Bifurcated convolutional network for specular highlight removal*

XU Jingting, LIU Sheng, CHEN Guanzhou, and LIU Qianxi**

College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

(Received 20 February 2023; Revised 6 June 2023)

Specular highlight usually causes serious information degradation, which leads to the failure of many computer vision algorithms. We have proposed a novel bifurcated convolution neural network to tackle the problem of high reflectivity image information degradation. A two-stage process is proposed for the extraction and elimination of the specular highlight features, with the procedure starting at a coarse level and progressing towards a finer level, to ensure the generated diffuse images are less affected by visual artifacts and information distortions. A bifurcated feature selection module is designed to remove the specular highlight features, thereby enhancing the detection capability of the network. The experiments on two types of challenging datasets demonstrate that our method outperforms state-of-the-art approaches for specular highlight detection and removal. The effectiveness of the proposed bifurcated feature selection module and the overall network is also verified.

Document code: A **Article ID:** 1673-1905(2023)12-0756-6

DOI https://doi.org/10.1007/s11801-023-3029-6

Specular highlight, as a common phenomenon in the digital images, often presents as bright spots on the surface of high-reflective materials. Specular highlights usually conceal essential image features, such as colors, textures and structures. Due to the proximity of the camera light source and object surfaces, digital images always suffer from strong specular highlights, which can both negatively affect the visual quality and extremely degrade the subsequent tasks of computer vision algorithms, such as object detection and tracking, image segmentation and stereo reconstruction[1,2]. Therefore, specular highlight detection and removal technology plays an important role in the downstream tasks of digital image processing. It is further desirable to remove specular highlights while preserving the original color, structure and texture details of the objects for better perform.

In recent years, numerous specular highlight removal methods have been proposed in the literatures. Early methods usually perform color or shape segmentation, but are not robust to complex backgrounds and illumination conditions[3,4]. Subsequently, several methods based on the dichromatic reflection model have been presented. They are convenient to implement, but do not have the ability to distinguish the specular highlight regions from the white objects in the theory of color homeostasis[1,5]. To cope with the issues, people leverage deep learning based approaches and construct large-scale real-world datasets. In 2021, WU et al[6] presented a novel generative adversarial network (GAN) for specular highlight removal in a single image. In 2022, WANG et al[7] proposed a fully convolutional network for single image highlight removal with a real-world dataset. Compared with traditional approaches, deep learning based approaches perform better without the constraints of the traditional models. The state-of-the-art methods, such as WU et al[6], have made great progress but still have some shortcomings, as illustrated in Fig.1. The ability of specular highlight detection in the high-brightness and low-saturation regions is weak. Indicated in the first row of Fig.1, the specular highlights cannot be detected when there is a white background in the environment. The high-brightness and low-saturation areas in the image are mistakenly recognized as specular highlights generated by WU et al[6], resulting in the overall darkening of the image. However, our method can distinguish the low-saturation and highlight areas. The removal of specular highlights is incomplete, as shown by the red box of the second row in Fig.1. The large specular highlight regions are not completely removed. The compensated areas in the generated diffuse images are damaged by visual artifacts and information distortion in color, structure, and texture. As shown in the red box of the third row in Fig.1, there are visual artifacts in the compensation area generated, but our results are much more realistic.

* This work has been supported by the National Key R&D Program of China (No.2018YFB1305200).

** E-mail: edliu@zjut.edu.cn

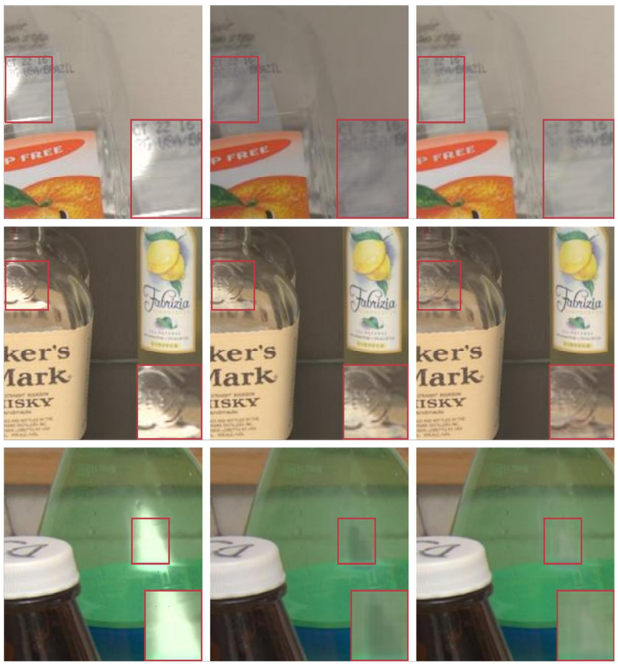

Fig.1 Visual comparison of specular highlight removal on SHIQ dataset[8]

We reported the preliminary research results as the oral presentation of the international conference on intelligent robotics and applications (ICIRA)[9]. Now in this paper, we have made some expansions and modifications.

