International Journal of Computer Vision (2024) 132:95–117
https://doi.org/10.1007/s11263-023-01845-3

Towards High-Resolution Specular Highlight Detection

Gang Fu[1] · Qing Zhang[2] · Lei Zhu[3,4] · Qifeng Lin[5] · Yihao Wang[1] · Siyuan Fan[1] · Chunxia Xiao[1]

Received: 2 February 2022 / Accepted: 3 July 2023 / Published online: 23 August 2023

Abstract
Specular highlight detection is an essential task with various applications in computer vision. This paper aims to detect specular highlights in single high-resolution images using deep learning while avoiding excessive GPU memory consumption. To achieve this, we present a high-resolution specular highlight detection dataset with manual annotations of specular highlights. Given our dataset, we propose a patch-level bidirectional refinement network for high-resolution specular highlight detection. The main idea is to utilize both the pathway from small-scale patch to large-scale patch and its reverse pathway to progressively refine the detection results of adjacent-scale specular highlight patches. Moreover, based on our detection network, we propose a modified inpainting framework for specular highlight removal as an application. Lastly, we provide ten potential research directions for specular highlight detection, inspiring researchers for further study.

Keywords Specular highlight detection · High-resolution image processing · Specular highlight removal

Communicated by Shaodi You.

✉ Chunxia Xiao
cxxiao@whu.edu.cn

Gang Fu
xyzgfu@gmail.com

Qing Zhang
zhangqing.whu.cs@gmail.com

Lei Zhu
leizhu@ust.hk

Qifeng Lin
linqf@fzu.edu.cn

Yihao Wang
harriswang@whu.edu.cn

Siyuan Fan
whdxfsy@whu.edu.cn

1 School of Computer Science, Wuhan University, Wuhan 430072, Hubei, China

2 School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou 510006, Guangdong, China

3 ROAS Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511400, Guangdong, China

4 Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

5 College of Computer and Data Science, Fuzhou University, Fuzhou 350116, Fujian, China

1 Introduction

A specular highlight is a bright and concentrated region of light that appears on a surface due to the reflection of incident light rays. As a common physical phenomenon observed in natural scenes, it occurs when light hits a smooth and shiny surface, such as plastic or leather, and reflects off at a particular angle known as the specular reflection angle. Specular highlights often appear as a distinct shape, such as curved or elongated strip, and thus provide visual cues about the reflective properties, surface curvature, and shape of the material or object. In computer vision, detecting and analyzing specular highlights can be useful for various applications such as glossy object recognition (Osadchy et al., 2003; Netz and Osadchy, 2012), light source estimation (Zhang et al., 2017), and intrinsic image decomposition (Shi et al., 2017). In this paper, we mainly focus on strong (i.e., high-intensity) specular highlights with clear boundaries to address a tractable problem, like existing specular highlight detection methods (Fu et al., 2020; Li et al., 2019; Zhang et al., 2018).

Most existing traditional specular highlight detection methods are based on either the premise that specular highlights in a scene are sparse in distribution and small in size (Zhang et al., 2018) or the premise that specular highlights in a scene have the highest intensities (Li et al., 2019). However, these traditional methods often suffer from one or both of the following two drawbacks. First, they fail to eliminate the semantic ambiguity between specular highlights

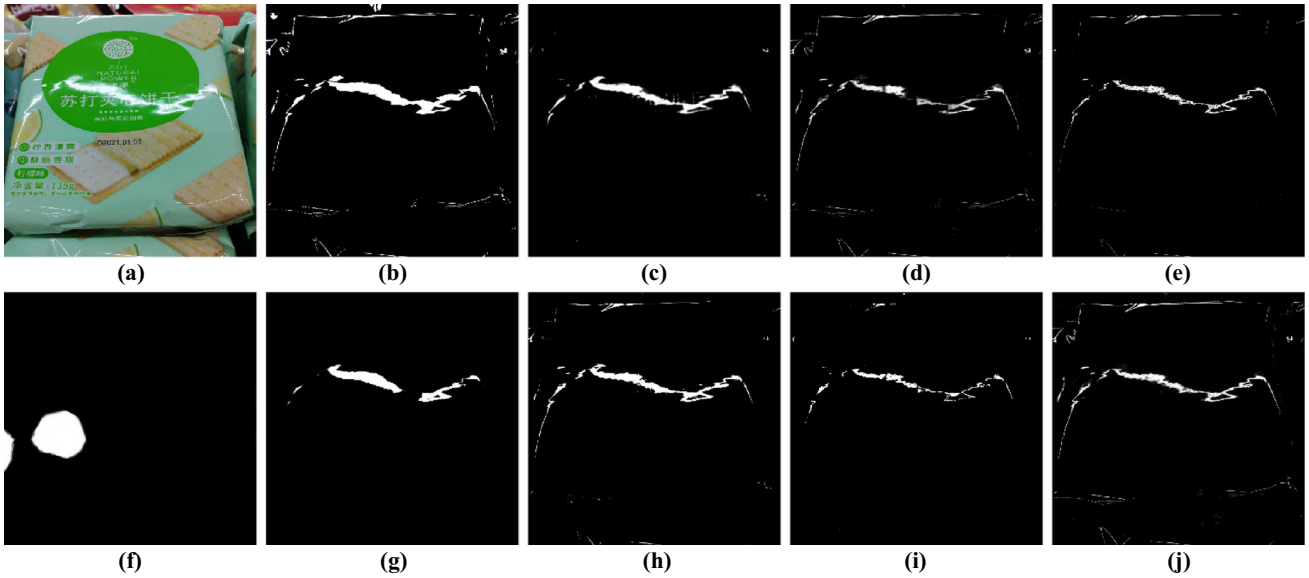

Fig. 1 Visual comparison of our PBRNet against state-of-the-art detection and segmentation methods. **a** Input. **b** Ground truth. **c** Downsampling. **d** Patch processing. **e** SHDNet (Fu et al. 2020). **f** DSC (Hu et al., 2019). **g** DSS (Hou et al., 2019). **h** HRSODNet (Zeng et al., 2019). **i** MagNet (Huynh et al., 2021). **j** Ours

and white material surfaces. Second, they often have significantly different optimal thresholding values for images in which specular highlights have a wide range of intensities. To overcome these limitations, Fu et al. (2020) recently proposed a deep learning-based method by leveraging multi-scale context contrasted features to detect specular highlights of various sizes and shapes. However, it is less able to finely detect specular highlight boundaries and very small-sized specular highlights in high-resolution images (see Fig. 1e).

Images taken by modern electronic products like smartphones or cameras are often of high resolution (e.g., 4K or higher). However, to the best of our knowledge, there is currently no work specifically focused on addressing specular highlight detection in high-resolution images. This line of work is crucial, since it can significantly benefit the applications aiming to produce high-quality results for such images. A common application is specular highlight removal, where specular highlight detection provides it with the location information of specular highlights (Fu et al., 2021). Inaccurate detection may lead to noticeable artifacts in removal results, thereby negatively affecting subsequent image manipulation and processing by users. To this end, we in this paper aim to develop a deep learning-based method that can accurately locate specular highlights in high-resolution images without overloading GPU memory.

There are generally two straightforward workarounds to address high-resolution images. One (downsampling) is to downsample the input high-resolution image to its low-resolution version for processing, while the other (patch processing) is to divide the input high-resolution image into a series of patches for separate processing. However, the former disregards local information from high-resolution specular highlights, and thus fail to locate local specular highlights (see Fig. 1c). Meanwhile, the latter disregards global information from the whole image, and thus may incur errors from adverse factors such as white material surfaces and overexposure (see Fig. 1d). Furthermore, researchers have proposed to make full use of both global and local information for high-resolution saliency detection (Zheng et al., 2019; Zhang et al., 2021) and semantic segmentation (Chen et al., 2019; Cheng et al., 2020). Although they have achieved good results, their performance is somewhat limited due to the huge scale gap between global information from the whole image and local information from local patches. Recently, Huynh et al. (2021) proposed a progressive semantic segmentation method that performs the coarse-to-fine information propagation at multiple processing stages. However, this method does not fully explore information integration from different pathways.

Since there is currently no high-resolution specular highlight detection dataset, we build a High-Resolution Specular Highlight Detection dataset (HRSHD) to support network training and evaluation. It consists of 3619 real images, each with a manually annotated specular highlight mask, which covers various materials and illumination conditions in our daily life. Based on our dataset, we also propose a high-resolution specular highlight detection method, Patch-level Bidirectional Refinement Network (PBRNet). Specifically, given an input image, we progressively refine the detection results of its local specular highlight patches of adjacent scales in two pathways simultaneously. One pathway is from small-scale patch to large-scale patch, while the other is from

large-scale patch to small-scale patch. Then, we fuse the last output refined results from these two pathways to yield the final detection results for a minimum-scale specular highlight patch. Finally, we combine the detection results of all specular highlight patches into the whole specular highlight mask. Figure 1 presents an example for visual comparison.

To sum up, our main contributions are as follows:

- We present a high-resolution specular highlight detection dataset, consisting of 3619 real images, each with a manually annotated specular highlight mask.
- We propose a patch-level bidirectional refinement network for high-resolution specular highlight detection, which is able to eliminate the huge scale gap between global information from the whole image and local information from local patches.
- We propose a modified inpainting framework for specular highlight removal as an application of our detection method, which is able to recover missing colors and texture details underneath high-intensity and large-area specular highlights.
- We provide ten potential research directions for specular highlight detection. We find that specular highlight detection is still far from being solved, leaving substantial room for improvement in this field.

2 Related Work

This section first presents single-image specular highlight detection methods, and single-image detection/segmentation methods in closely related fields including saliency detection, shadow detection, and semantic segmentation. Next, we briefly review specular highlight removal and intrinsic image decomposition methods, which can benefit from our detection method (see Sect. 7). Finally, we summarize and discuss the available datasets for specular highlight detection.

Specular Highlight Detection In the early days, Brelstaff and Blake (1988) proposed a method based on the characterization of Lambertian surfaces. Bajcsy et al. (1990) proposed a thresholding-based method to detect specular highlights by observing variations in saturation. Later, Park and Kak (2003) proposed a truncated least squares method for specular highlight detection. Tian and Clark (2013) proposed a method based on an unnormalized version of the Wiener Entropy which is commonly used in audio spectral analysis. Angelopoulou (2007) proposed a physics-based method depending on the Fresnel term of specular highlight. Based on an observation that specular highlights are often small in size, Zhang et al. (2018) formulated specular highlight detection as Non-negative Matrix Factorization (NMF) (Hoyer, 2004). Li et al. (2019) proposed a thresholding-based method to detect specular highlights in HSV channels. However, these traditional methods fail to disambiguate specular highlights from adverse background regions such as white material surfaces and overexposure. Fu et al. (2020) proposed a deep learning-based method that utilizes multi-scale context contrasted features to locate specular highlights of different scales. However, it only works well on low-resolution images. In contrast, our method can produce high-quality detection results with finer locations and boundaries of specular highlights for high-resolution images under memory constraints.

Saliency Detection It aims to locate the most visually distinctive objects/regions in a scene. Recently, Qin et al. (2019) proposed a boundary-aware salient object detection network with a new hybrid loss. Hou et al. (2019) proposed a deeply supervised salient object detection network with short connections, making full use of multi-level and multi-scale features. Zhou et al. (2020) discussed the correlation between saliency and contour, and proposed an interactive two-stream network. However, these methods are primarily designed for handling low-resolution images. Recently, researchers have proposed some high-resolution saliency detection methods. Zeng et al. (2019) proposed a method incorporating global semantic information and local high-resolution details. Xie et al. (2022) proposed a pyramid grafting network that utilizes Transformer and CNN backbone to extract features from images of different resolutions independently and then graft the features from Transformer branch to CNN branch. However, these two methods do not explore the huge scale gap between global information from the whole image and local information from local patches.

Shadow Detection It aims to detect cast shadows in a scene. Recently, Hu et al. (2018) proposed leveraging direction-aware spatial context features to detect shadows. Zheng et al. (2019) proposed a distraction-aware shadow detection network by learning and integrating the semantics of visual distraction regions. Zhu et al. (2018) proposed a detection framework that integrates global context features in deep layers and local context features in shallow layers in a deep convolutional neural network. Wang et al. (2020) presented a new task of instance shadow detection aiming to identify shadow instances associated with object instances, and further proposed a light-guided instance shadow-object association framework. However, these methods may not be suitable for our task, since specular highlights have distinct imaging conditions and appearances compared to shadows.

Semantic Segmentation It aims to classify and label each pixel of an image into meaningful semantic categories. Recently, Lin et al. (2020) proposed a graph-guided architecture search framework to automatically search segmentation networks. Yang and Soatto (2020) proposed a method based on Fourier Transform. Sun et al. (2020) treated semantic

Fig. 2 Visual comparison between our HRSHD and WHU-Specular (Fu et al., 2020). **a** Image pairs from WHU-Specular. **b** Image pairs from HRSHD. Please zoom in to compare the detailed difference between them

segmentation as a sequence-to-sequence prediction using Transformers. In addition, various high-resolution semantic segmentation methods have been proposed. Zhou et al. (2020) proposed to refine the boundaries in high-resolution images by leveraging coarse low-resolution masks. Wu et al. (2020) introduced a patch proposal network that adaptively selects critical patches to further refine. Cheng et al. (2020) proposed a network that refines and corrects local boundaries whenever possible, without relying on high-resolution training data. Huynh et al. (2021) proposed a progressive refinement network involving multiple processing stages, where the coarse result in the former stage is fed into the next stage for refinement in a coarse-to-fine way. Although this method effectively integrates information from large-scale patch to small-scale patch, it does not consider the refinement in its reverse pathway (i.e., from small-scale patch to large-scale patch). In comparison, our method can integrate information in both the two pathways simultaneously, resulting in improved performance.

