\textit{M}^2\text{-Net: Multi-stages Specular Highlight Detection and Removal in Multi-scenes}

Zhaoyangfan Huang, Kun Hu, Xingjun Wang

Tsinghua Shenzhen International Graduate School, Shenzhen, China
{hcyf20, Wang.Xingjun}@mails.tsinghua.edu.cn

Fig. 1: The results of this paper show the comparison between the original highlight images and the removed highlight images for a variety of scenes (including synthetic images, face images, text images and natural images) for the network. The 1 in the upper left corner of the image represents the original highlight image, and the 2 represents the image with the highlights removed using the network architecture of this paper. There are also zoomed-in comparison images, text OCR results shown in the scenes and video highlight removal result.

**Abstract.** In this paper, we propose a novel uniformity framework for highlight detection and removal in multi-scenes, including synthetic images, face images, natural images, and text images. The framework consists of three main components, highlight feature extractor module, highlight coarse removal module, and highlight refine removal module. Firstly, the highlight feature extractor module can directly separate the highlight feature and non-highlight feature from the original highlight image. Then highlight removal image is obtained using a coarse highlight removal network. To further improve the highlight removal effect, the refined highlight removal image is finally obtained using refine highlight removal module based on contextual highlight attention mechanisms. Extensive experimental results in multiple scenes indicate that the proposed framework can obtain excellent visual effects of highlight removal and achieve state-of-the-art results in several quantitative evaluation metrics. Our algorithm is applied for the first time in video highlight removal with promising results. Our code is available at: https://github.com/hzzzyf/specular-removal

**Keywords:** Specular Highlight Detection, Specular Highlight Removal, Highlight Feature Extractor, Contextual Highlight Attention
1 Introduction

Specular highlight is a common phenomenon in the real world, and it has been a long-standing problem in computer vision and image processing. An image taken by a camera usually exhibits specular highlights caused by the shiny material surface of the object when illuminated. Specular highlight appear frequently in many scenes, such as objects surface [1,2,3] and human faces [4], which affects the visual effect while posing great challenges to computer vision tasks, such as image segmentation [5], edge detection [6], text recognition [7,8], etc. The previous algorithm can only produce satisfactory highlight removal results for one of the scenes, but not uniformly for all scenes. The algorithm proposed in this paper can handle highlight removal in multiple scenes simultaneously and can enhance subsequent computer vision tasks, such as optical character recognition (OCR). The highlight removal effect cases of the proposed method is shown in Fig. 1 for different scenes and video.

Fig. 2: The algorithm proposed in this paper differs from the existing algorithms. The previous algorithms for pink background can only process images for specific scenes, such as natural images, face images, etc. While our algorithm can handle highlight removal for multiple scenes at the same time. And our algorithm for green background adds a multi-stage refinement highlight removal network to the previous single-stage algorithm. This ensures that the obtained highlight removal images are closer to the real scenes.

Early works typically remove highlight based on different constraints or assumptions, such as using colour reflection model [9], dichromatic reflection model [10], assuming special [11,12], and adopting optimization [5,13,14]. Recently, deep learning approaches have emerged in the field of specular highlight removal and have achieved remarkable improvements. However, both traditional algorithms and deep learning-based algorithms only use a single-stage algorithm and remove highlight in a single scene, resulting in unsatisfactory results in other scenes. The comparison in Fig. 2 demonstrates that the two most significant dif-
ferences between the algorithm in this paper and other existing algorithms are multi-scenes and multi-stages.

In this paper, we propose a novel highlight removal model for multi-scenes specular highlight removal. It includes a highlight feature extraction (HFE) module to extract highlight regions, a coarse highlight removal module, and a refine highlight removal module with a contextual feature attention (CHA) mechanism. Specifically, the highlight extraction module uses multi-scale information to extract regions with different intensity highlights. The coarse highlight removal module removes the underlying highlights with Gated Convolutions and Dilated Convolutions [15]. Inspired by image inpainting [16], we use a refine highlight removal model with a contextual attention mechanism to further improve the effect of the coarse highlight results. Our method achieves highlight removal with satisfactory quality in most real-world situations and outperforms existing learning-based techniques. To summarize, three main contributions of this paper as follows:

– We propose a unified multi-scene highlight removal framework capable of handling synthetic images, face images, text images and natural images.
– We propose a highlight feature extractor module and a contextual feature attention mechanism, which can effectively detect highlight locations and perform highlight removal. Moreover, two-stage (coarse and refine stages) highlight removal algorithm makes highlight removal more satisfying.
– Experimental results on multiple image datasets outperform existing state-of-the-art algorithms and perform very well in video highlight removal.

