Hierarchical Adaptive Filtering Network for Text Image Specular Highlight Removal

Zhi Jiang[1] Jingbo Hu[1] Ling Zhang[2] Gang Fu[3] Chunxia Xiao[1*]
[1] School of Computer Science, Wuhan University, China
[2] School of Computer Science and Technology, Wuhan University of Science and Technology, China
[3] Department of Computing, The Hong Kong Polytechnic University, Hong Kong SAR, China
zzz1203685136@whu.edu.cn, zhling@wust.edu.cn, gangfu@polyu.edu.hk, cxxiao@whu.edu.cn

Abstract

Despite significant advances in the field of specular highlight removal in recent years, existing methods predominantly focus on natural images, where highlights typically appear on raised or edged surfaces of objects. These highlights are often small and sparsely distributed. However, for text images such as cards and posters, the flat surfaces reflect light uniformly, resulting in large areas of highlights. Current methods struggle with these large-area highlights in text images, often producing severe visual artifacts or noticeable discrepancies between filled pixels and the original image in the central high-intensity highlight areas. To address these challenges, we propose the Hierarchical Adaptive Filtering Network (HAFNet). Our approach performs filtering at both the downsampled deep feature layer and the upsampled image reconstruction layer. By designing and applying the Adaptive Comprehensive Filtering Module (ACFM) and Adaptive Dilated Filtering Module (ADFM) at different layers, our method effectively restores semantic information in large-area specular highlight regions and recovers detail loss at various scales. The required filtering kernels are pre-generated by a prediction network, allowing them to adaptively adjust according to different images and their semantic content, enabling robust performance across diverse scenarios. Additionally, we utilize Unity3D to construct a comprehensive large-area highlight dataset featuring images with rich texts and complex textures. Experimental results on various datasets demonstrate that our method outperforms state-of-the-art approaches.

1. Introduction

Specular highlights are commonly encountered in our daily lives and often appear in photographs. For instance, when photographing an ID card, specular highlights may obscure important information. Additionally, images with highlights can interfere with various computer vision tasks such as image segmentation [37], text detection [18], and object detection [21]. Consequently, specular highlight removal is a critical and challenging task in computer vision.

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

Figure 1. Specular highlight removal results on text images. (a)(c) Specular highlight images. (b)(d) Highlight removal results using our proposed method.

Traditional methods for highlight removal predominantly analyze the physical and statistical properties of images, employing a variety of techniques such as color space analysis [32], optimization [20], filtering [43], polarization information [34, 38], and illumination estimation [13, 22]. However, these methods often perform poorly on text images, leading to issues such as color tone deviation, incomplete highlight removal, and black color block. The primary limitation of these approaches is their inability to capture high-level semantic information and leverage useful information from both weak highlight regions and non-highlight regions.

*Chunxia Xiao is the corresponding author.

Deep learning-based methods for specular highlight removal have achieved remarkable results in medical images [7], natural object images [46], and specific object images [47, 48, 50]. These highlights are typically small, sparse, and similar in color to the light source. However, for text images with rich texts and complex textures (*e.g.*, cards and posters), existing methods often perform poorly, leading to large area detail loss, or color distortion. Besides method limitations, a significant issue is the lack of datasets containing text images with large-area highlights.

To address the aforementioned issues, we have constructed a large-area specular highlight dataset (LSH) featuring images with rich texts and complex textures. Each image pair includes a highlight image and its highlight-free image, covering a variety of items such as bank cards, game cards, and posters. The highlights in these images vary in shape, intensity, and coverage area. We also propose a novel method for removing large-area highlights from images. By designing corresponding filtering modules at different feature layers, we aim to achieve a larger receptive field to recover semantic information under extensive strong highlights while better restoring detailed information, as shown in Figure 1.

When features are downsampled to the deep feature layer, the resulting features contain rich high-level semantic information. We propose to apply the Adaptive Comprehensive Filtering Module (ACFM) at this layer, which not only removes large-area highlights but also effectively recovers semantic information. In ACFM, we predict a small kernel to adaptively preserve local information, while simultaneously predicting a large kernel to capture a larger receptive field. Additionally, to address the limitation of large kernels in covering global regions, we utilize fast Fourier transform to process them in the frequency domain, thereby achieving a global receptive field. However, predicting large kernels requires significant parameter consumption. To address this, we introduce a Parameter-Optimized Filtering Module (POFM), which reduces the number of parameters from ck^2 to $2ck$ using outer product operations. This reduction allows us to achieve a large receptive field with fewer parameters. To better refine details and reduce artifacts, we perform the Adaptive Dilated Filtering Module (ADFM) at the upsampled image reconstruction layer to recover detail loss at various scales.

In summary, our contributions are:

- We propose a novel network for specular highlight removal, dubbed HAFNet. This network is effective for handling large-area specular highlights in text images.
- We propose the Adaptive Comprehensive Filtering Module (ACFM), which preserves local structures while obtaining a global receptive field, thereby better restoring the semantic information in highlight regions. Additionally, the Parameter-Optimized Filtering Module (POFM) maintains a low parameter count. Furthermore, we propose the Adaptive Dilated Filtering Module (ADFM) to refine details and reduce artifacts.
- We construct a large-area highlight dataset featuring images with rich texts and complex textures. The image pairs are precisely aligned, and non-highlight regions maintain consistent color tones.