Our network is divided into two stages to remove specular highlights from coarse to fine. The proposed specular highlight reflection model is expressed as

$$I(X)=D_{\text{fine}}(X)+S_{\text{fine}}(X)= D_{\text{coarse}}(X)\otimes\alpha\left(S_{\text{coarse}}(X)\,\|\,S_{\text{fine}}(X)\right)+S_{\text{fine}}(X), \quad (1)$$

where $D_{\text{fine}}(X)$ and $S_{\text{fine}}(X)$ denote the fine diffuse image and specular highlight feature mask image, $D_{\text{coarse}}(X)$ and $S_{\text{coarse}}(X)$ denote the coarse diffuse image and specular highlight feature mask image, $\otimes$, $\|$ and α denote convolution, concatenation and convolution operation, respectively. The architecture of our network is illustrated in Fig.2.

We take a real image r as the input. In the first stage of the network, we propose a bifurcated feature selection module combining an attention mechanism of convolutional block attention module (CBAM)[10], to generate $D_{\text{coarse}}(X)$ and $S_{\text{coarse}}(X)$. The specular highlight feature extraction module (SFE) is denoted as

$$F_{\text{diff}}=F-F_{\text{spec}}=\rho(r)-\gamma\left(\rho(r)\right), \quad (2)$$

which subtracts highlight feature map F_{spec} from original feature map F to generate diffuse feature map F_{diff}. ρ and γ are the processes of image preprocessing and highlight features extraction, respectively.