Specular Highlight Removal It aims to remove specular highlights in the scene while recovering colors and texture details underneath specular highlights. Early traditional methods often rely on intensity or chromaticity analysis, as well as optimization techniques. Tan and Ikeuchi (2005) proposed a method solely based on chromaticity, without explicit color segmentation sensitive to edges. Shen and Zheng (2013) proposed a real-time method using intensity ratio. Kim et al. (2013) proposed a method that utilizes the dark channel prior for natural images. Yang et al. (2015) formulated specular highlight removal as an iterative bilateral filtering process. Based on the dichromatic reflection model (Shafer, 1985), Akashi and Okatani (2015) formulated specular highlight removal as non-negative matrix factorization. Recently, researchers proposed various deep learning-based methods. Fu et al. (2021) built a real dataset

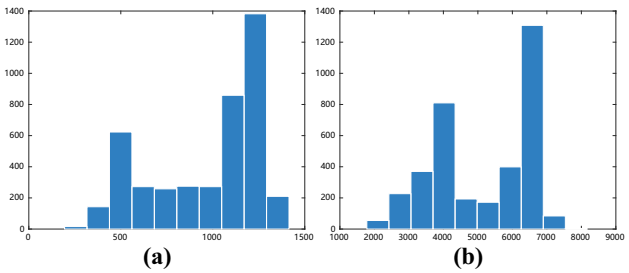

Fig. 3 Comparison of resolution distribution. **a** The histogram of diagonal length ($\leq$ 1415p) on WHU-Specular (Fu et al., 2020). **b** The histogram of diagonal length ($\geq$ 1781p) on our HRSHD

simultaneously for specular highlight detection and removal based on a multi-illumination dataset IIW (Murmann et al., 2019), and proposed a multi-task network for joint specular highlight detection and removal. Meanwhile, Wu et al. (2021) built a real specular highlight removal dataset using cross-polarization photography, and proposed a GAN-based specular highlight removal network. Despite achieving good results, they fail to effectively recover the missing information, particularly underneath high-intensity and large-area specular highlights, due to the lack of meaningful and reliable context information. To overcome it, we propose a modified inpainting framework for specular highlight removal, which is effective to address previously challenging images.

Intrinsic Image Decomposition It aims to separate an image into its intrinsic reflectance (also referred as albedo) and shading components. Early methods often formulate the decomposition problem as an optimization framework by integrating physical priors or constraints on reflectance and shading. Shen et al. (2008) utilized non-local texture constraints in conjunction with gradient separation cues from color-based Retinex (Grosse et al. 2009) for the formulation. Shen et al. (2011) proposed a method based on the

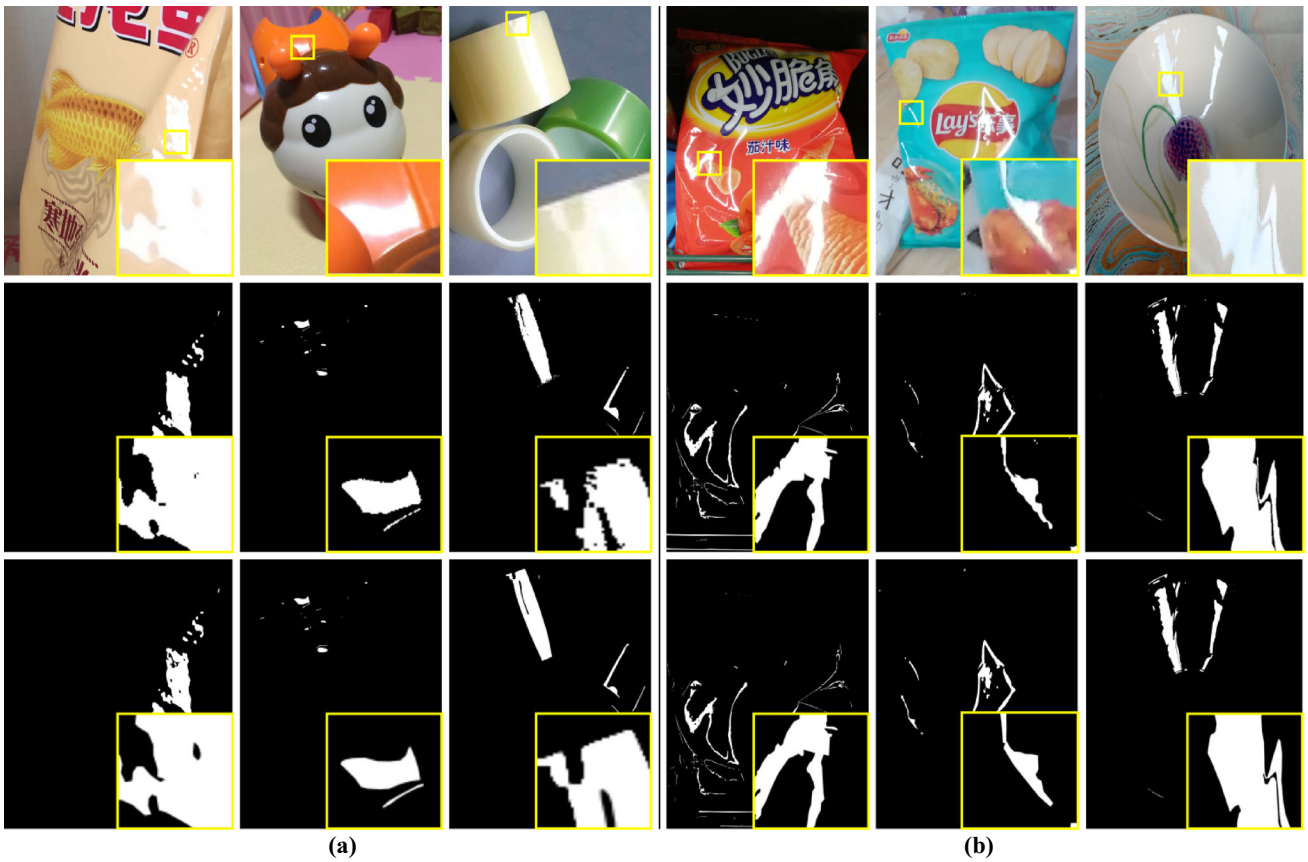

Fig. 4 Visual comparison of annotation quality. **a** WHU-Specular (Fu et al., 2020). **b** Our HRSHD. From the first row to the last row: specular highlight images, corresponding specular highlight masks in the datasets, and newly annotated specular highlight masks by additional experts as references, respectively

assumption that neighboring pixels with similar intensity values in a local image window should have similar reflectance values. Bell et al. (2014) developed a dense CRF-based method by exploring long-range material interactions for images in the wild. In addition, Grosse et al. (2009) presented an object-level benchmark dataset, and performed the quantitative evaluation on it for early baseline methods. Subsequently, some deep learning-based methods were proposed. Li and Snavely (2018) proposed an unsupervised framework that allows the network to learn without ground truth images, and to instead leverage information available from multiple images of the same scene but with different lighting conditions. Liu et al. (2020) also proposed an unsupervised framework by directly learning the latent features of reflectance and shading. Although these methods can achieve good intrinsic images, they fail to handle non-Lambertian surfaces, for example, those with specular highlights.

Datasets for Specular Highlight Detection Although the WHU-Specular dataset built by Fu et al. (2020) covers a diversity of real scenes with specular highlights, its images are of low resolutions. Networks trained on this dataset often fail to produce satisfactory detection results for high-resolution images, due to the huge scale gap between training and testing images. To overcome this issue, in this paper, we present a large-scale dataset with manual annotations of specular highlight regions for high-resolution specular highlight detection. Our dataset would facilitate network training and evaluation for the high-resolution specular highlight detection task.

3 High-Resolution Specular Highlight Detection Dataset

The WHU-Specular dataset (Fu et al., 2020) is currently the only publicly available large-scale dataset in the field of specular highlight detection, which consists of 4310 real image pairs (i.e., specular highlight images and corresponding ground truth specular highlight masks). However, the images in this dataset are of low resolutions (i.e., the diagonal length is not more than 1415p; see Fig. 3 for details), making it unsuitable for the high-resolution specular highlight detection task. In addition, it is essential to produce high-quality specular highlight masks with very high accu-

racy in boundaries for other related tasks such as specular highlight removal (Fu et al., 2021) and relighting (related to the movement of specular highlights) (Murmann et al., 2019).

To address the issues as mentioned above, we created a High-Resolution Specular Highlight Detection dataset (HRSHD) with manual annotations of specular highlight regions. Specifically, we first collected a total of 3619 specular highlight images. Among them, about 1800 images were taken by us with a mobile phone, and the remaining images were downloaded from popular image sharing websites including *Flickr*,[1] and *Pinterest*,[2] as well as image searching engines including *Google Images*[3] and *Bing Images*.[4] Note that images taken by our mobile phone have a fixed resolution of 3840 × 5120p. Then, we used the same annotation pipeline as described in (Fu et al., 2020) to generate ground truth masks for our dataset. Figure 2 presents the visual comparison of image pairs between HRSHD and WHU-Specular, showing that the distributions and shapes of specular highlights in our dataset are more complex and changeable than WHU-Specular. Moreover, Fig. 3 presents the resolution comparison of images between HRSHD and WHU-Specular, showing that the resolutions of the images in our dataset are considerably greater than in WHU-Specular. To the best of our knowledge, HRSHD is the first real dataset for high-resolution specular highlight detection. Finally, we randomly split the images in our dataset into two subsets: 3119 image pairs for training and 500 image pairs for testing.

Comparison of Annotation Quality Although absolute ground truth specular highlight masks are typically unavailable for real-world images, we quantitatively evaluated the annotation quality by treating annotated specular highlight masks by additional experts as ground truths. Specifically, we first randomly selected 100 trial images from HRSHD and WHU-Specular respectively, and recruited three additional experts from a commercial company for careful annotations. Then, we let all of them annotate all the trial images simultaneously for fair comparison, and their annotated masks were treated as ground truths. Finally, we employed four metrics, including MAE (Borji et al., 2015), F_m (Lang et al., 2016), S_m (Fan et al., 2017), and E_m (Fan et al., 2018) (see Sect. 5.1 for details) for quantitative evaluation. Table 1 reports the quantitative comparison result. As shown, the annotations of the 100 trial images from HRSHD have achieved lower errors than those from WHU-Specular in most cases, indicating that the overall quality of the annotations in our dataset is better than that in WHU-Specular, at least to some extent.

[1] https://www.flickr.com.

[2] https://www.pinterest.com.

[3] https://images.google.com.

[4] https://www.bing.com/images.

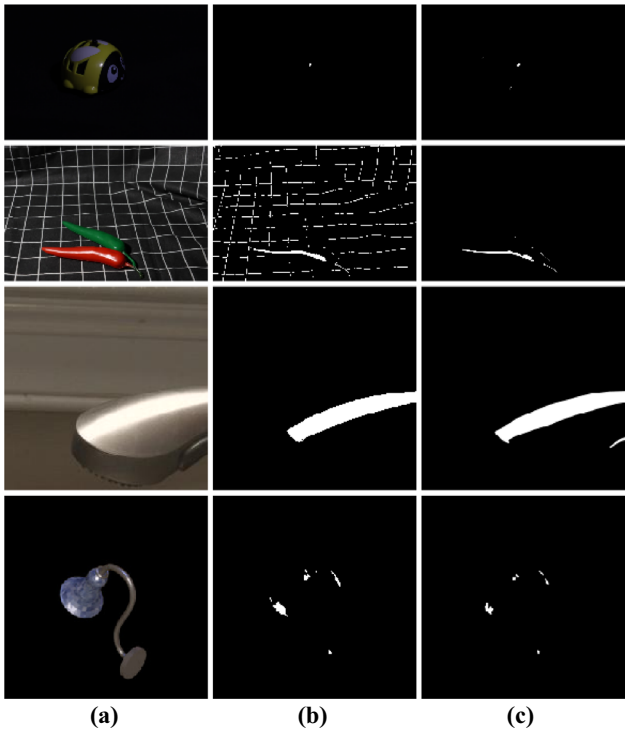

Fig. 5 Visual comparison between annotations by the manual and physics-based automatic annotation pipelines. **a** Input. **b** Physics-based automatic annotations. **c** Our manual annotations. First two rows: PSD-100; thrid row: SHIQ-100; forth row: ShapeNet-Intrinsic-100. Note that the physics-based automatic annotation pipeline may yield large errors (see **b** in the second row)

Moreover, Fig. 4 presents the visual comparison result. As shown, specular highlight masks in our dataset provide more accurate boundary locations of specular highlights compared to WHU-Specular.

Physical Correctness of the Manual Annotation Strategy To validate the physical correctness of the manual annotation strategy, we compare our manual annotations with the physics-based automatic annotations on our three newly prepared small datasets. Specifically, we first randomly selected 100 trial images from the PSD (Wu et al. 2021), SHIQ (Fu et al., 2021), and ShapeNet-Intrinsic (Shi et al., 2017) datasets, respectively. They are denoted as PSD-100, SHIQ-100, and ShapeNet-Intrinsic-100, respectively. Then, we binarized their grayscale specular residual[5] images to produce the specular highlight masks treated as ground truths. Finally, we employed four metrics, including MAE, F_m, S_m, and E_m, for quantitative evaluation. Table 2 reports the quantitative results, and Fig. 5 presents the visual comparison of our manual annotations with the physics-based automatic annotations. As can be seen, our manual annotations are highly

[5] We use the same term *specular residual* as in (Shi et al., 2017) to represent the remaining component obtained by subtracting the diffuse reflection component from an observed image.