2. Related Work

2.1. Traditional Methods

Early traditional specular highlight removal methods primarily relied on analyzing pixel brightness and color information in images [30, 33]. The dichromatic reflection model proposed by [29], significantly contributed to highlight removal research. Methods based on this model, such as those in works [3, 27, 43], perform well in handling complex images. Additionally, illumination estimation-based methods [2, 13] are particularly suitable for complex lighting conditions, effectively removing highlights while maintaining image detail and color consistency. However, these traditional methods often fail to capture high-level semantic information and do not leverage useful information from weak highlight and non-highlight areas, frequently resulting in illumination residues, black spots, and other visual artifacts.

2.2. Deep Learning-Based Methods

In recent years, deep learning-based methods for image highlight removal have advanced significantly. Lin et al. [19] introduced multi-class adversarial losses, enabling the model to more accurately identify and remove different types of highlights. Muhammad et al. [25] designed Spec-Net and Spec-CGAN specifically for facial image highlight. Wu et al. [40] created a real-world dataset and proposed a GAN-based highlight removal method. Fu et al. [4] improved model efficiency and accuracy through joint detection and removal tasks. Hou et al. [9] focused on highlight removal in text images, incorporating a text detection module to ensure text clarity and readability.

Fu et al. [5] used large-scale synthetic data for training and proposed a three-stage network based on the dichromatic reflection model to eliminate highlights while maintaining color consistency. Wu et al. [41] combined UNet and Transformer architectures, leveraging Transformer's global characteristics and UNet's local features to enhance removal accuracy. Hu, Huang, and Wang [10] proposed a method based on an improved dichromatic reflection model, and a coarse-to-fine network structure was used to remove highlights. However, these methods primarily target medical images, natural images, and specific object images. Their performance on large-area highlight images with rich texts and complex textures remains unsatisfactory.

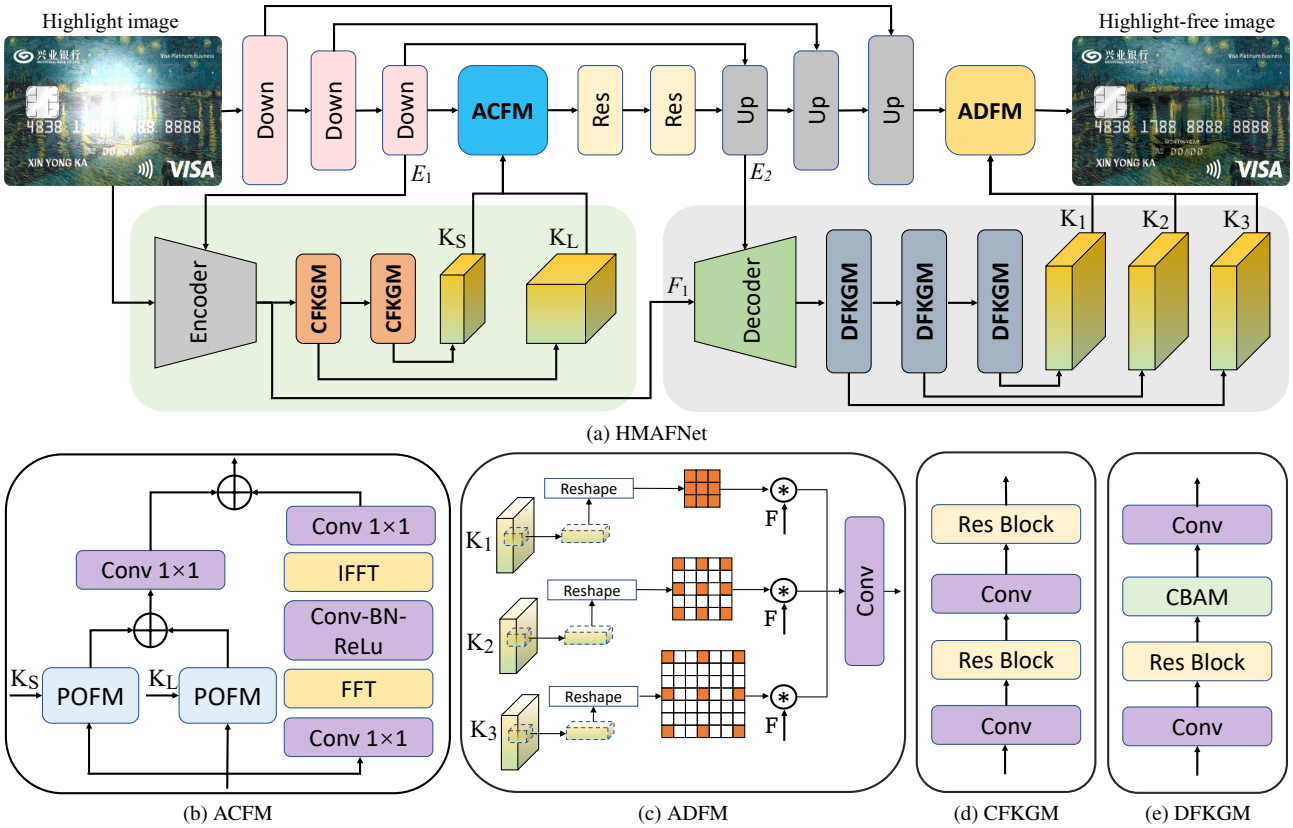

Figure 2. (a) The framework of our HAFNet. Down represents the Down-sampling Block, Up represents the Up-sampling Block and Res represents the Residual Block. (b) The Adaptive Comprehensive Filtering Module (ACFM). K_S and K_L are dynamically predicted by the network as the small and large kernels, respectively. (c) The Adaptive Dilated Filtering Module (ADFM). (d) The Comprehensive Filtering Kernel Generation Modul (CFKGM). (e) The Dilated Filtering Kernel Generation Modul (DFKGM).