2 Related Work

Highlight detection. Under the assumption that the highlight area is only small, the most used method in highlight detection is the threshold detection method [17]. Although threshold-based highlight detection algorithms are efficient in detecting highlight areas, most of them are very sensitive to the threshold value [18,19]. Yang et al. [20] treat highlight pixels as noise and use a low-pass bilateral filter to smooth out the highlights. By using multi-scale context contrasted features, Fu et al. [3] create a deep learning-based specular highlight detection network. There have been encouraging advances in highlight detection, but they are all algorithms for individual scenes, and there are no highlight detection algorithms for all scenes.

Highlight removal. Highlight removal algorithms can be classified as single-image based and multi-image based. In this paper, we focus on highlight removal for single images, and the highlight removal algorithm for multiple images can be referred to [21,22,23]. Tan et al. [24] separated the two reflection components based on the distribution of specular and diffuse points in a two-dimensional maximum chromaticity-intensity space. Yang et al. [25] utilized the HSI color space to separate diffuse and specular reflection components for color images and proposed an approach to adjust saturation of specular pixels to the values of diffuse-only pixels with the same diffuse chromaticity. Some methods are to
derive a synthetic diffuse image that exhibits the specular highlight \cite{26,27}. Yang et al. \cite{28} presented an effective real-time specular highlight removal method by utilizing the bilateral filtering. A sparse and low-rank reflection model has been proposed for highlight removal and detection \cite{29}, by regarding the task as a nuclear norm and $l_1$-norm minimization problem been solved by Lagrange multiplier method. Recently, deep learning-based methods are emerging in the field of highlight removal. Muhammad et al. \cite{4} designed two network models (Spec-Net and Spec-CGAN), which are utilized for removing specular highlight of facial Images. All of the above methods have achieved remarkable performance in highlight removal. However, they are difficult to implement in large-scale highlight regions, and they are unable to restore highlight images with colored illuminations or complicated textures.

**Image inpainting.** Unlike highlight removal, the purpose of image inpainting is to fill in the missing areas of images with reasonable content. The methods proposed by \cite{30} typically complete the missing part by leveraging patch similarity and transferring the contents from the background of the image to the missing region. Recently, deep learning-based approaches have obtained significant attention owing to the capabilities of reasoning and extracting semantic information. CNN-based methods have been designed for image inpainting and achieve remarkable performance \cite{31}. Yu et al. \cite{32} proposed the contextual attention and gated convolution \cite{15} to acquire the textual of inpainting images, which led to significant progress in image inpainting. When employing these methods for highlight removal, a highlight mask is required to remove highlight regions that might otherwise result in the loss of textual content. Therefore, these methods are not applicable for highlight removal.

## 3 Proposed Method

### 3.1 Motivation

Fu et al. \cite{3} observed that the intensity of highlights in natural images is high, but the distribution is sparse. Based on the above observations, a multi-scale contextual contrast feature module is proposed to detect highlights. In this paper, we are inspired to design a highlight removal model based on multiscale contextual features.

In image inpainting, a network structure with coarse extraction and refined extraction is used by Yu et al. \cite{15}. In this paper, we modify the structure based on this and use it for inpainting after highlight detection.

