CDFF-Net: Cumulative Dense Feature Fusion for Single Image Specular Highlight Removal

Shijian XU
xsj13260906215@gmail.com

Abstract

Specular highlights are everywhere in our daily lives. However, they are often undesirable in the photography community, as they can severely bury the details of objects in the scene and degrade the image qualities. Existing highlight removal methods primarily rely on strict assumptions and can easily fail in scenes with complex backgrounds and illumination conditions. Although the success of deep learning techniques has been witnessed in many low-level vision areas, such as image denoising and image super-resolution, there is still few explorations of deep learning in the highlight removal area. This is partly due to the fact that there exists no large scale datasets of high-quality training data. In this paper, we propose to address the single image highlight removal problem from the following aspects. First, to facilitate the learning process for highlight removal, we construct a large scale synthetic dataset, which contains 11000 *pairs of images with/without highlights. It will be made publicly available for future research. Second, we propose a novel Cumulative Dense Feature Fusion network, called CDFF-Net, to take full advantages of low-level features for producing high-quality highlight-free images. Moreover, a cascaded learning scheme is applied to first learn the residual specular layer, and then remove the corresponding specular highlights according to the specular prediction. We conduct extensive evaluations on both synthetic and real data to verify the superiority of the proposed method against the state-of-the-art highlight removal methods. We also demonstrate the potential of the proposed network for other low-level vision tasks such as single image deraining.*

1. Introduction

Specular highlights can be easily observed in our daily lives, as long as there is a light source. However, as they will bury the object details and distort the object colors, highlights are often annoying to photographers. The degraded images suffering highlights can further fail existing computer vision tasks, such as object detection, semantic segmentation, *etc*. Hence, it is necessary to remove the highlights from the input images.

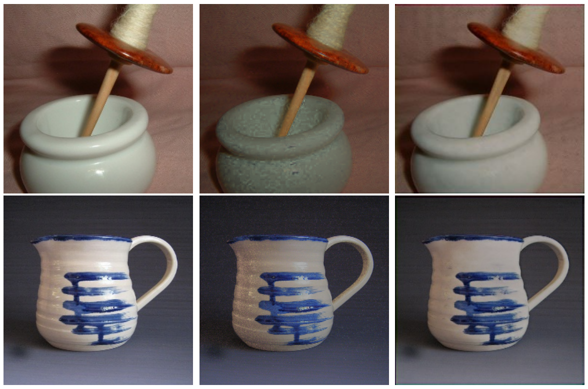

(a) Input (b) Guo *et al*. [11] (c) Our CDFF-Net

Figure 1: State-of-the-art specular highlight removal method [11] still under/over remove the highlights from the objects, which produces images with disrupted object appearances and noisy background. Our CDFF-Net can accurately remove the highlights and preserve the images with clean background.

As it is an extremely ill-posed problem, a line of previous works [27, 49, 48, 28, 40, 42] typically take advantage of multiple input images so that additional constraints can be imposed. Other methods [47, 51, 50, 43, 21, 1, 46, 39, 25, 11] focus on the more challenging single image specular highlight removal problem. However, while they are developed based on strict assumptions/priors such as color space analysis and sparse matrix decomposition, they can easily fail in scenes with complex background or illumination conditions, where their assumptions (*e.g*., uniform surface colors) are violated. Figure 1 shows two examples that the state-of-the-art method [11] can not accurately remove the highlights from the object textures. This under/over removal can significantly disrupt object appearances and produce noisy backgrounds.

On the other hand, the success of deep learning has been witnessed not only in high-level semantic vision tasks [29, 9, 37, 10], but also in various low-level visual tasks, such as image dehazing [7, 38], deraining [35, 52] and shadow removal [36, 20]. Compared with traditional methods, deep learning has the superior ability to exploit the contextual information for recovering the images from degradation. Nonetheless, due to the lack of large scale high-quality training data in the highlight removal area, deep learning is still less explored. Although there is a recent method from Shi *et al.* [44], which proposes to use deep learning to decompose the image into diffuse albedo, shading and specular highlight components. However, the highlight removal problem is not addressed, as directly subtracting the decomposed specular highlight from the original image will cause severe color distortion.

In this paper, we propose to leverage deep neural networks to learn discriminative features in an end-to-end manner, for the single image highlight removal problem. Following the dichromatic reflection model [41], we formulate the highlight removal problem as a signal separation problem as: $I_d(x) = I(x) - I_s(x)$, where x is the pixel, $I_d(x)$ and $I_s(x)$ are diffuse reflection and specular reflection, respectively. Instead of directly learning the complex mapping between the input image $I(x)$ and the desired highlight-free image $I_d(x)$, we propose to first estimate the residual specular image $\hat{I_s}(x)$, and then obtain the highlight-removed $\hat{I_d}(x)$ based on the specular estimation. To this end, we construct a large-scale synthetic dataset consists of highlight/highlight-free image pairs via [16] to facilitate the learning. We also collect a real highlight dataset from ImageNet [5] for evaluation. We then propose a novel CDFF-Net that fully exploit the low-level features in a top-down aggregation manner, for accurately preserving the object appearances and image background while removing the highlights.

To summarize, this work has the following contributions:

- We propose a novel CDFF-Net that can learn discriminative features for highlight removal, from the proposed cumulative dense feature fusion strategy as well as the cascade learning scheme.
- We construct a large scale synthetic dataset of 11000 image pairs with/without highlights, which is proved to be able to generalize well on real world data.
- Extensive experiments show that the proposed method plays favorably against not only state-of-the-art highlight removal methods, but also the single image deraining methods.

