, C.; Xiao, Y.; Lu, T.; Pal, U.; Lopresti, D.; Yang, Z. Arbitrarily-oriented text detection in low light natural scene images. *IEEE Transactions on Multimedia* Vol. 23, 2706–2720, 2021.

[7] Osadchy, M.; Jacobs, D. W.; Ramamoorthi, R. Using specularities for recognition. In: Proceedings of the 9th IEEE International Conference on Computer Vision, 1512–1519, 2003.

[8] Li, R. Y.; Pan, J. J.; Si, Y. Q.; Yan, B.; Hu, Y.; Qin, H. Specular reflections removal for endoscopic image sequences with adaptive-RPCA decomposition. *IEEE Transactions on Medical Imaging* Vol. 39, No. 2, 328–340, 2020.

[9] Artusi, A.; Banterle, F.; Chetverikov, D. A survey of specularity removal methods. *Computer Graphics Forum* Vol. 30, No. 8, 2208–2230, 2011.

[10] Shafer, S. A. Using color to separate reflection components. *Color Research & Application* Vol. 10, No. 4, 210–218, 1985.

[11] Shen, H. L.; Zhang, H. G.; Shao, S. J.; Xin, J. H. Chromaticity-based separation of reflection components in a single image. *Pattern Recognition* Vol. 41, No. 8, 2461–2469, 2008.

[12] Brainard, D. H.; Freeman, W. T. Bayesian color constancy. *Journal of the Optical Society of America A* Vol. 14, No. 7, 1393–1411, 1997.

[13] Finlayson, G. D.; Hordley, S. D.; HubeL, P. M. Color by correlation: A simple, unifying framework for color constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 23, No. 11, 1209–1221, 2001.

[14] Tan, R. T.; Nishino, K.; Ikeuchi, K. Color constancy through inverse-intensity chromaticity space. *Journal of the Optical Society of America A* Vol. 21, No. 3, 321–334, 2004.

[15] Shi, J.; Dong, Y.; Su, H.; Yu, S. X. Learning non-Lambertian object intrinsics across ShapeNet categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5844–5853, 2017.

[16] Lin, J.; El Amine Seddik, M.; Tamaazousti, M.; Tamaazousti, Y.; Bartoli, A. Deep multi-class adversarial specularity removal. In: *Image Analysis. Lecture Notes in Computer Science, Vol. 11482.* Felsberg, M.; Forssén, P. E.; Sintorn, I. M.; Unger, J. Eds. Springer Cham, 3–15, 2019.

[17] Muhammad, S.; Dailey, M. N.; Farooq, M.; Majeed, M. F.; Ekpanyapong, M. Spec-Net and Spec-CGAN: Deep learning models for specularity removal from faces. *Image and Vision Computing* Vol. 93, 103823, 2020.

[18] Xu, Y. F.; Wei, H. P.; Lin, M. X.; Deng, Y. Y.; Sheng, K. K.; Zhang, M. D.; Tang, F.; Dong, W.; Huang, F.; Xu, C. Transformers in computational visual media: A survey. *Computational Visual Media* Vol. 8, No. 1, 33–62, 2022.

[19] Liu, Z.; Lin, Y. T.; Cao, Y.; Hu, H.; Wei, Y. X.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 9992–10002, 2021.

[20] Maloney, L. T.; Wandell, B. A. Color constancy: A method for recovering surface spectral reflectance. *Journal of the Optical Society of America A* Vol. 3, No. 1, 29–33, 1986.

[21] Park, J. B.; Kak, A. C. A truncated least squares approach to the detection of specular highlights in color images. In: Proceedings of the IEEE International Conference on Robotics and Automation, 1397–1403, 2003.

[22] Meslouhi, O.; Kardouchi, M.; Allali, H.; Gadi, T.; Benkaddour, Y. Automatic detection and inpainting of specular reflections for colposcopic images. *Central European Journal of Computer Science* Vol. 1, No. 3, 341–354, 2011.