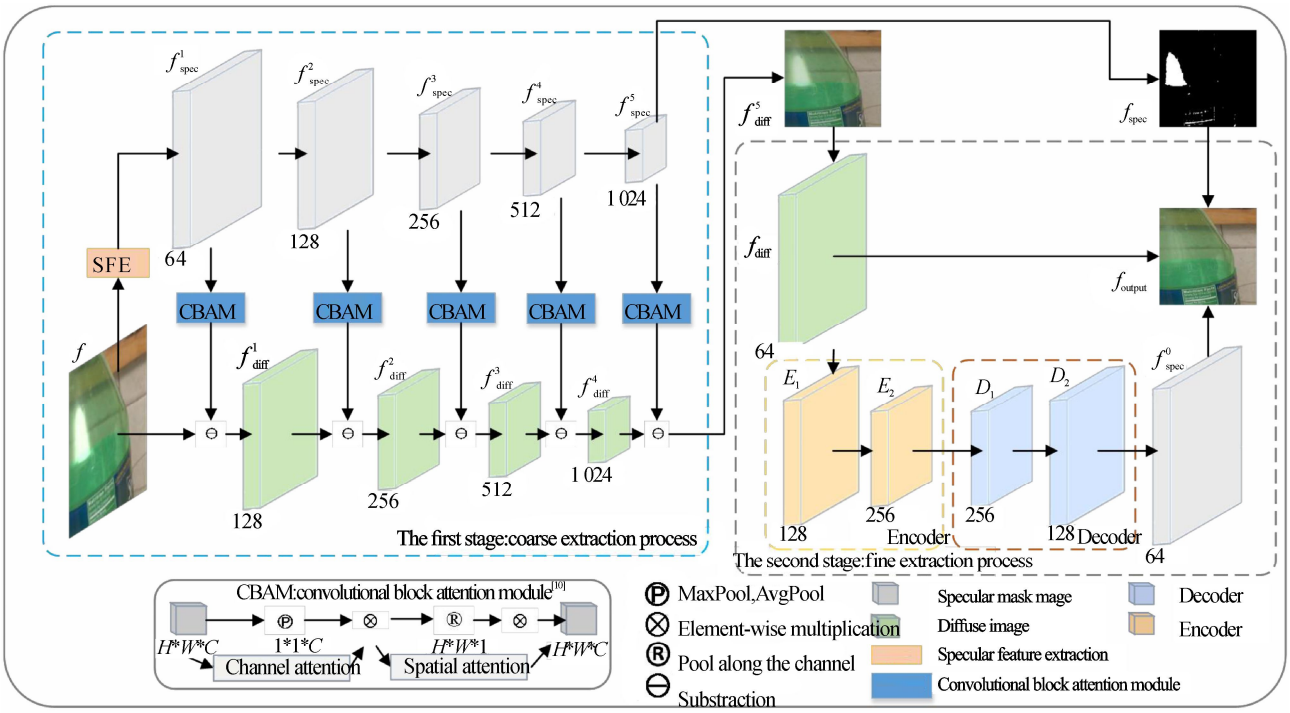

Fig.2 Architecture of the proposed network

We extract adjacent pixels with large changes in pixel values as highlight features, and roughly remove them in the original features with explicitly leveraging subtraction operation. Based on the output of SFE, we use bifurcated feature selection module combined with CBAM to extract five levels of highlight features, expressed as

$$f_{\text{spec}}^{i}=\begin{cases}\gamma\left(\rho(r)\right), & i=1\\ \beta\left(f_{\text{spec}}^{i-1}\right), & i=2,3,4,5\end{cases}, \tag{3}$$

where β represents the convolution. Specular highlight feature f_{spec}^{i} is derived from the convolution operation of f_{spec}^{i-1}. As shown in Fig.2, for example, we perform convolution calculations on f_{spec}^{1} to obtain f_{spec}^{2}. And then the result of processing CBAM on f_{spec}^{2} is subtracted from f_{diff}^{1} to get f_{diff}^{2}. Like this, we filter out specular highlight features from original features layer by layer and finally obtain the coarse diffuse image f_{diff}^{5} and specular highlight feature mask image f_{spec}.

The detailed information in low-level features is essential to generate fine diffuse images. In contrast, the global context information contained in high-level features is conducive to locate specular highlights. Inspired by this, we use the details of low-level features to supplement high-level features, to improve the effectiveness and robustness of specular highlights removal.

In the second stage of the network, we make full use of the information acquisition capability of the fully convolutional neural network. We design the encoder and decoder structures to remove specular highlights in the coarse diffuse image f_{diff} and output the fine diffuse image f_{output}. The core components of the encoder are the dilated convolution layers. The decoder is a combination of the up-sampling and convolutional layers. After obtaining the detailed highlight mask image, we assign different weights to different features, and retain useful features through gated convolution and residual calculation. The features of each layer are fused without information pollution between highlight and diffuse regions.

We exploit two prediction losses to calculate the total loss L of the training process, which is given by

$$L=\lambda_1 L_{\text{dete}}+\lambda_2 L_{\text{rem}}, \tag{4}$$

where we empirically set λ_1=1.0 and λ_2=0.5. L_{dete} and L_{rem} denote the training loss of specular highlight detection and removal, which can be formulated as

$$\begin{aligned}L_{\text{dete}}\left(S_{\text{coarse}},S_{\text{detailed}},S_0\right)&=w_1 L_{\text{focal}}\left(S_{\text{coarse}},S_{\text{detailed}}\right)+\\&\quad w_2 L_{\text{focal}}\left(S_{\text{detailed}},S_0\right),\\ L_{\text{rem}}\left(D_{\text{coarse}},D_{\text{detailed}},D_0\right)&=w_1 L_{\text{pixel}}\left(D_{\text{coarse}},D_{\text{detailed}}\right)+\\&\quad w_2 L_{\text{pixel}}\left(D_{\text{detailed}},D_0\right).\end{aligned} \tag{5}$$

We set w_1=0.5, w_2=1.0 according to existing studies[11,12]. Given the input image with ground-truth of specular highlight mask image S_0 and diffuse image D_0, the network outputs two specular highlight mask images S_{coarse}, S_{detailed} and two diffuse images D_{coarse}, D_{detailed}. We use the focal loss[9] L_{focal} to train the network to detect specular highlights, which maintains good performance even when processing high-brightness, low-saturation pixels. In order to minimize visual artifacts and information distortion problems of the image caused by highlight removal, we also use a pixel loss[9] L_{pixel}.