2. Related Work

In this section, we briefly review previous works in specular highlight removal field, and one related work from intrinsic image decomposition field.

Multi-Image Based Methods. Since the specular highlights are direction-dependent, many previous works resort to use multiple images as input to impose additional constraints on this ill-posed highlight removal problem. Some works take multiple images with one specific scene from different point of views [27, 49, 48], while others [28, 40] use a series of images taken under the light source at different positions. Recently, Shah *et al.* [42] propose to leverage the feature correspondences across video frame sequences to remove the specular reflections.

While these methods use additional constraints from the multi-image inputs to help highlight removal, in this paper, we aim to address the more challenging single image highlight removal problem.

Single Image Based Methods. The pioneering work of Tan and Ikeuchi [47] proposes to first estimate pseudo specular-free images, and then iteratively remove the specular components via comparing the intensity logarithmic differentiation of the generated pseudo specular-free images and input images. Yang *et al.* [51, 50] extend this method and propose to use the bilateral filtering method for the comparisons of pseudo specular-free images and input images, to their goal of real-time inference. Since these pseudo specular-free images can significantly affect the final specular removal results, a lot of methods are then developed varying in the way of generating better pseudo specular-free images, based on the dark channel priors [21], intensity ratio based specular fraction computation [43], and ℓ_2 chromaticity [46].

Other assumptions/priors are also developed to alleviate the heavy reliance on the quality of pseudo specular-free images. Ren *et al.* [39] introduce the global color-lines constraints into the dichromatic reflection model and use this model for specular and diffuse reflection recovery. Li *et al.* [25] propose to leverage physical and statistical priors from specialized domain knowledge to remove the specular highlights in facial images. Akashi and Okatani [1] formulate the specular highlight removal problem as a sparse non-negative matrix factorization problem. Guo *et al.* [11] generalize the idea from [1] and propose a sparse and low-rank reflection model for highlight removal, which can be efficiently optimized by the augmented Lagrange multiplier method.

Nonetheless, as these methods require strict prior assumptions, real scenes with complex background and illumination conditions can easily fail these methods and cause the under/over highlight removal problem. On the contrary, we propose in this paper to address this problem by using the Cumulative Dense Feature Fusion network with a large scale training dataset, which can adaptively learn discriminative features for detecting and removing highlights.

Intrinsic Image Decomposition. Previous intrinsic image decomposition works [31, 32] do not consider the specular reflection while factorizing the input image into an albedo

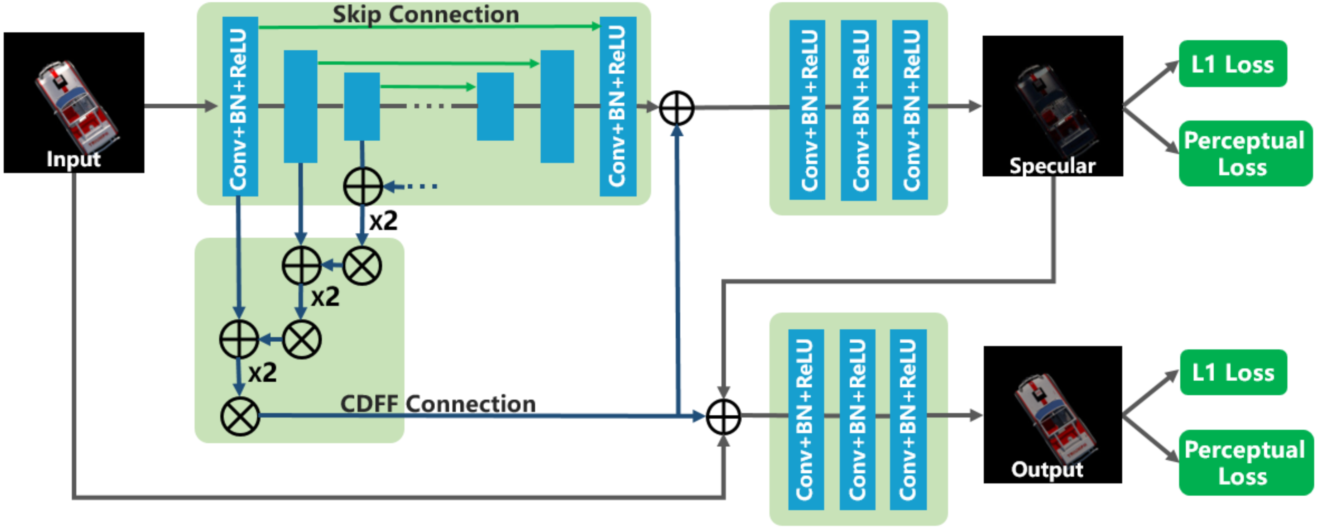


Figure 2: The architecture of our proposed CDFF-Net. Given an input image with highlights, it first predicts a specular layer via an encoder-decoder with skip connections (upper part). The estimated specular layer is then combined with the input image for producing the final specular-free image (bottom-right part). The cumulative dense feature fusion connections are introduced to the encoder-decoder to facilitate the feature propagation in a top-down manner for enriching the low-level features with high-level contextual information (bottom-left part).

image and a shading image. A recent work from Shi *et al*. [44] proposes to use deep network to decompose the images into three components, *i.e*., albedo, shading, and specular reflection.

