Towards High-Quality Specular Highlight Removal by Leveraging Large-Scale Synthetic Data

Gang Fu[1], Qing Zhang[2], Lei Zhu[3,4], Chunxia Xiao[5], and Ping Li[1,*]

[1]The Hong Kong Polytechnic University, Hong Kong
[2]Sun Yat-sen University, Guangzhou, China
[3]The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China
[4]The Hong Kong University of Science and Technology, Hong Kong
[5]School of Computer Science, Wuhan University, Wuhan, China

Abstract

This paper aims to remove specular highlights from a single object-level image. Although previous methods have made some progresses, their performance remains somewhat limited, particularly for real images with complex specular highlights. To this end, we propose a three-stage network to address them. Specifically, given an input image, we first decompose it into the albedo, shading, and specular residue components to estimate a coarse specular-free image. Then, we further refine the coarse result to alleviate its visual artifacts such as color distortion. Finally, we adjust the tone of the refined result to match that of the input as closely as possible. In addition, to facilitate network training and quantitative evaluation, we present a large-scale synthetic dataset of object-level images, covering diverse objects and illumination conditions. Extensive experiments illustrate that our network is able to generalize well to unseen real object-level images, and even produce good results for scene-level images with multiple background objects and complex lighting.

1. Introduction

Specular highlights are very common in the real world, but they are usually undesirable in photographs, since they can degrade the image quality. In daily life, users often want to achieve the specular-free image from an image. For example, specular highlights in facial or document images sweep away skin details or meaningful texture patterns which are very important to users. Removing specular highlights from a single image enables recovering visual content with better perceptibility. Moreover, it has many related applications such as recoloring [1], light source estimation [11], recognition of specular objects [18], and intrinsic image decomposition [31]. Thus, specular highlight removal is a long-standing and challenging problem in computer vision and computer graphics.

To address this problem, researchers have proposed various specular highlight methods. They can be roughly divided into two categories: traditional methods [25, 29, 13, 24] based on intensity and chromaticity analysis as well as optimization, and deep learning-based methods [30, 4, 28]. However, the traditional methods often produce unsatisfactory or even poor results with visual artifacts such as black color block and detail missing; see Figure 1(b). The main reason is that they fail to capture high-level semantic information to recover the missing colors and details underneath specular highlights using those meaningful and reliable information from the non-highlight region. In addition, although the deep learning-based methods have achieved certain performance improvement, they may still produce unsatisfactory results with visual artifacts such as illumination residue and color distortion; see Figure 1(c)(e). It is partly attributed to the fact that they are trained on relatively simple images in which materials and illumination conditions are not diverse enough, leading to their limited generalization to unseen images.

We in this paper propose a three-stage specular highlight removal network, consisting of (i) physics-based specular highlight removal, (ii) specular-free refinement, and (iii) tone correction. In the first stage, based on a physics-based image formation model, we decompose an input image into its albedo, shading, and specular residue components, and then estimate a coarse specular-free image. In the second stage, we further refine the coarse result to alleviate visual artifacts for improving the quality. In the third stage, we adjust the tone of the refined result to produce the final result with the similar tone of the input. In addition, to facilitate network training and quantitative evaluation, we build a

---

*Corresponding author.

Figure 1. Visual comparison of our method against state-of-the-art methods on a challenging image with nearly white material surfaces. (a) Input. (b) Yang *et al.* [29]. (c) Fu *et al.* [4]. (d) Wu *et al.* [28]. (e) Ours.

large-scale synthetic dataset rendered by software using diverse 3D models and real HDR environment maps. Figure 1 presents the visual comparison on a real image. As shown, our method is able to produce high-quality specular-free images without noticeable artifacts encountered by previous methods. Below, we summarize the major contributions of our work.

- We propose a three-stage specular highlight removal network to progressively eliminate multiple types of visual artifacts such as color distortion and tone inconsistency.

- We present a large-scale synthetic dataset of object-level images, in which each specular highlight image has corresponding ground truth albedo, shading, specular residue, and specular-free images.

- We conduct extensive experiments on existing datasets and our new dataset, and demonstrate that our method achieves better quantitative and qualitative results than state-of-the-art methods.

2. Related Work

Single-Image methods. Early methods are mostly based on chromaticity propagation or optimization. Tan and Ikeuchi [25] proposed to remove specular highlights via iteratively comparing the intensity logarithmic differentiation of an input image and its specular-free image. Yang *et al.* [29] proposed to use the bilateral filter to propagate information from the diffuse region to the specular highlight region. Kim *et al.* [13] formulated specular highlight removal as a MAP optimization problem based on the priors of specular highlights in the real world. However, these methods may produce unsatisfactory results with visual artifacts such as black color block, resulting in unrealistic appearances. To alleviate the issue, Liu *et al.* [16] proposed a two-step method in which an over-saturated specular-free image is first produced by global chromaticity propagation, and then recovered its saturation via an optimization framework. Guo *et al.* [6] proposed a sparse and low-rank reflection model for specular highlight removal. However, they may fail to effectively recover the missing content underneath specular highlights.

