mathematics

MDPI

Article

Weakly Supervised Specular Highlight Removal Using Only Highlight Images

Yuanfeng Zheng [1], Guangwei Hu [1], Hao Jiang [1,*], Hao Wang [1] and Lihua Wu [2]

[1] School of Electronic Information, Wuhan University, Wuhan 430072, China; zhengyuanfeng@whu.edu.cn (Y.Z.); huguangwei@whu.edu.cn (G.H.); whuwanghao@163.com (H.W.)
[2] Wuhan Second Ship Design and Research Institute, Wuhan 430064, China; wlhcheers@126.com
* Correspondence: jh@whu.edu.cn

Abstract: Specular highlight removal is a challenging task in the field of image enhancement, while it can significantly improve the quality of image in highlight regions. Recently, deep learning-based methods have been widely adopted in this task, demonstrating excellent performance by training on either massive paired data, wherein both the highlighted and highlight-free versions of the same image are available, or unpaired datasets where the one-to-one correspondence is inapplicable. However, it is difficult to obtain the corresponding highlight-free version of a highlight image, as the latter has already been produced under specific lighting conditions. In this paper, we propose a method for weakly supervised specular highlight removal that only requires highlight images. This method involves generating highlight-free images from highlight images with the guidance of masks estimated using non-negative matrix factorization (NMF). These highlight-free images are then fed consecutively into a series of modules derived from a Cycle Generative Adversarial Network (Cycle-GAN)-style network, namely the highlight generation, highlight removal, and reconstruction modules in sequential order. These modules are trained jointly, resulting in a highly effective highlight removal module during the verification. On the specular highlight image quadruples (SHIQ) and the LIME datasets, our method achieves an accuracy of 0.90 and a balance error rate (BER) of 8.6 on SHIQ, and an accuracy of 0.89 and a BER of 9.1 on LIME, outperforming existing methods and demonstrating its potential for improving image quality in various applications.

Keywords: image processing; specular highlight removal; non-negative matrix factorization; weakly supervised learning; GANs

MSC: 68U10

Citation: Zheng, Y.; Hu, G.; Jiang, H.; Wang, H.; Wu, L. Weakly Supervised Specular Highlight Removal Using Only Highlight Images. *Mathematics* **2024**, *12*, 2578. https://doi.org/10.3390/math12162578

Academic Editor: Shuai Liu

Received: 8 July 2024
Revised: 10 August 2024
Accepted: 19 August 2024
Published: 21 August 2024

1. Introduction

Upon illumination, surfaces composed of highly reflective materials inevitably become obscured by light spots, leading to a marked deterioration in the quality of the highlighted regions. Specifically, this phenomenon is particularly conspicuous in image content, where light spots engender increased interference and uncertainty, especially under complex lighting conditions. The effective removal of these light spots is therefore crucial, as it can greatly impact the performance of computer vision tasks that necessitate high-quality inputs, encompassing applications such as object detection [1], semantic segmentation [2], and object tracking [3].

The formation of specular highlights is a complex physical phenomenon, resulting in significant challenges in establishing effective physical models. Despite the availability of numerous illumination models, the specular reflections generated by diverse materials and lighting conditions introduce substantial uncertainty. Conventional methods for addressing this issue can be broadly classified into two primary categories: multi-image and single-image techniques. Multi-image highlight removal methods typically utilize viewpoint

dependence, leveraging several images to identify and separate specular and diffuse pixels [4]. While these methods yield superior performance, they are computationally intensive. Single-image highlight removal confronts more complex situations, necessitating the establishment of specific assumptions or conditions. They mainly consist of specific types of images [5–8], and the estimates of illumination, reflectance, geometry, and surface material properties [9–19]. However, the current methods for image highlight removal are susceptible to the erroneous classification of white regions as highlights, particularly in complex real-world situations.

Following the proposal of convolutional neural networks (CNNs), numerous recent methods for highlight removal have pivoted towards deep learning, resulting in substantially enhanced performance when benchmarked against traditional approaches.

In the majority of deep learning-based methods, network performance is critically contingent upon a training set that necessitates laborious annotations. Additionally, these methods mandate that the training data be either paired or aligned. For example, in the acquisition of a highlight dataset, it is necessary to obtain both the highlight image and its corresponding highlight-free image at an identical location. However, capturing such paired images under natural conditions proves to be a formidable challenge, considering specular highlights tend to be produced readily in the presence of light. While acquiring such paired images in a controlled laboratory setting may weaken these issues, it can compromise the generalizability of the model to real-world scenarios. This reliance on paired data has become an significant obstacle in the development of large-scale and robust highlight removal models.

One potential strategy, such as employing unsupervised methods [20], involves the utilization of unpaired data for training purposes. The collection of highlight-free images does not directly correspond to the highlight set. However, these unsupervised methodologies necessitate a sufficient statistical similarity between the two image sets. In practice, the challenge of capturing highlight-free images with good variety in natural environments remains unresolved. A recent method [21] proposes a feasible solution to mitigate the data dependency issue in shadow removal. This approach capitalizes on the observation that an image encompassing highlight regions inevitably contains highlight-free regions as well, both of which can be harnessed for highlight removal. Specifically, different regions within an image can be excerpted to serve as unpaired data for training. Due to the fact that the data are derived from identical images, the statistical similarity between the two sets of data is well guaranteed.

In order to separate highlight areas from those devoid of highlights, it is initially necessary to obtain the pertinent highlight masks. These masks can be approximated by applying highlight detection methods. In this paper, we estimate the highlight masks by integrating the NMF method, a procedure also employed in [22,23] for analogous tasks. Once the corresponding masks are acquired, we can excerpt the highlight regions to compose the highlight training set. Moreover, we utilize the masks from other images for extracting datasets devoid of highlights from the highlight-free regions, thereby facilitating the training of the model using these two datasets.

Employing the above idea, we aim to achieve the task of specular highlight removal via weak supervision, given the absence of ground truth data. This solution draws inspiration from the recently introduced G2R-ShadowNet [24], which is trained with weak supervision using solely shadow images and their associated masks. Before generating the masks, we first detect the existence of highlight conditions in the images using a highlight detection method. This ensures that our proposed method is specifically applied to images that contain specular highlights. Once highlights are detected, we generate the highlight-free dataset by extracting the highlight-free region from the original image, guided by a random mask calculated by the NMF method applied to a separate image.

The network we propose comprises three principal modules that are jointly trained: highlight generation, highlight removal, and reconstruction. The highlight generation module generates artificial highlights, paired with the corresponding highlight-free regions

from the input highlight image, thereby constructing a paired training dataset. Subsequently, the highlight removal module leverages this paired dataset to train its capability for effective highlight removal. However, it is important to note that the output image from the highlight removal module may still exhibit discrepancies from the ground truth in terms of color and illumination. To address this, we introduce a reconstruction module that utilizes contextual information from the surrounding area to refine the image, resulting in a more realistic processing outcome. Extensive experiments conducted on the Specular Highlight Image Quadruples (SHIQ) [25] and LIME [26] datasets have demonstrated the superiority of our proposed method in comparison to existing techniques for highlight removal in natural images. The main contributions of this article include the proposal of a weakly supervised learning framework for specular highlight removal, the development of a novel network architecture comprising three jointly trained modules, and the achievement of superior performance over existing methods.