However, directly applying their method by subtracting the decomposed specular reflection from the input is not reliable as it may cause severe color distortion. This is mainly due to the fact that they do not consider the reconstruction consistency after decomposition. On the contrary, we propose to learn the highlight/highlight-free mapping via a sub-stage of estimating the “residual highlight layer”, which helps us to preserve better object details and colors after removing highlights.

3. Proposed CDFF-Net

To address the over/under removal of highlights problem for images with complex backgrounds and illumination conditions, we propose a novel Cumulative Dense Feature Fusion network (CDFF-Net) to augment the low-level features with rich contextual information by recursively aggregating high-level features in a top-down manner. A cascaded learning strategy is further applied to effectively decouple the highlight/highlight-free mapping learning process into two inter-dependent stages, *i.e*., learning the specular layer and learning the highlight-free image reconstruction.

3.1. Overview

As shown in Fig. 2, given one normalized image as input, the proposed CDFF-Net first predicts a specular layer via an encoder-decoder with skip connections (upper part of Fig. 2). The estimated specular layer is then combined with the input image for producing the final specular-free image (bottom-right part of Fig. 2). The cumulative dense feature fusion connections are introduced into the encoder-decoder to facilitate the feature propagation in a top-down manner for enriching the low-level features with high-level contextual information (bottom-left part of Fig. 2).

3.2. Cumulative Dense Feature Fusion

While separating the specular reflection from the objects is extremely challenging, we observe that low-level features play an important role in recovering the object details during highlight removal process. However, as highlights can severely bury the local details, causing few meaningful features can be extracted by the shallower convolutional layers, we therefore propose the cumulative dense feature fusion (CDFF) strategy to further exploit the low-level features, which is a generalization of the dense connection idea.

Dense connection was first proposed in [14] for image classification. It creates short paths from early layers to later layers in a network. The idea is similar to ResNet [13]. Dense connections can greatly alleviate the vanishing gradient problem and strengthen the feature propagation. The original dense connection is implemented via Dense Block, which is essentially a different network architecture and can not be directly applied to our pre-trained encoder network. To exploit the ability of dense connection, one possible solution is first simply extracting the low-level features from different layers in the encoder, then upsampling them to the same size and finally concatenating them together. As shown in Fig. 3(a), we call this naive dense feature fusion. However, due to the pooling layers in the encoder, the feature map from the center layer of our network is 32 times smaller than the original image and simply upsampling the feature map by a large factor will produce a very coarse feature map, which is not efficient for feature fusion and sometimes will cause severe checkerboard artifacts. To avoid this problem, we propose the cumulative dense feature fusion (CDFF) strategy, as shown in Fig. 3(b).

Suppose an encoder network has multiple downsampling layers. We extract k features with different scales from the encoder after each downsampling layer, denoted as $\psi_i(x)$. Usually, the number of channels increases as the layer gets higher, hence directly concatenating all the features together will not only consume a lot of memories, but also make the high-level information dominate the fused feature. To address this problem, we use a simple convolution operation to reduce the channels of them to a fixed number. In this case, each feature $\psi_i(x)$ will contribute the same number of channels in the fused feature. This fusion strategy will implicitly give more weights to the low-level features since while the low-level features have fewer channels, the ratio of contributed channels to the original channels is larger than high-level features. If each $\psi_i(x)$ contributes ℓ channels, the total number of channels in the final fused feature is $k \times \ell$.

Let $Up(\cdot)$, $Cat(\cdot)$ and $Conv(\cdot)$ denote the upsampling, concatenation and convolution operations. The CDFF strategy can be formulated as follow:

$$\begin{aligned} \tilde{\psi_k}(x) &= Conv(Up(\psi_k(x))), \\ \tilde{\psi_i}(x) &= Conv(Up(Cat(\psi_i(x), \tilde{\psi}_{i+1}(x)))), \\ &\quad i = k-1, \ldots, 1. \end{aligned} \tag{1}$$

Here, the channel number of $\tilde{\psi_i}(x)$ is $(k+1-i) \times \ell$. Each $Up(\cdot)$ will upsample the i–th feature $\psi_i(x)$ to the same size of its previous feature $\psi_{i-1}(x)$. The final fused feature is $\tilde{\psi_1}(x)$ whose channel number is $k \times \ell$.

In our network, the encoder has 5 max-pooling layers ($k = 5$) and each of them will downsample the input feature by the factor of 2. Empirically, to save the memories required in training, we set the number of channels of the fused feature as 60 and each $\psi_i(x)$ will contribute 12 channels. The illustration is shown in Fig. 3(b).

We use the fused feature in two parts. Let's denote the output feature of the encoder-decoder as $F(x)$. $f_{specular}(\cdot)$ represents the *Conv + BN + ReLU* layers used for specular reflection image generation and $f_{removal}(\cdot)$ represents the *Conv + BN + ReLU* layers used for specular-free image generation. By using CDFF, the prediction of the specular reflection layer and the specular-free image can be written like this:

$$\begin{aligned} \hat{I_s}(x) &= f_{specular}(F(x), \tilde{\psi_1}(x)), \\ \hat{I_d}(x) &= f_{removal}(I(x) - \hat{I_s}(x), \tilde{\psi_1}(x)). \end{aligned} \tag{2}$$

3.3. Two-stage Residual Learning for Specular Highlight Removal

Directly learning to reconstruct the highlight-free images with recovered object details and colors is still challenging for deep neural networks, as the highlight intensities are very difficult to be predicted. To address this problem, we adopt the deep residual learning strategy to first explicitly estimate the highlight intensities via predicting a specular reflection layer, and then obtain a specular-free image based on the predicted specular reflection. This strategy allows us to explicitly learn to detect the highlights while preserving the global properties of background objects (*i.e.*, colors and structures), before reconstructing the specular-free images. It also allows additional supervision to be imposed to facilitate the learning process of our network.