Table 1 Quantitative comparison of annotation quality between our HRSHD and WHU-Specular (Fu et al., 2020)

Dataset	Expert	MAE↓	F_m↑	S_m↑	E_m↑
WHU-Specular	Expert 1	0.003	0.953	0.904	0.763
	Expert 2	0.007	0.973	0.903	0.838
	Expert 3	**0.001**	0.962	**0.932**	0.885
	Average	0.004	0.963	0.906	0.829
HRSHD	Expert 1	**0.002**	**0.984**	**0.956**	**0.930**
	Expert 2	**0.001**	**0.987**	**0.919**	**0.892**
	Expert 3	0.003	**0.983**	0.914	**0.902**
	Average	**0.002**	**0.985**	**0.933**	**0.908**

↑ and ↓ indicate that larger and smaller values are better respectively. The best results are marked in bold

Table 2 Quantitative evaluation of the physical correctness of manual annotations on the three small datasets

Dataset	MAE↓	F_m↑	S_m↑	E_m↑
PSD-100	0.016	0.707	0.768	0.765
SHIQ-100	0.005	0.969	0.952	0.986
ShapeNet-Intrinsic-100	0.002	0.827	0.838	0.935

consistent with the physics-based ground truths in most cases. Notably, our manual annotations on PSD-100 receive the worst results. This quality drop is primarily attributed to the errors in estimating specular highlight mask (see the second row of Fig. 5), resulting from the physical inconsistency of colors and luminosity between specular highlight images and corresponding diffuse images in PSD-100. In comparison to the physics-based annotation pipeline, our manual annotation pipeline has two advantages. First, it allows flexibly and effectively annotating complex specular highlights of spatially-varying intensities and various shapes. Second, it does not rely on auxiliary images (such as diffuse or specular residual images) captured by cross-polarization photography or generated by rendering software.

4 High-Resolution Specular Highlight Detection Network

4.1 Overall Architecture

Figure 6 illustrates the proposed Patch-level Bidirectional Refinement Network (PBRNet) for high-resolution specular highlight detection. It consists of four main ingredients: Specular Highlight Patch Selector (SHPS), Low-resolution Specular Highlight Detection Network (LSHDNet), Refinement Module (RM), and Attentional Fusion Module (AFM). Specifically, given an input high-resolution image, we first utilize SHPS to determine which patches contain specular highlights and which patches do not in it. Note that patches are at multiple scales, typically set to 256p, 512p, 1024p, and 2048p (indicated by the four boxes with solid blue lines at the top of Fig. 6) in our paper. Then, we feed those specular highlight patches at different scales into LSHDNet to produce their specular highlight masks as the coarse detection results. Next, we apply two series of RMs to progressively refine the detection results of adjacent-scale specular highlight patches in two pathways simultaneously. One pathway is from small-scale patch to large-scale patch (i.e., 256p $\rightarrow$ 512p $\rightarrow$ 1024p $\rightarrow$ 2048p, denoted as "small-to-large"), while the other pathway is from large-scale patch to small-scale patch (i.e., 256p $\leftarrow$ 512p $\leftarrow$ 1024p $\leftarrow$ 2048p, denoted as "large-to-small"). Finally, we fuse the last refined detection result from the small-to-large pathway and one from the large-to-small pathway using AFM to produce the final detection result for a minimum-scale specular highlight patch. The whole specular highlight mask of the input can be obtained by integrating the specular highlight mask of all minimum-scale specular highlight patches according to their position indexes.

4.2 Specular Highlight Patch Selector

Specular highlights in real-world scenes are often sparse in distribution (Akashi and Okatani, 2015), which implies that most regions of an image do not contain specular highlights. Therefore, considering the efficiency of the network, it is not necessary to carry out a series of refinement processing steps on those non-specular-highlight regions. Based on this observation, we propose a Specular Highlight Patch Selector (SHPS) to identify the patches that contain specular highlights for subsequent refinement processing.

Figure 7 presents the pipeline of our SHPS. Specifically, given an input high-resolution image, we take its low-resolution version produced by downsampling as input into our low-resolution specular highlight detection network, LSHDNet (see Sect. 4.5 for details), to output the low-resolution specular highlight mask. Then, we upsample this mask to be of the same resolution as the input to produce the high-resolution mask M. Finally, using M, we can identify those specular highlight patches at different scales. Formally, let $\mathrm{id}_{i,j}$ be the specular highlight index of a patch $\Omega_{i,j}$ whose location index and scale index are i and j, respectively. Here, $\mathrm{id}_{i,j}$ is 1 or 0, where "1" and "0" indicate that $\Omega_{i,j}$ contains and does not contain specular highlights, respectively. The number of scales is typically set to 4 ($1 \leq j \leq 4$) in our paper, where $j = 1, 2, 3, 4$ corresponds to the scales of 256p, 512p, 1024p, and 2048p, respectively. $\Omega_{i,j}$ contains specular highlights, when one or more of its corresponding four patches $\Omega_{f_k(i),j-1}(k = 1, 2, 3, 4)$ contain specular highlights (as shown by the solid and dashed blue lines in Fig. 7); otherwise, it does not contain. $f_k(k = 1, 2, 3, 4)$ are

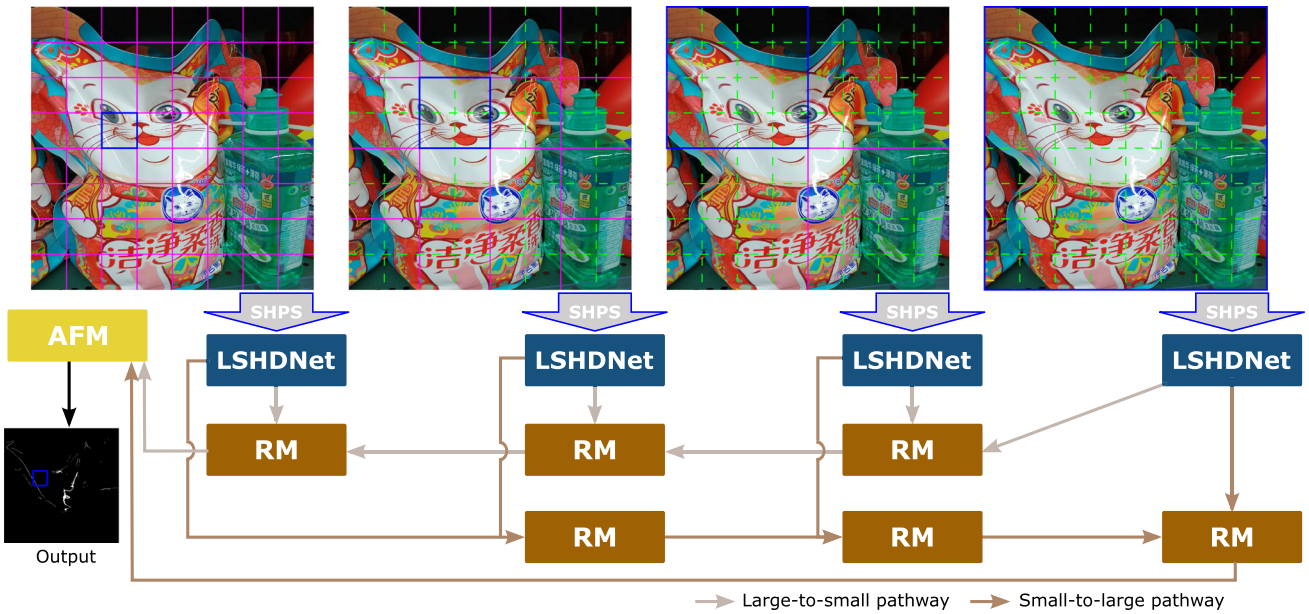

Fig. 6 The schematic illustration of the proposed PBRNet. Given an input high-resolution image, we first utilize SHPS to identify its local specular highlight patches. Then, we apply two series of RMs to progressively refine the detection results (estimated by LSHDNet) of adjacent-scale specular highlight patches in both the small-to-large and large-to-small pathways simultaneously. Next, we fuse the last refined detection result from the small-to-large pathway and one from the large-to-small pathway using AFM to produce the final detection result for a minimum-scale patch. Finally, we integrate the refined specular highlight masks of all specular highlight patches into the whole specular highlight mask for the input

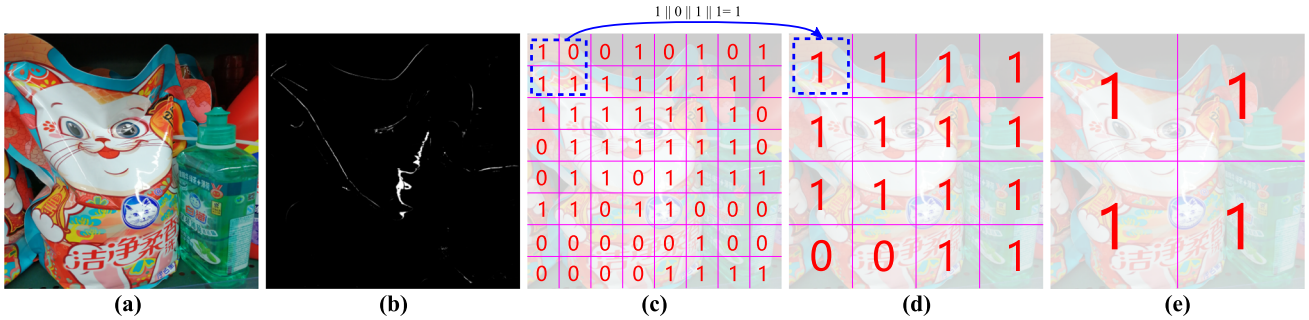

Fig. 7 The schematic illustration of the proposed SHPS. Given a input high-resolution image in (**a**), we first take its low-resolution version as input into LSHDNet to output the low-resolution detection result, and upsample it to be of the same resolution with the input to yield its high-resolution version in (**b**). Then, we can estimate the specular highlight index map at the minimum scale in (**c**). Finally, we can further sequentially estimate the specular highlight index maps at larger scales in (**d**) at 512p and (**e**) at 1024p

four position index transformation functions. The iterative process can be mathematically expressed as

$$\mathrm{id}_{(i,j)} = \begin{cases} 1, \sum_{k=1}^{4} \mathrm{id}_{f_k(i),j-1} >= 1, \; j = 2, 3, 4\,, \\ 0, \text{otherwise}\,. \end{cases} \tag{1}$$

The specular highlight index of a minimum-scale patch ($j = 1$) can be directly estimated using M:

$$\mathrm{id}_{(i,1)} = \begin{cases} 1, & \sum_{(p,q)\in\Omega_{i,1}} M^{(p,q)} > 0\,, \\ 0, & \text{otherwise}\,. \end{cases} \tag{2}$$

Here, (p, q) denotes a pixel location whose horizontal and vertical coordinates are p and q, respectively.

4.3 Refinement Module

We propose a Refinement Module (RM) to refine the detection results of adjacent-scale specular highlight patches. Figure 8(a) presents its detailed architecture. Specifically, we first concatenate the detection results of two adjacent-scale specular highlight patches, and take them as input into RM. Then, we sequentially apply two 3×3 convolution layers, two residual blocks (He et al., 2016), two 3×3 convolution

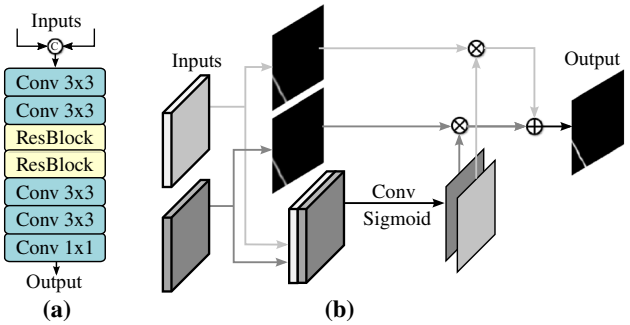

Fig. 8 The detailed architectures of the proposed RM and AFM, as shown in (**a**) and (**b**), respectively

layers, and a 1×1 convolution layer on it. Finally, we obtain a finer detection result.

4.4 Attentional Fusion Module

We further propose an Attentional Fusion Module (AFM) to refine the detection results from the small-to-large and large-to-small pathways. Figure 8(b) presents its detailed architecture. Specifically, we first take the two last refined detection results from the small-to-large and large-to-small pathways as inputs (denoted as F_1 and F_2, respectively) into AFM. Then, we sequentially apply a 3×3 convolution layer and a 1×1 convolution layer, followed by a sigmoid function on the concatenation of F_1 and F_2 to obtain their respective attention maps. Finally, we compute the element-wise products between F_1 and its attention map and between F_2 and its attention map, and add them together to yield the final detection result. Most previous attentional fusion schemes (Zhu et al., 2018) focus on fusing high-level features at the image level. In contrast, our AFM is designed to directly refine specular highlight masks at the patch level for high-resolution image processing.

4.5 Low-resolution Specular Highlight Detection Network

Figure 9 illustrates the proposed Low-resolution Highlight Detection Network (LSHDNet). It takes a single low-resolution image as input, and outputs its specular highlight mask in an end-to-end way. We build our LSHDNet on FPN (Lin et al., 2017), and incorporate the proposed Context Mixer (CM) to extract abundant context information of specular highlights, which significantly benefits to detect complex-shaped specular highlights across multiple scales. Specifically, we first utilize a convolution neural network to generate the five feature maps with varying spatial resolutions. Then, we embed a CM to generate context features for each layer, and construct the context feature pyramid by combining context features from all layers. Finally, giving the

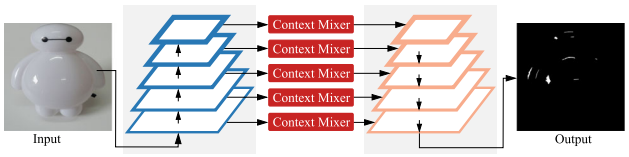

Fig. 9 The Schematic Illustration of the proposed LSHDNet

supervision at each layer, LSHDNet produces the specular highlight mask at each layer. Note that the specular highlight mask with the largest spatial resolution is considered as the final detection result. Our context mixer consists of three branches: Spatial CNN (SCNN) branch, residual branch, and Squeeze-and-Excitation (SE) branch. We now describe them in detail.

SCNN Branch We have observed that specular highlights in the real world often appear as straight or curved lines (see Fig. 2), mainly attributed to both the curved surfaces of objects and the direction of lighting. Inspired by Pan et al. (2018), we adopt SCNN to specially locate those specular highlights of straight and curved line shapes. SCNN with slice-by-slice convolutions within feature maps is able to encourage message passing between pixels across rows and columns in a layer, which can effectively capture long continuous shape regions (Pan et al., 2018). Based on SCNN, we further propose to use multi-scale SCNN, which benefits accurately detecting specular highlights of various scales in a scene simultaneously.