The remainder of the paper is organized as follows. Section 2 reviews the related work. Section 3 presents details of our innovations. Section 4 is the experiments part and further discussions. Section 5 is our conclusion part.

2. Related Works

2.1. Model-Based Methods

Traditional approaches to highlight removal have relied on optimization techniques, clustering algorithms, and filtering methods to remove the highlight. Multi-image approaches use multiple input images and generally exploit viewpoint dependence. Lee et al. [27] introduced a model that incorporates multiple color images captured from different viewing directions. Guo et al. [28] exploited the correlations between transmission layers in multiple images to successfully separate these layers. Single-image methods, which rely solely on a single input image, face a more formidable challenge. Shen et al. [29] addressed this by conducting an error analysis of chromaticity and employing a meticulous selection of body colors for each pixel, thereby distinguishing specular highlights from the color image. For single-image approaches, traditional methods often demand additional priors. Tan et al. [10] introduced a method that leverages chromaticity analysis to examine highlights. Based on the dichromatic reflection model [14,30], the removal of specular highlights from natural images is rendered feasible and effective. Shen and Zheng [12] analyzed the distribution of diffuse and specular reflectance components in color space, segregating them based on their distinct distributions. Yang et al. [14] employed an edge-preserving low-pass filter to remove highlights identified as noise originating from specular pixels within the HSI color space. Liu et al. [15] approached the challenge by initially generating a supersaturated specular-free image, followed by a two-step process that restores saturation based on diffuse chromaticity and specular reflection. Akashi et al. [22] proposed a modified model rooted in the sparse NMF method, facilitating highlight removal without reliance on spatial priors. While these methods for specular highlight removal have demonstrated significant advancements and faster processing speeds, the complex ambient illumination and varied content of scenes of real-world images often prevent them from producing satisfactory results.

2.2. Deep-Learning-Based Methods

Recently, deep learning-based methods utilizing CNNs have emerged as the frontier technology in highlight removal, outperforming traditional techniques by a significant margin. These data-driven approaches alleviate the need for laborious searches for features and priors, which might not even be associated with a wide variety of possible scenarios. Shi et al. [31] proposed a method based on an encoder–decoder CNN architecture, trained on specular and ground truth image pairs. Fu et al. [25] presented a multi-task network tailored for specular highlight removal, utilizing a vast dataset comprising natural images along with their corresponding ground truth information. However, the performance of these methods is strongly dependent on the quality and quantity of the training data,

and the task of compiling suitable datasets is always arduous, thus posing limitations on such methods. Additionally, another challenge faced by data-driven approaches is the generalization capability, which becomes even more intricate when training with synthetic data. A potential solution to this challenge lies in the employment of the Generative Adversarial Network (GAN) framework. Lin et al. [32] introduced a GAN-based approach, where a generator network is trained to remove specularities and generate diffuse images. To ascertain the effectiveness of specularity removal, a discriminator network is also incorporated, solely for the purpose of training.

Muhammad et al. [33] proposed Spec-Net, a network that utilizes an intensity channel as input to mitigate high-intensity specularities in images with low chromaticity. Furthermore, they proposed Spec-CGAN that considers RGB images as input and generates diffuse images. Wu et al. [34] proposed a novel GAN approach for specular highlight removal, guided by an innovative detection mechanism for specular reflection information. Their method also leverages the attention mechanism to establish a direct mapping between diffuse regions and specular highlight regions. Nonetheless, these methods remain inherently dependent on data, and the acquisition of paired highlight and highlight-free images remains a practical challenge.

To address the above-mentioned problem, recent research has paid considerable attention to the utilization of unpaired data. Yi et al. [20] introduced an unsupervised fine-tuning framework for deep neural networks, focusing on extracting facial highlights and tracing their reflections back to the scene to reconstruct the environment map. Yi et al. [35] further extended this work by introducing an unsupervised method based on local color distribution image representation. This approach leverages the synergistic benefits of specular separation and intrinsic image decomposition, requiring only unpaired highlight and highlight-free images. Fu et al. [36] proposed a novel three-stage framework utilizing physics-based models and deep learning techniques to progressively eliminate specular highlights from images, resulting in high-quality specular-free images that are visually consistent with the original inputs. Xu et al. [37] proposed a bifurcated convolutional neural network to tackle specular highlight removal. However, it is crucial for unsupervised methods to ensure sufficient statistical similarity between the highlight and highlight-free images. Furthermore, even acquiring unpaired highlight-free images can pose difficulties in certain cases.

In this paper, we propose a weakly supervised specular highlight removal method that only requires highlight images, addressing the challenges associated with acquiring paired or unpaired highlight-free images. Our method leverages non-negative matrix factorization (NMF) to estimate masks for generating highlight-free images, which are then used to train a Cycle-GAN-style network consisting of highlight generation, highlight removal, and reconstruction modules. This approach leads to a highly effective highlight removal module, as demonstrated through extensive experiments on the specular highlight image quadruples (SHIQ) and the LIME datasets. Our proposed method outperforms existing approaches for processing natural images, highlighting its potential for improving image quality in various applications.

3. Proposed Method

The architecture of our proposed network is illustrated in Figure 1, which contains three jointly trained modules: highlight generation, highlight removal, and reconstruction.

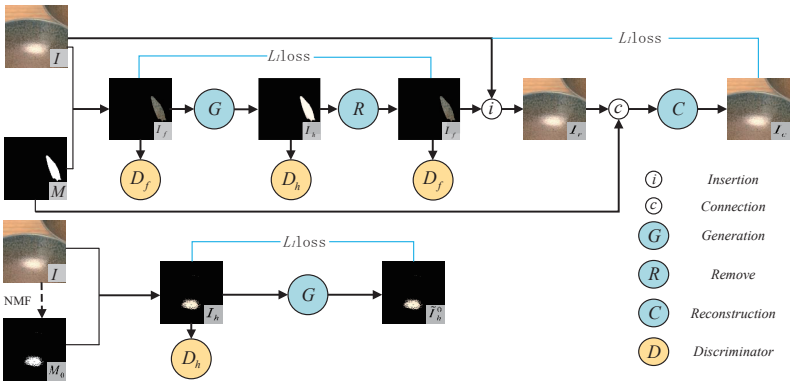


Figure 1. Training process of our proposed network, including highlight generation G, highlight removal R, and reconstruction C. The NMF method estimates the highlight mask, which is used to generate a highlight-free dataset. Consistency between the original and reconstructed images ensures dataset similarity.