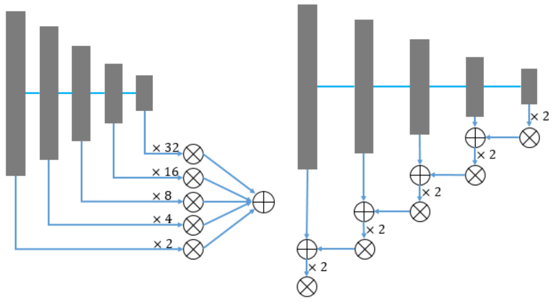
Formally, we first learn a mapping $f(\theta_s)$, from the input image $I(x)$ to the specular reflection image $I_s(x) = I(x) - I_d(x)$. The predicted specular-free image $\hat{I_d}(x)$ is then obtained via learning a mapping $f(\theta_d)$, based on $I(x)$ and the predicted specular reflection image $\hat{I_s}(x)$, where $\hat{I_d}(x) = f(I(x), \hat{I_s}(x); \theta_d)$. θ_s and θ_d are the parameters in the corresponding mappings.

For training our network, we use the ℓ_1 loss instead of the commonly used ℓ_2 loss for producing sharpener images:

$$\begin{aligned} L_{img} = &\frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|I_s(x)_{ij} - \hat{I_s}(x)_{ij}\|_1 \\ &+ \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \|I_d(x)_{ij} - \hat{I_d}(x)_{ij}\|_1, \end{aligned} \tag{3}$$

where, $\hat{I_s}(x) = f(I(x); \theta_s)$ and $\hat{I_d}(x) = f(I(x), \hat{I_s}(x); \theta_d)$.

Additionally, to train the network to be more sensitive to the differences between the predicted images and the ground-truth images in the semantical level, we use two feature losses to compare the images in feature space, also known as perceptual losses [18, 4, 8, 24]. Specifically, we obtain the features by feeding the predicted images and the ground-truth images to a VGG-16 [45] network ϕ pre-trained on the ImageNet dataset [5], and extract the features from the first max-pooling layer. The feature losses are computed via the ℓ_1 difference as:



(a) (b)

Figure 3: Illustration of the dense feature connection. (a): A naive dense feature fusion. (b): Proposed cumulative dense feature fusion (CDFF). $\otimes$ denotes upsampling and convolution. $\oplus$ denotes concatenation.

$$\begin{aligned} L_{feat} = & \frac{1}{N}\sum_{i=1}^{N}\|\phi(I_s(x))_i - \phi(\hat{I_s}(x))_i\|_1 \\ & + \frac{1}{N}\sum_{i=1}^{N}\|\phi(I_d(x))_i - \phi(\hat{I_d}(x))_i\|_1. \end{aligned} \tag{4}$$

Setting the total parameters as $\Theta = (\theta_s, \theta_d)$, the objective of this network is to find the best parameter Θ that minimizes the loss function:

$$\Theta^* = \arg\min_{\Theta} L_{img} + L_{feat}. \tag{5}$$

3.4. Implementation and Training

We implement our network with PyTorch [33]. The structure of the encoder-decoder network is similar to SegNet [2], which was proposed for semantic segmentation. We use a pre-trained VGG-16 [45] as the encoder. The convolutional layers in the rest parts are initialized with *kaiming_normal* [12]. We train the network for 60 epochs with batch size 6 on an Nvidia GTX 1080Ti GPU. The weights are optimized using the Adam optimizer[22] with weight decay as 0.0005. The initial learning rate is 1e-4 and decreased by a factor of 10 every 20 epochs.

4. Dataset

4.1. Synthetic Data

Most of the traditional methods for specular highlight removal are based on color space analysis or matrix optimization. These methods do not require a large training dataset. To the best of our knowledge, there is no public large-scale dataset for single image specular highlight removal. While [44] creates a large synthetic dataset for intrinsic image decomposition which contains specular components, currently the dataset is not available.

To create synthetic images with specular highlights, we use more than 10,000 3D objects with albedo texture from ShapeNet, a richly-annotated, large-scale dataset of 3D shapes [3]. Similar to the rendering method in [44], we use the modified Phong reflectance model [23, 34] to generate the specular reflection and diffuse reflection for the 3D objects. Each object is rendered with random view point and different highlight intensity. The renderer we use is Mitsuba[16], a research-oriented rendering system in the style of PBRT.

Different from the method in [44], which renders the specular reflection and diffuse reflection separately and then uses ImageMagick [15] to synthesize the specular highlight image, we render a specular highlight image by rendering the specular and diffuse reflections together. Besides, to make the rendered image looks more like a real image, we set the emitter as sun and sky. Some synthetic images are shown in Fig. 5. While the images are rendered without backgrounds, extensive experiments show that our network trained with these data can handle the background highlights properly.

4.2. Real Data

To make comprehensive analyses of our method and test its generalization performance, we also create a small dataset containing images in the wild for qualitative comparison. These images are collected from ImageNet [5]. We select the images that contain objects with notable specular highlights. Due to the diversity of ImageNet dataset, the collected images have various backgrounds and complex illuminations.