In the SCNN branch (see the middle of Fig. 10), an input feature map f passes through four spatial CNNs with different kernel sizes (i.e., typically $w_1 = 1$, $w_2 = 2$, $w_3 = 3$, and $w_4 = 4$). The generated feature maps are then concatenated and fed into a deconvolution layer to output the feature map f^{SCNN}:

$$f^{\text{SCNN}} = \Phi^{\text{SCNN}}\left(\text{cat}(\Phi^{w_1}(f); \cdots; \Phi^{w_4}(f))\right), \tag{3}$$

where $\Phi^{w_i}(\cdot)(i = 1, 2, 3, 4)$ denotes functions learned by the SCNN layers; $\text{cat}(\cdot)$ denotes the concatenation operation; and $\Phi^{\text{SCNN}}(\cdot)$ denotes the function learned by the deconvolution layer.

Residual Branch In the residual branch (see the bottom of Fig. 10), an input feature map f are fed into a 1×1 convolutional layer to output the feature map f^{Res}:

$$f^{\text{Res}} = \Phi^{\text{Res}}(f), \tag{4}$$

where Φ^{Res} denotes the function learned by the 1×1 convolution layer.

SE Branch We also utilize the Squeeze-and-Excitation (SE) scheme (Hu et al., 2018) to balance the weights of four feature maps produced by SCNN with different kernel sizes,

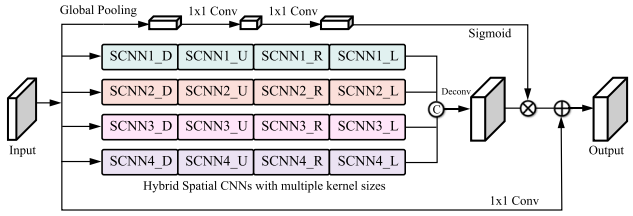

Fig. 10 The detailed architecture of the proposed CM. From top to down: SE branch, SCNN branch, and residual branch

enhancing the ability of the network to locate multi-scale specular highlights. Specifically, in the SE branch (see the top of Fig. 10), an input feature map f with the resolution of $H \times W \times C$) is first passed through a global pooling layer. Then, its output feature vector is fed into a 1×1 convolution layer, reducing its dimension to be $1 \times 1 \times C/4$. Next, another 1×1 convolution layer is applied to increase the dimension to be $1 \times 1 \times C$. Finally, we use a sigmoid function σ to squeeze the features f^{SE} into the range [0, 1], expressed as

$$\lambda^{\mathrm{SE}} = \sigma\left(\Phi^{\mathrm{SE}}(\mathrm{GloPool}(f))\right), \tag{5}$$

where GloPool is the global pooling operation; and $\Phi^{\mathrm{SE}}(\cdot)$ denotes the function learned by the two 1×1 convolutional layers.

The output feature map of the whole context mixer is finally expressed as

$$f^{\mathrm{CM}} = f^{\mathrm{SCNN}} \odot \lambda^{\mathrm{SE}} + f^{\mathrm{Res}}, \tag{6}$$

where $\odot$ denotes channel-wise multiplication.

4.6 Implementation details

We propose a separate learning strategy to train our LSHDNet, six RMs, and AFM. Specifically, we first train LSHDNet using training images (resized to 256×256p) in both our HRSHD dataset and the WHU-Specular dataset (Fu et al., 2020). The used loss function $\mathcal{L}^{\mathrm{LSHDNet}}$ is defined as

$$\mathcal{L}^{\mathrm{LSHDNet}}(X, Y) = \sum_{m=1}^{M} \Phi_{\mathrm{BCE}}(X^m, Y^m), \tag{7}$$

where Φ_{BCE} denotes the Binary Cross-Entropy (BCE) function; X and Y are the predicted specular highlight map and its corresponding ground truth, respectively; and M is the number of the side-outputs set to 5. After completing the training of LSHDNet, we use it to generate the detection results for all cropped specular highlight patches from our training images in HRSHD. Note that these detection results, along with their corresponding ground truths, are used as initial training pairs for the sequential training of RMs and AFM. Then, we sequentially train three RMs from left to right in the small-to-large pathway (see Fig. 6). This ordering is necessary, because the training of the current RM (except for the leftmost RM) depends on the outputs generated by the preceding RM. Meanwhile, we sequentially train three RMs from right to left in the large-to-small pathway. Finally, we train AFM using the output results from the rightmost RM in the small-to-large pathway and those from the leftmost RM in the large-to-small pathway, along with their corresponding ground truths. Here, the BCE loss function is also adopted for the training of RMs and AFM.

We implement our network in PyTorch and train it for 50 epochs with a mini-batch size of 8 on an NVIDIA 2080 Ti GPU. The entire network is optimized by the Adam optimizer (Kingma and Ba, 2014) with the initial learning rate of 10^{-4}, divided by 10 after 15 epochs. Besides, the input resolutions of LSHDNet, RM, and AFM are fixed to 256×256p.

5 Experiments

5.1 Experimental Settings

5.1.1 Datasets and Evaluation Metrics

We choose the WHU-Specular dataset (Fu et al., 2020) and our HRSHD dataset for evaluation. WHU-Specular is a low-resolution dataset, consisting of 3017 image pairs for training and 1293 image pairs for testing. To the best of our knowledge, these two datasets are the only two large-scale real datasets for specular highlight detection.

Following the commonly-used metrics in the fields of saliency detection and shadow detection, we adopt four metrics: Mean Absolute Error (MAE) (Borji et al., 2015), F-measure (F_m) (Lang et al., 2016), S-measure (S_m) (Fan et al., 2017), and E-measure (E_m) (Fan et al., 2018) to evaluate the performance of all methods. In general, a lower MAE value and higher values of F_m, S_m, and E_m indicate better results. We now describe these four metrics in detail.

MAE (Borji et al., 2015) evaluates the average difference between a predicted specular highlight mask X and its ground truth Y, formulated as

$$MAE = \frac{1}{W_X \times H_X} \sum_{x=1}^{H_X} \sum_{y=1}^{W_X} |X(p, q) - Y(p, q)|, \tag{8}$$

where W_X and H_X are the width and height of X, respectively; and (p, q) denotes a pixel location whose horizontal and vertical coordinates are p and q, respectively.

Table 3 Quantitative comparisons between our method and others on our HRSHD and WHU-Specular (Fu et al., 2020)

Method	HRSHD				WHU-Specular			
	MAE↓	F_m↑	S_m↑	E_m↑	MAE↓	F_m↑	S_m↑	E_m↑
NMFB (Zhang et al., 2018)	0.034	0.324	0.652	0.586	0.021	0.383	0.521	0.602
HSVT (Li et al., 2019)	0.013	0.359	0.698	0.632	0.012	0.413	0.712	0.598
SHDNet (Fu et al., 2020)	0.005	0.596	0.735	0.857	0.006	0.736	0.793	0.813
DSC (Hu et al., 2019)	0.187	0.411	0.404	0.299	0.091	0.114	0.412	0.357
DSD (Zheng et al., 2019)	0.076	0.490	0.464	0.331	0.049	0.490	0.482	0.512
DSS (Hou et al. 2019)	0.006	0.525	0.650	0.781	0.053	0.382	0.491	0.450
HRSODNet (Zeng et al., 2019)	0.005	0.594	0.719	0.848	0.005	0.636	0.768	0.839
Deeplab v3+ (Chen et al., 2018a)	0.009	0.332	0.608	0.772	0.019	0.501	0.619	0.657
MagNet (Huynh et al., 2021)	0.006	0.506	0.627	0.622	0.005	0.483	0.618	0.630
Ours	**0.004**	**0.621**	**0.803**	**0.892**	**0.004**	**0.701**	**0.812**	**0.860**

↑ and ↓ indicate that larger and smaller values are better respectively. The best results are marked in bold

F-measure (Lang et al., 2016) is a harmonic mean of average precision and average recall, formulated as

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}, \tag{9}$$

where β^2 is typically set to 0.3 to weigh precision more than recall, as done in (Cheng et al., 2014). Here, precision and recall denote the ratios of detected specular highlight pixels in the predicted specular highlight mask and the ground truth specular highlight mask, respectively. They are computed on binary images.

S-measure (Fan et al., 2017) simultaneously evaluates region-aware and object-aware structural similarity between a predicted specular highlight map and its ground truth, formulated as

$$S_m = \gamma S_0(X, Y) + (1-\gamma) S_r(X, Y), \tag{10}$$

where S_0 and S_r denotes the region-aware and object-aware structure similarity, respectively; and γ is typically set to 0.5, as done in (Fan et al., 2017).

E-measure (Fan et al., 2018) compares a predicted specular highlight mask and its ground truth by considering both the local pixel matching and the global means of the whole image, formulated as

$$E_m = \frac{1}{W_X \times H_X} \sum_{p=1}^{p=W_X} \sum_{q=1}^{q=H_X} \mathbf{A}(p, q) \tag{11}$$

where $\mathbf{A}$ is an enhanced alignment matrix, which indicates the correlation between X and Y. For the computation of $\mathbf{A}$ and further details, please refer to (Fan et al., 2018).

5.1.2 Data Augmentation

To enhance the robustness of our method to colored lighting and white noise, we leverage two data augmentations (i.e., mixing colored illumination and adding white Gaussian noise) for network training. We now describe them in detail.

Mixing Colored Illumination Colored illumination in a scene can lead to the ambiguity between colored specular highlights and colored material surface, which is rarely discussed and not addressed in the literature. To overcome the issue, we proposed to mix various colored illuminations with white-balanced images to generate more synthetic images for network training, as done in (Fu et al., 2020). It is mathematically formulated as

$$I^c = I * \mathbb{C}, \tag{12}$$

where I is a white-balanced image optionally processed by the white-balanced method (Barron and Tsai, 2017); I^c is an augmented image with new color illumination; $*$ denotes element-wise multiplication; and $\mathbb{C} = [R, G, B]$ is the mixed illumination color vector.

Mixing Gaussian White Noise There also exists the ambiguity between specular highlights and white noise, from which previous methods often suffer. To overcome this issue, we propose to add white noise of various densities with original training images to produce more synthetic images for network training. It is mathematically formulated as

$$I^w = I + N(m, var), \tag{13}$$

where I^w is an augmented image with Gaussian white noise; N denotes Gaussian white noise; and m and var are the mean value and variable value of Gaussian white noise, respectively. For each image, ten augmented images with white

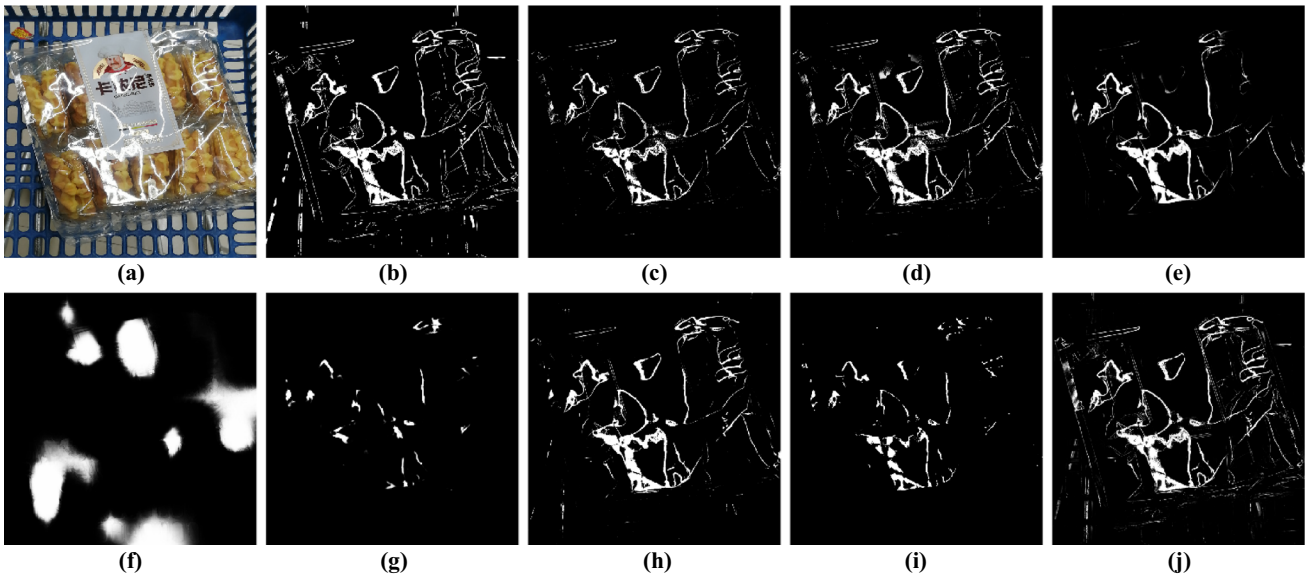

Fig. 11 Visual comparison of our method against state-of-the-art specular highlight detection methods on our datasets. **a** Input. **b** Ground truth. **c** NMFB (Zhang et al., 2018). **d** HSVT (Li et al., 2019). **e** SHDNet (Fu et al., 2020). **f** DSC (Hu et al., 2019). **g** DSS (Hou et al., 2019). **h** HRSODNet (Zeng et al., 2019). **i** MagNet (Huynh et al., 2021). **j** Ours

Gaussian noise of different densities and ten augmented images with different color illuminations are generated as additional training images.

5.1.3 Compared Methods

For qualitative and quantitative evaluation, we compare our method with the following three state-of-the-art specular highlight detection methods: two traditional methods of NMFB (Zhang et al., 2018) and HSVT (Li et al., 2019), and a deep learning-based method of SHDNet (Fu et al., 2020). Additionally, we compare our method with the following six state-of-the-art deep learning-based detection and segmentation methods from related fields: two shadow detection methods of DSC (Hu et al., 2019) and DSD (Zheng et al., 2019), two saliency detection methods of DSS (Hou et al., 2019) and HRSODNet (Zeng et al., 2019), and two semantic segmentation methods of Deeplab v3+ (Chen et al., 2018a) and MagNet (Huynh et al., 2021). Note that HRSODNet and MagNet fall into the category of high-resolution image processing. For our PBRNet and compared networks, we resize input original high-resolution images to be of their respective acceptable input resolutions for training. At the testing stage, we resize the output masks back to be of their corresponding specified resolutions for evaluation.