[23] Zhang, W. M.; Zhao, X.; Morvan, J. M.; Chen, L. M. Improving shadow suppression for illumination robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 41, No. 3, 611–624, 2019.

[24] Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research* Vol. 5, 1457–1469, 2004.

[25] Fu, G.; Zhang, Q.; Lin, Q. F.; Zhu, L.; Xiao, C. X. Learning to detect specular highlights from real-world images. In: Proceedings of the 28th ACM International Conference on Multimedia, 1873–1881, 2020.

[26] Tan, P.; Quan, L.; Lin, S. Separation of highlight reflections on textured surfaces. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1855–1860, 2006.

[27] Shen, H.-L.; Cai, Q.-Y. Simple and efficient method for specularity removal in an image. *Applied Optics* Vol. 48, No. 14, 2711, 2009.

[28] Shen, H. L.; Zheng, Z. H. Real-time highlight removal using intensity ratio. *Applied Optics* Vol. 52, No. 19, 4483–4493, 2013.

[29] Yang, J. W.; Liu, L. X.; Li, S. Z. Separating specular and diffuse reflection components in the HSI color space. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 891–898, 2013.

[30] Yang, Q. X.; Tang, J. H.; Ahuja, N. Efficient and robust specular highlight removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 37, No. 6, 1304–1311, 2015.

[31] Akashi, Y.; Okatani, T. Separation of reflection components by sparse non-negative matrix factorization. In: *Computer Vision – ACCV 2014. Lecture Notes in Computer Science, Vol. 9007.* Cremers, D.; Reid, I.; Saito, H.; Yang, M. H. Eds. Springer Cham, 611–625, 2015.

[32] Guo, J.; Zhou, Z. J.; Wang, L. M. Single image highlight removal with a sparse and low-rank reflection model. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11208.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 282–298, 2018.

[33] Fu, G.; Zhang, Q.; Song, C. F.; Lin, Q. F.; Xiao, C. X. Specular highlight removal for real-world images. *Computer Graphics Forum* Vol. 38, No. 7, 253–263, 2019.

[34] Nayar, S. K.; Fang, X. S.; Boult, T. Separation of reflection components using color and polarization. *International Journal of Computer Vision* Vol. 21, No. 163–186, 1997.

[35] Umeyama, S.; Godin, G. Separation of diffuse and specular components of surface reflection by use of polarization and statistical analysis of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 26, No. 5, 639–647, 2004.

[36] Wang, F.; Ainouz, S.; Petitjean, C.; Bensrhair, A. Specularity removal: A global energy minimization approach based on polarization imaging. *Computer Vision and Image Understanding* Vol. 158, 31–39, 2017.

[37] Wen, S.; Zheng, Y.; Lu, F. Polarization guided specular reflection separation. *IEEE Transactions on Image Processing* Vol. 30, 7280–7291, 2021.

[38] Sapiro, G. Color and illuminant voting. *IEEE Trans-*

actions on Pattern Analysis and Machine Intelligence Vol. 21, No. 11, 1210–1215, 1999.

[39] Imai, Y.; Kato, Y.; Kadoi, H.; Horiuchi, T.; Tominaga, S. Estimation of multiple illuminants based on specular highlight detection. In: *Computational Color Imaging. Lecture Notes in Computer Science, Vol. 6626*. Schettini, R.; Tominaga, S.; Trémeau, A. Eds. Springer Berlin Heidelberg, 85–98, 2011.

[40] Forsyth, D. A. A novel algorithm for color constancy. *International Journal of Computer Vision* Vol. 5, No. 1, 5–35, 1990.

[41] Hansen, T.; Olkkonen, M.; Walter, S.; Gegenfurtner, K. R. Memory modulates color appearance. *Nature Neuroscience* Vol. 9, No. 11, 1367–1368, 2006.

[42] Joze, H. R. V.; Drew, M. S. Exemplar-based color constancy and multiple illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 36, No. 5, 860–873, 2014.