We train and test our network on two NVIDIA TITAN RTX graphics processing units (GPUs). To train our model, we randomly divide the dataset into two sets: 10k for training and 3k for testing. We resize the input images to 256 by 256 pixels. The initial learning rate is set to 10^{-4}, and is multiplied by 0.2 after every 5 epochs in the first 10 epochs. With batch size of 4, the whole training process requires nearly 22 h.

To validate the effectiveness of our algorithm, we conduct experiments on two datasets (PSD[6] and SHIQ[8]) that have ground-truths. We compare with six state-of-the-art highlight detection and removal methods (Multi-class generative adversarial network (GAN)[13], Spec-CGAN[14], SHEN et al[1], YAMAMOTO et al[15], FU et al[8], and WU et al[6]). For quantitative evaluation, we adopt three commonly used metrics including mean-squared error (*MSE*), structural similarity index (*SSIM*), and peak signal to noise ratio (*PSNR*). The statistical results are shown in Tab.1. In general, lower *MSE*, higher *PSNR* and *SSIM* scores indicate better removal results. The best and second best results are highlighted in bold and underlined formats, respectively.

Apparently, our method outperforms all others in *MSE* and *PSNR*. Since the method from WU et al[6] has strong adaptability to PSD dataset they proposed, it performs slightly better than ours on PSD dataset in *SSIM*. However, its performance on SHIQ dataset is obviously inferior to ours in *MSE*, *PSNR* and *SSIM*. According to the experimental findings, our proposed network has exhibited a higher level of performance compared to all the other methods that were evaluated.

Tab.1 Quantitative comparison on PSD[6] and SHIQ[8]

Dataset	PSD[6]			SHIQ[8]		
Metric / Method	*MSE*/10^{-2}↓	*PSNR*↑	*SSIM*↑	*MSE*/10^{-2}↓	*PSNR*↑	*SSIM*↑
Full	**0.08**	**32.95**	0.97	**0.11**	**31.68**	**0.97**
[6]	0.14	3	**0.99**	0.24	28.23	0.94
[8]	0.22	0.46 29.32	0.92	0.35	28.17	0.86
[13]	0.50	23.52	0.91	0.48	27.63	0.88
[14]	0.36	25.70	0.86	0.42	26.44	0.85
[1]	1.07	20.62	0.88	1.12	21.90	0.80
[15]	8.46	11.85	0.62	4.76	19.54	0.63

Then we conduct comparison on the real-world images from SHIQ in terms of visual inspection. Since PSD dataset has no corresponding ground-truth mask images of specular highlight detection and is composed of laboratory images, we do not evaluate on PSD dataset. To fully perform the comparison, we choose SHIQ dataset with ground-truth mask and diffuse images. Compared with the most advanced method presented by WU et al[6], the promising results produced by our method indicate the effectiveness of ours. We evaluate our method on

SHIQ dataset (with ground-truth mask and diffuse images) and select three images as shown in Fig.3. Our method is capable of producing high-quality results that have fewer occurrences of visual artifacts and information distortions in the compensated area, as is evident. The final specular highlight mask images generated by our method successfully detect most of the regions of specular highlights, even in challenging cases such as high reflectivity and low saturation. Additionally, the diffuse images produced by our method are more accurate compared to existing advanced methods, and they closely resemble the ground-truth.

To further analyze how each component contributes to the final performance of our designed network, the networks N1, N2, N3 and N4 composed of different components are retrained. N1 represents a new network that replaces the subtraction operation in our network with the addition operation. N2 abandons our bifurcated feature selection module of our complete network. N3 extracts features from one layer less. Similarly, N4 takes from one layer more. From Tab.2 (the best result of each measurement is highlighted in bold), we can see that our complete network achieves a quantity superior over N1, N2, N3 and N4. The higher *PSNR*, higher *SSIM* performance and lower *MSE* value of our network prove it has the best perform ance of specular highlight detection and removal.