Subsequently, researchers have proposed various deep learning-based methods. Shi *et al.* [23] proposed a unified framework that can simultaneously estimate the albedo, shading, and specular residue components from a single object-level image. However, it fails to generalize well to real images with complex specular highlights. Yi *et al.* [30] proposed to leverage multi-view image sets (*i.e.*, customer product photos) to perform specular highlight removal in an unsupervised way. Fu *et al.* [4] proposed a multi-task network for joint specular highlight detection and removal based on a region-aware specular highlight image formation model. Wu *et al.* [28] proposed a GAN-based network for specular highlight removal using specular highlight detection map as guidance. Jin *et al.* [12] proposed to estimate the reflectance layer from a single image with shadows and specular highlights. Although these methods achieve good results, their performance is often limited, particularly for real images with adverse factors such as achromatic material surfaces and complex illumination conditions. In contrast, our three-stage method is able to effectively address previous challenging images.

Multi-Image and Normal-Based Methods. Researchers have proposed various multi-image and normal-based methods to more robustly remove specular highlights. Guo *et al.* [7] proposed to remove specular highlights for superimposed multiple images. Wei *et al.* [27] proposed a unified framework of specular highlight removal and light source position estimation by assuming that surface geometry is known. Li *et al.* [14] proposed a method for specular highlight removal in facial images that may contain varying illumination colors, with the help of facial surface normals. Although these methods can produce promising results, the requirement of multiple images or extra auxiliary cues limits their applicability.

Benckmark Datasets. Grosse *et al.* [5] presented the MIT intrinsic images dataset, including 20 object-level images and their corresponding ground truth intrinsic images. However, these images are not sufficient to support network training. Shi *et al.* [23] rendered a large-scale synthetic dataset for non-Lambertian intrinsic image decomposition by software. Although this dataset includes a large amount of images, many of them do not have obvious and meaningful specular highlights for our task. Recently, Fu *et al.* [4] presented a real dataset simultaneously for specular high-

light detection and removal, produced by a series of image processing algorithms on the multi-illumination dataset IIW [17]. At the same time, Wu *et al.* [28] also built a real paired specular-diffuse image dataset via the cross-polarization photography technique. However, objects and illumination conditions in these two datasets are somewhat limited for network training, leading to the unsatisfactory generalization to unseen images. In contrast, we present a large-scale synthetic dataset of object-level images, which covers diverse objects and illumination conditions, and thus contains various appearances of specular highlights.

3. Methodology

3.1. Overview

Figure 2 presents the pipeline of our three-stage framework. It consists of three stages: (i) physics-based specular highlight removal; (ii) specular-free refinement; and (iii) tone correction. Specifically, in the first stage (see (a)), we decompose an input image into its albedo and shading using two encoder-decoder networks (E_a-D_a for albedo, and E_s-D_s for shading). Then, the specular-free image can be estimated by multiplying the albedo and shading. In the second stage (see (b)), we feed the coarse result along with the input into an encoder-decoder network (E_r-D_r) to further refine it to alleviate visual artifacts. In the third stage (see (c)), we feed the refined result along with the input and its specular residue image into an encoder-decoder network (E_c-D_c) to adjust its tone so that it has the similar tone as the input as much as possible. Figure 5 validate the effectiveness of each stage in our framework.

3.2. Physics-Based Specular Highlight Removal

According to the dichromatic reflection model [21], an input image I can be decomposed into its intrinsic images [1], expressed as

$$I = A \times S + R,\qquad(1)$$

where A, S, and R are albedo, shading, and specular residue, respectively. Based on the physical image formation model in Eq. (1), we propose the Physics-based Specular Highlight Removal stage (PSHR) to recover the intrinsic images from an input image. Figure 2(a) illustrates the mechanism of PSHR. Specifically, given an input image, we use an encoder-decoder network (E_a-D_a) to estimate albedo, and another one (E_s-D_s) to estimate shading. The specular-free (*i.e.*, diffuse) image D is estimated by

$$D = A \times S,\qquad(2)$$

---

[1]Throughout the paper we use the terms *albedo* and *shading* loosely for simplicity. Actually, *albedo* and *shading* refer to diffuse albedo and diffuse shading, respectively.
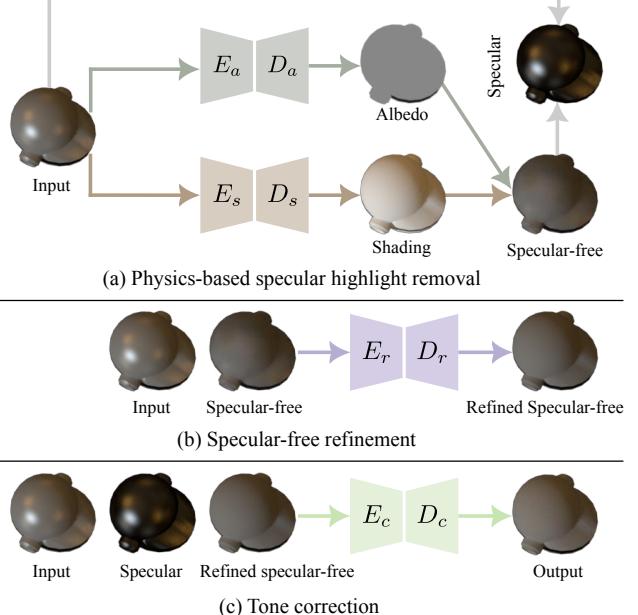

(a) Physics-based specular highlight removal

(b) Specular-free refinement

(c) Tone correction

Figure 2. The pipeline of our three-stage specular highlight removal framework.