### 3.2 Highlight Feature Extractor Module (HFE)

Given a highlight image $I(x, y)$ with size of $m \times n$, we first adopt a series of resnet \cite{33} to acquire $U$ layers of $N$ feature maps $F^u_i(x, y), i = 1, 2, ..., N, u = 1, 2, ..., U$ with size of $m/(2^{u-1}) \times n/(2^{u-1})$. In this paper, we set $N = 3, U = 4$. For each $v = 1, 2, ..., N - 1$, we consider a transposed convolutional layer with
Stage 1: Coarse Network and Highlight Feature Extractor (HFE)

Stage 2: Refine Network and Contextual Highlight Attention (CHA)

Fig. 3: An overview of the proposed network. Stage 1 includes highlight feature detection network and coarse highlight removal network. Stage 2 includes contextual highlight attention and refined highlight removal network. (a) highlight image \( I(3 \times m \times n) \); (b) highlight feature extractor; (c) highlight feature \( F(3 \times m \times n) \) of the highlight location obtained by the highlight detection extractor module (HFE); (d) Coarse highlight removal network; (e) Highlight-free image \( D_1(3 \times m \times n) \) obtained using coarse removal network; (f) Contextual highlight attention (CHA) module; (g) Refinement highlight removal of the coarse highlight removal image \( D_1(3 \times m \times n) \) and the highlight feature image \( F(3 \times m \times n) \); (h) Highlight-free image \( D_2(3 \times m \times n) \).

For each \( v = 1, 2, ..., N-1 \), \( K \) of filters \( W_k(x,y), k = 1, 2, ..., K \) are made to generate the highlight feature by using \( HF_i(x,y) = \sum_{k=1}^{K} W_k(x,y) \cdot (F_i^v(x,y) + \text{Up}F_i^{v+1}(x,y)) \). The role of the highlight extraction module is to detect highlight and non-highlight regions. Unlike existing highlight detection algorithms that detect highlight regions with 0 or 1 results, we use intensity coefficients to represent the likelihood of belonging to highlight regions, shown in Fig. 3 (c). The larger the coefficient, the greater the intensity of the highlights. Different from the existing single-channel highlight masks, the highlight feature extraction module proposed in this paper can extract highlight features from each channel, which means it can better handle real-world highlights. This is because real-world highlights behave differently in the three-channel, rather than simply as a result of single-channel reflections. Fig. 4 (b) shows the results of highlight extraction. It is clear that the higher the highlight intensity, the higher the highlight feature coefficients. And the extraction effect is obvious, most of the highlight regions can be extracted.
Fig. 4: Images of the algorithmic process results in this paper. (a) Input highlight image; (b) Extracted highlight feature images; (c) Coarse highlight removal images; (d) Refine the image for highlight removal.

3.3 Coarse Network and Refine Network

Architecture designs based on coarse and refined networks are widely used in various computer vision tasks. The network structure has the obvious advantage of being able to balance the whole and the details, such as 3D point cloud completion [34], rain removal [35], shadows removal [36], semantic segmentation [37]. Since the highlight removal task in this paper has many similarities with the shadow removal task, the network structure of this paper has been slightly modified from [36]. Fig. 4 (c) and (d) shows the results of coarse highlight removal and refine highlight removal. The effect of coarse removal mainly removes the highlight region in a base-level, and the highlight-filled effect is unnatural. By refining the removal network, the highlight-filled pixels are more closely matched to the surrounding pixels and more natural.

3.4 Contextual Highlight Attention Mechanism (CHA)

The contextual attention was widely used in the image inpainting field [31,38] to generate the missing region from the background information. While the previous contextual attention mechanism focuses on using the background information to fill in the missing parts of the image, our CHA not only leverage the background contextual features but also the highlight region features. In particular, we divide the contextual features into highlight features and background features, and piecemeal them at multiple scales shown in Fig. 5.

For highlight feature image $HF(x, y)$ with size of $m \times n$, we first divide the $HF$ into a number of small patches ($SP_r, r = 1, 2, ..., R; R = m/l \times n/l$) of length $l = 2$. The $SP_r$ is divided into highlight regions ($HP_s, i = 1, 2, ..., S; S \in [1, m/l \times n/l]$) and non-highlight regions ($BP_t, t = 1, 2, ..., T, T \in [1, m/l \times n/l]$) by highlight features $HF$, where $SP_r = HP_s \cup BP_t$. To retrieve the matched patches, we consider calculating the cosine similarity of two different patches to acquire the attention scores $C(s, t) = softmax(\frac{HP_s \cdot BP_t}{||HP_s|| ||BP_t||})$. After calculating
Multi-stages Specular Highlight Detection and Removal in Multi-scenes