3.1. Mask and Training Data Generation

In order to generate the dataset, we first perform highlight detection as a pre-stage to identify the potential highlight regions in the image. Subsequently, we perform an approximate estimation of the position of the highlight mask, which serves as a guide for cropping out the corresponding highlight regions. After that, we adopt the reflection model proposed by Phong [38] in our work as follows:

$$L_{diffuse} + L_{specular} = S_d E_d (n \cdot l) + S_s E_s (v \cdot r)^{\gamma} \tag{1}$$

where S_d and S_s denote the material diffuse and specular reflections, E_d and E_s represent the illumination intensity, n and l represent the normal vector and the direction of the incident light, v indicates the observation direction, $r = 2(n \cdot l) - l$ refers to the fully reflected light direction, and γ indicates the shininess of the surface material. This model can be widely applied to represent real-life specular highlights, where S_s in the highlight region is set to 1 since it can be considered approximately equal to the spectrum of the light source. Based on the widely used dichromatic reflection model [30], we can combine the NMF method proposed by [23] to denote the highlight mask by the following method. A color image I can be defined as $I = D + S$, where D is similar with $L_{diffuse}$ and indicates diffuse components, and S denotes specular reflection components and is analogous to $L_{specular}$. In the RGB response of an image, the above combination can be decomposed as follows:

$$\begin{bmatrix} R(x) \\ G(x) \\ B(x) \end{bmatrix} = \begin{bmatrix} R_d & R_s \\ G_d & G_s \\ B_d & B_s \end{bmatrix} \times \begin{bmatrix} k_d(x) \\ k_s(x) \end{bmatrix} \tag{2}$$

where the first term on the right side of the equation expressed as a 3×2 matrix is the RGB channel parameter, the last term contains the diffuse and specular reflection coefficients and is a $2 \times N$ matrix, and N is the total number of pixels in I.

It is assumed that the image under analysis is already registered, as it is captured from a consistent viewpoint and under controlled lighting conditions, allowing us to focus on modeling the RGB response without the need for additional registration steps. $k(x)$ encapsulates the diffuse and specular reflection coefficients, which are combined with the RGB channel parameters (represented as a 3×2 matrix) to model the RGB response of an image pixel x. Specifically, $k(x)$ is a $2 \times N$ matrix where each column corresponds to the diffuse and specular reflection coefficients for a given pixel, and N is the total number of pixels in the image. This matrix is non-negative and often sparse due to the localized nature of specular reflections. We solve Equation (2) using the NMF [39] method. Since

the matrix is 3×2, we fix the internal dimension of the decomposition to 2 and constrain $k_s(x)$ sparsely by minimizing the ℓ_1 norm while holding the ℓ_2 norm (see Figure 2).

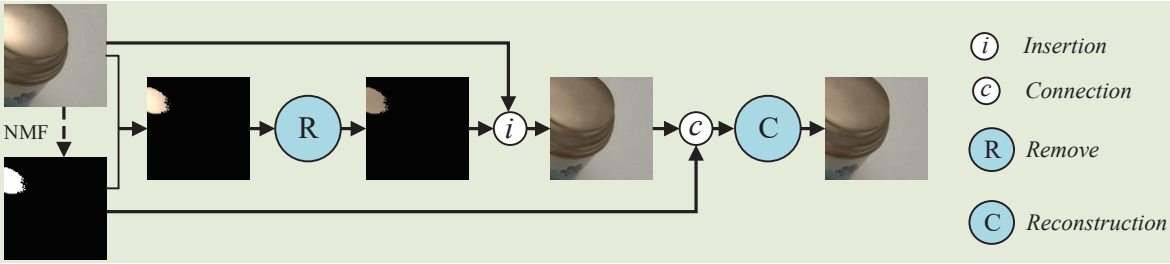


Figure 2. Testing process overview. The highlight removal module transforms the input into a highlight-free image, which is compared with the ground truth. The reconstruction output is also evaluated.

With the guidance of the highlight mask, we crop the highlight region I_h from the original image I , while the rest of the image I_h is set to 0. Then, from the obtained highlight masks, we pick a random mask that is larger than the set area and apply it to crop the highlight-free image I_f from the image I , ensuring that the areas of the two images do not overlap. Figure 3 illustrates the calculated highlight masks, as well as examples of cropped highlighted and highlight-free images captured under different lighting conditions.

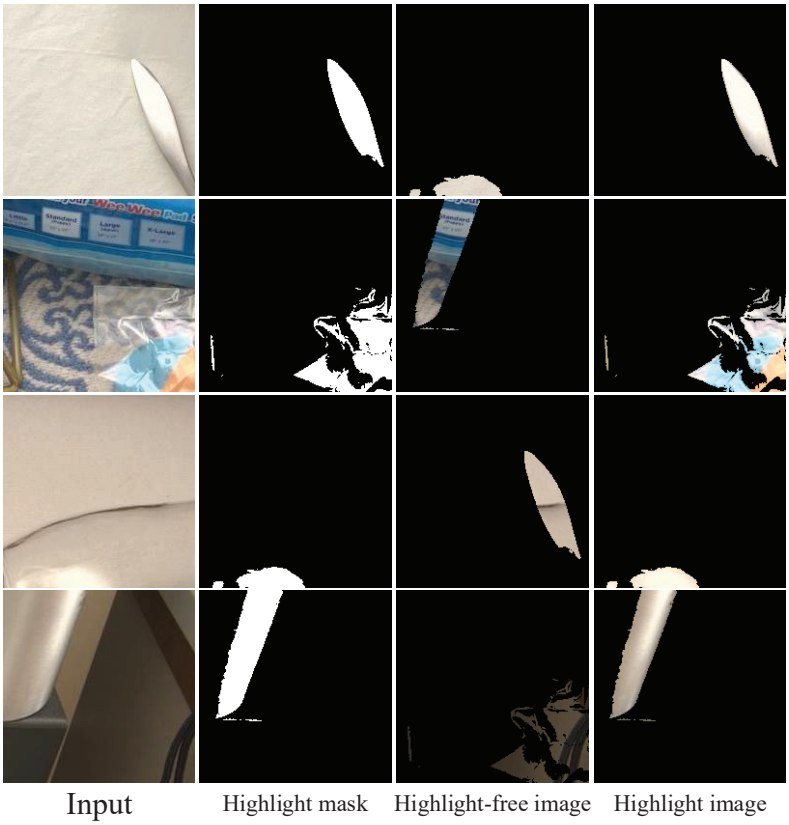


Figure 3. Highlight masks (**second column**) calculated by NMF from natural images (**first column**). Highlight-free images (**third column**) and highlight images (**last column**) are obtained by cropping with the guidance of random masks and corresponding masks crops, respectively.

3.2. Network Architecture

A significant number of unpaired data, generated via the Non-negative Matrix Factorization (NMF) method, are utilized to enhance the performance of the generator G. Through adversarial training, the generator is tasked with producing increasingly realistic

synthetic highlight images $\tilde{I}_h$, which are then paired with the corresponding highlight-free regions of the original image I. We employ a discriminator Dh to accurately discern between false highlight $\tilde{I}_h$ and a real sampled highlight I_h, thereby ensuring the statistical similarity between these two categories of images.