5. Experiments

To verify the effectiveness and robustness of our method, we evaluate it on both synthetic dataset and collected real images. We compare our method with some previous works: Ren *et al.* [39], Shen and Zheng [43], Tan *et al.* [47] and Guo *et al.* [11]. The similarity evaluation metrics are PSNR and SSIM. We also compute the RMSE as error metric. The comparison among different parts of our network are also presented in ablation study.

5.1. Results on Synthetic Dataset

We compare quantitative and qualitative performance of different methods on the synthetic dataset. Table 1 shows the quantitative comparisons between our method and the previous works. It can be clearly observed that the proposed CDFF-Net outperforms the previous methods greatly.

To visually demonstrate the effectiveness of our method, results on some sample images are presented in Fig. 4. Compared with other methods, the proposed CDFF-Net can remove the specular highlights effectively without causing severe color distortions.

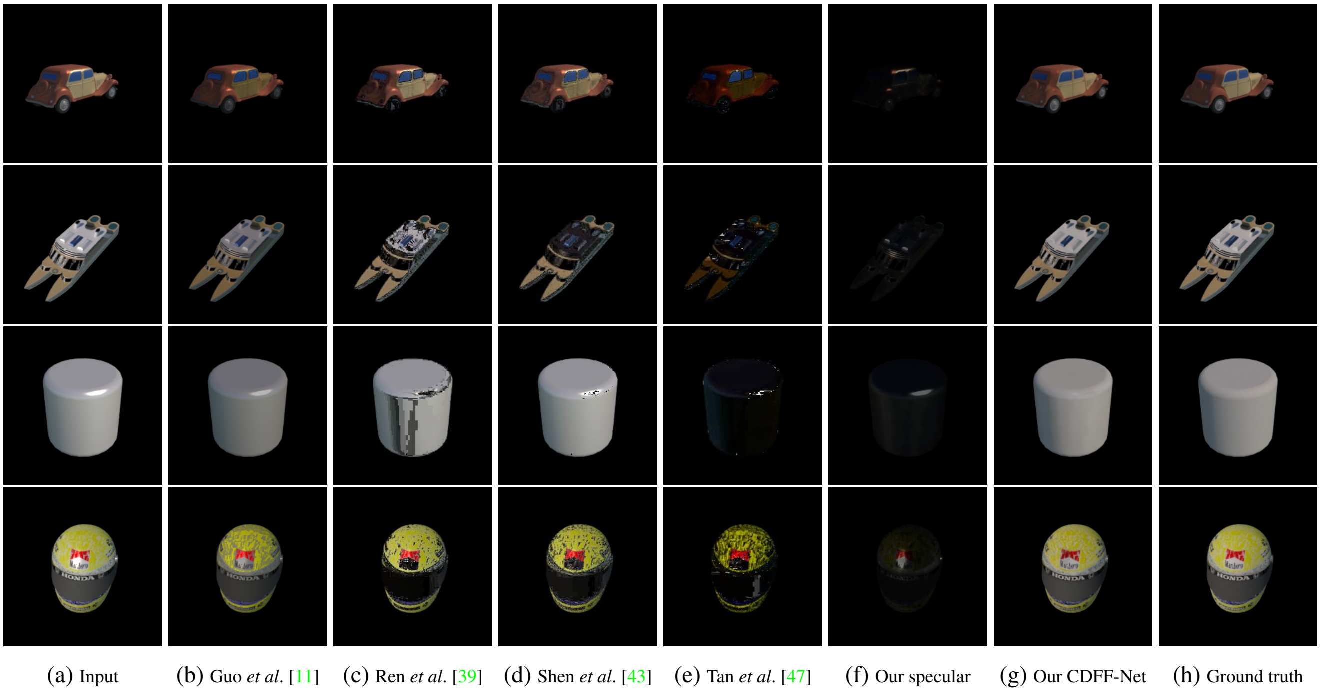


Figure 4: Results of comparing our method with state-of-the-art methods on the synthetic dataset. From left to right: input, Guo *et al.* [11], Ren *et al.* [39], Shen and Zheng [43], Tan *et al.* [47], generated specular by our method, our output, ground-truth. Our method can remove the specular highlights without changing the original color greatly.

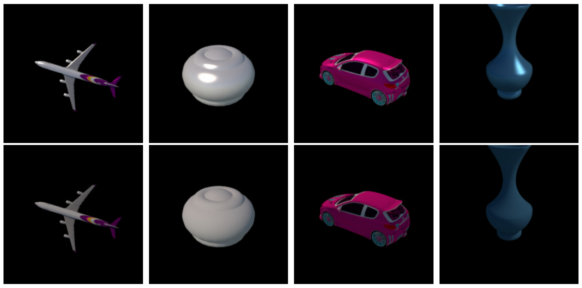


Figure 5: Samples of our synthetic dataset. Top row: Images with specular highlights. Bottom row: Corresponding ground-truth images without specular highlights.

Method	PSNR ↗	SSIM ↗	RMSE ↘
Tan *et al.* [47]	21.578	0.865	3.547
Shen and Zheng [43]	27.977	0.935	2.982
Ren *et al.* [39]	28.925	0.939	3.060
Guo *et al.* [11]	29.743	0.971	3.213
Ours	**40.673**	**0.990**	**2.112**

Table 1: Quantitative results on the synthetic dataset.