Fig. 5: Illustration of our contextual highlight attention mechanism. The highlight feature image in (a) is divided into $s$ highlight patches (HP) and $t$ background patches (BP) (non-highlight patches) according to the result of the highlight feature extraction module, and the total number of patches is $r$ in (b). Then, we calculate the cosine similarity $C(s, t)$ between each of the highlight patches and each of the background patches in (c). Then the corresponding coefficients are multiplied between the highlight and background patches to obtain the highlight attention (HA) in ($d_1$) and background attention (BA) in ($d_2$). After, the two are connected and the patches in the background patches that are similar to the highlight patches are found by the convolutional matching module in (e). Finally we get the output highlight attention feature map in (f).

the attention scores, the attention scores will be applied on the feature map extracted from non-highlight regions to fill the highlight regions. The highlight regions is filled using the non-highlight regions, and the highlight regions after filling are

$$HA_s = \sum_{t=1}^{T} BP_t \cdot C(s, t).$$

In order to make the image after filling closer to the surrounding pixels, this paper uses non-highlight regions to fill the highlight regions while retaining some pixels of the original highlight regions by using

$$BA_t = \sum_{s=1}^{S} HP_s \cdot C(s, t).$$

$HA_s$ and $BA_t$ are convolutionally matched to obtain the final output new feature image.

3.5 Loss Functions

For acceptable performance, we design the loss function of our network, which consists of an adversarial loss and a removal loss. For adversarial loss, we adopt the hinge loss for the discriminator to identify whether the input image is generated. it is expressed as:

$$L_d = E_{x \sim P_{data}(x)}[\text{max}(0, 1 - D(x, HF))] + E_{z \sim P_{g}(z)}[\text{max}(0, 1 + D(G(z), HF))]$$

where $HF$ is highlight feature extracted by proposed HFE module. $D(.)$ denotes the discriminator and $G(.)$ is the generator. $x$ represents GT images and $z$ is input highlight images. $P_{data}(x)$ and $P_{g}(z)$ are distribution of $x$ and $z$.

The removal loss contains three parts including a gan loss $L_g$, a content loss $L_{content}$, and a perceptual loss $L_{per}$. The Gan loss is written as:

$$L_g = E_{z \sim P_{g}(z)}[D(G(z), HF)]$$
The content loss $L_{\text{content}}$ is used to maintain the visual content between the GT images and highlight removal images generated by the proposed network.

$$L_{\text{content}} = \|G_{\text{coarse}}(z) - x\|_1 + \|G_{\text{refine}}(z) - x\|_1$$  \hspace{1cm} (3)

where $G_{\text{coarse}}(.)$ and $G_{\text{refine}}(.)$ denote the output of coarse network and refine network.

Finally, the perceptual loss plays an important role in ensuring content similarity of the real non-highlight image and generated images. We employ a pre-trained VGG-16 model [39] to extract low-level feature maps, and then the perceptual loss can be expressed as the $\ell_1$ norm between feature maps of GT and those highlight removal images.

$$L_{\text{per}} = \|\phi(G(z)) - \phi(x)\|_1$$  \hspace{1cm} (4)

where $\phi(.)$ represents the output of the pre-trained VGG16 network.

The overall removal loss function for training the generator is thus expressed as:

$$L_{\text{rem}} = \lambda_g L_g + \lambda_{\text{content}} L_{\text{content}} + \lambda_{\text{per}} L_{\text{per}}$$  \hspace{1cm} (5)

where $\lambda_g$, $\lambda_{\text{content}}$, $\lambda_{\text{per}}$ represent the weighs of gan loss, content loss, and perceptual loss respectively. In this paper, we experimentally set it to $\lambda_{\text{content}} = 10$, $\lambda_{\text{per}} = 1$, $\lambda_g = 1$.

4 Experiment

4.1 Setup

We use the Specular Highlight Image Quadruples (SHIQ) dataset [1] and SD1 [8]. SHIQ is a high quality real-world highlight dataset including 10k training quadruples and 1k test quadruples and SD1 is a text highlight images datasets including 12k training data and 2k test data. Our model is implemented in PyTorch on a GPU with NVIDIA GeForce 3090 and the input size of image is $224 \times 224$. We use the Adam Optimizer to train our network with batch size of 16 and epochs of 100 which will require 12 hours. In our experiments, we set the learning rate to $2 \times 10^{-4}$ and then reduce it by half every 10 epochs.