2.3. Kernel Prediction Filtering

kernel prediction filtering is an advanced image processing method where a kernel prediction network dynamically predicts kernels for each pixel to filter the input image and generate high-quality output images [11]. In recent years, this method has seen various optimizations and enhancements in multiple studies [23, 49]. This technology has been widely applied to various computer vision tasks, including denoising [15, 24], super-resolution [1, 36, 42], video interpolation [26], image inpainting [8, 17], and shadow removal [6]. The core advantage of kernel prediction filtering lies in its ability to better preserve local structures of the image and effectively eliminate artifacts. However, it often faces the challenge of limited receptive fields, leading to suboptimal performance in some tasks. The method proposed in this paper employs the strengths of kernel prediction filtering. Furthermore, our approach addresses the issue of limited receptive fields by employing comprehensive filtering at the deep feature layer.

3. Dataset

3.1. Background

Natural Images. Currently, some publicly available high-quality natural image highlight datasets, such as SHIQ [4], PSD [40], and SSHR [5], have been widely used for experimental training and evaluation. In these images, highlights usually appear on protrusions, edges, or folds of the object's surface, typically small in area and sparsely distributed. Networks trained on these datasets are unable to effectively remove large-area highlights in text images and fail to recover the information obscured by highlights.

Text Images. Hou et al. [9] introduced three finely annotated text image highlight datasets (RD, SD1, SD2), but they have some limitations. RD was captured under controlled conditions by switching lights, and the highlights were generated by adding a plastic film to the object's surface rather than being inherent to the object itself, as shown in the first row of Figure 3. Collecting high-quality highlight images requires significant human and material resources due to manual adjustments of objects and lights. Consequently,

Figure 3. Example image pairs from the RD, SD1, and SD2 datasets are shown in the first, second, and third rows, respectively.

the dataset size is limited to only 1800 images for training. Additionally, this dataset features simple lighting conditions and tone deviations between image pairs. As shown in the second and third rows of Figure 3, SD1 and SD2 are synthetic datasets where the highlights in the images exhibit simple shapes and straightforward positional distribution. The highlights in SD1 and SD2 are simply added as distinct shapes to the images, without simulating realistic lighting conditions and material properties to produce authentic highlights. Additionally, some image pairs exhibit severe tone inconsistencies, posing significant challenges for model training.

3.2. LSH Dataset Construction

We obtained highlight-free images through camera photography and online downloads. Next, we leveraged Unity3D's advanced rendering and designed custom Physically Based Rendering (PBR) shaders to enable different shapes and intensities of highlights under various lighting conditions.

Our approach employs parameterized shader functions to dynamically adjust PBR parameters—Metallic, Smoothness, and more. Beyond basic PBR, we incorporated advanced material properties such as Normal Mapping, Ambient Occlusion, and Anisotropic Reflection to ensure nuanced highlight realism. Additionally, we implemented an adaptive lighting strategy, adjusting directional lights, point lights, area lights, and spot lights based on environment-specific models that simulate real-world illumination scenarios. Our pipeline leverages physics-informed lighting adjustments to generate highlights with high variability. As shown in Figure 4, this approach allows us to generate a dataset with realistic highlight effects under various lighting

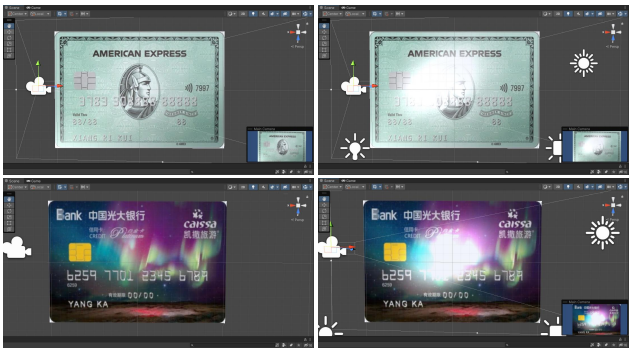

Figure 4. The collection pipelines of our dataset.

conditions, thereby enhancing the training and generalization capabilities of our highlight removal model.

Our dataset contains 13.6k pairs for training and 1.8k pairs for testing, with each pair consisting of a highlight image and its highlight-free image. The main types include bank cards, bus cards, identity cards and posters. Image pairs are perfectly aligned and have consistent tonal values in non-highlight regions. Our dataset is characterized by large highlight areas, rich text information, and complex textures, with a variety of highlight shapes and intensities.

4. Method

4.1. Framework Overview

The overall pipeline is illustrated in Figure 2 (a). The input image is downsampled and filtered at the deep feature layer using the Adaptive Comprehensive Filtering Module (ACFM), which removes large-area specular highlights while restoring semantic information. The kernels required for ACFM are predicted by a network that takes the original image as input along with E_1 features for guidance. The filtered features are then upsampled to the image level, where the Adaptive Dilated Filtering Module (ADFM) performs filtering to refine details and reduce artifacts. The kernels required for ADFM are predicted by a network, which receives feature F_1 as input, guided by E_2.