5.1.4 Fairness Setting

Since traditional specular highlight detection methods are sensitive to parameters, we varied their key parameters (i.e., inner dimension of factorization for (Zhang et al., 2018), and two weighting parameters for (Li et al., 2019), and tried to produce their best results. For other compared CNN-based networks, we retrained them using their publicly available codes on our HRSHD dataset and the WHU-Specular dataset (Fu et al., 2020), and fine-tuned their key parameters to produce better results as much as possible. Moreover, for a fair comparison of memory footprint and inference time, we tested all methods on the same PC equipped with Xeon(R) CPU E5-2630 v4 @ 2.20GHz and NVIDIA 2080Ti GPU.

Table 4 Ablation study for the number of scales

Scale level	MAE↓	F_m↑	S_m↑	E_m↑
256	0.011	0.503	0.530	0.562
256, 512	0.008	0.550	0.591	0.634
256, 1024	0.010	0.531	0.586	0.621
256, 2048	0.011	0.526	0.585	0.619
256, 512, 2048	0.006	0.612	0.727	0.810
256, 1024, 2048	0.007	0.603	0.714	0.798
256, 512, 1024, 2048	**0.004**	**0.621**	**0.803**	**0.892**

5.2 Comparison with State-of-the-art Methods

Quantitative Comparison Table 3 reports the quantitative comparison results on our HRSHD dataset and the WHU-Specular dataset. As can be seen, our method achieves the best quantitative results on all four metrics, including MAE, F-measure (F_m), S-measure (S_m), and E-measure (E_m).

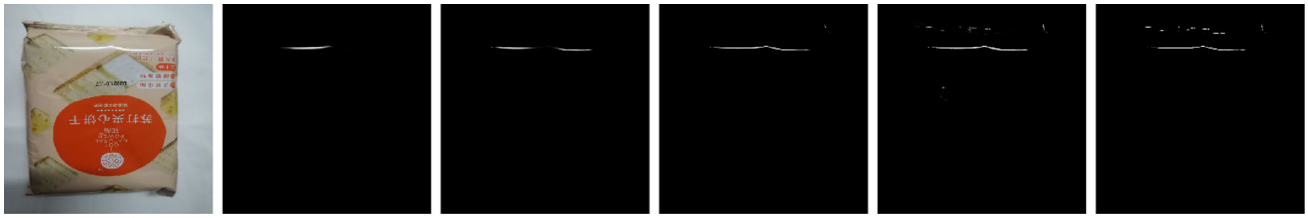

Fig. 12 Ablation study for the number of scales. **a** Input. **b–e** Specular highlight masks produced by our PBRNet using the scale levels of "256p", "256p, 512p", "256p, 512p, 1024p", and "256p, 512p, 1024p, 2048p", respectively. **f** Ground truth

Notably, two traditional methods of NMFB (Zhang et al., 2018) and HSVT (Li et al., 2019) achieve unsatisfactory or even poor results, since they strictly depend on a premise that specular highlights have the highest intensities in a white-balanced scene.

Visual Comparison Fig. 11 presents the visual comparison on a randomly selected image from our dataset. As shown, two traditional methods of NMFB (Zhang et al., 2018) and HSVT (Li et al., 2019) fail to accurately detect both strong and soft specular highlights while disregarding white material surfaces, due to the inherent ambiguity between specular highlights and white material surfaces. Although deep learning-based methods are able to alleviate the ambiguity issue, they still suffer from various drawbacks. Specifically, for low-resolution detection/segmentation methods, LSHDNet (Fu et al., 2020), DSS (Hou et al., 2019), and DeepLab V3+ (Chen et al. 2018b) fail to locate detailed specular highlights in high-resolution images well. DSC (Hu et al., 2019) may produce messy detection results with large errors. In addition, two high-resolution detection/segmentation methods of HRSODNet (Zeng et al., 2019) and MagNet (Huynh et al., 2021) fail to accurately detect weak and small-sized specular highlights, since their networks are designed mainly for capturing object-level regions rather than small regions like specular highlights. In comparison, our method is able to effectively overcome the above issues, and thus produce more accurate detection results with finer locations and boundaries of specular highlights.

6 Ablation Experiments and More Analysis

we first conduct ablation experiments on our testing dataset to evaluate various variants of our framework, including our multiple scale levels, bidirectional refinement, LSHDNet, and the IoUE loss. Then, we perform comparison of memory footprint and inference time for our method and others. Finally, we analyze the generalization of our method to unseen in-the-wild images and grayscale images, as well as the robustness to color lighting and white noise.

Table 5 Ablation study for our bidirectional refinement

Pathway	MAE↓	F_m↑	S_m↑	E_m↑
Small-to-large	0.005	0.593	0.794	0.841
Large-to-small	0.006	0.521	0.798	0.810
Bidirectional	**0.004**	**0.621**	**0.803**	**0.892**

Multiple Scale Levels Table 4 reports the quantitative comparison of our PBRNet with different scale levels. As shown, the performance of our PBRNet becomes better and better as the number of scales increases. Notably, our PBRNet using four scales achieves the maximum performance gain with MAE: 0.07, F_m: 0.118, S_m: 0.273, and E_m: 0.330, compared to using only one scale of 256p on our testing dataset. In addition, Fig. 12 presents the visual comparison on an image. As shown, our PBRNet with more intermediate scales is able to capture finer locations of specular highlights. The quantitative and visual comparison results clearly illustrate the effectiveness of employing multiple scales in our network. It is worth noting that when using only one scale, our method degenerates into the patch processing method, and thus it does not require AMs and AFM to refine the detection results.

Bidirectional Refinement Table 5 reports the quantitative comparison of our PBRNet with the refinement in different pathways. As shown, our PBRNet with both the small-to-large and large-to-small pathways achieves better results than with only one of them. This illustrates the effectiveness of our bidirectional refinement scheme.

LSHDNet We evaluated our PBRNet with different low-resolution detection backbones (i.e., our default LSHDNet and FPN (Lin et al., 2017)). Note that their input resolutions are typically set to 256 × 256p. Table 6 reports the quantitative comparison result. From it, we can draw two main observations. First, two baseline methods of down-sampling and patch processing, and our PBRNet, using our LSHDNet as the backbone network, achieve lower errors than using FPN, since our LSHDNet has a superior performance over FPN. Second, whether using FPN or our LSHDNet as the backbone network, our PBRNet consistently achieves lower

Table 6 Ablation study for our LSHDNet

Method	MAE↓	F_m↑	S_m↑	E_m↑
Backbone: FPN				
Downsampling	0.009	0.497	0.584	0.624
Patchprocessing	0.007	0.510	0.625	0.683
PBRNet	**0.006**	**0.564**	**0.751**	**0.843**
Backbone: LSHDNet				
Downsampling	0.007	0.502	0.649	0.702
Patchprocessing	0.005	0.533	0.702	0.786
PBRNet	**0.004**	**0.621**	**0.803**	**0.892**

Table 7 Ablation study for the IoUE loss

Method	MAE↓	F_m↑	S_m↑	E_m↑
Ours-w/-IoUE	**0.004**	0.619	**0.812**	0.891
Ours-w/o-IoUE	**0.004**	**0.621**	0.803	**0.892**

errors than the two baseline methods. The main reason is that PBRNet is able to effectively integrate local and global information, whereas the two baseline methods are limited to leveraging only one of these two types of information.

IoUE loss Actually, we do not additionally include the IoUE loss (Fu et al., 2020) for our default network training, since it fails to bring a noticeable performance gain. To demonstrate this, we retrained our network with the additional IoUE loss, and tested it on our HRSHD dataset. Table 7 reports the quantitative evaluation result. By observing the results of our network only with the BCE loss (denoted as "Ours-w/o-IoUE") and with both the BCE loss and the IoUE loss (denoted as "Ours-w/-IoUE"), we can see that the IoUE loss introduces a very slight change in the final results of our network. Note that to more accurately locate line-shaped specular highlights, we actually have utilized spatial CNN in our network, which also has the same function of accurately locating boundaries of specular highlights as the IoUE loss.

Memory Footprint and Inference Time Table 8 reports the memory footprint and inference time for our method and others. As shown, our network takes up about 2314Mb memory and takes about 2.14s to process an image with the resolution of 2048 × 2048p on a single NVIDIA 2080 Ti GPU card. Note that two traditional optimization- or thresholding-based methods of NMFB (Zhang et al., 2018) and HSVT (Li et al., 2019), are implemented on CPU, and we accordingly report their results on it. From Table 8, we found that compared to deep learning-based methods, these traditional methods often occupy smaller memory, while their detection performance is much lower. In addition, our method achieves comparable results in terms of memory footprint and inference time among deep learning-based methods. Meanwhile, our method outperforms other state-of-the-art detection and segmentation methods in terms of detection accuracy on our HRSHD dataset and the WHU-Specular dataset (Fu et al., 2020) (see Table 3). Reducing memory usage and speeding up our inference process are left as our future work.

Table 8 Comparison of memory usage and inference time between our method and other detection and segmentation methods on our testing dataset

Method	PT	MS↓	IT↓
NMFB (Zhang et al., 2018)	CPU	79	1.988
HSVT (Li et al., 2019)	CPU	121	1.230
SHDNet (Fu et al., 2020)	GPU	3189	0.049
DSC (Hu et al., 2019)	GPU	1835	0.053
DSD (Zheng et al., 2019)	GPU	1333	**0.034**
DSS (Hou et al. 2019)	GPU	3048	4.931
HRSODNet (Zeng et al., 2019)	GPU	2608	0.392
Deeplab v3+ (Chen et al., 2018a)	GPU	1702	0.039
MagNet (Huynh et al., 2021)	GPU	2007	2.921
Ours	GPU	2314	2.140

Here, PT: processor type (CPU/GPU); MS: memory size (MB); and IT: inference time (s)

Generalization to In-the-wild Images Our method has a good generalization to in-the-wild images featuring achromatic surfaces, heavy texture, complex lighting, and other challenging conditions. Figure 13 presents the visual comparison on an in-the-wild image. As shown, traditional methods sometimes wrongly detect high-intensity white material surfaces as specular highlights. This is because they are often based on intensity analysis with a strict assumption that specular highlights are those pixels with the highest intensities. In contrast, deep learning-based methods (Fu et al. (2020); Chen et al. (2018a); Hu et al. (2019); Hou et al. (2019); Zeng et al. (2019); Huynh et al. (2021)) have a good ability to obtain high-level semantic information of specular highlights, enabling effectively disambiguate them from white material surfaces. Nevertheless, they may either partially (even entirely) lose the locations of small-sized specular highlights, or fail to accurately detect the boundaries of complex-shaped specular highlights. These limitations are mainly attributed to their limited performance on high-resolution images. Compared to them, our method is able to effectively address these issues, and generate more accurate results with fine locations of specular highlights.

Generalization to Grayscale Images Our method also has a good generalization to grayscale images. Figure 14 presents the visual comparison on a grayscale image. As shown, NMFB (Zhang et al., 2018) wrongly detects high-intensity non-specular-highlight regions as specular highlights, while HSVT (Li et al., 2019) fails to effectively detect soft specular

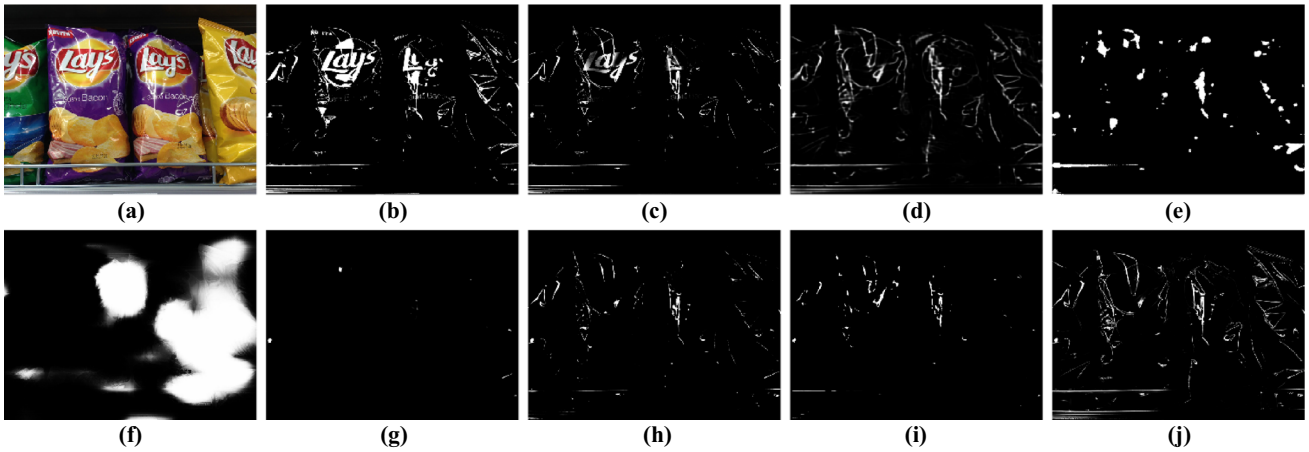

Fig. 13 Visual comparison of our method against state-of-the-art specular highlight detection methods on an in-the-wild image. **a** Input. **b** NMFB (Zhang et al., 2018). **c** HSVT (Li et al., 2019). **d** SHDNet (Fu et al., 2020). **e** DeepLabV3+ (Chen et al., 2018b). **f** DSC (Hu et al., 2019). **g** DSS (Hou et al., 2019). **h** HRSODNet (Zeng et al., 2019). **i** MagNet (Huynh et al., 2021). **j** Ours