The objective function derived from the standard generative adversarial network [40], is formulated to optimize the generator G and its corresponding discriminator D_h , as defined below:

$$L_{GAN}^{g}(G, D_h) = \mathbb{E}_{I_h \sim p(I_h)}[\log(D_h(I_h))] + \mathbb{E}_{I_h \sim p(I_h)}\left[\log\left(1 - D_h\left(G\left(I_f\right)\right)\right)\right] \tag{3}$$

where p denotes the data distribution.

Meanwhile, to ensure the authenticity of the highlights generated by the generator G, we incorporate a real highlight I_h into the generator G to generate $\tilde{I}_h^0$. It should theoretically be the same as I_h. Then, we apply the identity loss [41] to optimize the generator G via this restriction, given by

$$L_{\text{iden}}(G) = \mathbb{E}_{I_h \sim p(I_h)}[\|G(I_h), I_h\|_1] \tag{4}$$

where $\| \cdot \|_1$ denotes the ℓ_1 norm, which can reflect the pixel-wise deviations.

Then, we use a highlight removal module R trained by the pairs of I_f and $\tilde{I}_h$ to remove the fake highlight, which is the output of the generator G, i.e., $\tilde{I}_h$. According to the classic architecture of Cycle-GAN [42], the highlight removal module R shares the same architecture with the generator G , while it obtains a highlight-free image $\tilde{I}_h$ that should be identical to the original image I_f after processing. R is trained by formulating the following consistency loss [42], which can also be applied to train generator G,

$$L_{\text{cycle}}(G, R) = \mathbb{E}_{I_f \sim p(I_f)}\left[\left\|R\left(G\left(I_f\right)\right), I_f\right\|_1\right] \tag{5}$$

Similarly, another discriminator D_f is employed to train the highlight removal module R, producing more realistic highlight-free images. We apply adversarial loss [40] to train both R and D_f, which is defined as:

$$L_{GAN}^{r}\left(R, D_f\right) = \mathbb{E}_{I_f \sim p(I_f)}\left[\log\left(D_f\left(I_f\right)\right)\right] + \mathbb{E}_{I_f \sim p(I_f)}\left[\log\left(1 - D_f(R(\tilde{I}_h))\right)\right] \tag{6}$$

Since the resulting highlight-free image $\tilde{I}_f$ is confined to a particular region of the image, R is likely not to take into account the surrounding information of the highlight areas. Consequently, the processed highlighted areas may exhibit substantial differences in color and detail compared to the original image I_f. To address this limitation, we use the reconstruction module C to reconstruct the original image I, the fake highlight-free image $\tilde{I}_f$, and the mask M, and then synthesize the image I_r by the reconstructed image. Image I_r is defined as:

$$I_r = \left(I + \tilde{I}_f - I_f\right) \oplus M \tag{7}$$

where M covers the regions of I_f that have content, and $\oplus$ represents concatenation operation. The reconstruction module C also adopts the same network architecture as G and R, except that the input I_r is a 4-channel tensor containing an additional 1 channel for the mask M. Subsequently, we use the pixel loss to encourage the final output image I_c to be consistent with I,

$$L_{pixall} = \mathbb{E}_{I_f \sim p(I_f)}\left[\left\|C\left(R\left(G\left(I_f\right)\right)\right), I\right\|_1\right] \tag{8}$$

3.3. Module Details

The overall architecture of our proposed three modules is based on Hu et al. [43], as depicted in Figure 4. This framework commences with three convolutional layers responsible for downscaling operations, followed by nine residual blocks [44] that serve to

extract intricate features. The network concludes with three distinct convolutional layers, which are tasked with generating the output image. Notably, instance normalization [45] is consistently applied after each convolutional layer to preserve the preservation of unique image details. The discriminator component is a direct realization of the PatchGAN model [46], ensuring precise discrimination of the authenticity of the generated images. Additionally, the notation (P1) represents the first pixel-wise operation within our network architecture. This operation occurs after the initial convolutional layer and plays a crucial role in processing the input features before they are passed through the subsequent layers of the network. For clarity, the filter bank sizes for all convolutional layers are as follows: first convolutional layer: 64 filters, second convolutional layer: 128 filters, third convolutional layer: 256 filters, and fourth convolutional layer: 512 filters (with a 7×7 kernel size). We have intentionally used a 7×7 kernel size for this layer. The reason behind this choice is to capture a larger receptive field in the deeper layers of the network, which can be beneficial for extracting more complex and abstract features from the input data. This larger kernel size allows the network to learn richer representations by considering a broader context in the input feature maps.

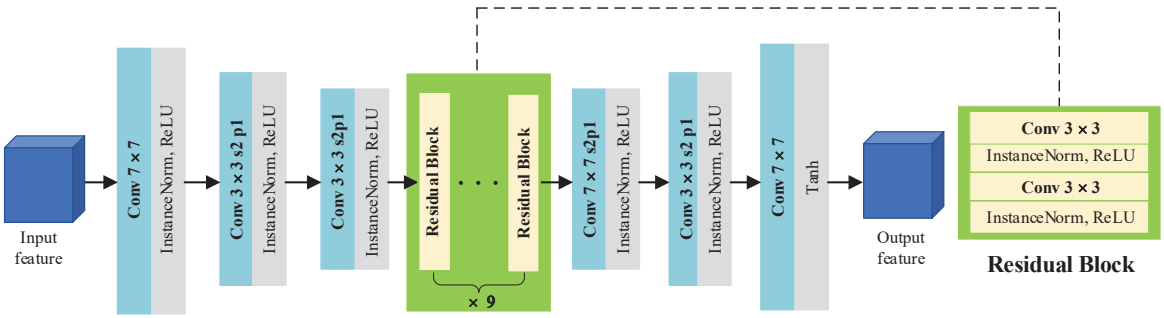


Figure 4. The overall architecture of the modules, including three convolutional layers, nine residual blocks, and three different deconvolutional layers.

3.4. Loss Function

As aforementioned, our proposed approach incorporates four distinct losses: adversarial loss L_{GAN} [40], identity loss L_{iden} [41], cycle consistency loss L_{cycle} [42], and pixel loss L_{pixel}. These losses are jointly optimized for the three interdependently trained modules. The final loss function $\mathcal{L}$ is a weighted sum of the above loss functions:

$$\mathcal{L} = \omega_1(L^g_{GAN} + L^r_{GAN}) + \omega_2(L_{cycle}) + \omega_3(L_{iden}) + \omega_4(L_{pixel}). \quad (9)$$

Although we adopt G2R-Net [24] as a baseline for our network architecture, we encountered a unique challenge due to the significant difference in size between the highlight mask and the shadow mask when generating the training dataset. Specifically, the highlight region tends to be considerably smaller during the removal process. To address this issue, we introduce a pixel loss function, analogous to the identity loss, which effectively constrains the output image I_c. Following extensive experimental validation, we empirically determined the optimal weights ω_1, ω_2, ω_3, and ω_4 to 1, 1, 20, and 10, respectively, in order to achieve a balanced optimization of the various loss terms. These ablation studies are detailed in Section 4.3.