4.2 Comparison with Traditional Methods

We compared the performance of our network with current state-of-the-art traditional methods [40,41,42]. As shown in the Fig. 6, Souza [40] can remove some of the highlights, and the metal surface is poorly removed with a large number of black patches. In contrast, the Akashi [41] and Arnold [42] algorithms remove better results. However, Arnold [42] can not handle the boundary of the highlights and tend to preserve some weaker highlights. The highlight part is not natural with the non-highlight part. And the algorithm in this paper uses the CHA, which makes the filled pixels after highlight removal closer to the pixels in the non-highlight part, and the effect is better.
Fig. 6: Visual results of specular highlight removal and comparison with the traditional state-of-the-art methods [40,41,42], and ours method on SHIQ [1] dataset.

Table 1: Quantitative comparison results between our method with previous methods on SHIQ [1] and SD1 [8] datasets. The evaluation metrics include PSNR, SSIM, and the best results are marked in red.

| Datasets     | SHIQ [1] | SD1 [8] |
|--------------|----------|---------|
| Method/Metric | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| Multi-class [43] | - | - | 26.29 | 0.89 |
| SPEC [4] | 19.56 | 0.69 | 15.61 | 0.69 |
| TA [8] | - | - | 22.65 | 0.88 |
| DeepFillv2 [15] | 32.19 | 0.84 | - | - |
| JSHDR [1] | 34.30 | 0.86 | 24.59 | 0.85 |
| **Ours** | **35.72** | **0.91** | **33.44** | **0.92** |

4.3 Comparison with learning-based Methods

The results of quantitative evaluation on both SHIQ (Real-world highlight images) [1] and SD1 (Text highlight images) [8] dataset are presented in Table 1. To effectively evaluate highlight removal performance, we adopt the metrics including PSNR and SSIM. Through comparing the evaluation metrics of our method with existing methods, it can be observed that our proposed methods significantly outperforms other learning-based methods in terms of both metrics. In specific, our method outperforms state-of-the-art highlight removal methods by more than 1.42 dB in PSNR and 0.05 in SSIM on the SHIQ dataset and 7.15 dB in PSNR and 0.07 in SSIM on the SD1 dataset.

4.4 Comparison on Natural Datasets

Figure 7 shows the results of proposed algorithm compared with the existing state-of-the-art algorithms on the natural image dataset. The algorithms DeepFillv2 [15] and JSHDR [1] leave artifacts in the highlight area, causing significant
differences between the pixels in the highlight section and the surrounding pixels. The algorithm SPEC [4] does not have the problem of image artifacts, but it causes a deepening of the color of the whole image and a large visual difference from the original image. Compared with the existing state-of-the-art algorithms, the proposed algorithm removes the highlights while maintaining the integrity of the image and can handle the reflections on the metal surface well for all three different light illumination cases.

Fig. 7: Compare with state-of-the-art algorithms (SPEC [4], DeepFillv2 [15] and JSHDR [1]) on natural images dataset [44].

Fig. 8: Compare with state-of-the-art algorithms on text images data. The first row show the comparative results of the different algorithms, and the last row shows the results of the recognition of the highlight removal image using the PaddleOCR (https://github.com/PaddlePaddle/PaddleOCR).
4.5 Comparison on Text Datasets

Fig. 8 shows the results of proposed algorithm compared with the existing state-of-the-art algorithms on the text image datasets. It can be seen that the algorithm SPEC [4] also appears as a color deepening effect on text images. The areas with high light pollution do not get a substantial effect enhancement. The effect of algorithm JSHDR [1] is slightly better than that of algorithm SPEC [4], but it is still unsatisfactory. Our algorithm removes highlights well and has a more visually pleasing effect. At the same time, the text in the image can be recognized by the naked eye without the assistance of the OCR algorithm.