4.2. Kernel Prediction Filtering for Highlight Removal

To address the issue of visual artifacts commonly produced by previous methods after removing specular highlights, we leverage Kernel Prediction Filtering (KPF) [35]. Artifacts often arise from inaccurate predictions or excessive smoothing. One of the primary design goals of KPF is to minimize these artifacts. By employing fine-grained local structure restoration and dynamic adjustment, KPF effectively mitigates common artifacts.

The kernel prediction network can dynamically distinguish highlight areas of varying intensities, predicting ap-

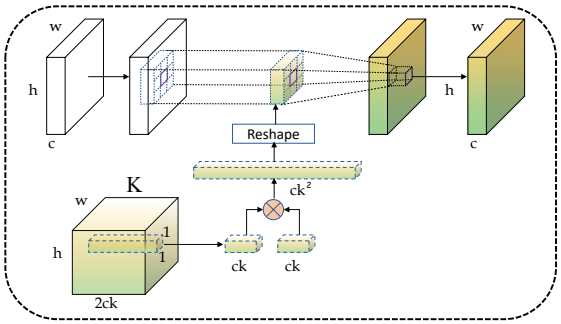

Figure 5. The Parameter-Optimized Filtering Module (POFM).

propriate kernels accordingly, as shown in Figure 6 (a). For non-highlight areas, the network effectively preserves the original pixels; for strong highlight edge areas, it fully utilizes the surrounding useful pixels; and for highlights of varying intensities, it performs intelligent processing. The result of applying kernel prediction filtering to highlight image is shown in Figure 6 (c). While the kernel prediction filtering demonstrates its advantages in highlight removal, it also reveals a significant issue. Due to insufficient receptive fields, the network struggles to restore information in the central regions after removing large-area specular highlights.

4.3. Large-Area Highlight Removal

To address the issue of insufficient receptive fields in kernel prediction filtering and the inability to restore semantic information under large-area highlights, we considered two intuitive approaches. The first approach involves increasing the size of the predicted kernel. For example, Niklaus, Mai, and Liu [26] used neural networks to generate two 41×41 kernels for each output pixel, requiring 26GB of memory for 1080p video frames due to the quadratic increase in memory with kernel size, which also significantly complicates training. Figure 6 (d) shows the result of this strategy, where the central region exhibits severe visual artifacts. This occurs because large kernels can cause uneven filter responses, resulting in over-processing in some areas and under-processing in others during highlight removal.

The second approach involves designing a recurrent network that iteratively filters the highlight removal results multiple times, using the pixels filtered in the previous round to reconstruct the missing regions. Figure 6 (e) shows the result of this strategy. The details in the central region remain unrecovered, and some areas become blurred. This is mainly because extensive highlight regions disrupt the local structure, causing reconstruction errors to accumulate during the iterative filtering process.

To address the aforementioned challenges, we propose a filtering strategy at the deep feature layer. This extension allows for a larger receptive field, which not only effectively removes extensive highlights but also restores semantic information. However, this strategy has notable limitations: in prioritizing a larger receptive field, it neglects local information, leading to a loss of fine details. Additionally, when the feature size is large, the predicted kernel fails to fully cover all essential information, resulting in incomplete filtering. To address these issues, we propose the Adaptive Comprehensive Filtering Module (ACFM).

Adaptive Comprehensive Filtering Module. As shown in Figure 2 (b), the ACFM effectively preserves fine-grained local details crucial for maintaining texture integrity by using a 3x3 predicted kernel (K_S). Simultaneously, the use of a 15x15 predicted kernel (K_L) offers extensive coverage to handle large-scale highlight regions. Our predicted kernels are generated through the prediction network, which adaptively adjusts according to different images and their semantic information, thereby adapting to various scenarios.

When dealing with large-area highlights, a global receptive field is essential for understanding the image and effectively restoring semantic information obscured by highlights. To achieve this, we utilize Fast Fourier Transform (FFT) to convert features from the spatial domain to the frequency domain. After processing in the frequency domain, we apply the inverse Fourier Transform to convert the features back to the spatial domain. Processing in the frequency domain allows us to achieve a global receptive field that enhances our ability to recover critical semantic information under large-area highlights. Even in the presence of large, intense highlights in the central region, the semantic-level understanding of the image enables the reasonable restoration of information obscured by the highlights.

In the ACFM, we propose the Parameter-Optimized Filtering Module (POFM) to reduce the parameter count of the predicted kernels. We incorporate the concept of separable kernel estimation to further reduce the parameter count:

$$K^s(x,y) = k(x,y)^1 \otimes k(x,y)^2, \tag{1}$$

where $k(x,y)^1$ and $k(x,y)^2$ are from K, as shown in Figure 5. $\otimes$ represents the outer product operation. $K^s(x,y)$ is the predicted kernel with size s at feature (x,y). This method can reduce the number of parameters from s^2 to $2s$. Specifically, when we filter at deep feature (x,y):

$$F(x,y) = K^s(x,y) * P^s(x,y), \tag{2}$$

where $*$ represents the filtering operation. $P^s(x,y)$ is the patch centered at (x,y) with size s in the feature.