With Eqs. (1) and (2), we can yield the specular residue R by

$$R = I - D.\qquad(3)$$

To facilitate the network training of PSHR, we present a large-scale synthetic dataset of object-level images for specular highlight removal (named SSHR). Now, we detail it.

Dataset. To the best of our knowledge, SHIQ [4] and PSD [28] are only two publicly available real datasets for specular highlight removal. However, they suffer from the following three issues. First, the quantity of objects is quite small, and the images were captured in controllable laboratory environments with limited illumination conditions. Second, a pair of specular highlight and specular-free images may not be aligned well, since the camera shakes caused by itself or hand touch during the process of capturing data. Third, even a well-aligned pair of specular highlight and specular-free images may have inconsistent color and shading, since the environmental lighting may have a subtle fluctuation over time and the camera exposure may vary. In addition, Shi *et al.* [23] presented a large-scale synthetic dataset for non-Lambertian intrinsic image decomposition. However, most input images in it are not with obvious and meaningful specular highlights, and thus are not well-suited for our task. Note that this dataset is currently not publicly available.

To this end, we built a large-scale synthetic dataset tailored for specular highlight removal. Specifically, to render the data, we first picked up 1500 3D models with their albedo texture maps from several common categories (such as car, bus, container, and sofa) of the large-scale 3D shape

Figure 3. Example environment maps for our rendering. Top: indoor scenes. Bottom: outdoor scenes.
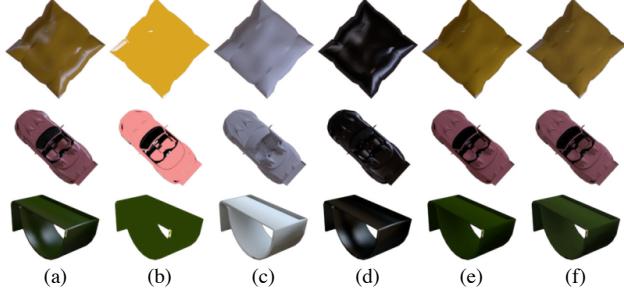

Figure 4. Example image groups in our dataset. (a) Input. (b) Albedo. (c) Shading. (d) Specular residue. (e) Ground truth. (f) Tone correction version of (e).

dataset ShapeNet [2]. Then, we collected 90 HDR environment maps from the Internet [2], which includes indoor and outdoor scenes with diverse material surfaces and illumination conditions. Figure 3 presents example environment maps. Finally, we used an open-source render software Mitsuba [10] and adopted the modified Phong reflection model [19] to render object models with various environment maps to generate photo-realistic shading and specular residue appearance. According to the rendered results, the specular-free and input images can be obtained via Eqs. (2) and (1), respectively. Finally, we randomly split the collected 1500 models into 1300 models for training and 200 for testing. In total, we have 117,000 training images and 18,000 testing images. Figure 4 shows example image groups in our dataset.

Loss Function. The total loss for physics-based specular highlight removal $\mathcal{L}^{\text{PSHR}}$ is defined as

$$\mathcal{L}^{\text{PSHR}} = ||A - \hat{A}||^2 + ||S - \hat{S}||^2 + ||I - D_1 - \hat{R}||^2 , \quad (4)$$

where $\hat{A}$, $\hat{S}$, and $\hat{R}$ are the ground truths of the estimated albedo A, shading S, and specular residue R, respectively; and $D_1 = A \times S$ is the estimated specular-free image. The rightmost term of Eq. (4) is to encourage the estimated specular residue image (*i.e.*, $I - D_1$) to be similar with its ground truth as much as possible.