5.2. Results on Real-World Images

The performance of the proposed method is also evaluated on the real-world images collected from ImageNet. Some results are shown in Fig. 6.

The comparison demonstrates that our method achieves favorable performance over previous methods. [39], [43] and [47] are very sensitive to the backgrounds and color deviations. [11] is robust but our method can generate images with more coherent contents and fewer noises.

5.3. Ablation Study

The first ablation study is conducted to demonstrate the improvements obtained by the CDFF strategy compared with the naive dense feature fusion (NDFF). The two models are trained and tested with the same configuration on the same dataset. Results are shown in Table 2. It can be observed that CDFF strategy has better performance than NDFF.

Method	PSNR ↗	SSIM ↗	RMSE ↘
w/ CDFF	40.673	0.990	2.112
w/ NDFF	40.405	0.990	2.164

Table 2: Comparisons between CDFF and NDFF.

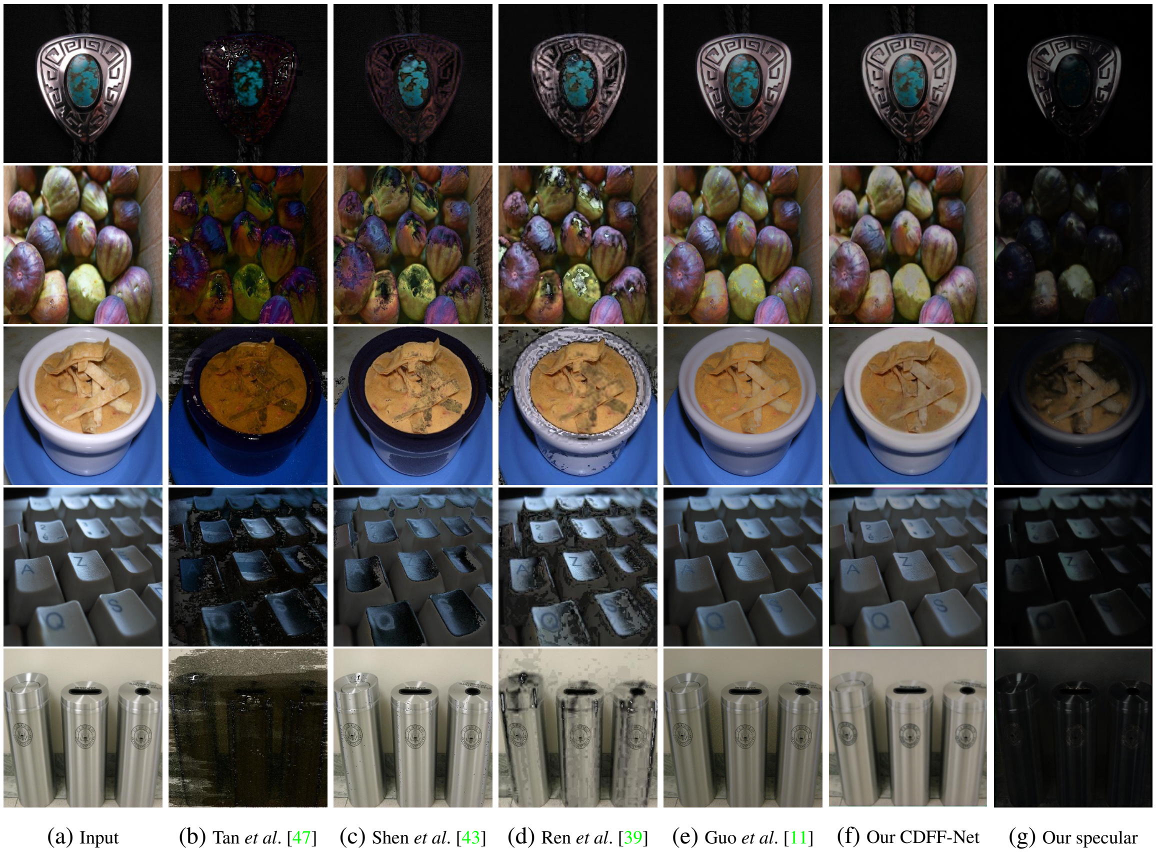


Figure 6: Results of comparing the methods on the collected real images. From left to right: input, Tan *et al.* [47], Shen and Zheng [43], Ren *et al.* [39], Guo *et al.* [11], our output, generated specular by our method. Results show that our methods can dealing with complex backgrounds and illuminations more effectively.

In the second ablation study, we demonstrate the effectiveness of different parts in our network. Specifically, we evaluate the effectiveness of the perceptual loss and CDFF strategy. The results are shown in Table. 3. Visual comparisons on two real images are shown in Fig. 7.

From the comparison we can see that without CDFF, the qualities of the generated specular-free images degrade greatly and there will be severe color distortions.

Method	PSNR ↗	SSIM ↗	RMSE ↘
w/o CDFF	38.174	0.989	2.754
w/o L_{feat}	39.568	0.988	2.307
Ours complete	40.673	0.990	2.112

Table 3: Quantitative comparisons on synthetic images among multiple ablated models of our method. “w/o CDFF” denotes our method trained without CDFF strategy. “w/o L_{feat}” denotes our method trained without perceptual loss.