The result of using ACFM at the deep feature layer is shown in Figure 6 (f). Compared to image-level prediction filtering, our method removes extensive highlights and restores semantic information. Although the main structure

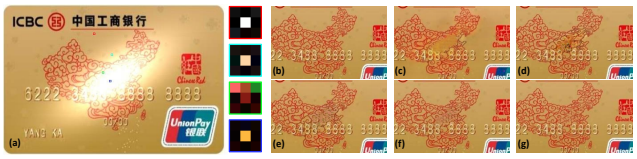

Figure 6. (a) Highlight images and predicted kernels for different specular highlight regions. (b) Ground truth (GT). (c) Result after image-level filtering using KPF [35]. (d) Result after applying large-kernel filtering. (e) Result from multiple filtering iterations with a recurrent network. (f) Result after filtering with ACFM at deep feature layer. (g) Result obtained with our proposed method.

has been recovered, some details are still lost. After removing strong highlights from central areas, issues such as color distortion and structural blurring persist.

4.4. Detail Refinement for Highlight Regions

Comprehensive filtering at the deep feature layer captures high-level semantic information, ensuring more contextually coherent restoration of large highlight areas. However, it has limitations in recovering fine details, potentially leading to blurring or imprecision. To address this, we propose incorporating finer-grained filtering at the image reconstruction layer, enhancing detail fidelity in the final output.

We predict the required kernels for filtering at the image reconstruction layer, incorporating CBAM [39] in the kernel prediction process. By applying attention mechanisms across both channel and spatial dimensions, CBAM enhances feature maps by emphasizing critical features and suppressing less relevant ones. This attention-driven refinement is essential for generating accurate adaptive kernels within the kernel prediction module, enabling precise detail recovery and producing natural images free from highlight artifacts.

Adaptive Dilated Filtering Module. Upon observation, we found that the results after comprehensive filtering often exhibit missing details at various scales. Inspired by comprehensive experiments, as discussed in [16], multi-scale information is crucial for restoring high-quality images. A straightforward solution would be to directly predict multi-scale kernels [23], but this approach incurs additional parameter and time costs. To address this, we adopt the concept of dilated convolutions [45], expanding each 3×3 predicted kernel to different scales:

$$\hat{I}_i(\mathbf{x}) = \sum_{\mathbf{t},\mathbf{y}=\mathbf{x}+l\mathbf{t}} K_i^s(\mathbf{t}) I(\mathbf{y}), \tag{3}$$

where $\mathbf{x}$ and $\mathbf{y}$ denote pixel coordinates, where the range of $\mathbf{t}$ is from $(-\frac{S-1}{2}, -\frac{S-1}{2})$ to $(\frac{S-1}{2}, \frac{S-1}{2})$. K_i is obtained through the DFKGM. and $i = 1, 2, 3$. l is the dilation factor, l=i. Then simply fuse these three features through a convolution:

$$I_{free} = \text{Conv}(\text{Concat}(\hat{I}_1, \hat{I}_2, \hat{I}_3)), \tag{4}$$

This enables handling details at different scales while reducing the number of parameters.

4.5. Loss Functions

Adversarial Loss. We adopt the relativistic average adversarial loss, which not only considers the discriminator's scores for generated and real images but also their relative differences.

$$\begin{aligned}\mathcal{L}_{adv} = 0.5 \cdot (&\text{BCE}(\sigma(D(I_{free}) - D(I_{gt})), y') \\ &+ \text{BCE}(\sigma(D(I_{free}) - D(I_{gt})), y)),\end{aligned} \tag{5}$$

where σ is the sigmoid function, and BCE$(*)$ measures the binary cross-entropy. For the generator, (y', y) is set to $(1, 0)$, and for the discriminator, it is set to $(0, 1)$. D is our discriminator. This relative scoring strategy captures subtle differences better, promoting the generation of higher-quality images.

Perceptual Loss. We use the perceptual loss defined in [12], utilizing the VGG-19 network pre-trained on the ImageNet dataset [28]. This loss captures high-level features of the images, ensuring that the estimated highlight-free images retain important content and structural information, making the final generated images more semantically consistent with the original ones. The perceptual loss is formulated as:

$$\mathcal{L}_{perc} = \frac{1}{C_i H_i W_i} \|\phi_i(I_{free}) - \phi_i(I_{gt})\|_1, \tag{6}$$

where $\phi_i(\cdot)$ denotes the features from the i-th layer of the pre-trained VGG-19 network, and C_i, H_i, W_i are the dimensions of the features.

Style Loss. We use the style loss, which calculates the differences between the Gram matrices of the generated and target images, effectively capturing and preserving the texture.

$$\mathcal{L}_{style} = \sum_i \|G(\phi_i(I_{free})) - G(\phi_i(I_{gt}))\|_1, \tag{7}$$

where ϕ_i denotes the features from the i-th layer of the same pre-trained network used for perceptual loss, and $G(\cdot)$ is the Gram matrix, which computes the correlations between feature maps' channels. This loss helps to recover texture details lost due to highlights, making the highlight-free images visually more natural and coherent.

The total loss of our method is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{perc} + \lambda_4 \mathcal{L}_{style}, \tag{8}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are weight parameters. In our experiments, we set $\lambda_1 = 1$, $\lambda_2 = 0.1$, $\lambda_3 = 0.1$, and $\lambda_4 = 0.5$.

(a) Input (b) TASHR (c) JSHDR (d) Wu (e) TSHR (f) IDRHR (g) Ours (h) GT

Figure 7. Visual comparison results on our LSH dataset.