3.3. Specular-Free Refinement

The first stage, PSHR, has two drawbacks. First, it tends to overly remove specular highlights and produce visual ar-
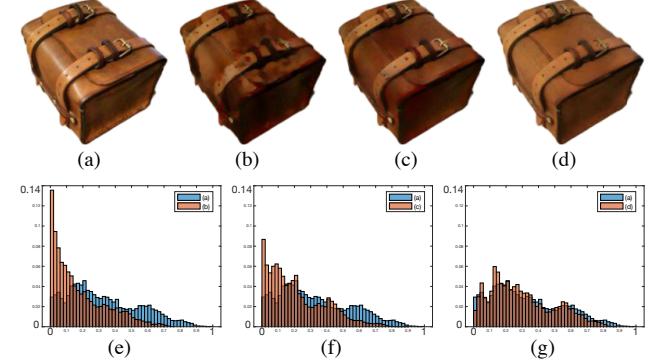

Figure 5. Ablation study that demonstrates the effectiveness of each stage in our framework. (a) Input. (b)-(d) Resulting specular-free images produced by the first, second, and third stages in our framework, respectively. (e)-(f) Histogram comparison between (a) and (b)-(d), respectively. Note that the abscissa and ordinate axes indicate the pixel intensity value and the ratio of the number of target pixels and the total number of pixels in an image.

tifacts such as color distortion and black color block; see Figure 5(b). Second, the estimation of specular-free image by Eq. (2) in low dynamic range has a certain amount of error, while that in high dynamic range is correct and accurate. In our dataset, the rendering of shading and specular residue images, as well as the estimation of specular-free and specular highlight images, is carried out in high dynamic range. And all generated images are converted to be of low dynamic range for network training.

To overcome the above issues, we propose the Specular-free Refinement stage (SR) to further refine the result from PSHR. Figure 2(b) illustrates the mechanism of SR. As shown, the coarse specular-free image, along with the input, is fed into an encoder-decoder network (E_r-D_r) to produce a refined result. Compared to PSHR, SR is able to produce better results in terms of detail preserving and natural appearances; see Figure 5(c). Furthermore, the histogram comparisons of Figure 5(e)(f) also validate the performance improvement. As shown, compared to the coarse result, the intensity distribution of the refined result is more consistent with that of the input image over the non-highlight region.

Loss Function. The loss for specular-free refinement $\mathcal{L}^{\text{SR}}$ is defined as

$$\mathcal{L}^{\text{SR}} = ||D_2 - \hat{D}||^2 , \quad (5)$$

where D_2 and $\hat{D}$ are the refined specular-free image and its ground truth, respectively.

3.4. Tone Correction

Although the specular-free image from PSHR is further refined by SR, its overall tone is sometimes noticeably different from the input, and thus looks somewhat unreal; see Figure 5(c). The main reason is that the specular-free images in our training data are of slightly lower brightness than

the input images, due to the inherent defect of software rendering; see Figure 4. To overcome this issue, we propose the Tone Correction stage (TC) to adjust the tone of the refined result to match that of the input as closely as possible. Figure 2(c) illustrates the mechanism of TC. As shown, the refined result, along with the input and specular residue images, is fed into an encoder-decoder network (E_c-D_c) to produce a tone-corrected result. Figure 5 validate the effectiveness of TC in terms of tone preservation. From it, we can see that the overall tone of the tone-corrected result by TC is significantly closer to that of the input than the results by PSHR and SR.

The key idea of TC is to correct the tone of the ground truth specular-free images in our dataset as new supervisions for network training. Figure 6 illustrates this mechanism. Formally, given an input image I, and its ground truth specular-free image $\hat{D}$ and specular residue image $\hat{R}$, we first use Otsu's method on $\hat{R}$ to separate all pixels of I into two types of regions, specular highlight region M_h and non-highlight region M_n. Then, we find tone correction function T that minimizes the tone correction error E between the specular-free and input images over the non-highlight region:

$$E = |T(\hat{D}) - I|_{\Omega_{M_n}}^2 , \qquad (6)$$

where Ω_{M_n} denotes all pixels of M_n. We formulate T as the following linear transformation:

$$T = \mathbf{M} * (p_h \ \ p_s \ \ p_v \ \ 1)' , \qquad (7)$$

where p denotes a pixel in $\hat{D}$, whose intensity value in HSV color space is (p_h, p_s, p_v); $\mathbf{M}$ is a 3×4 matrix which stores the parameters in the tone correction function; $*$ denotes matrix multiplication; and $(\cdot)'$ denotes matrix transpose. The above operation in HSV instead of RGB benefits obtaining a robust solution, because specular highlights mainly cause variations in the saturation and value channels. We can solve the problem in Eq. (6) using the least-squares method. Finally, we utilize T to correct all pixels of $\hat{D}$ for each training group in our dataset, and use them as new supervisions for network training.

Loss Function. The loss for tone correction $\mathcal{L}^{\text{TC}}$ is defined as

$$\mathcal{L}^{\text{TC}} = ||D_3 - \tilde{D}||^2 , \qquad (8)$$

where D_3 and $\tilde{D}$ are the tone-corrected specular-free image and its ground truth, respectively.