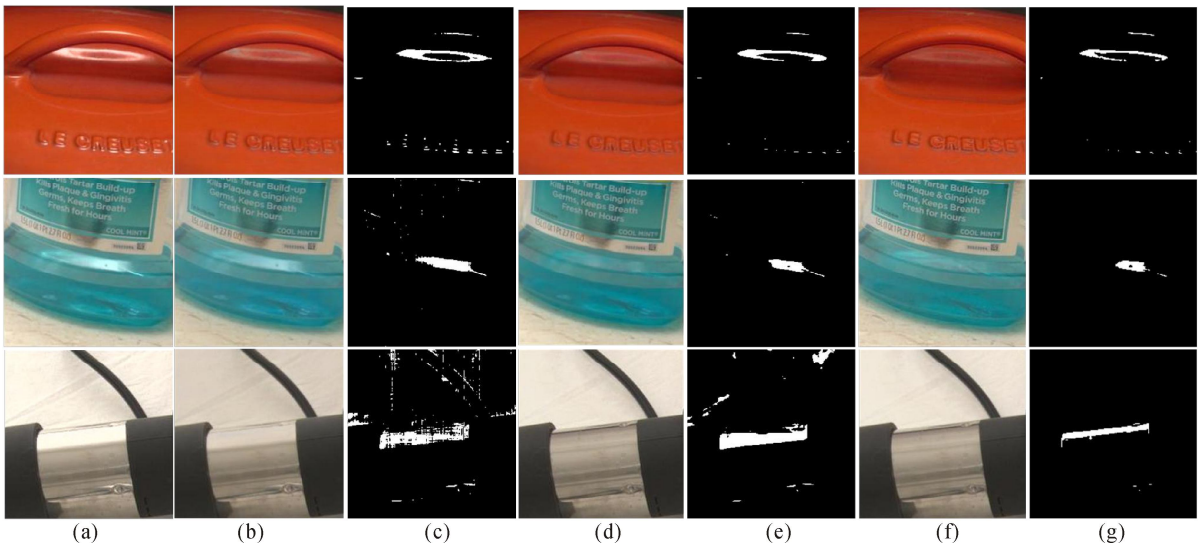

Fig.3 Specular highlight removal results on SHIQ dataset[8]: (a) Input specular highlight images; (b) Generated diffuse images by WU et al[6]; (c) Generated specular highlight mask images by WU et al[6]; (d) Generated diffuse images by our method; (e) Generated specular highlight mask images by our method; (f) Ground-truths of diffuse images; (g) Ground-truths of specular highlight mask images

Tab.2 Quantitative comparison on SHIQ dataset[8]

Method	*MSE*/10^{-2}↓	*PSNR*↑	*SSIM*↑
Ours	**0.08**	**32.95**	**0.97**
N1	0.29	25.22	0.93
N2	0.20	26.20	0.94
N3	0.11	29.91	0.95
N4	0.16	30.14	0.96

N1 replaces the subtraction operation with addition operation in the first stage. Addition operation is commonly used in salient feature detection to enhance feature representation[16]. This paper argues that subtraction operation can filter out highlight features. This view is proved by the experimental results of N1. As shown in Tab.2 and Fig.4(b), the highlight removal effect of N1 is not good, which reversely verifies the effectiveness of our design. Addition operation cannot remove the highlight well. Subtraction operation is more suitable for highlight removal applications, which represents the design idea of roughly removing highlight features to generate preliminary diffuse images.

N2 abandons the coarse extraction process of specular highlight features based on our proposed bifurcated feature selection module. In this module, we extract highlight features at different levels. High-level features in an image provide rich information of the overall context, which is helpful in identifying specular highlight areas. On the other hand, low-level features carries a lot of detailed information, which is great for produce refined diffuse images[16]. To take advantage of both, we use the detailed information from low-level features to enhance the high-level features and gradually separate out the specular highlight features from the original features layer by layer. Assuming that the coarse extraction process is removed, we will get the experimental results as shown in Fig.4(c), which are far from the ground-truths. The results in Tab.2 also illustrate that N2 has weaker performance than the complete network.

N3 sets the coarse feature extraction module of our bifurcated-convolutional neural network (CNN) to a

four-layer structure. In the first stage of our network, the highlight features in the original features are removed layer by layer. As the number of layers increases, the size of the image becomes smaller. Although the specular highlights in the high-level feature map are obviously removed, a lot of detail information is lost. The low-level feature map retains the detailed features completely, but there are obvious specular highlight residues. Therefore, the number of layers should be set to the appropriate value.

N4 sets the coarse feature extraction module of our bifurcated-CNN to a six-layer structure. We determined the appropriate number of layers through comparative experiments. The results of Tab.2 demonstrate that the number of layers should be five. Therefore, our method uses a five-layer structure, which is better than the four-layer structure N3 and the six-layer structure N4.