4. Experiments

4.1. Implementation Details

Datasets: In this paper, we conduct experiments utilizing two recent datasets to validate the efficacy of our proposed approach.

(1) SHIQ: The SHIQ dataset is specifically designed for the purpose of highlight detection and removal, and provides comprehensive annotations including ground truth, highlight images, and corresponding highlight masks. Each of these components

comprises approximately 12,000 images, with a resolution of 200 × 200 pixels and 36,000 images in all. Notably, we utilize only the highlight images from this triplet in our experiments, employing the NMF method to generate the corresponding highlight masks. The SHIQ dataset was captured in natural scenes, exhibiting a diverse range of illumination conditions, object materials, and scenarios. For our experiments, we split the datasets into a training set of 24,000 images, a validation set of 6000 images, and a test set of 6000 images. We ensure that the ground truth, highlight images, and corresponding highlight masks are evenly distributed among them. Specifically, each subset contains an equal number of images from these three components, with approximately one-third of the total images allocated to each subset. This balanced distribution allows us to effectively train, validate, and test our proposed approach.

(2) LIME: The LIME dataset comprises images of diverse materials, including specular reflection images representative of these materials, and each comprises approximately 25,000 images. Notably, our experiments solely utilize the highlight images within this dataset. We split the datasets into a training set of 15,000 images, a validation set of 5000 images, and a test set of 5000 images.

Metrics and Comparison: To evaluate the effectiveness of our proposed method, we conduct a comparative analysis with various existing approaches, including traditional methods [10,14,15,22,47] and deep learning-based methods [25,31,35]. Furthermore, we adopt the widely utilized peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as performance evaluation metrics, in accordance with the current works [48]. For both of these metrics, a higher value indicates superior performance. Accuracy is a metric used to measure the performance of a classification model. It is calculated as the ratio of correct predictions to the total number of predictions made. Balance Error Rate (BER) is a performance metric specifically designed for imbalanced datasets. It is calculated as the average of the false positive rate and the false negative rate. The BER provides a balanced view of the model's performance, considering both types of errors. The Peak Signal-to-Noise Ratio (PSNR) measures the quality of image compression or reconstruction by comparing the signal power to the noise power. The Structural Similarity Index (SSIM) assesses the visual similarity between two images, considering structural information, luminance, and contrast.

Experimental Details: Our proposed network is implemented in PyTorch, inspired partly by the architecture of Cycle-GAN. During the initial stage of the generator, the model was initialized using a Gaussian distribution with a mean of 0 and a standard deviation of 0.02. The model is trained for 100 epochs, with an initial learning rate of 2×10^{-4} for the first 50 epochs, followed by a linear decay to zero over the subsequent 50 epochs. The batch size was consistently set at 1. For data generation, we produced an equivalent number of cropped non-highlighted images as the original highlight images in the SHIQ dataset to serve as input for model training. All three modules were involved in the training process of the network and underwent joint optimization. During the testing process, only the highlight removal and reconstruction models were utilized to generate the final results.

4.2. Comparison Results

We conduct a comparative analysis of our proposed method with eight existing techniques, containing Tan [10], Yang [11], Shen [12], Akashi [22], Yamamoto [19], Shi [31], Yi [35], and Fu [25], utilizing the SHIQ dataset. Among these, Yi [35] represents an unsupervised approach that solely relies on unpaired highlight and highlight-free images. Shi [31] and Fu et al. [25] employ paired highlight and highlight-free images, in addition to corresponding highlight masks and specular reflections for training. The remaining methods operate on a single image basis, thus eliminating the need for a training set. Furthermore, we have slightly modified the architecture of our network to facilitate training on paired highlight and highlight-free images. This adapted version is illustrated in Figure 5, where we have eliminated the highlight generation module while directly comparing the images generated by the highlight removal and reconstruction modules with paired ground truth.

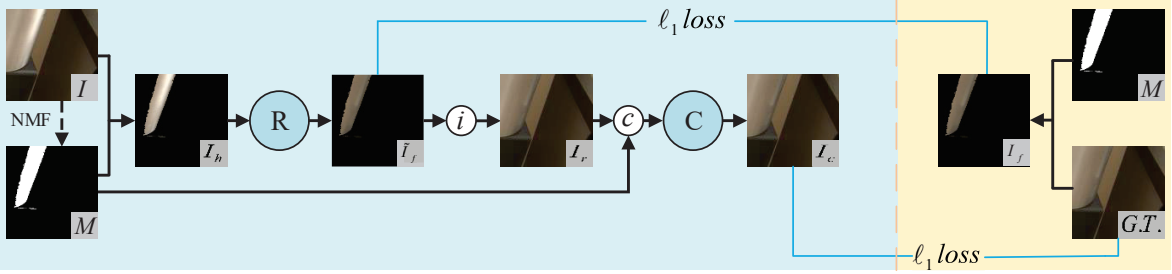


Figure 5. The architecture of the slightly adapted network trained on paired data. I and M denotes the input image and mask, respectively. I_f obtained from the ground truth is the corresponding highlight-free image with $\tilde{I}_f$; I_c is the output of the network, which ought to be consistent with the ground truth.The other letters have been depicted in Figure 2.

As shown in Figure 5, the blue region represents the workflow of data processing and reconstruction, encompassing input, transformation, and output stages. The yellow region highlights the core modules involved in the segmentation task, focusing on generating accurate segmented outputs.

Qualitative Evaluation: In Figure 6, we present the qualitative results of the aforementioned highlight removal methods evaluated on the SHIQ dataset. Our analysis reveals that traditional methods based on statistical and chromaticity analysis exhibit suboptimal performance in highlight removal and are susceptible to introducing color distortion. Since traditional methods cannot semantically differentiate between highlights and reflected light from white materials, white regions in results are often processed as highlights and turn black. Regarding the deep learning-based methods, Fu et al. [25] and our proposed method perform well in the treatment of white regions and yield relatively high-quality images.