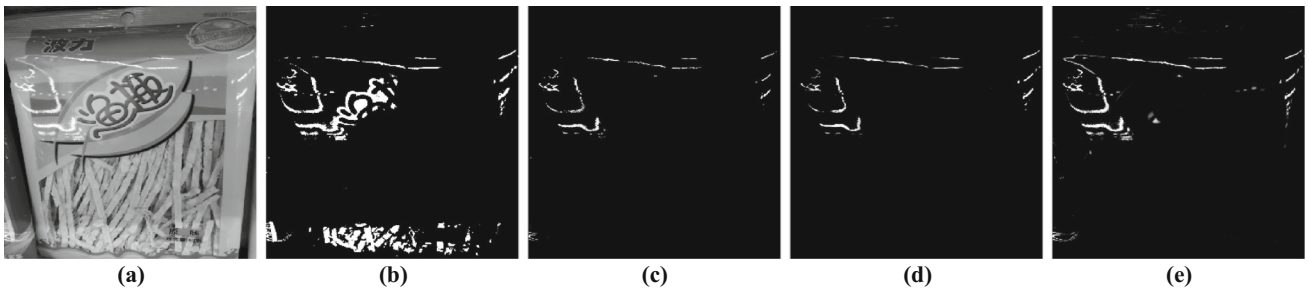

Fig. 14 Visual comparison on a grayscale image. **a** Input. **b** NMFB (Zhang et al., 2018). **c** HSVT (Li et al., 2019). **d** SHDNet (Fu et al., 2020). **e** Ours

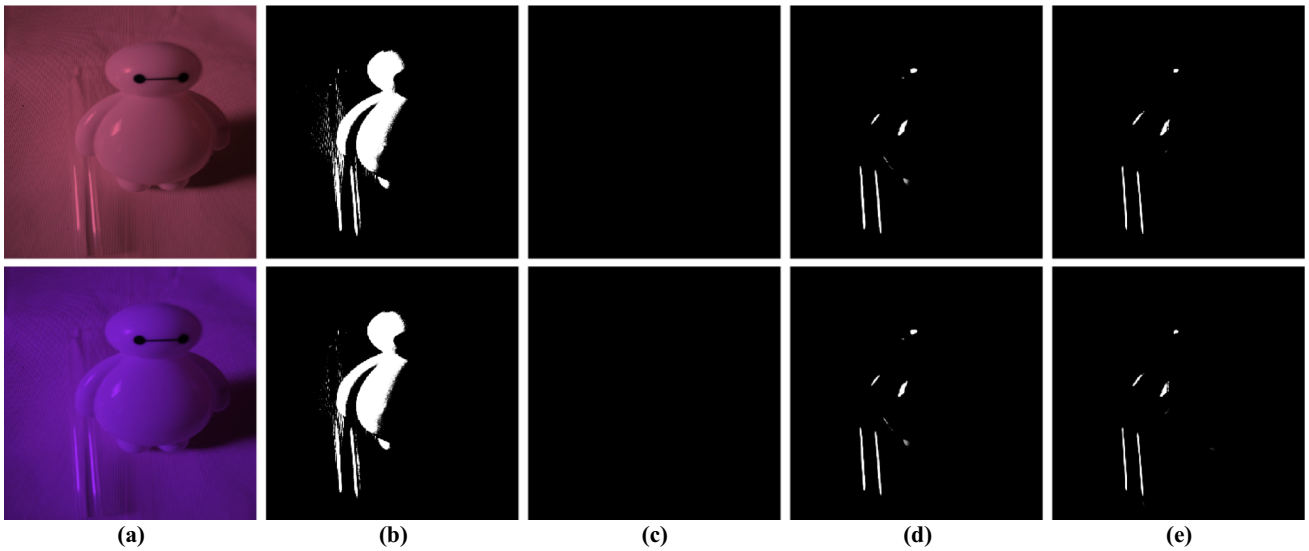

Fig. 15 Visual comparison on multi-illumination image sequences from the same scene. **a** Input. **b** NMFB (Zhang et al., 2018). **c** HSVT (Li et al., 2019). **d** SHDNet (Fu et al., 2020). **e** Ours

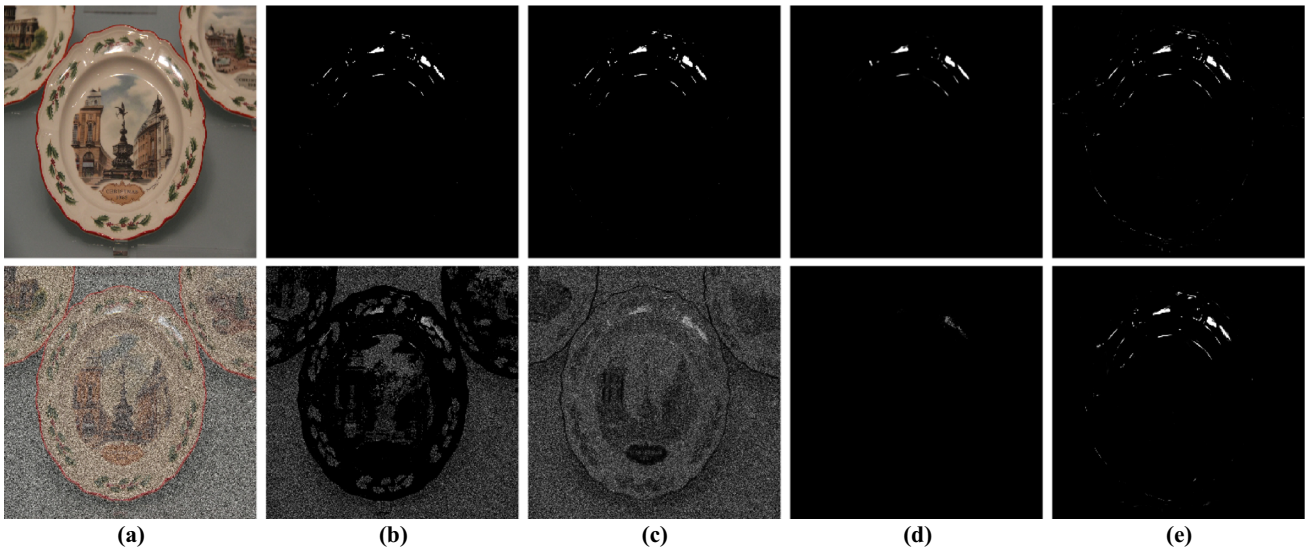

Fig. 16 Visual comparison on images with different noise intensity levels from the same scene. **a** Input. **b** NMFB (Zhang et al., 2018). **c** HSVT (Li et al., 2019). **d** SHDNet (Fu et al., 2020). **e** Ours

highlights. The main reason is twofold. First, these traditional methods are built on color (e.g., RGB and HSV) channels, each of which specular highlights often exhibit a specific distribution closely related to. Second, as mentioned earlier, they fail to overcome the semantic ambiguity between specular highlights and white material surfaces. Furthermore, despite achieving good detection results, SHDNet sometimes fails to accurately detect small-sized specular highlights, as it is mainly designed to work with low-resolution images. In comparison, our method achieves high-quality detection results comparable to those obtained on color images, even though it is trained on color images.

Robustness to Colored Illumination Our method has a good capability in effectively handling color illumination variations in images. Figure 15 presents the visual comparisons on two images with different illumination colors. As shown, HSVT (Li et al., 2019) fails to detect colored specular highlights due to the violation of its assumption regarding white balance caused by colored illumination. NMFB (Zhang et al., 2018) suffers from the semantic ambiguity between colored specular highlights and material surfaces of the same color. In addition, the low-resolution detection method of SHDNet (Fu et al., 2020) fails to produce sufficiently accurate detection results, although it is not sensitive to colored illumination. In comparison, our method is able to effectively address these challenging images with colored illumination, mainly attributed to the data augmentation of mixing color illumination.

Robustness to White Noise Our method is also robust to white noise. Figure 16 presents the visual comparisons on two images with white Gaussian noise at different density levels.

Fig. 17 A failure case. **a** Input with a burning candle (i.e., light source). **b** Ours

As shown, two traditional methods, NMFB (Zhang et al., 2018) and HSVT (Li et al., 2019), fail to eliminate the semantic ambiguity between specular highlights and white noise, resulting in unsatisfactory or even poor detection results. In addition, SHDNet (Fu et al., 2020) has a performance deterioration due to white noise. In comparison, our method is able to effectively disregard white noise and produce more accurate detection results, mainly attributed to the data augmentation of adding white noise.

Limitations Our method has two limitations. First, our method may wrongly detect high-intensity light sources as specular highlights, due to their similar appearance in terms of shape, intensity, and color. Figure 17 presents an example, where the flame generated by a candle is wrongly detected as a specular highlight. We believe that more training data like this image is needed for our network training to learn to disambiguate specular highlights from light sources. Second, our specular highlight patch selector, SHPS, could miss out those patches with very weak and small-sized specular high-

lights. As a result, any errors introduced by it would become part of the final detection result. This is actually caused by the trade-off between speed and accuracy for our network. Also, refer to Sect. 8 for further analysis and discussion.

7 Application for Specular Highlight Removal

Background Most specular highlight removal methods may fail to recover colors and texture details particularly underneath strong (i.e., high-intensity and large-area) specular highlights. To address this issue, a straightforward solution is to utilize existing inpainting networks to help restore missing information underneath strong specular highlights. However, we argue that an inpainting network may have the potential to undesirably generate the appearance of specular highlights in resulting specular-free images. This can occur (see Fig. 18) due to the network's learning on a large-scale dataset of real images that often contain various specular highlights. To this end, we propose to exclude specular highlight regions from training images when training inpainting networks. As our default choice, we adopt PConvN (Liu et al., 2018) as the inpainting network, known for its conciseness and effectiveness. Note that other inpainting networks should also perform well. In the following, we will detail our modified inpainting framework.

Modified Inpainting Framework To train PConvN, we built a Specular Highlight Inpainting (SHI) dataset by collecting 12,199 real images with specular highlights from the Internet (e.g., well-known *Flickr* and *Pinterest* websites). It covers most of the material surfaces and illumination conditions in our daily life. For each image, we use our PBRNet to produce its specular highlight mask M_s, while generating a random mask M_h with irregular holes as done in (Liu et al., 2018). Thus, we yield the valid mask M by excluding specular highlight regions from M_h:

$$M = M_h \times (1 - M_s). \tag{14}$$

Using the above equation, we can obtain masks for all images in SHI, referred to as SHI-mask. Then, we retrain PConvN using images in SHI and their corresponding masks in SHI-mask. The purpose of using M instead of M_h is to encourage the network to generate only diffuse (i.e., specular-free) appearances for our application.

At the testing stage, given an input image, we use our PBRNet to produce its specular highlight mask, and then use our retrained PConvN to generate the inpainted result (treated as specular-free image in our application). Figure 18 illustrates the effectiveness of our modified inpainting framework. As shown, compared to the original PConvN, our modified PconvN exhibits better color and texture detail recovery in the presence of strong specular highlight, while avoiding the generation of undesirable appearances of specular highlights.

User Study To verify the effectiveness of our method, we further conducted a user study, as done in (Wang et al., 2019). We compare our method with a recent deep learning-based method, JSHDR (Fu et al., 2021), as well as three traditional methods: HRIR (Shen and Zheng, 2013), HRBF (Yang et al., 2015), and SNMF (Akashi and Okatani, 2015). We first downloaded additional 500 testing images (not included in SHI) from *Flickr* and *Pinterest* websites by searching with keywords "bottle", "snacks", "doll", "glass", and so on. Then, we produced the specular-free images for each test image by our method and others, and recruited 50 participants from a campus to rate each group of results. For each

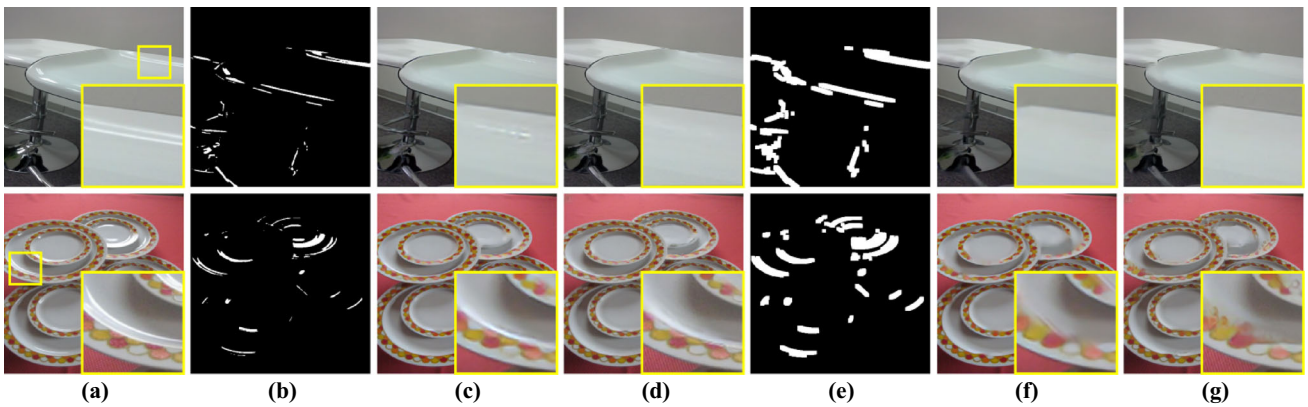

Fig. 18 Ablation study for our modified inpainting framework. **a** Input. **b** Our specular highlight mask. **c**, **d** Inpainted results (using the masks in **b**) produced by PConV trained without and with our strategy of excluding specular highlight regions in training images, respectively. **e** The dilated version of (**b**). **f, g** The same as (**c, d**), but using the mask in (**e**). We can observe a significant improvement in preventing the generation of specular highlight artifacts in the process of inpainting specular highlights