Method	LSH		RD		SD1	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
Shen	15.479	0.771	18.705	0.765	18.489	0.855
Yang	11.473	0.405	15.625	0.531	13.890	0.479
TASHR	22.996	0.869	21.474	0.793	24.664	0.917
JSHDR	22.771	0.899	21.546	0.807	21.866	0.891
Wu	24.564	0.907	22.989	0.845	24.324	0.918
TSHR	25.066	0.905	22.793	0.818	25.147	0.930
IDRHR	24.584	0.914	23.903	0.840	24.824	0.938
Ours	**26.593**	**0.937**	**26.217**	**0.885**	**28.952**	**0.947**

Table 1. Quantitative comparison of our method with state-of-the-art specular highlight removal methods on our LSH, RD [9], and SD1 [9]. The best results are marked in bold.

5. Experiments

5.1. Implementation Details

We implemented the entire network in PyTorch and trained it for 100 epochs on a PC equipped with NVIDIA GeForce GTX 3090. The entire network is optimized using the Adam optimizer [14]. The initial learning rate is set to 1×10^{-4}, with a batch size of 8. The input image size to our network is consistent with the image sizes in datasets.

5.2. Datasets and Evaluation Metrics

We evaluated our network on three datasets, including our LSH, RD [9], and SD1 [9]. We adopt two commonly used metrics in the highlight removal task to quantitatively evaluate the performance of our method: Structural Similarity Index (SSIM) and Peak Signal-to-Noise Ratio (PSNR).

5.3. Comparisons with State-of-the-art Methods

We compared our method with some state-of-the-art methods: Shen [31], Yang [44], TASHR [9], JSHDR [4], Wu [40], TSHR [5], IDRHR [10]. Shen and Yang are traditional methods, while the others are state-of-the-art deep learning-based methods. For a fair comparison, we used the publicly available code and the best parameter settings described in the papers provided by the authors for training and evaluation, selecting the best results among them.

Quantitative Comparison. Table 1 presents the quantitative evaluation of our method compared with the aforementioned methods on three datasets. From the table, we can observe that our method achieves higher SSIM and PSNR scores on all three datasets compared to the other methods, indicating superior performance. Furthermore, the higher SSIM and PSNR scores in the RD validate the effectiveness of our method in real-world scenarios. Observing the data in the table, it is evident that traditional highlight removal methods perform poorly on our large-area highlight images compared to learning-based methods.

Qualitative Comparison. Figure 7 shows the visual comparison results on our LSH dataset. TASHR [9] and JSHDR [4] can remove large-area highlights, but they often produce black patches and illumination residues, as shown in Figure 7 (b) and (c). Wu [40] can more effectively remove large-area highlights, but they often produce color distortions and other visual artifacts, as shown in Figure 7 (d). TSHR [5] and IDRHR [10] largely avoid some visual artifacts, but they cannot recover the detailed information under highlights, as shown in Figure 7 (e) and (f). In contrast, our method: (1) generates more natural and high-fidelity images; (2) effectively restores information in areas with strong highlights; and (3) recovers finer details in regions where the information under highlights is not completely lost. Due to space constraints, we provide visual comparisons for the RD and SD1 datasets in the supplementary material.

To further verify the robustness and generalization ca-

Input Wu TSHR IDRHR GT

Figure 8. Visual results for real highlight images.

pability of our method on real images, we compare it with three recent state-of-the-art methods Wu [40], TSHR [5], IDRHR [10]. Figure 8 presents highlight removal results for real highlight images, which were captured using smartphones or downloaded from the Internet, and for which ground truth data is not available. Although we cannot perform quantitative comparisons due to the lack of ground truth, visually, our method effectively removes highlights while restoring the information obscured by the highlights and avoiding visual artifacts.

5.4. Ablation Studies

To demonstrate the impact of each component of our method on the experimental results, we conducted a series of ablation experiments. We compare our network with four variants to assess the impact of each component. The variants are (1) Filtering is not performed at the deep feature layer, meaning the ACFM is removed; (2) Filtering is not performed at the image reconstruction layer, meaning the ADFM is removed; (3) replace ACFM with standard KPF and (4) replace ADFM with standard KPF. We train the variants on Our LSH. From the table 2, we can observe: (1) our HAFNet with all components gets the best results; (2) the proposed ACFM and ADFM can help improve the performance of the network, and the combination leads to the best performance.

To further investigate the proposed Module, we tested different Large kernel sizes in ACFM and different predicted kernel numbers in ADFM. As shown in Table 3, for the large predicted kernel of ACFM, the larger the size of the adaptive kernel, the better the performance. The performance improvement is significant from $k = 11$ to $k = 15$ but less so beyond that, while the parameter amount increases significantly. The use of POFM significantly reduces the parameter amount for larger kernels. We predict 3×3 kernels and tested different Kenel nums with different dilation factors in ADFM, As shown in Table 4. we balanced computational complexity, model size, and performance, and finally determined the large kernel size for ACFM as $k = 15$, and numbers for ADFM as $i = 3$ with dilation factors $l = i$.

Methods	PSNR	SSIM
Without ACFM	22.017	0.849
Without ADFM	25.711	0.907
Replace ACFM with Standard KPF	25.148	0.912
Replace ADFM with Standard KPF	26.019	0.921
Our HMAFNet	26.593	0.937

Table 2. Quantitative results of ablation study on our LSH.

Kernel Size	K=11	K=13	K=15	K=17	K=19
PSNR	25.698	26.284	26.593	26.603	26.611
SSIM	0.929	0.931	0.937	0.933	0.932
Pre-Params(M)	87.228	121.831	162.201	208.338	260.243
Params(M)	15.859	18.743	21.626	24.510	27.394

Table 3. Results of different sizes of large predicted kernels in ACFM. Predicted kernel parameters without applying POFM. Predicted kernel parameters in the proposed method.