3.5. Network Training

The total loss $\mathcal{L}$ for the training of our whole network includes $\mathcal{L}^{\text{PSHR}}$, $\mathcal{L}^{\text{SR}}$, and $\mathcal{L}^{\text{TC}}$, written as

$$\mathcal{L} = \lambda_1 \mathcal{L}^{\text{PSHR}} + \lambda_2 \mathcal{L}^{\text{SR}} + \lambda_3 \mathcal{L}^{\text{TC}} . \qquad (9)$$

Here, λ_1, λ_2, and λ_3 are the weighting balance parameters, which are experimentally set to 1.
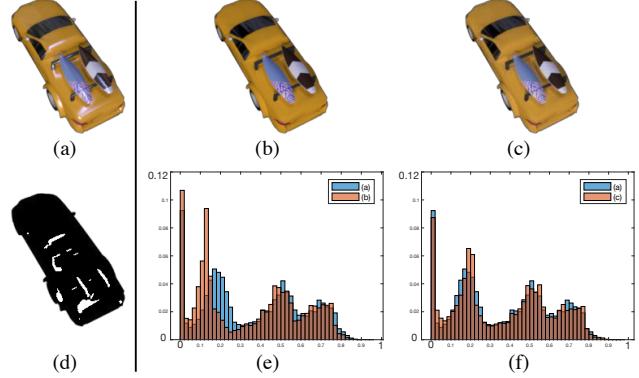

Figure 6. Tone correction for ground truth specular-free images in our dataset. (a) Input. (b) Ground truth specular-free image. (c) Tone correction version of (b). (d) Specular highlight mask of (a). (e) Histogram comparison between (a) and (b). (f) Histogram comparison between (a) and (c). Note that the specular highlight pixels are excluded using (d) for plotting histograms of (a)-(c).

3.6. Implementation Details

The four encoder-decoder networks in our three-stage framework have the same architecture. We adopt the U-Net architecture [20] as the default choice, known for its conciseness and effectiveness. We implement our whole network in PyTorch and train it for 60 epochs on a PC with NVIDIA GeForce GTX 3090Ti. The whole network is optimized using the Adam optimizer. The initial learning rate is set to 1×10^{-4}, divided by 10 after every 10 epochs, and the batch size is set to 16. Moreover, we also adopt horizontal flip, and specular highlight attenuation and boosting editing [4] for data augmentation.

4. Experimental Results

4.1. Datasets and Evaluation Metrics

We evaluate our network on three datasets, including our SSHR, SHIQ [4], and PSD [28]. We adopt two commonly-used metrics (*i.e.*, PSNR and SSIM) to quantitatively evaluate the performance of our network, as in [4, 29]. In general, higher PSNR and SSIM values indicate better results.

4.2. Comparison with State-of-the-Art Methods

We compare our method against four traditional methods [25, 22, 29, 3] and two recent deep learning-based methods [4, 28]. For fair comparison, we produce removal results for four traditional methods using publicly available implementation provided by the authors with optimal parameter setting. Besides, if necessary, we re-train two deep learning-based methods, and fine-tune their key parameters to produce better results as much as possible. We note that our network fails to be trained on SHIQ and PSD, since they do not include ground truth intrinsic images. To train and evaluate our network on them, we modify the first stage of our
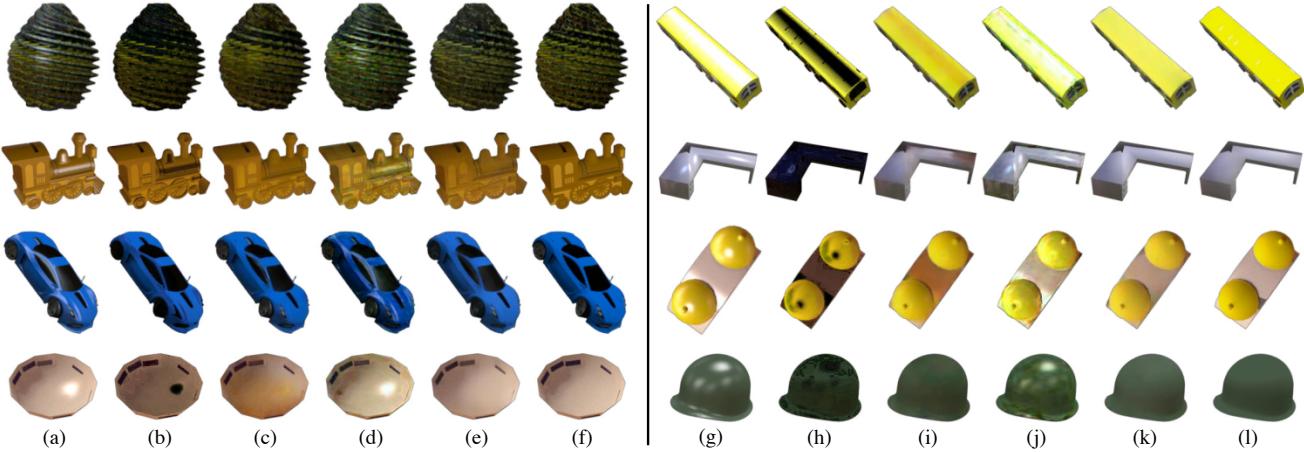
Figure 7. Visual comparison of our method against state-of-the-art methods on our synthetic testing images. (a)(g) Input. (b)(h) Yang *et al.* [29]. (c)(i) Fu *et al.* [4]. (d)(j) Wu *et al.* [28]. (e)(k) Ours. (f)(l) Ground truth.

method to estimate the specular-free and specular residue instead of the original albedo and shading.