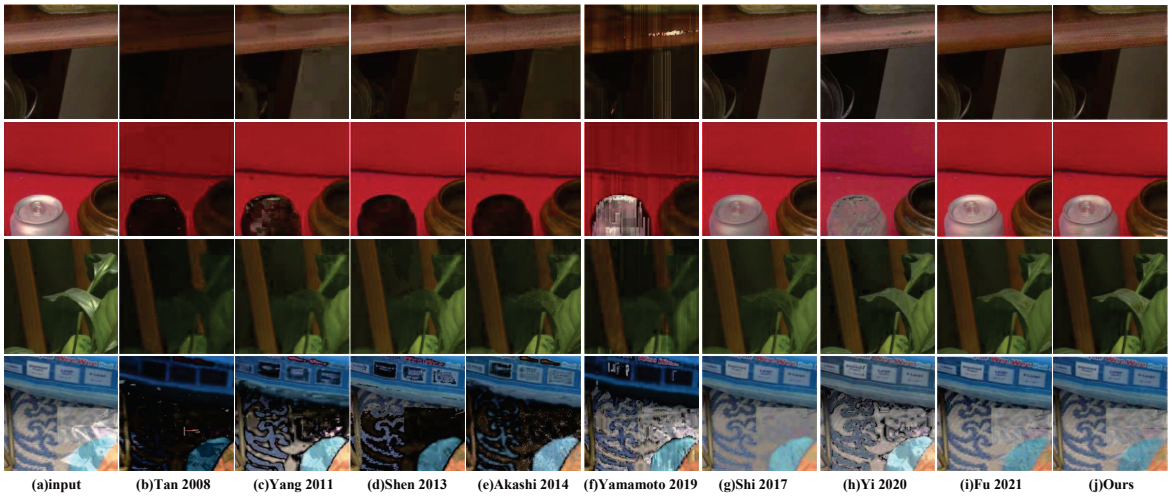


Figure 6. A visual comparison between our method and the state-of-the-art method is performed on the SHIQ dataset. It can be seen that our method has the advantage of recovering highlight-free images more clearly [10–12,19,22,25,31,35].

Quantitative Evaluation: The quantitative results of the various methods are shown in Table 1. It can be seen that our proposed method achieves a satisfactory performance. Table 2 reveals that the accuracy and equilibrium error rates achieved by our method are better than other methods, especially in SSIM with the highest score, and it outperforms the recently proposed methods using unpaired images [35] and based on a single image [19]. Compared with [25] that utilizes paired data and more additional information, our method achieves competitive results in PSNR and better results in SSIM on both datasets. As mentioned in [48], PSNR does not match well with the perceived visual quality and SSIM is more effective in characterizing the visual similarity between images. This phenomenon

can also be seen in the last row of the image obtained by [25], as depicted in Figure 6. The blue pattern in the figure deviates somewhat compared to the ground truth. We can also discern that the effectiveness of our proposed method can be improved significantly by incorporating corresponding highlight-free images and training our network with paired datasets. Nonetheless, in the absence of corresponding ground truth, our approach still exhibits commendable performance, effectively minimizing the need for constructing paired datasets at the cost of a marginal decrement in accuracy.

Table 1. Quantitative comparison of the proposed method with state-of-the-art highlight removal methods. 'N/A' indicates that the method relies only on a single image for processing. 'Hig.' represents highlight and 'S' denotes specular highlight. 'Unpaired' indicates that there is no correspondence between the classes of the training set.

Method	Training Data	SHIQ		LIME	
		PSNR (dB)	SSIM	PSNR (dB)	SSIM
Tan [10]	N/A	11.04	0.40	13.21	0.52
Yang [11]	N/A	14.31	0.50	17.64	0.58
Shen [12]	N/A	13.90	0.42	14.08	0.51
Akashi [22]	N/A	14.01	0.52	16.13	0.55
Yamamoto [19]	N/A	19.54	0.63	19.89	0.63
Shi [31]	Hig.Free + Hig.Mask + S (Paired)	18.21	0.61	24.21	0.76
Yi [35]	Hig.Free (Unpaired)	21.32	0.72	26.77	0.79
Fu [25]	Hig.Free + Hig.Mask + S (Paired)	34.13	0.86	37.01	0.91
Ours	N/A	30.94	0.96	32.86	0.97
	Hig.Free (Paired)	31.86	0.97	33.43	0.98

Table 2. We have used accuracy and balance error rates (BERs) for comparison, where our results are marked in bold.

Dataset Method	SHIQ		LIME	
	Accuracy	BER	Accuracy	BER
Tan [10]	0.62	17.8	0.70	19.9
Akashi [22]	0.69	24.1	0.59	21.2
Fu [25]	0.85	10.7	0.88	11.3
Ours	**0.90**	**8.6**	**0.89**	**9.1**

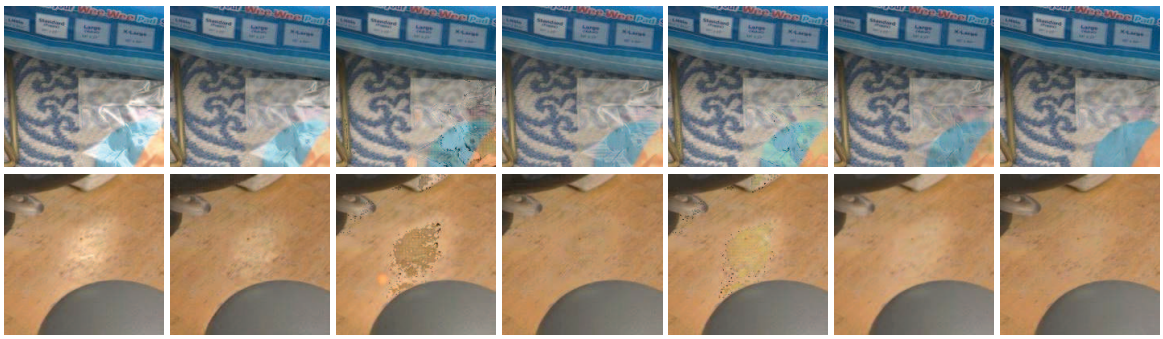
4.3. Ablation Study

To investigate the efficacy of each component within our proposed network, we conduct an ablation study to assess the impact of their absence on the experimental results on the SHIQ dataset. Firstly, we experimentally examine the value of introducing an additional discriminator D_f by comparing two configurations: one with a discriminator specifically designed for highlight-free images, while the other without. Furthermore, we evaluate the influence of the reconstruction module C through a similar experimental setup. The quantitative results are summarized in Table 3, which reveals that the inclusion of the new discriminator and reconstruction module indeed enhances the overall performance of the network.

Table 3. Ablation study to validate the effectiveness of discriminator D_h on the SHIQ.

Methods	Metrics	
	PSNR (dB)	SSIM
Without discriminator D_h	30.41	0.959
Without reconstruction module C	27.35	0.939
Full structure	30.92	0.964

Subsequently, our loss function is comprised of adversarial loss, identity loss, cycle consistency loss, and pixel loss. The adversarial loss L_{GAN} facilitates the generators to produce more realistic images, while the cycle consistency loss L_{cycle} serves a similar purpose, ensuring that the generated image aligns closely with the ground truth in terms of color and texture. The identity loss L_{iden} constrains the highlight generator G to focus exclusively on the generation of highlights, preventing it from producing non-highlight content. Additionally, the pixel loss L_{pixel} aids the network in further enhancing the processed highlight areas. To demonstrate the effectiveness of each loss component, we conducted a series of experiments, and the quantitative results are presented in Table 4. Notably, for the situation that the adversarial loss L_{GAN} or the identity loss L_{iden} is absent, the generation module fails to generate high-quality highlights, significantly impacting subsequent processing steps and resulting in a considerable decline in performance metrics. In contrast, when the cycle consistency loss L_{cycle} is removed, the reconstruction module C, which is trained alongside the generator G, can partially fulfill the task of highlight removal, leading to a relatively smaller decline in performance. Finally, the removal of the pixel loss L_{pixel} contributes to a degradation in performance, highlighting the advantage of utilizing the entire image as a constraint. We also present qualitative results in Figure 7, which are generally in alignment with the quantitative findings described above, further validating the effectiveness of each loss component.