Kernel nums	i=1	i=2	i=3	i=4
PSNR	26.019	26.532	26.593	26.607
SSIM	0.921	0.929	0.937	0.930

Table 4. Results for different predicted kernel numbers in ADFM.

5.5. Limitations

When the highlight area is too large and its intensity too strong, the lack of reliable information for learning makes highlight removal challenging. Additionally, when the highlights are complex, even after removal, accurately restoring the underlying detail information is difficult, often resulting in blurriness. Due to space constraints, specific examples are provided in the supplementary materials.

6. Conclusion

In this paper, we have proposed the HAFNet for highlight removal. The goal is to address text images with large-area highlights. We achieved high-fidelity highlight removal results by employing HAFNet. Additionally, we constructed a high-quality dataset to facilitate network training and quantitative evaluation. Image pairs in our dataset are perfectly aligned and have consistent tonal values in non-highlight regions. Our dataset includes text images such as bank cards, posters, etc., characterized by large-area highlights with high intensity and containing rich text and complex texture. Extensive experiments demonstrate the state-of-the-art performance of our method.

Acknowledgments

This work is partially supported by the National Natural Science Foundation of China (No.61972298, No.62372336 and No.62402324).

References

[1] Wooyeong Cho, Sanghyeok Son, and Dae-Shik Kim. Weighted multi-kernel prediction network for burst image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 404–413, 2021. 3

[2] Graham D. Finlayson, Steven D. Hordley, and Paul M. Hubel. Color by correlation: A simple, unifying framework for color constancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1209–1221, 2001. 2

[3] Gang Fu, Qing Zhang, Chengfang Song, Qifeng Lin, and Chunxia Xiao. Specular highlight removal for real-world images. In *Computer graphics forum*, pages 253–263. Wiley Online Library, 2019. 2

[4] Gang Fu, Qing Zhang, Lei Zhu, Ping Li, and Chunxia Xiao. A multi-task network for joint specular highlight detection and removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7752–7761, 2021. 2, 3, 7

[5] Gang Fu, Qing Zhang, Lei Zhu, Chunxia Xiao, and Ping Li. Towards high-quality specular highlight removal by leveraging large-scale synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12857–12865, 2023. 2, 3, 7, 8

[6] Lan Fu, Changqing Zhou, Qing Guo, Felix Juefei-Xu, Hongkai Yu, Wei Feng, Yang Liu, and Song Wang. Auto-exposure fusion for single-image shadow removal. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10571–10580, 2021. 3

[7] Isabel Funke, Sebastian Bodenstedt, Carina Riediger, Jürgen Weitz, and Stefanie Speidel. Generative adversarial networks for specular highlight removal in endoscopic images. In *Medical imaging 2018: Image-guided procedures, robotic interventions, and modeling*, pages 8–16. SPIE, 2018. 2

[8] Qing Guo, Xiaoguang Li, Felix Juefei-Xu, Hongkai Yu, Yang Liu, and Song Wang. Jpgnet: Joint predictive filtering and generative network for image inpainting. In *Proceedings of the 29th ACM International conference on multimedia*, pages 386–394, 2021. 3

[9] Shiyu Hou, Chaoqun Wang, Weize Quan, Jingen Jiang, and Dong-Ming Yan. Text-aware single image specular highlight removal. In *Pattern Recognition and Computer Vision: 4th Chinese Conference, PRCV 2021, Beijing, China, October 29–November 1, 2021, Proceedings, Part IV 4*, pages 115–127. Springer, 2021. 2, 3, 7

[10] Kun Hu, Zhaoyangfan Huang, and Xingjun Wang. Highlight removal network based on an improved dichromatic reflection model. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2645–2649. IEEE, 2024. 2, 7, 8

[11] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 3

[12] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 6

[13] Hamid Reza Vaezi Joze and Mark S Drew. Exemplar-based color constancy and multiple illumination. *IEEE transactions on pattern analysis and machine intelligence*, 36(5): 860–873, 2013. 1, 2

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[15] Junyong Lee, Hyeongseok Son, Jaesung Rim, Sunghyun Cho, and Seungyong Lee. Iterative filter adaptive network for single image defocus deblurring. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2034–2042, 2021. 3

[16] Juncheng Li, Faming Fang, Kangfu Mei, and Guixu Zhang. Multi-scale residual network for image super-resolution. In *Proceedings of the European conference on computer vision (ECCV)*, pages 517–532, 2018. 6

[17] Xiaoguang Li, Qing Guo, Di Lin, Ping Li, Wei Feng, and Song Wang. Misf: Multi-level interactive siamese filtering for high-fidelity image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1869–1878, 2022. 3

[18] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11474–11481, 2020. 1

[19] John Lin, Mohamed El Amine Seddik, Mohamed Tamaazousti, Youssef Tamaazousti, and Adrien Bartoli. Deep multi-class adversarial specularity removal. In *Image Analysis: 21st Scandinavian Conference, SCIA 2019, Norrköping, Sweden, June 11–13, 2019, Proceedings 21*, pages 3–15. Springer, 2019. 2

[20] Stephen Lin, Yuanzhen Li, Sing Bing Kang, Xin Tong, and Heung-Yeung Shum. Diffuse-specular separation and depth recovery from image sequences. In *Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, May 28–31, 2002 Proceedings, Part III 7*, pages 210–224. Springer, 2002. 1