[43] Lin, P.; Quan, L.; Shum, H.-Y. Highlight removal by illumination-constrained inpainting. In: Proceedings of the 9th IEEE International Conference on Computer Vision, 164–169, 2003.

[44] Tan, R. T.; Ikeuchi, K. Separating reflection components of textured surfaces using a single image. In: *Digitally Archiving Cultural Objects*. Springer Boston MA, 353–384, 2008.

[45] Tan, T. T.; Nishino, K.; Ikeuchi, K. Illumination chromaticity estimation using inverse-intensity chromaticity space. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, I, 2003.

[46] Funke, I.; Bodenstedt, S.; Riediger, C.; Weitz, J.; Speidel, S. Generative adversarial networks for specular highlight removal in endoscopic images. In: Proceedings of the SPIE 10576, Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling, 1057604, 2018.

[47] Wu, Z. Q.; Zhuang, C. Q.; Shi, J.; Xiao, J.; Guo, J. W. Deep specular highlight removal for single real-world image. In: Proceedings of the SIGGRAPH Asia 2020 Posters, Article No. 34, 2020.

[48] Wu, Z. Q.; Zhuang, C. Q.; Shi, J.; Guo, J. W.; Xiao, J.; Zhang, X. P.; Yan, D.-M. Single-image specular highlight removal via real-world dataset construction. *IEEE Transactions on Multimedia* Vol. 24, 3782–3793, 2022.

[49] Yi, R. J.; Tan, P.; Lin, S. Leveraging multi-view image sets for unsupervised intrinsic image decomposition and highlight separation. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 12685–12692, 2020.

[50] Fu, G.; Zhang, Q.; Zhu, L.; Li, P.; Xiao, C. X. A multi-task network for joint specular highlight detection and removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7748–7757, 2021.

[51] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351*. Navab, N.; Hornegger, J.; Wells, W.; Frangi, A. Eds. Springer Cham, 234–241, 2015.

[52] Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint* arXiv:2105.05537, 2021.

[53] Wei, K. X.; Yang, J. L.; Fu, Y.; Wipf, D.; Huang, H. Single image reflection removal exploiting misaligned training data and network enhancements. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8170–8179, 2019.

[54] Gatys, L. A.; Ecker, A. S.; Bethge, M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2414–2423, 2016.

[55] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556, 2014.

[56] Murmann, L.; Gharbi, M.; Aittala, M.; Durand, F. A dataset of multi-illumination images in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4079–4088, 2019.

[57] Guo, X. J.; Cao, X. C.; Ma, Y. Robust separation of reflection from multiple images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2195–2202, 2014.

[58] Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. In: Proceedings of the 3rd International Conference for Learning Representations, 2015.

[59] Yamamoto, T.; Kitajima, T.; Kawauchi, R. Efficient improvement method for separation of reflection components based on an energy function. In: Proceedings of the IEEE International Conference on Image Processing, 4222–4226, 2017.

[60] Hou, S.; Wang, C.; Quan, W.; Jiang, J.; Yan, D. M. Text-aware single image specular highlight removal. In: *Pattern Recognition and Computer Vision. Lecture Notes in Computer Science, Vol. 13022*. Springer Cham, 115–127, 2021.

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

Jianwei Guo is an associate professor in the National Laboratory of Pattern Recognition, 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.

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

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

Dong-Ming Yan is a professor in the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences. 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.

Xiaopeng Zhang received his Ph.D. degree in computer science from the Institute of Software, Chinese Academic of Sciences in 1999, where he is a professor. He received a National Scientific and Technological Progress Prize (second class) in 2004 and a Chinese Award of Excellent Patents in 2012. His main research interests include image processing, computer graphics, and computer vision.

Other papers from this open access journal are available free of charge from http://www.springer.com/journal/41095. To submit a manuscript, please go to https://www.editorialmanager.com/cvmj.