(a) (b) (c) (d) (e) (f) (g)

Figure 7. Ablation study of loss functions on SHIQ dataset. (**a**): input images; (**b**–**e**) are the results of experiments without L_{GAN}, L_{iden}, L_{cycle}, and L_{pixel}, respectively; (**f**): with full loss; (**g**): ground truth.

Table 4. Ablation study of loss functions on the SHIQ.

Methods	Metrics	
	PSNR (dB)	SSIM
Without L_{GAN}	20.17	0.872
Without L_{iden}	22.35	0.908
Without L_{cycle}	30.26	0.958
Without L_{pixel}	28.72	0.949
Full loss	30.92	0.964

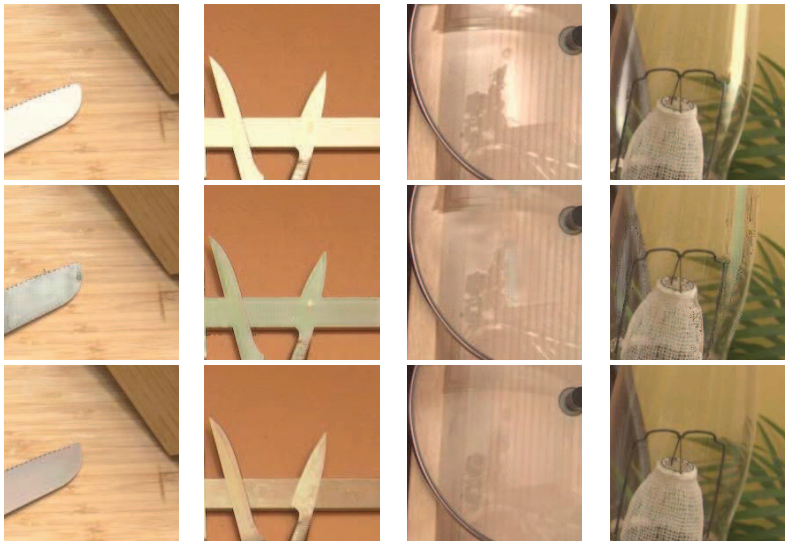
To further evaluate the impact of different weights for the identity loss L_{iden} and pixel loss L_{pixel}, we conducted additional experiments. Specifically, we varied the value of ω_3 to 1, 5, 10, 20, and 30, respectively. The quantitative results are presented in Table 5, and indicate that appropriately increasing the weight of the identity loss results in improved performance. Similarly, pixel loss L_{pixel} shares functional similarities with identity loss L_{iden}, and our experiments suggest that a weight of 10 for L_{pixel} achieves the optimal performance.

Table 5. Effectiveness of various identity loss L_{iden} changes to the proposed method on SHIQ.

L_{iden}	1	5	10	20	30
PSNR (dB)	21.85	26.37	29.95	30.92	30.20
SSIM	0.898	0.936	0.959	0.964	0.957

4.4. Limitations

While our proposed approach demonstrates satisfactory performance in numerous cases, it encounters challenges in certain cases, such as the example depicted in Figure 8. When highlight regions partially cover the surface material as illustrated in Figure 6, our method typically performs well. However, when applied to tasks involving the entire material, the performance is poor. As shown in Figure 8, where specular reflection covers the entire knife, the resulting color differs slightly from the actual image. This discrepancy may be attributed to our reconstruction module, which incorporates surrounding information of the highlighted regions to optimize these regions. Since the highlight encompasses the entire object, there is a lack of relevant information in the surrounding area, leading to the decreased performance of our method.



(a) (b) (c) (d)

Figure 8. Failure cases of our method on the SHIQ dataset. Top row: input specular highlight images. Middle row: our removal methods. Bottom row: ground truth. (**a**,**b**) are the cases where the specular highlight covers the entire area. (**c**,**d**) are the cases where the specular highlight is caused by the glass.

Furthermore, certain failures arise due to reflections originating from the glass, as observed in Figure 8, particularly when there are diverse backgrounds beneath the glass. In such cases, our highlight removal and reconstruction module may struggle to discern whether the reflection stems from the background or the glass itself, given the limited number of specular reflections caused by glass in our training set. Consequently, accurately reproducing the background information beneath the glass in the highlighted areas becomes challenging in the final result.

5. Conclusions

In this paper, we have introduced a novel approach for specular highlight removal through weak supervision, which operates without reliance on ground truth data. Our proposed network comprises three key modules: highlight generation, highlight removal, and reconstruction, all of which are trained jointly. Compared to state-of-the-art methods

that require additional preprocessed and annotated data, including highlight-free images, specular highlights, and highlight masks, our method achieves competitive performance both qualitatively and quantitatively. Extensive experiments conducted on the SHIQ and LIME datasets have validated the efficacy of our proposed approach.

For future work, we plan to further explore the potential of weak supervision in specular highlight removal tasks and aim to enhance the generalization ability of our network to tackle more complex and diverse real-world scenarios. Additionally, we will investigate how to integrate our method with other image processing techniques for more efficient and comprehensive image quality improvement.

Author Contributions: Conceptualization, Y.Z.; Software, G.H.; Formal analysis, H.W.; Investigation, L.W.; Resources, H.J.; Writing—original draft, H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported National Natural Science Foundation of China [grant number U19B2004].