4.6 Comparison on Face Datasets

![Comparison on Face Datasets](image)

Fig. 9: Compare with state-of-the-art algorithms (SPEC[4] and JSHDR [1]) on face image data. The first row is a natural face image collected from the Internet and the second row is a synthetic face highlight image.

Fig. 9 shows the results of proposed algorithm compared with the existing state-of-the-art algorithms on the face dataset. It can be seen that the algorithm in this paper has the best result in removing highlights on both the real face data and the synthetic face data, and the removal effect is more natural. In contrast, SPEC [4] and JSHDR [1] algorithms will bring artifacts on the image while removing some highlights. The SPEC [4] algorithm also results in black circles in the chin area on natural images of the face.

4.7 Results on Synthetic Image

Fig. 10 shows the two sets of highlight images and the corresponding highlight removal images. It can be seen that the algorithm in this paper can achieve
satisfactory results regardless of the highlight removal of small blocks (Fig. 10 (a) and (b)) or large blocks (Fig. 10 (c) and (d)).

![Image](image.png)

Fig. 10: Highlight removal results of synthetic images. The first row is the synthetic highlight image and the second row is the output. The images were collected from the Internet.

4.8 Application on Video Datasets

As far as we know, there is no algorithm to remove video highlights. In order to demonstrate the superiority of the proposed algorithm, we use the image highlight removal our algorithm directly for the video highlight processing. Specifically, the video is exported on a per-frame basis, and then processed for each image frame of the video. Fig. 11 shows the video highlight removal effect. We can see that the highlight regions are well removed, and many small highlight regions are also precisely removed.

![Image](image.png)

Fig. 11: Ours algorithm is applied on highlight video. The video resolution selected is $1920 \times 1080$. A frame for every 0.5 seconds. Red dashed line represents the contrast of the main highlight region.
5 Discussion

5.1 Ablation Study

To verify the effectiveness of the proposed components of our network, we perform the ablation experiment by modifying the components. Table 2 show the quantitative comparison results of ablation study for specular highlight removal on SHIQ [1] and SD1 [8] datasets. The best and second best results are marked in red and blue, respectively. Baseline: only Coarse Network and Refine Network, excluding our proposed HFE and CHA. DenseUnet: use only DenseUnet to remove highlight directly. CHA consists of BA and HA.

| Datasets | Method/Metric | SHIQ [1] | SD1 [8] |
|----------|---------------|---------|---------|
|          | PSNR↑ | SSIM↑ | PSNR↑ | SSIM↑ |
| Baseline | 31.08 | 0.86 | 23.35 | 0.82 |
| w/ HFE w/o CHA | 32.64 | 0.84 | 25.67 | 0.87 |
| w/ HFE&BA | 31.71 | 0.87 | 28.46 | 0.89 |
| w/ HFE&HA | 33.79 | 0.89 | 33.06 | 0.90 |
| DenseUnet | 29.18 | 0.83 | 22.38 | 0.84 |
| Ours | 35.72 | 0.91 | 33.44 | 0.92 |

Fig. 12: Ablation results for our proposed components. The first row is the natural highlight image, the second row is the text highlight image, and the third row is the face highlight image.

PSNR and SSIM of the ablation study and Fig. 12 show the visual results. To be fair, we use the same dataset to train the network.
From the results, we can draw the following conclusions: (1) Our proposed HFE module plays an essential role in the task of highlight removal. Without HFE module, it is impossible for the generator to produce a satisfactory result. (2) The CHA module can improve the performance of highlight removal. As shown in Fig. 12 (c), without the CHA, the network may not ensure the colors and can’t address the large scale highlight region. (3) From table and Fig. 12 (d-e), it can be observed that BA tends to keep more background information, HA tends to keep more highlight information. The two attention work together to ensure the best performance. (4) The poor results of highlight removal directly using DenseUnet prove the validity of our model, not for fitting on just the training dataset.

5.2 Limitation

Although our proposed method has achieved an excellent result in highlight removal, it does have some limitations. For unsaturated highlights, our model can effectively remove them. However, for saturated highlights, our model cannot recover reliable textures. The effect of our algorithm on the endoscopic images is shown in Fig. 13 (a). It can be seen that for the large red endoscope dataset, the proposed algorithm does not measure the highlight removal well and produces patches that are very different from the surrounding cells. Besides, for extra-large region highlight images removal, our proposed algorithm is not yet able to handle it well in Fig. 13 (b). It is mainly because the extra-large highlight region severely affects the contextual attention mechanism and can not fill the large region of missing pixels with non-highlight regions.