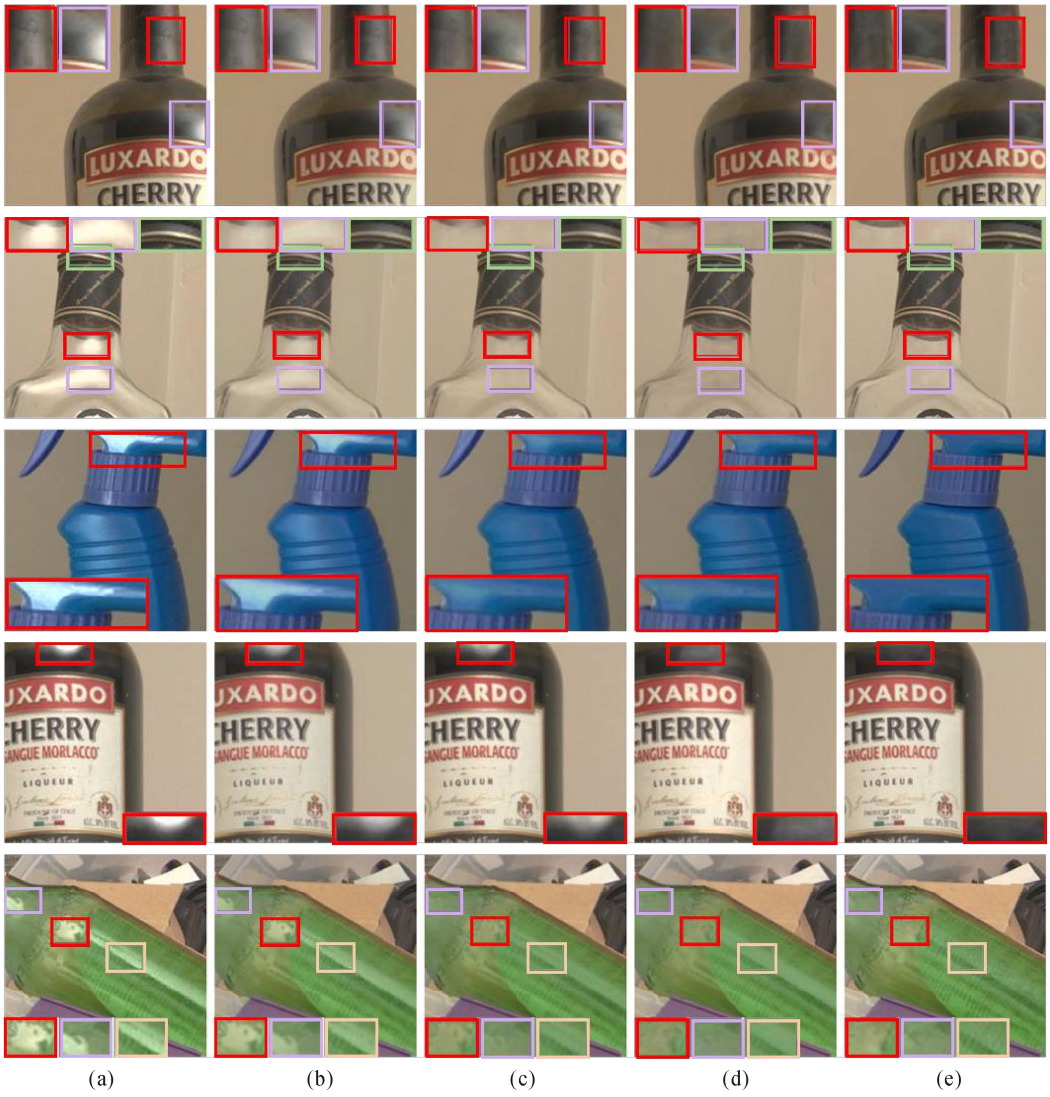

Fig.4 Visual comparison on each key component of our designed network (We enlarge several highlight regions in the figure): (a) Input specular highlight images; (b) Generated diffuse images by N1; (c) Generated diffuse images by N2; (d) Generated diffuse images by our complete network; (e) Ground-truths of diffuse images

As shown in Tab.2 and Fig.4, we can see that the complete implementation of our network performs better. The ablation studies based on N1, N2, N3 and N4 validate that each component contributes greatly to the final performance of our designed network. As a complete structure, our network achieves the best results and has the best practicality and reliability for highlight removal applications.