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Borji, A.; Cheng, M.M.; Jiang, H.; Li, J. Salient object detection: A benchmark. *IEEE Trans. Image Process.* **2015**, *24*, 5706–5722. [CrossRef] [PubMed]
2. Wang, S.; Wang, Y. Weakly supervised semantic segmentation with a multiscale model. *IEEE Signal Process. Lett.* **2014**, *22*, 308–312. [CrossRef]
3. Gao, J.; Zhang, T.; Xu, C. Graph convolutional tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4649–4659.
4. Jachnik, J.; Newcombe, R.A.; Davison, A.J. Real-time surface light-field capture for augmentation of planar specular surfaces. In Proceedings of the 2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Atlanta, GA, USA, 5–8 November 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 91–97.
5. Weyrich, T.; Matusik, W.; Pfister, H.; Bickel, B.; Donner, C.; Tu, C.; McAndless, J.; Lee, J.; Ngan, A.; Jensen, H.W.; et al. Analysis of human faces using a measurement-based skin reflectance model. *ACM Trans. Graph. (ToG)* **2006**, *25*, 1013–1024. [CrossRef]
6. Li, C.; Zhou, K.; Lin, S. Intrinsic face image decomposition with human face priors. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 218–233.
7. Suo, J.; An, D.; Ji, X.; Wang, H.; Dai, Q. Fast and high quality highlight removal from a single image. *IEEE Trans. Image Process.* **2016**, *25*, 5441–5454. [CrossRef] [PubMed]
8. Li, C.; Lin, S.; Zhou, K.; Ikeuchi, K. Specular highlight removal in facial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3107–3116.
9. Tan, P.; Quan, L.; Lin, S. Separation of highlight reflections on textured surfaces. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; IEEE: Piscataway, NJ, USA, 2006; Volume 2, pp. 1855–1860.
10. Tan, R.T.; Ikeuchi, K. Separating reflection components of textured surfaces using a single image. In *Digitally Archiving Cultural Objects*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 353–384.
11. Yang, Q.; Wang, S.; Ahuja, N. Real-time specular highlight removal using bilateral filtering. In Proceedings of the European Conference on Computer Vision, Crete, Greece, 30 August 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 87–100.
12. Shen, H.L.; Zheng, Z.H. Real-time highlight removal using intensity ratio. *Appl. Opt.* **2013**, *52*, 4483–4493. [CrossRef] [PubMed]
13. Kim, H.; Jin, H.; Hadap, S.; Kweon, I. Specular reflection separation using dark channel prior. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1460–1467.
14. Yang, Q.; Tang, J.; Ahuja, N. Efficient and robust specular highlight removal. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1304–1311. [CrossRef] [PubMed]
15. Liu, Y.; Yuan, Z.; Zheng, N.; Wu, Y. Saturation-preserving specular reflection separation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3725–3733.
16. Ren, W.; Tian, J.; Tang, Y. Specular reflection separation with color-lines constraint. *IEEE Trans. Image Process.* **2017**, *26*, 2327–2337. [CrossRef] [PubMed]
17. Souza, A.C.; Macedo, M.C.; Nascimento, V.P.; Oliveira, B.S. Real-time high-quality specular highlight removal using efficient pixel clustering. In Proceedings of the 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, 29 October–1 November 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 56–63.

18. Guo, J.; Zhou, Z.; Wang, L. Single image highlight removal with a sparse and low-rank reflection model. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 268–283.

19. Yamamoto, T.; Nakazawa, A. General improvement method of specular component separation using high-emphasis filter and similarity function. *ITE Trans. Media Technol. Appl.* **2019**, *7*, 92–102. [CrossRef]
20. Yi, R.; Zhu, C.; Tan, P.; Lin, S. Faces as lighting probes via unsupervised deep highlight extraction. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 317–333.
21. Le, H.; Samaras, D. From shadow segmentation to shadow removal. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 264–281.
22. Akashi, Y.; Okatani, T. Separation of reflection components by sparse non-negative matrix factorization. In Proceedings of the Asian Conference on Computer Vision, Singapore, 1–5 November 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 611–625.
23. Zhang, W.; Zhao, X.; Morvan, J.M.; Chen, L. Improving shadow suppression for illumination robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 611–624. [CrossRef] [PubMed]
24. Liu, Z.; Yin, H.; Wu, X.; Wu, Z.; Mi, Y.; Wang, S. From Shadow Generation to Shadow Removal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 4927–4936.
25. Fu, G.; Zhang, Q.; Zhu, L.; Li, P.; Xiao, C. A Multi-Task Network for Joint Specular Highlight Detection and Removal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 7752–7761.
26. Meka, A.; Maximov, M.; Zollhoefer, M.; Chatterjee, A.; Seidel, H.P.; Richardt, C.; Theobalt, C. Lime: Live intrinsic material estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6315–6324.
27. Lee, S.W.; Bajcsy, R. Detection of specularity using color and multiple views. In Proceedings of the European Conference on Computer Vision, Santa Margherita Ligure, Italy, 19–22 May 1992; Springer: Berlin/Heidelberg, Germany, 1992; pp. 99–114.
28. Guo, X.; Cao, X.; Ma, Y. Robust separation of reflection from multiple images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2187–2194.
29. Shen, H.L.; Zhang, H.G.; Shao, S.J.; Xin, J.H. Chromaticity-based separation of reflection components in a single image. *Pattern Recognit.* **2008**, *41*, 2461–2469. [CrossRef]
30. Shafer, S.A. Using color to separate reflection components. *Color Res. Appl.* **1985**, *10*, 210–218. [CrossRef]
31. Shi, J.; Dong, Y.; Su, H.; Yu, S.X. Learning non-lambertian object intrinsics across shapenet categories. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1685–1694.
32. Lin, J.; Seddik, M.E.A.; Tamaazousti, M.; Tamaazousti, Y.; Bartoli, A. Deep multi-class adversarial specularity removal. In Proceedings of the Scandinavian Conference on Image Analysis, Norrkoping, Sweden, 11–13 June 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 3–15.
33. Muhammad, S.; Dailey, M.N.; Farooq, M.; Majeed, M.F.; Ekpanyapong, M. Spec-Net and Spec-CGAN: Deep learning models for specularity removal from faces. *Image Vis. Comput.* **2020**, *93*, 103823. [CrossRef]
34. Wu, Z.; Zhuang, C.; Shi, J.; Guo, J.; Xiao, J.; Zhang, X.; Yan, D.M. Single-Image Specular Highlight Removal via Real-World Dataset Construction. *IEEE Trans. Multimed.* **2021**, *24*, 3782–3793. [CrossRef]
35. Yi, R.; Tan, P.; Lin, S. Leveraging Multi-View Image Sets for Unsupervised Intrinsic Image Decomposition and Highlight Separation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12685–12692.
36. Fu, G.; Zhang, Q.; Zhu, L.; Xiao, C.; Li, P. Towards High-Quality Specular Highlight Removal by Leveraging Large-Scale Synthetic Data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; pp. 12857–12865.
37. Xu, J.; Liu, S.; Chen, G.; Liu, Q. Bifurcated convolutional network for specular highlight removal. *Optoelectron. Lett.* **2023**, *19*, 756–761. [CrossRef]
38. Phong, B.T. Illumination for computer generated pictures. *Commun. ACM* **1975**, *18*, 311–317. [CrossRef]
39. Hoyer, P.O. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* **2004**, *5*, 9.
40. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Montreal, 8–13 December 2014; pp. 2672–2680.
41. Taigman, Y.; Polyak, A.; Wolf, L. Unsupervised cross-domain image generation. *arXiv* **2016**, arXiv:1611.02200.
42. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
43. Hu, X.; Jiang, Y.; Fu, C.W.; Heng, P.A. Mask-ShadowGAN: Learning to remove shadows from unpaired data. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, South Korea, 27 October–2 November 2019; pp. 2472–2481.
44. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
45. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. *arXiv* **2016**, arXiv:1607.08022.
46. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.

47. Li, R.; Pan, J.; Si, Y.; Yan, B.; Hu, Y.; Qin, H. Specular reflections removal for endoscopic image sequences with adaptive-RPCA decomposition. *IEEE Trans. Med Imaging* **2019**, *39*, 328–340. [CrossRef] [PubMed]
48. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]