Quantitative Comparison. Tables 1 reports the quantitative comparison result on three datasets. As shown, overall, our method achieves higher PSNR and SSIM values, indicating that our method is superior to state-of-the-art methods. In addition, four traditional methods [25, 22, 29, 3] achieve much higher PSNR and SSIM values on our synthetic dataset than real SHIQ and PSD datasets. The reason is two-fold. First, these methods are based on the dichromatic reflection model, and so does the rendering of our synthetic dataset. As a result, they are capable of addressing our synthetic images. Second, real specular highlights in SHIQ and PSD may not be well characterized by an idealized image formation model, while images in them are often with adverse factors such as white material surfaces and heavy texture.

Visual Comparison. Figure 7 presents the visual comparison on our testing images. We can see that for images with nearly white material surfaces, traditional methods often produce unrealistic results with severe visual artifacts such as color distortion (see the 4th row in (b)) and black color block (see the 1st row in (h)). Although the deep learning-based method [4] is able to effectively remove specular highlights and recover the missing details, it sometimes suffers from color distortion artifacts (see the 4th in (c)). Besides, the deep learning method [28] fails to effectively remove specular highlights (see the 1st row in (d)), and may produce unreasonable texture details (see the 2nd row in (d)). In comparison, our method is able to produce high-quality photo-realistic removal results without noticeable visual artifacts caused by previous methods. Due to space limit, the visual comparisons on SHIQ and PSD are provided in our supplementary material.

Table 1. Quantitative comparison of our method with state-of-the-art specular highlight removal methods on our SSHR, SHIQ [4], and PSD [28]. The best results are marked in **bold**, while the second-best results are underlined. Ours-A, Ours-B, and Ours-C denote our network without the specular-free refinement stage, the tone correction stage, and both these two stages, respectively.

Dataset	SSHR		SHIQ		PSD	
Metric	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Tan [25]	24.281	0.874	11.041	0.403	11.581	0.560
Shen [22]	24.388	0.904	13.923	0.428	13.886	0.610
Yang [29]	23.243	0.894	14.310	0.502	12.866	0.611
Fu [3]	23.270	0.881	15.746	0.723	14.400	0.665
Fu [4]	26.979	0.895	**34.131**	0.860	21.516	0.883
Wu [28]	25.731	0.894	23.420	0.920	21.801	0.880
Ours-A	26.083	0.918	24.930	0.896	21.263	0.897
Ours-B	**28.903**	**0.945**	22.019	0.843	18.932	0.830
Ours-C	24.231	0.893	20.309	0.805	18.301	0.801
Ours	28.633	0.940	25.575	**0.933**	**22.759**	**0.903**

User Study. We further conducted a user study to evaluate the robustness and generalization capability of our method on real images. Here, three recent state-of-the-art methods [3, 4, 28] are compared. We first randomly downloaded 200 images from the Internet by searching the keywords "chair", "statue", "storage bag", and "decoration". Figure 8(a) presents several example images. Then, we produced specular-free images for all downloaded images using our method and other compared methods, and recruited 20 participants from a school campus for rating. Finally, we asked the participants to score all results in a random order using a 1(worst)-to-4(best) scale (as done in [26, 8]) on the three questions: (1) Is the result free of highlights? (denoted as Q1); (2) Are the missing details recovered? (denoted as Q2); and (3) Is the result visually realistic? (denoted as Q3).

Figure 9 summarizes the user study results, where the

Figure 8. Visual comparison of our method against state-of-the-art methods on real object-level images. (a)(f) Input and its grayscale version, respectively. (b)(g) Fu *et al.* [3]. (c)(h) Fu *et al.* [4]. (d)(i) Wu *et al.* [28]. (e)(j) Ours.
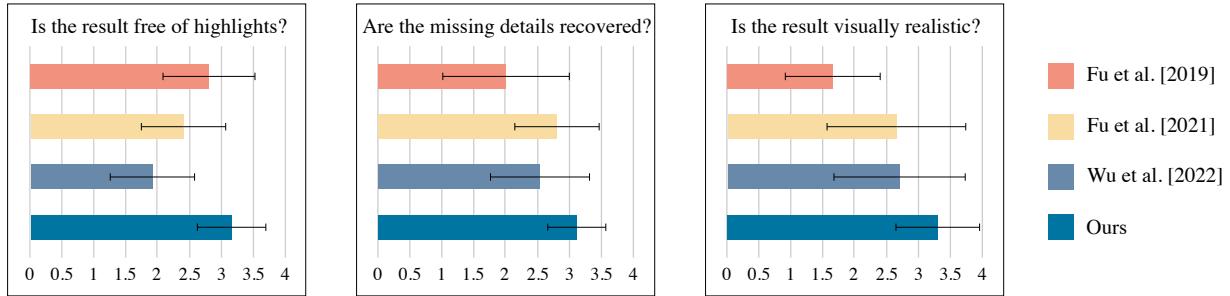

Figure 9. User study results on the three questions.

average and standard deviation values of scores received by each method are presented. As shown, our method achieves higher average scores and lower standard deviations, indicating that our results are more preferred by the participants with lower subjective bias. Figure 8 presents the visual comparison on example images. As shown, our method is able to effectively address real images and produce high-quality results with natural appearances, although it is trained on the synthetic data.