This paper introduces a specular highlight removal network that aims to address the issue of degraded image information caused by high reflectivity materials. The network has a two-stage implementation, with a bifurcated feature selection module in the first stage that enhances the accuracy and robustness of highlight detection and removal. In addition, multi-scale highlight features are extracted to improve the ability of the network to handle specular highlights at different scales. The second stage achieves refinement removal by processing the specular highlight features roughly extracted in the first stage. To prevent information pol-

lution between the highlight and diffuse regions, the network utilizes gated convolution and residual calculation operations. The resulting diffuse images have less visual artifacts and information distortions. Experimental results show that the proposed method outperforms the state-of-the-arts.

Ethics declarations

Conflicts of interest

The authors declare no conflict of interest.

References

[1] SHI J, DONG Y, SU H, et al. Learning non-Lambertian object intrinsics across ShapeNet categories[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York：IEEE, 2017：1063-6919.

[2] WEI X, XU X B, ZHANG J W, et al. Specular highlight reduction with known surface geometry[J]. Computer vision and image understanding, 2018, 168：132-144.

[3] YANG Q X, TANG J H, AHUJA N. Efficient and robust specular highlight removal[J]. IEEE transaction on pattern analysis and machine intelligence, 2014, 37(6)：1304-1311.

[4] WEI X, XU X B, ZHANG J W, et al. Specular highlight reduction with known surface geometry[J]. Computer vision and image understanding, 2018, 168：132-144.

[5] LI R Y, PAN J J, SI Y Q, et al. Specular refmections removal for endoscopic image sequences with adaptive-RPCA decomposition[J]. IEEE transactions on medical imaging, 2020, 39(2)：328-340.

[6] WU Z Q, ZHUANG C Q, SHI J, et al. Single-image specular highlight removal via real-world dataset construction[C]//Proceedings of the IEEE Transactions on Multimedia, August 27, 2021. New York：IEEE, 2021：3782-3793.

[7] WANG X C, TAO C N, TAO X, et al. SIHRNet：a fully convolutional network for single image highlight removal with a real-world dataset[J]. Journal of electronic imaging, 2022, 31：033013.

[8] FU G, ZHANG Q, ZHU L, et al. A multi-task network for joint specular highlight detection and removal[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, Nashville, TN, USA. New York：IEEE, 2021：7748-7757.

[9] XU J T, LIU S, CHEN G Z, et al. Highlight detection and removal method based on bifurcated-CNN[C]//Proceedings of the EI Conference on Intelligent Robotics and Applications, August 10, 2022, Harbin, China. New York：EI, 2022：307-318.

[10] WOO S, PARK J, LEE J Y, et al. CBAM：convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision, August 22-25, 2018, Munich, Germany. Berlin, Heidelberg：Springer-Verlag, 2018：3-19.

[11] ZHANG X, NG R, CHEN Q. Single image reflection separation with perceptual losses[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 18-22, 2018, Salt Lake City, UT, USA. New York：IEEE, 2018：4786-4794.

[12] WEI K, YANG J, FU Y, et al. Single image reflection removal exploiting misaligned training data and network enhancements[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, July 16-20, 2019, Long Beach, CA, USA. New York：IEEE, 2019：8178-8187.

[13] LIN J, EL AMINE SEDDIK M, TAMAAZOUSTI M, et al. Deep multi-class adversarial specularity removal[C]//Proceedings of the 21st Scandinavian Conference on Image Analysis, June 11-13, 2019, Norrköping, Sweden. Berlin, Heidelberg：Springer-Verlag, 2019：11482.

[14] FUNKE I, SEBASTIAN B, CARINA R, et al. Generative adversarial networks for specular highlight removal in endoscopic images[C[//Proceedings of the SPIE, March, 2018, Houston, Texas, USA. Washington：SPIE, 2018：10576(9).

[15] YAMAMOTO T, KITAJIMA T, KAWAUCHI R. Efficient improvement method for separation of reflection components based on an energy function[C]//Proceedings of the IEEE International Conference on Image Processing, September 17-20, 2017, Beijing, China. Beijing：IEEE, 2017：4222-4226.

[16] FAN D P, ZHAI Y G, ALI B, et al. BBS-Net：RGB-D salient object detection with a bifurcated backbone strategy network[C]//Proceedings of the European Conference on Computer Vision, October 23-27, 2020, Tel Aviv, Israel. Berlin, Heidelberg：Springer-Verlag, 2020：275-292.