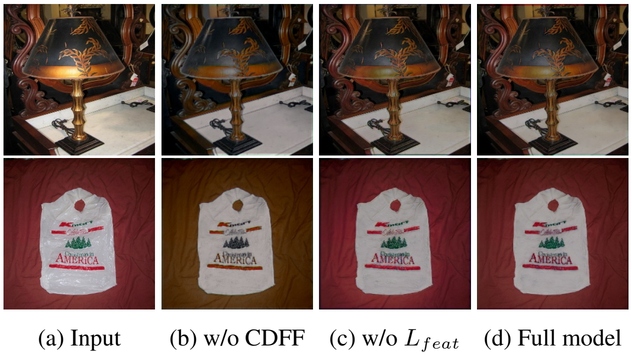


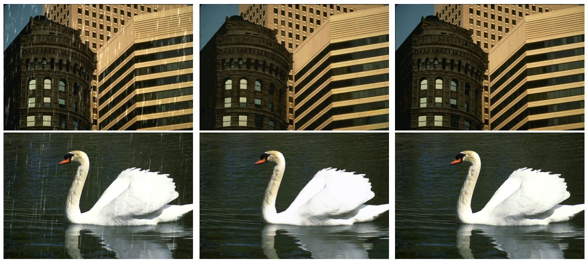
Figure 7: Visual comparisons on two real images among multiple ablated models of our method. “w/o CDFF” denotes our method trained without CDFF strategy. “w/o L_{feat}” denotes our method trained without perceptual loss.

5.4. Extensions

To demonstrate the generality of our approach, we also train our network for de-rain. Typically, the rain-streak in a rainy image is modeled as an additive residual component. Our network can be directly applied to this task. We train and test our network on the *Rain100L* dataset created by [52]. We compare our method with several state-of-the-art methods mentioned in [52]. The quantitative results are shown in Table 4. From the table we can see that our method can achieve comparable results compared with the state-of-the-art methods. Qualitative results on *Rain100L* and real images are shown in Fig. 8 and Fig. 9 respectively.

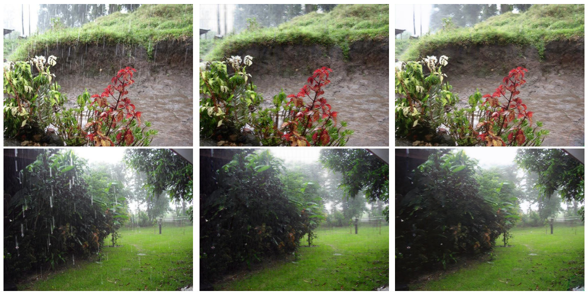
Method	PSNR ↗	SSIM ↗
ID [19]	23.13	0.70
DSC [30]	24.16	0.87
LP [26]	29.11	0.88
CNN [6]	23.70	0.81
SRCNN[17]	32.63	0.94
JORDER[52]	36.11	0.97
Our network	31.76	0.96

Table 4: Quantitative comparison on *Rain100L* dataset. The results of the state-of-the-art methods are reported in [52].



(a) Rain image (b) Prediction (c) Ground-truth

Figure 8: Results of our method on *Rain100L*. Left: rain image. Middle: output image. Right: ground-truth image.



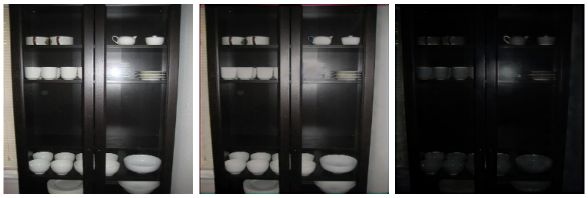
(a) Rain image (b) Yang *et al.* [52] (c) Our CDFF-Net

Figure 9: Results on real rain images. Left: rain images. Middle: results of [52]. Right: our results.

6. Conclusion and Future Work

In this paper, we have presented a novel deep learning method for the single image highlight removal problem. Specifically, we have constructed a large scale synthetic dataset of 11000 pairs of images with and without highlights, for training and testing. It will be made publicly available for future research. We have also collected a real dataset that contains images of highlights for evaluation. We have designed a novel cumulative dense feature fusion (CDFF) network that effectively refines low-level features from the higher layers in a top-down manner for producing high-quality highlight-free images. Moreover, a two-stage residual learning strategy is applied to explicitly decouple the highlight removal problem into two interdependent stages, *i.e.*, highlight detection and highlight removal. Extensive evaluations on both synthetic and real data demonstrate the superiority of the proposed method against the state-of-the-art highlight removal methods. We also show the potential of the proposed network by applying it to the single image deraining task.

Our method does have limitation. As illustrated in Fig. 10, our method may fail in scenes that the highlight region is large and overexposed, as our method can not recover the original object colors after removing the specular. A possible solution may be to incorporate the generative methods, such as GAN [10], for generating visually pleasing background.



(a) Input (b) Our CDFF-Net (c) Our specular

Figure 10: A challenging case with overexposed highlight regions. Our method can remove most of the specular highlights, but can not fully recover the original color of this region.