4.3. Discussions

Ablation Study. Besides the visual comparison results shown in Figure 5, we also quantitatively validate the effectiveness of each stage of our method (denoted as "Ours") by constructing the following three variants:

- Ours-A: ours without *specular-free refinement*.
- Ours-B: ours without *tone correction*.
- Ours-C: ours without both *specular-free refinement*

and *tone correction* (*i.e.*, only with *physics-based specular highlight removal*).

Table 1 reports the quantitative results of our method and its variants on our SSHR, SHIQ, and PSD. From the results, we can observe that the PSNR and SSIM scores of our method and its three variants overall follow the relationship: Ours > Ours-A > Ours-B > Ours-C, except for a special case: Ours-B > Ours > Ours-A > Ours-C on our dataset. From it, we can draw two conclusions. First, as the number of the used stages increases, the performance of our method overall gets better and better, illustrating the effectiveness of each stage of our method. Second, the tone correction stage leads to a performance drop on our dataset, due to the domain gap between our synthetic data and its tone correction version. However, it further improves the performance on SHIQ and PSD. This is because the resulting errors from the differences between them and their tone correction versions can be fully offset by the performance gain brought by further learning of the network.
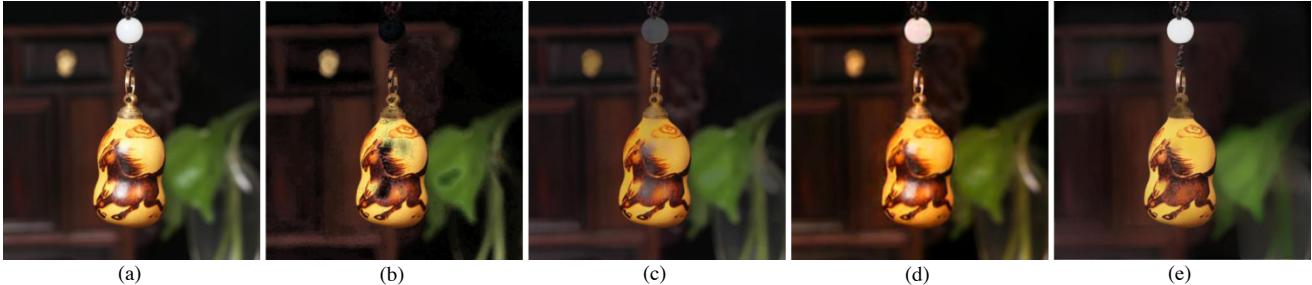
Figure 10. Visual comparison of our method against state-of-the-art methods on a challenging image with multiple background objects. (a) Input. (b) Yang *et al.* [29]. (c) Fu *et al.* [4]. (d) Wu *et al.* [28]. (e) Ours.

Generalization to Grayscale Images. Figure 8 presents the visual comparison on color images (see the left column) and their grayscale version (see the right column). As can be seen, the traditional method [3] suffers from leaking a small amount of specular highlights into the specular-free images (see the 1st and 2nd rows in (g)). For two deep learning-based methods, the method [4] sometimes fails to effectively remove specular highlights (see the 3rd row in (h)). The method [28] produces unsatisfactory or even poor results with visual artifacts such as severe color distortion and disharmonious color block. In comparison, our method trained on our synthetic data is able to generalize well to real grayscale images, which have almost the same performance as on color images.

Generalization to Scene-Level Images. Figure 10 presents the visual comparison on scene-level images. As can be seen, the traditional method [29] often mistakes white material surfaces (see the circular jade in (b)) as specular highlights to be removed, and undesirably produce black color block artifacts. For the two deep learning-based methods, the method [4] fails to effectively recover the missing color underneath specular highlights. The method [28] often produces unsatisfactory results with color distortion artifacts. In comparison, our method produces good results with realistic color and clear texture details. This illustrates that our method is able to generalize to scene-level images with multiple background objects to a certain extent.