[21] Yudong Liu, Yongtao Wang, Siwei Wang, TingTing Liang, Qijie Zhao, Zhi Tang, and Haibin Ling. Cbnet: A novel composite backbone network architecture for object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11653–11660, 2020. 1

[22] Satya P Mallick, Todd E Zickler, David J Kriegman, and Peter N Belhumeur. Beyond lambert: Reconstructing specular surfaces using color. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 619–626. Ieee, 2005. 1

[23] Talmaj Marinč, Vignesh Srinivasan, Serhan Gül, Cornelius Hellge, and Wojciech Samek. Multi-kernel prediction networks for denoising of burst images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2404–2408. IEEE, 2019. 3, 6

[24] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE con-*

ference on computer vision and pattern recognition, pages 2502–2510, 2018. 3

[25] Siraj Muhammad, Matthew N Dailey, Muhammad Farooq, Muhammad F Majeed, and Mongkol Ekpanyapong. Spec-net and spec-cgan: Deep learning models for specularity removal from faces. *Image and Vision Computing*, 93:103823, 2020. 2

[26] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE international conference on computer vision*, pages 261–270, 2017. 3, 5

[27] Weihong Ren, Jiandong Tian, and Yandong Tang. Specular reflection separation with color-lines constraint. *IEEE Transactions on image processing*, 26(5):2327–2337, 2017. 2

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 6

[29] Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985. 2

[30] Hui-Liang Shen and Qing-Yuan Cai. Simple and efficient method for specularity removal in an image. *Applied optics*, 48(14):2711–2719, 2009. 2

[31] Hui-Liang Shen and Zhi-Huan Zheng. Real-time highlight removal using intensity ratio. *Applied optics*, 52(19):4483–4493, 2013. 7

[32] Ping Tan, Long Quan, and Stephen Lin. Separation of highlight reflections on textured surfaces. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 1855–1860. IEEE, 2006. 1

[33] Robby T Tan and Katsushi Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE transactions on pattern analysis and machine intelligence*, 27(2):178–193, 2005. 2

[34] Shinji Umeyama and Guy Godin. Separation of diffuse and specular components of surface reflection by use of polarization and statistical analysis of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(5):639–647, 2004. 1

[35] Thijs Vogels, Fabrice Rousselle, Brian McWilliams, Gerhard Röthlin, Alex Harvill, David Adler, Mark Meyer, and Jan Novák. Denoising with kernel prediction and asymmetric loss functions. *ACM Transactions on Graphics (TOG)*, 37 (4):1–15, 2018. 4, 6

[36] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 606–615, 2018. 3

[37] Zhengyang Wang, Na Zou, Dinggang Shen, and Shuiwang Ji. Non-local u-nets for biomedical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6315–6322, 2020. 1

[38] Sijia Wen, Yinqiang Zheng, and Feng Lu. Polarization guided specular reflection separation. *IEEE Transactions on Image Processing*, 30:7280–7291, 2021. 1

[39] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 6

[40] Zhongqi Wu, Chuanqing Zhuang, Jian Shi, Jianwei Guo, Jun Xiao, Xiaopeng Zhang, and Dong-Ming Yan. Single-image specular highlight removal via real-world dataset construction. *IEEE Transactions on Multimedia*, 24:3782–3793, 2021. 2, 3, 7, 8

[41] Zhongqi Wu, Jianwei Guo, Chuanqing Zhuang, Jun Xiao, Dong-Ming Yan, and Xiaopeng Zhang. Joint specular highlight detection and removal in single images via unet-transformer. *Computational Visual Media*, 9(1):141–154, 2023. 2

[42] Yu-Syuan Xu, Shou-Yao Roy Tseng, Yu Tseng, Hsien-Kai Kuo, and Yi-Min Tsai. Unified dynamic convolutional network for super-resolution with variational degradations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12496–12505, 2020. 3

[43] Qingxiong Yang, Shengnan Wang, and Narendra Ahuja. Real-time specular highlight removal using bilateral filtering. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 87–100. Springer, 2010. 1, 2

[44] Qingxiong Yang, Jinhui Tang, and Narendra Ahuja. Efficient and robust specular highlight removal. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1304–1311, 2014. 7

[45] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 6

[46] Ling Zhang, Yidong Ma, Zhi Jiang, Weilei He, Zhongyun Bao, Gang Fu, Wenju Xu, and Chunxia Xiao. Highlightremover: Spatially valid pixel learning for image specular highlight removal. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10046–10054, 2024. 2

[47] Hongsheng Zheng, Wenju Xu, Zhenyu Wang, Xiao Lu, and Chunxia Xiao. Facial highlight removal with cross-context attention and texture enhancement. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2

[48] Hongsheng Zheng, Zhongyun Bao, Gang Fu, Xuze Jiao, and Chunxia Xiao. Phr-diff: Portrait highlight removal via patch-aware diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. 2

[49] Jingkai Zhou, Varun Jampani, Zhixiong Pi, Qiong Liu, and Ming-Hsuan Yang. Decoupled dynamic filter networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6647–6656, 2021. 3

[50] Wentao Zou, Xiao Lu, Zhilv Yi, Ling Zhang, Gang Fu, Ping Li, and Chunxia Xiao. Eyeglass reflection removal with joint learning of reflection elimination and content inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 2