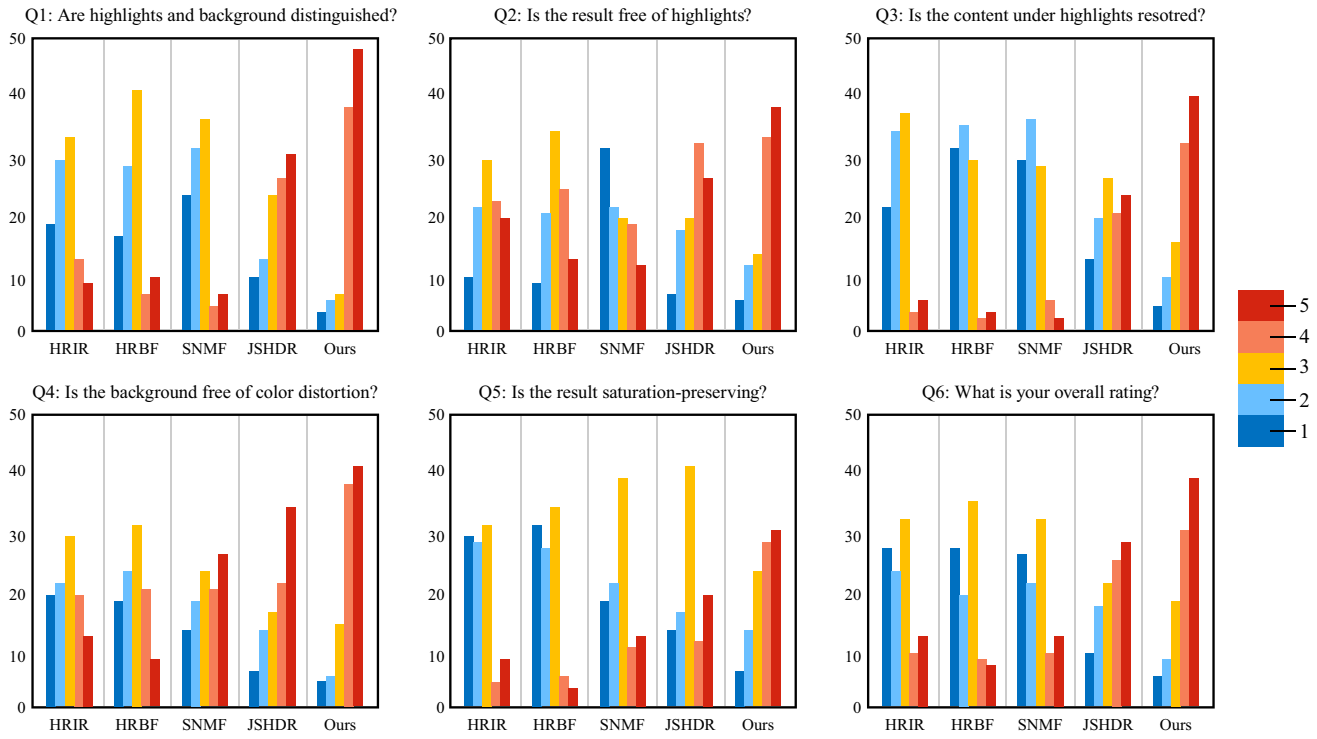

Fig. 19 Rating distribution for various specular highlight removal methods on the six questions in the user study. The ordinate axis denotes the frequency of ratings received by the methods from the participants

result, the participants were asked to give a rating score on a 5-point Likert scale ranging from 1 (worst) to 5 (best) for each of the six questions shown in Fig. 19. From the statistical results of the user study, we can see that our method receives more "red" and far fewer "blue" ratings compared to previous methods, illustrating that our results are more preferred by the participants.

Figure 20 presents the visual comparison on two images. As shown, the three traditional methods (Shen and Zheng, 2013; Yang et al., 2015; Akashi and Okatani, 2015) often suffer from the hue-saturation ambiguity (Kim et al., 2013), and thus produce low-quality results with noticeable artifacts such as color and illumination distortion. The deep learning-based method of JSHDR (Fu et al., 2021) is less able to recover missing colors and texture details underneath strong specular highlights. In contrast, our method can effectively overcome these issues and produce high-quality results without noticeable artifacts.

Intrinsic image decomposition has the function of removing specular highlights, since it aims to separate shading (including specular highlights) from reflectance. Therefore, to further illustrate the superior performance of our method, we additionally compare it with state-of-the-art intrinsic image decomposition methods, including three traditional methods (Grosse et al., 2009; Shen et al., 2011; Bell et al., 2014) and a deep learning-based method (Liu et al., 2020). Note that for several implemented algorithms by (Grosse et al., 2009), we chose the color Retinex algorithm for comparison due to its superior performance. Figure 21 presents the visual comparison on two images. As shown, the compared methods often suffer from the leakage of specular highlights from the shading into the reflectance, and thus fail to satisfactorily recover the missing material information, especially underneath strong specular highlights. This limitation arises from their reliance on an idealized Lambertian assumption. In contrast, our method produces better specular-free images with natural-looking appearances again.

8 Potential Research Directions

Although our method, as well as other state-of-the-arts, has achieved good performance for specular highlight detection, there are still some unresolved issues that require careful attention and further study. In this regard, we propose ten potential research directions to inspire further investigation and advancement in this field:

(1) Pursuing Higher-precision and Real-time Performance Our method, PBRNet, may fail to accurately specular highlights in some challenging scenes due to achromatic material surfaces, chromatic illumination, noise, overexposure, light sources, and so on. Besides, it does not achieve real-time performance to deal with high-resolution images. Similarly, previous state-of-the-art methods also suffer from

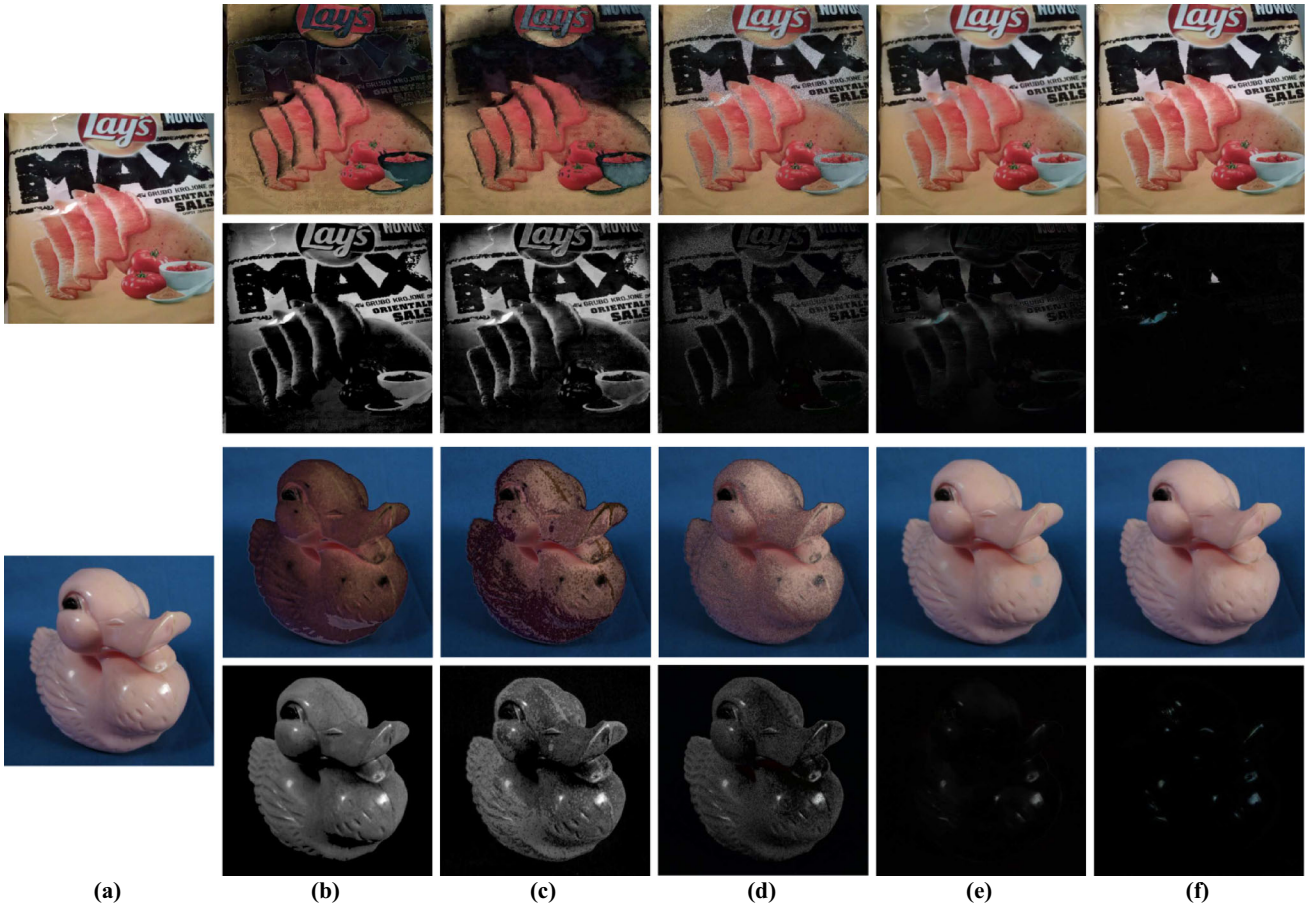

Fig. 20 Visual comparison of our inpainting-based specular highlight removal method against state-of-the-art specular highlight removal methods. **a** Input. **b** Shen and Zheng (2013). **c** Yang et al. (2015). **d** Akashi and Okatani (2015). **e** Fu et al. (2021). **f** Ours. Odd rows: specular-free image; even rows: specular residual image. Here, the specular residual image is obtained by subtracting the estimated specular-free image from the input image

these two issues. Higher-precision and real-time specular highlight detection would open up opportunities for real-time analysis, processing, and interaction with specular highlights in related fields such as augmented reality and virtual reality.

(2) Dataset Extension Most of the images in our HRSHD dataset were captured in controlled indoor environments, and the variety of material types is somewhat limited (mainly focusing on plastic and porcelain). To overcome this limitation, researchers can extend HRSHD by incorporating a more diverse range of objects, materials, and lighting conditions. By doing so, the extended dataset encompasses a wider range of specular highlight appearances commonly encountered in our daily life. This extension will facilitate more efficient network training and enable a more comprehensive evaluation of specular highlight detection methods.

(3) Soft Specular Highlight Detection Soft specular highlight detection is a very challenging problem due to two main factors. First, soft specular highlights tend to incorporate texture details from non-specular-highlight regions, thus suffering from the well-known semantic ambiguity issue. Second, they have a wider range of intensity variations than strong specular highlights. To overcome the above issues, researchers can explore low-level features, such as gradients, with other relevant cues to achieve accurate detection results. Additionally, it is essential to develop more appropriate evaluation metrics tailored for soft specular highlight detection to effectively evaluate the performance of detection algorithms.

(4) Focusing on Special Scene Images or Videos Most specular highlight detection methods are designed to deal with general natural scenes. However, they may not be well-suited for addressing data in special scenes such as remote sensing, medical, document, and facial images. As a result, they may yield somewhat unsatisfactory detection results in specific cases. Therefore, considering the distinct characteristics of specular highlights in these special cases, it is beneficial to develop specific detection methods that can provide improved results compared to a general detection method. In addition, to our knowledge, there is currently no

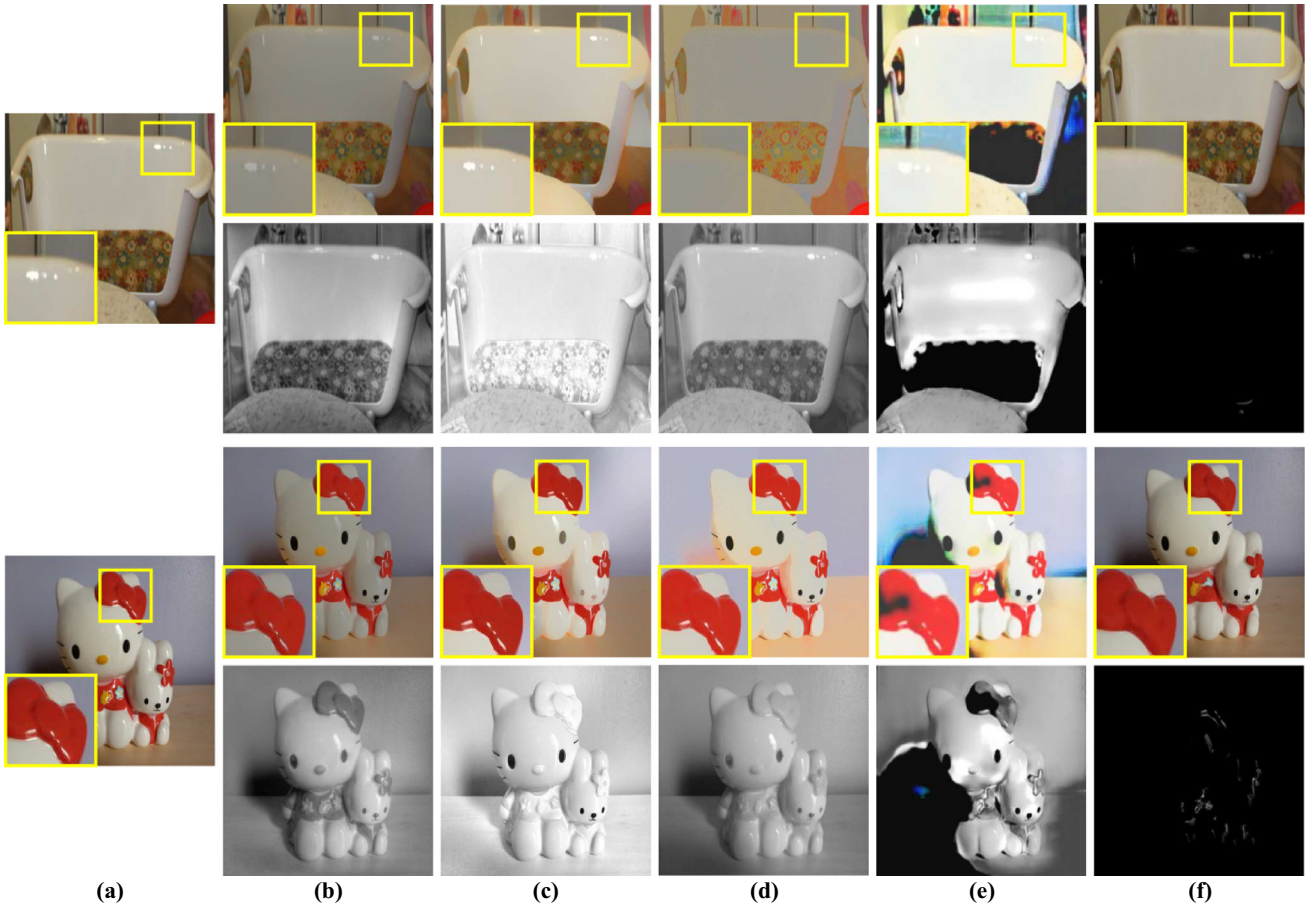

Fig. 21 Visual comparison of our inpainting-based specular highlight removal method against state-of-the-art intrinsic image decomposition methods. **a** Input. **b** Grosse et al. (2009). **c** Shen et al. (2011). **d** Bell et al. (2014). **e** Liu et al. (2020). **f** Ours. Odd rows: reflectance and specular-free (in the last column) images; even rows: shading and specular residual (in the last column) images

work of specular highlight detection for videos. Therefore, it would be highly meaningful and worth investing considerable effort to build a large-scale real video dataset and design a spatio-temporal and efficient network for specular highlight detection in videos.