Limitations. Our method has two limitations. First, our method, as well as previous methods, all fail to recover missing texture details and color underneath strong (*i.e.*, high-intensity and large-area) specular highlights. Figure 11 presents an example. As can be seen, the missing detailed patterns on the body of the wooden kitten underneath strong specular highlights (see the red boxes) are less able to be recovered very well. Second, although our method achieves good results for object-level images, it may produce unsatisfactory results, particularly for complex natural scenes often with achromatic material surfaces, color lighting, noise, and so on.
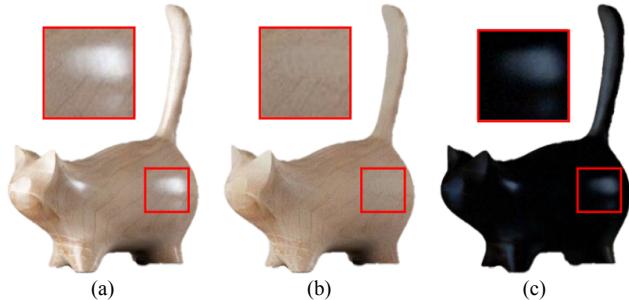

Figure 11. A failure case of our method. (a) Input. (b) Specular-free image. (c) Specular residue image.

5. Conclusion

We in this paper have proposed a three-stage method for object-level specular highlight removal. Our key idea is to progressively eliminate multiple types of visual artifacts to produce high-quality results with natural appearances. In addition, we have presented a large-scale synthetic dataset of object-level images to facilitate network training and quantitative evaluation. In our dataset, each input specular highlight image has corresponding ground truth albedo, shading, specular residue, and specular-free images. We have conducted extensive experiments to illustrate the superiority of our method over previous methods in terms of quantitative comparison (*i.e.*, higher PSNR and SSIM values), visual comparison, and a user study.

Our future work is to integrate features from inpainting [15] into our network to remove strong specular highlights while restoring the missing texture details and color underneath them. Another direction is to design more effective and complex backbone networks such as diffusion models [9] to further improve the performance of our method.

Acknowledgments

References

[1] Shida Beigpour and Joost Van De Weijer. Object recoloring based on intrinsic image estimation. In *ICCV*, pages 327–334, 2011. 1

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 4

[3] Gang Fu, Qing Zhang, Chengfang Song, Qifeng Lin, and Chunxia Xiao. Specular highlight removal for real-world images. *Computer Graphics Forum*, 38(7):253–263, 2019. 5, 6, 7, 8

[4] Gang Fu, Qing Zhang, Lei Zhu, Ping Li, and Chunxia Xiao. A multi-task network for joint specular highlight detection and removal. In *CVPR*, pages 7752–7761, 2021. 1, 2, 3, 5, 6, 7, 8

[5] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, pages 2335–2342, 2009. 2

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

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

[8] Mingming He, Jing Liao, Dongdong Chen, Lu Yuan, and Pedro V Sander. Progressive color transfer with dense semantic correspondences. *ACM Transactions on Graphics*, 38(2):1–18, 2019. 6

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NIPS*, pages 6840–6851, 2020. 8

[10] Wenzel Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 4

[11] Salma Jiddi, Philippe Robert, and Eric Marchand. Detecting specular reflections and cast shadows to estimate reflectance and illumination of dynamic indoor scenes. *IEEE Transactions on Visualization and Computer Graphics*, 28(2):1249–1260, 2020. 1

[12] Yeying Jin, Ruoteng Li, Wenhan Yang, and Robby T Tan. Estimating reflectance layer from a single image: Integrating reflectance guidance and shadow/specular aware learning. pages 1069–1077, 2022. 2

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

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

[15] Qiankun Liu, Zhentao Tan, Dongdong Chen, Qi Chu, Xiyang Dai, Yinpeng Chen, Mengchen Liu, Lu Yuan, and Nenghai Yu. Reduce information loss in transformers for pluralistic image inpainting. In *CVPR*, pages 11347–11357, 2022. 8

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

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

[18] Aaron Netz and Margarita Osadchy. Recognition using specular highlights. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):639–652, 2012. 1

[19] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 4

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015. 5

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

[22] Hui-Liang Shen and Zhi-Huan Zheng. Real-time highlight removal using intensity ratio. *Applied Optics*, 52(19):4483–4493, 2013. 5, 6

[23] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, pages 1685–1694, 2017. 2, 3

[24] Minjung Son, Yunjin Lee, and Hyun Sung Chang. Toward specular removal from natural images based on statistical reflection models. *IEEE Transactions on Image Processing*, 29:4204–4218, 2020. 1

[25] Robby T Tan and Katsushi Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(2):178–193, 2005. 1, 2, 5, 6

[26] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *CVPR*, pages 6849–6857, 2019. 6

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

[28] Zhongqi Wu, Chuanqing Zhuang, Jian Shi, Jianwei Guo, Jun Xiao, Xiaopeng Zhang, and Dong-Ming Yan. Single-image specular highlight removal via real-world dataset construction. *IEEE Transactions on Multimedia*, 24:3782–3793, 2022. 1, 2, 3, 5, 6, 7, 8

[29] Qingxiong Yang, Jinhui Tang, and Narendra Ahuja. Efficient and robust specular highlight removal. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(6):1304–1311, 2015. 1, 2, 5, 6, 8

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

[31] Qing Zhang, Jin Zhou, Lei Zhu, Wei Sun, Chunxia Xiao, and Wei-Shi Zheng. Unsupervised intrinsic image decomposition using internal self-similarity cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9669–9686, 2021. 1