Fig. 13: Scenarios and cases where the proposed algorithm does not work well. (a) Comparison of the removal effect of three endoscopic highlight images; (b) Comparison of the removal effect of extra-large region highlight under natural scene.

6 Conclusion

We propose a multi-scenes and multi-stages highlight detection and highlight removal network. In particular, the highlight extractor uses multi-scale image
information to achieve accurate detection of highlight regions. The contextual highlight attention mechanism that combines highlight information and background information enables the architecture to remove highlights while achieving a closer fill effect to the surrounding pixels. Extensive experimental results reveal that the proposed algorithm achieves better results on synthetic images, text images, face images, and natural images compared to the existing state-of-the-art algorithms. In particular, the proposed algorithm achieves satisfactory results on video highlight removal for the first time as well.

References

1. 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 Conference on Computer Vision and Pattern Recognition (CVPR). (2021) 7752–7761
2. Wu, Z., Zhuang, C., Shi, J., Xiao, J., Guo, J.: Deep specular highlight removal for single real-world image. In: SIGGRAPH Asia 2020 Posters. (2020) 1–2
3. Fu, G., Zhang, Q., Lin, Q., Zhu, L., Xiao, C.: Learning to detect specular highlights from real-world images. In: Proceedings of the 28th ACM International Conference on Multimedia (ACM MM). (2020) 1873–1881
4. Muhammad, S., Dailey, M.N., Farooq, M., Majeed, M.F., Ekpanyapong, M.: Specnet and spec-cgan: Deep learning models for specularity removal from faces. Image and Vision Computing 93 (2020) 103823
5. 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 (CVPR). (2013) 1460–1467
6. Feris, R., Raslan, R., Tan, K.H., Turk, M.: Specular reflection reduction with multi-flash imaging. In: Proceedings. 17th Brazilian symposium on computer graphics and image processing, IEEE (2004) 316–321
7. Wang, T., Zhu, Y., Jin, L., Luo, C., Chen, X., Wu, Y., Wang, Q., Cai, M.: Decoupled attention network for text recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Volume 34. (2020) 12216–12224
8. Hou, S., Wang, C., Quan, W., Jiang, J., Yan, D.M.: Text-aware single image specular highlight removal. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV), Springer (2021) 115–127
9. Huang, J., Jin, W., Zhao, D., Qin, N., Li, Q.: Double-trapezium cylinder codebook model based on yuv color model for foreground detection with shadow and highlight suppression. Journal of Signal Processing Systems 85 (2016) 221–233
10. Shafer, S.A.: Using color to separate reflection components. Color Research & Application 10 (1985) 210–218
11. Tan, P., Quan, L., Lin, S.: Separation of highlight reflections on textured surfaces. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Volume 2., IEEE (2006) 1855–1860
12. Koirala, P., Hauta-Kasari, M., Parkkinen, J.: Highlight removal from single image. In: International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS), Springer (2009) 176–187
13. 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 (CVPR). (2015) 3725–3733
14. Fu, G., Zhang, Q., Song, C., Lin, Q., Xiao, C.: Specular highlight removal for real-world images. In: Computer Graphics Forum. Volume 38., Wiley Online Library (2019) 253–263

15. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Free-form image inpainting with gated convolution. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019) 4471–4480

16. Chang, Y.L., Liu, Z.Y., Lee, K.Y., Hsu, W.: Free-form video inpainting with 3d gated convolution and temporal patchgan. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019) 9066–9075

17. Park, J.B., Kak, A.C.: A truncated least squares approach to the detection of specular highlights in color images. In: 2003 IEEE International Conference on Robotics and Automation. Volume 1., IEEE (2003) 1397–1403

18. Ortiz, F., Torres, F., Gil, P.: A comparative study of highlights detection and elimination by color morphology and polar color models. In: Iberian Conference on Pattern Recognition and Image Analysis, Springer (2005) 295–302