References

[1] Y. Akashi and T. Okatani. Separation of reflection components by sparse non-negative matrix factorization. In *ACCV*, pages 611–625. Springer, 2014. 1, 2

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 39(12):2481–2495, 2017. 5

[3] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv:1512.03012*, 2015. 5

[4] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, pages 1511–1520, 2017. 4

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 4, 5

[6] D. Eigen, D. Krishnan, and R. Fergus. Restoring an image taken through a window covered with dirt or rain. In *ICCV*, pages 633–640, 2013. 8

[7] D. Engin, A. Genc, and H. Kemal Ekenel. Cycle-dehaze: Enhanced cyclegan for single image dehazing. In *CVPR Workshops*, pages 825–833, 2018. 2

[8] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423, 2016. 4

[9] R. Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 2, 8

[11] J. Guo, Z. Zhou, and L. Wang. Single image highlight removal with a sparse and low-rank reflection model. In *ECCV*, pages 268–283, 2018. 1, 2, 5, 6, 7

[12] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 5

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 3

[15] L. ImageMagick Studio. Imagemagick, 2008. 5

[16] W. Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 2, 5

[17] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM Multimedia*, pages 675–678. ACM, 2014. 8

[18] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 4

[19] L.-W. Kang, C.-W. Lin, and Y.-H. Fu. Automatic single-image-based rain streaks removal via image decomposition. *IEEE TIP*, 21(4):1742–1755, 2012. 8

[20] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri. Automatic shadow detection and removal from a single image. *IEEE TPAMI*, 38(3):431–446, 2016. 2

[21] H. Kim, H. Jin, S. Hadap, and I. Kweon. Specular reflection separation using dark channel prior. In *CVPR*, pages 1460–1467, 2013. 1, 2

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5

[23] E. P. Lafortune and Y. D. Willems. Using the modified phong reflectance model for physically based rendering. 1994. 5

[24] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *CVPR*, pages 4681–4690, 2017. 4

[25] C. Li, S. Lin, K. Zhou, and K. Ikeuchi. Specular highlight removal in facial images. In *CVPR*, pages 3107–3116, 2017. 1, 2

[26] Y. Li, R. T. Tan, X. Guo, J. Lu, and M. S. Brown. Rain streak removal using layer priors. In *CVPR*, pages 2736–2744, 2016. 8

[27] S. Lin, Y. Li, S. B. Kang, X. Tong, and H.-Y. Shum. Diffuse-specular separation and depth recovery from image sequences. In *ECCV*, pages 210–224. Springer, 2002. 1, 2

[28] S. Lin and H.-Y. Shum. Separation of diffuse and specular reflection in color images. In *CVPR*, 2001. 1, 2

[29] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 2

[30] Y. Luo, Y. Xu, and H. Ji. Removing rain from a single image via discriminative sparse coding. In *ICCV*, pages 3397–3405, 2015. 8

[31] T. Narihira, M. Maire, and S. X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, pages 2992–2992, 2015. 2

[32] T. Narihira, M. Maire, and S. X. Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, pages 2965–2973, 2015. 2

[33] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 5

[34] B. T. Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 5

[35] R. Qian, R. T. Tan, W. Yang, J. Su, and J. Liu. Attentive generative adversarial network for raindrop removal from a single image. In *CVPR*, pages 2482–2491, 2018. 2

[36] L. Qu, J. Tian, S. He, Y. Tang, and R. W. Lau. Deshadownet: A multi-context embedding deep network for shadow removal. In *CVPR*, pages 4067–4075, 2017. 2

[37] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015. 2

[38] W. Ren, S. Liu, H. Zhang, J. Pan, X. Cao, and M.-H. Yang. Single image dehazing via multi-scale convolutional neural networks. In *ECCV*, pages 154–169. Springer, 2016. 2

[39] W. Ren, J. Tian, and Y. Tang. Specular reflection separation with color-lines constraint. *IEEE TIP*, 26(5):2327–2337, 2017. 1, 2, 5, 6, 7

[40] Y. Sato and K. Ikeuchi. Temporal-color space analysis of reflection. *JOSA A*, 11(11):2990–3002, 1994. 1, 2

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

[42] S. M. A. Shah, S. Marshall, and P. Murray. Removal of specular reflections from image sequences using feature correspondences. *Machine Vision and Applications*, 28(3-4):409–420, 2017. 1, 2

[43] H.-L. Shen and Z.-H. Zheng. Real-time highlight removal using intensity ratio. *Applied optics*, 52(19):4483–4493, 2013. 1, 2, 5, 6, 7

[44] J. Shi, Y. Dong, H. Su, and S. X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, pages 1685–1694, 2017. 2, 3, 5

[45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 4, 5

[46] J. Suo, D. An, X. Ji, H. Wang, and Q. Dai. Fast and high quality highlight removal from a single image. *IEEE TIP*, 25(11):5441–5454, 2016. 1, 2

[47] R. T. Tan and K. Ikeuchi. Separating reflection components of textured surfaces using a single image. *IEEE TPAMI*, 27(2):179–193, 2005. 1, 2, 5, 6, 7

[48] C. Wang, S.-i. Kamata, and L. Ma. A fast multi-view based specular removal approach for pill extraction. In *ICIP*, pages 4126–4130. IEEE, 2013. 1, 2

[49] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV*, volume 2, pages 68–75. IEEE, 2001. 1, 2

[50] Q. Yang, J. Tang, and N. Ahuja. Efficient and robust specular highlight removal. *IEEE TPAMI*, 37(6):1304–1311, 2015. 1, 2

[51] Q. Yang, S. Wang, and N. Ahuja. Real-time specular highlight removal using bilateral filtering. In *ECCV*, pages 87–100. Springer, 2010. 1, 2

[52] W. Yang, R. T. Tan, J. Feng, J. Liu, Z. Guo, and S. Yan. Deep joint rain detection and removal from a single image. In *CVPR*, pages 1357–1366, 2017. 2, 8