(5) Exploring Material Cues Dealing with specular highlights on various material surfaces is challenging due to the different reflection properties exhibited by different materials such as plastic, metal, and wood. To efficiently detect specular highlights on all materials in a scene simultaneously, researchers can develop material-specific methods that consider the specific reflection characteristics of different materials. This involves understanding the underlying physics of how light interacts with different surfaces and designing detectors tailored to each material type.

(6) Leveraging Multi-model Input Data It mainly involves two aspects. First, according to the classic Phong reflection model (Phong, 1975), the curvature of an object's surface plays an important role in the formation (e.g., shape, size, and intensity) of specular highlights. Therefore, specular highlights can be more accurately located by combining information from RGB images with their corresponding depth maps. Second, researchers can leverage motion information of specular highlights from multi-illumination image sequences to distinguish them from non-specular-highlight regions. Note that the position of specular highlights will shift with the variation of illumination conditions in a scene.

(7) Unsupervised or Self-supervised Learning Recently, researchers have proposed various unsupervised or self-supervised learning frameworks for tasks in computer vision, such as semantic segmentation (Van Gansbeke et al., 2021), foreground object segmentation (Croitoru et al., 2019), and saliency detection (Wang et al., 2022). These fields have seen significant achievements and have garnered considerable attention from the research community. Drawing inspiration from them, exploring unsupervised or self-supervised learning for specular highlight detection can overcome the dependence on large-scale annotated training images.

(8) Unified Detector Detecting illumination regions like shadow, specular highlight, and flare is an important prob-

lem in computer vision. Most methods in the literature are proposed to deal with just detecting a specific type of illumination region. Despite their impressive performance, these solutions are not universally applicable to all illumination detection problems, since the networks have to be trained separately for each task. The requirement of multiple models and the decision-making process for switching between multiple illumination detectors increases computational complexity, and thus hinders the adoption of them in real-time systems. Therefore, it is meaningful and practical to design a unified framework that can simultaneously detect multiple illumination regions in a scene.

(9) Integration with Other Tasks Specular highlight detection can be integrated into related computer vision tasks, thereby enhancing their capabilities and expanding their applicability. These tasks include but not limited to: (a) Object Recognition: Specular highlights on an object can be used to establish correspondence between its image and the 3D model, allowing to verify the hypothesized pose and the identity of the object (Netz and Osadchy, 2012). (b) Shape Estimation: By analyzing the position, size, and orientation of specular highlights on a highly reflective or glossy object under different viewing directions, it is possible to extract useful cues about the object's surface geometry and estimate its shape. (c) Image Understanding: Specular highlight detection can contribute to a deeper understanding of images by providing important information about the lighting conditions, material properties, and geometry structures of objects in the scene, leading to improved image interpretation and understanding.

(10) Various Applications We here underline three notable applications in production and life. First, it can be used for quality control and inspection tasks, aiding in the identification of surface defects, scratches, or imperfections by analyzing the presence or absence of expected specular highlights. Second, it is valuable in autonomous driving systems, helping vehicles understand and respond to road conditions, such as wet or icy surfaces. Lastly, it can be applied in human-computer interaction, such as gesture recognition and gaze tracking. It would enable accurate tracking of hand movements or eye reflections, and enhance interaction with computer systems. Thus, researchers can investigate the impact of specular highlight detection on these or more applications in the future.

The above ten research directions listed for specular highlight detection are still far from being solved. To study these directions, researchers can refer to numerous excellent works in related fields, such as object detection, semantic segmentation, and saliency detection. We believe that exploring these directions would be fruitful in this field.

9 Conclusion

We have presented a real dataset for high-resolution specular highlight detection, consisting of 3619 images with manual annotations. Building upon this dataset, we have proposed a high-resolution specular highlight detection network that utilizes a patch-level bidirectional refinement scheme to progressively integrate the detection results of adjacent-scale specular highlight patches to generate the refined results. Additionally, we have proposed a modified inpainting framework for specular highlight removal as an application of our detection method. Extensive experiments have demonstrated the superior performance of our methods in comparison to state-of-the-art detection and removal methods. While we have taken a small step towards specular highlight detection for high-resolution images, numerous challenges and future research opportunities remain on this problem. We hope that our work inspires researchers to unearth more research points and design new detectors for specular highlight detection.

Acknowledgements This work is partially supported by the National Natural Science Foundation of China under Grant (No. 61972298), CAAI-Huawei MindSpore Open Fund, and the Research Program for Young and Middle-Aged Teachers of Fujian Province under Grant (No. JAT210036).

References

Akashi, Y., & Okatani, T. (2015). Separation of reflection components by sparse non-negative matrix factorization. *Computer Vision and Image Understanding, 100*(146), 77–85.

Angelopoulou, E. (2007) Specular highlight detection based on the fresnel reflection coefficient, In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1–8).

Bajcsy, R., Lee, S. W., & Leonardis, A. (1990). Color image segmentation with detection of highlights and local illumination induced by inter-reflections. In *Proceedings of the IEEE International Conference on Pattern Recognition* (pp. 785–790).

Barron, J. T. & Tsai, Y.-T. (2017) Fast Fourier color constancy. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 886–894)

Bell, S., Bala, K., & Snavely, N. (2014). Intrinsic images in the wild. *Transactions on Graphics, 33*(4), 159.

Borji, A., Cheng, M.-M., Jiang, H., & Li, J. (2015). Salient object detection: A benchmark. *IEEE Transactions on Image Processing, 24*(12), 5706–5722.

Brelstaff, G., & Blake, A. (1988). Detecting specular reflections using Lambertian constraints. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 297–302).

Chen, W., Jiang, Z., Wang, Z., Cui, K., & Qian, X. (2019). Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8924–8933).

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018a). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision* (pp. 801–818).

Chen, L., Zhu, Y., Papandreou, G., Schroff, F. & Adam, H. (2018b). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision* (pp. 833–851).

Cheng, H. K., Chung, J., Tai, Y.-W., & Tang, C.-K. (2020). CascadePSP: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8890–8899).

Cheng, M.-M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S.-M. (2014). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(3), 569–582.

Croitoru, I., Bogolin, S.-V., & Leordeanu, M. (2019). Unsupervised learning of foreground object segmentation. *International Journal of Computer Vision, 127*, 1279–1302.

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., & Borji, A. (2017). Structure-measure: A new way to evaluate foreground maps, In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4548–4557).

Fan, D.-P., et al. (2018). Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421

Fu, G., Zhang, Q., Lin, Q., Zhu, L., & Xiao, C. (2020). Learning to detect spec1ular highlights from real-world images. In *Proceedings of the ACM International Conference on Multimedia* (pp. 1873–1881).

Fu, G., Zhang, Q., Zhu, L., Li, P., & Xiao, C. (2021). A multi-task network for joint specular highlight detection and removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7752–7761).

Grosse, R., Johnson, M. K., Adelson, E. H., & Freeman, W. T. (2009). Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2335–2342).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778).

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., & Torr, P. (2019). Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(4), 815–828.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research, 5*(9), 1457–1469.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141).

Hu, X., Zhu, L., Fu, C.-W., Qin, J., & Heng, P.-A. (2018). Direction-aware spatial context features for shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7454–7462).

Hu, X., Fu, C.-W., Zhu, L., Qin, J., & Heng, P.-A. (2019). Direction-aware spatial context features for shadow detection and removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(11), 2795–2808.

Huynh, C., Tran, A. T., Luu, K., & Hoai, M. (2021). Progressive semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 16755–16764).

Kim, H., Jin, H., Hadap, S., & Kweon, I. (2013). Specular reflection separation using dark channel prior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1460–1467).

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Lang, C., Feng, J., Feng, S., Wang, J., & Yan, S. (2016). Dual low-rank pursuit: Learning salient features for saliency detection. *IEEE Transactions on Neural Networks and Learning Systems, 27*(6), 1190–1200.

Li, Z., & Snavely, N. (2018). Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9039–9048).

Li, R., et al. (2019). Specular reflections removal for endoscopic image sequences with adaptive-rpca decomposition. *IEEE Transactions on Medical Imaging, 39*(2), 328–340.

Lin, T.-Y., et al. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2117–2125).

Lin, P., et al. (2020). Graph-guided architecture search for real-time semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4203–4212).

Liu, Y., Li, Y., You, S., & Lu, F. (2020). Unsupervised learning for intrinsic image decomposition from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3248–3257).

Liu, G., et al. (2018). Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision* (pp. 85–100).

Murmann, L., Gharbi, M., Aittala, M., & Durand, F. (2019). A dataset of multi-illumination images in the wild. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4080–4089).

Netz, A., & Osadchy, M. (2012). Recognition using specular highlights. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(3), 639–652.

Osadchy, M., Jacobs, D. W., & Ramamoorthi, R. (2003). Using specularities for recognition. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1512–1519).

Pan, X., Shi, J., Luo, P., Wang, X., & Tang, X. (2018). Spatial as deep: Spatial CNN for traffic scene understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 7276–7283).

Park, J. B., & Kak, A. C. (2003). 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* (pp. 1397–1403).

Phong, B. T. (1975). Illumination for computer generated pictures. *Communications of the ACM, 18*(6), 311–317.

Qin, X., et al. (2019). BASNet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7479–7489).

Shafer, S. A. (1985). Using color to separate reflection components. *Color Research and Application, 10*(4), 210–218.

Shen, L., Tan, P., & Lin, S. (2008). Intrinsic image decomposition with non-local texture cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–7).

Shen, J., Yang, X., Jia, Y. & Li, X., (2011). Intrinsic images using optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3481–3487).

Shen, H.-L., & Zheng, Z.-H. (2013). Real-time highlight removal using intensity ratio. *Applied Optics, 52*(19), 4483–4493.

Shi, J., Dong, Y., Su, H., & Yu, S. X. (2017). Learning non-Lambertian object intrinsics across shapenet categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1685–1290).

Sun, Z., Cao, S., Yang, Y., & Kitani, K. (2020). Rethinking Transformer-based set prediction for object detection. arXiv preprint arXiv:2011.10881

Tan, R. T., & Ikeuchi, K. (2005). Separating reflection components of textured surfaces using a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(2), 178–193.

Tian, Q., & Clark, J. J. (2013). Real-time specularity detection using unnormalized wiener entropy. In *Proceedings of the IEEE International Conference on Computer and Robot Vision* (pp. 356–363).

Van Gansbeke, W., Vandenhende, S., Georgoulis, S., & Van Gool, L. (2021). Unsupervised semantic segmentation by contrasting object

mask proposals. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 10052–10062).

Wang, T., Hu, X., Wang, Q., Heng, P.-A., & Fu, C.-W. (2020). Instance shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1880–1889).

Wang, Y., Zhang, W., Wang, L., Liu, T., & Lu, H. (2022). Multi-source uncertainty mining for deep unsupervised saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 11727–11736).

Wang, R., et al. (2019). Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6849–6857).

Wu, T., et al. (2020). Patch proposal network for fast semantic segmentation of high-resolution images. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 12402–12409).

Wu, Z., et al. (2021). Single-image specular highlight removal via real-world dataset construction. *IEEE Transactions on Multimedia, 24*, 3782–3793.

Xie, C., et al. (2022). Pyramid grafting network for one-stage high resolution saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 11717–11726).

Yang, Y., & Soatto, S. (2020). FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4085–4095).

Yang, Q., Tang, J., & Ahuja, N. (2015). Efficient and robust specular highlight removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 37*(6), 1304–1311.

Zeng, Y., Zhang, P., Zhang, J., Lin, Z., & Lu, H. (2019). Towards high-resolution salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 7234–7243).

Zhang, P., Liu, W., Zeng, Y., Lei, Y., & Lu, H. (2021). Looking for the detail and context devils: High-resolution salient object detection. *IEEE Transactions on Image Processing, 30*, 3204-3216.

Zhang, L., Yan, Q., Liu, Z., Zou, H., & Xiao, C. (2017). Illumination decomposition for photograph with multiple light sources. *IEEE Transactions on Image Processing, 26*(9), 4114–4127.

Zhang, W., Zhao, X., Morvan, J.-M., & Chen, L. (2018). Improving shadow suppression for illumination robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(3), 611–624.

Zheng, Q., Qiao, X., Cao, Y., & Lau, R. W. (2019). Distraction-aware shadow detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5167–5176).

Zhou, P., Price, B., Cohen, S., Wilensky, G., & Davis, L. S. (2020). DeepStrip: High-resolution boundary refinement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 10558–1350).

Zhou, H., Xie, X., Lai, J.-H., Chen, Z., & Yang, L. (2020). Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9141–9150).

Zhu, L., et al. (2018). Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *Proceedings of the European Conference on Computer Vision* (pp. 122–137).