19. Li, R., Pan, J., Si, Y., Yan, B., Hu, Y., Qin, H.: Specular reflections removal for endoscopic image sequences with adaptive-rpca decomposition. IEEE transactions on medical imaging 39 (2019) 328–340

20. Yang, Q., Tang, J., Ahuja, N.: Efficient and robust specular highlight removal. IEEE transactions on pattern analysis and machine intelligence 37 (2014) 1304–1311

21. Lee, S.W., Bajcsy, R.: Detection of specularity using color and multiple views. In: European Conference on Computer Vision (ECCV), Springer (1992) 99–114

22. Wang, H., Xu, C., Wang, X., Zhang, Y., Peng, B.: Light field imaging based accurate image specular highlight removal. Plos one 11 (2016) e0156173

23. Lei, C., Chen, Q.: Robust reflection removal with reflection-free flash-only cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2021) 14811–14820

24. Tan, R.T., Nishino, K., Ikeuchi, K.: Separating reflection components based on chromaticity and noise analysis. IEEE transactions on pattern analysis and machine intelligence 26 (2004) 1373–1379

25. Yang, J., Liu, L., Li, S.: Separating specular and diffuse reflection components in the hsi color space. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCVW). (2013) 891–898

26. Tan, R.T., Ikeuchi, K.: Reflection components decomposition of textured surfaces using linear basis functions. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Volume 1., IEEE (2005) 125–131

27. Tan, R.T., Ikeuchi, K.: Separating reflection components of textured surfaces using a single image. In: Digitally Archiving Cultural Objects. Springer (2008) 353–384

28. Yang, Q., Wang, S., Ahuja, N.: Real-time specular highlight removal using bilateral filtering. In: European conference on computer vision (ECCV), Springer (2010) 87–100

29. 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). (2018) 268–283

30. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. 28 (2009) 24
31. Song, Y., Yang, C., Lin, Z., Liu, X., Huang, Q., Li, H., Kuo, C.C.J.: Contextual-based image inpainting: Infer, match, and translate. In: Proceedings of the European Conference on Computer Vision (ECCV). (2018) 3–19 4, 6
32. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S.: Generative image inpainting with contextual attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). (2018) 5505–5514 4
33. 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 (CVPR). (2016) 770–778 4
34. Wang, X., Ang Jr, M.H., Lee, G.H.: Cascaded refinement network for point cloud completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 790–799 6
35. Wang, G., Sun, C., Sowmya, A.: Attentive feature refinement network for single rainy image restoration. IEEE Transactions on Image Processing 30 (2021) 3734–3747 6
36. Chen, Z., Long, C., Zhang, L., Xiao, C.: Canet: A context-aware network for shadow removal. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2021) 4743–4752 6
37. Ding, X., Shen, C., Che, Z., Zeng, T., Peng, Y.: Scarf: A semantic constrained attention refinement network for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2021) 3002–3011 6
38. Yi, Z., Tang, Q., Azizi, S., Jang, D., Xu, Z.: Contextual residual aggregation for ultra high-resolution image inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 7508–7517 6
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014) 8
40. Souza, A., Macedo, M., Nascimento, V., Oliveira, B.: Real-time high-quality specular highlight removal using efficient pixel clustering. In: Proceedings of the 31st Conference on Graphics, Patterns and Images (SIBGRAPI), IEEE (2018) 56–63 8, 9
41. Akashi, Y., Okatani, T.: Separation of reflection components by sparse non-negative matrix factorization. In: Asian Conference on Computer Vision (ACCV), Springer (2014) 611–625 8, 9
42. Arnold, M., Ghosh, A., Ameling, S., Lacey, G.: Automatic segmentation and inpainting of specular highlights for endoscopic imaging. EURASIP Journal on Image and Video Processing 2010 (2010) 1–12 8, 9
43. Lin, J., Amine Seddik, M.E., Tamaazousti, M., Tamaazousti, Y., Bartoli, A.: Deep multi-class adversarial specularity removal. In: Scandinavian Conference on Image Analysis, Springer (2019) 3–15 9
44. Murmann, L., Gharbi, M., Aittala, M., Durand, F.: A dataset of multi-illumination images in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). (2019) 4080–4089 10