Hue Guidance Network for Single Image Reflection Removal

Yurui Zhu, *Student Member, IEEE*, Xueyang Fu, *Member, IEEE*, Zheyu Zhang, Aiping Liu, *Member, IEEE*, Zhiwei Xiong, *Member, IEEE*, and Zheng-Jun Zha, *Member, IEEE*

***Abstract*— Reflection from glasses is ubiquitous in daily life, but it is usually undesirable in photographs. To remove these unwanted noises, existing methods utilize either correlative auxiliary information or handcrafted priors to constrain this ill-posed problem. However, due to their limited capability to describe the properties of reflections, these methods are unable to handle strong and complex reflection scenes. In this article, we propose a hue guidance network (HGNet) with two branches for single image reflection removal (SIRR) by integrating image information and corresponding hue information. The complementarity between image information and hue information has not been noticed. The key to this idea is that we found that hue information can describe reflections well and thus can be used as a superior constraint for the specific SIRR task. Accordingly, the first branch extracts the salient reflection features by directly estimating the hue map. The second branch leverages these effective features, which can help locate salient reflection regions to obtain a high-quality restored image. Furthermore, we design a new cyclic hue loss to provide a more accurate optimization direction for the network training. Experiments substantiate the superiority of our network, especially its excellent generalization ability to various reflection scenes, as compared with state-of-the-arts both qualitatively and quantitatively. Source codes are available at https://github.com/zhuyr97/HGRR**

***Index Terms*— Deep learning, hue guidance, reflection removal.**

I. Introduction

TRANSPARENT glass, e.g., windows and glass doors, is very common in daily life. When people shoot scenes behind glass, the captured image usually contains an undesirable reflection phenomenon, which degrades the image quality due to distortion, occlusion, or blurring of the background scenes. Therefore, single image reflection removal (SIRR), which aims to restore the clean background scene by removing the reflections, is of great practical significance for both visual perception and downstream vision systems.

Since the properties of the background and reflections are similar, it is hard to accurately model the generation process of glass reflection. The widely used model for SIRR is that the reflection-distorted image $\mathbf{I}$ comes from the linear combination of the background or transmission layer $\mathbf{T}$ and the reflection layer $\mathbf{R}$, as $\mathbf{I} = \alpha * \mathbf{T} + \beta * \mathbf{R}$. Obviously, SIRR is an ill-posed problem since the number of unknowns is twice the number of equations. To handle this challenging task, existing methods introduce additional auxiliary information and priors, e.g., handcrafted assistance, ghosting effect, absorption effect, and gradient sparsity prior, to constrain this problem [1], [2], [3], [4], [5], [6]. For instance, many methods use multiview images or user annotations to introduce new information. Recently, deep learning-based methods have achieved great success in various low-level vision tasks, such as image de-noising [7], [8], image de-raining [9], [10], and general image processing [11], [12], [13], [14]. For SIRR, CEILNet [15] designs a two-stage network to address this layer separation task with the aid of the edges information. Bidirectional network (BDN) [16], and iterative boost convolutional LSTM network (IBCLN) [17] construct cascaded deep networks, in which the previously generated results can be served as auxiliary information for the next sub-network. Zhang et al. [18] and Lei and Chen [5] further utilize the reflection-free flash-only cues to help remove reflection in specific scenes. Although the aforementioned methods perform well in certain scenarios, there is still much room for SIRR improvement. The reason lies in the complex and changeable reflection scenes and the limitations of auxiliary information designed by these methods. This causes the existing methods could not to be well generalized to various reflection scenes. Fig. 1 shows examples of our approach compared with two recent methods.

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

Fig. 1. Visual comparisons with two recent methods on the real-world reflection scenes. Note that both the strong reflection (top row) and color distortion (bottom row) can be effectively addressed by using our approach. (a) Inputs. (b) ICBLN [44]. (c) YTMT [17]. (d) Ours.

By following the previous research direction, i.e., exploring potential auxiliary information for SIRR, we rethink the properties of glass reflection. According to the Retinex theory

Manuscript received 14 February 2022; revised 12 October 2022 and 5 April 2023; accepted 19 April 2023. Date of publication 23 May 2023; date of current version 8 October 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0105702; and in part by the National Natural Science Foundation of China (NSFC) under Grant 62225207, Grant U19B2038, Grant 62121002, and Grant 62276243. *(Corresponding author: Xueyang Fu.)*

The authors are with the University of Science and Technology of China, Hefei 230022, China (e-mail: zyr@mail.ustc.edu.cn; xyfu@ustc.edu.cn; Zhangzy0@mail.ustc.edu.cn; aipingl@ustc.edu.cn; zwxiong@ustc.edu.cn; zhazj@ustc.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2023.3270938

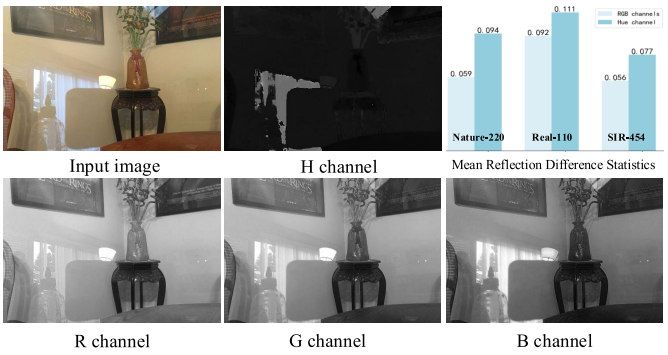


Fig. 2. Visual examples and statistics about channels of the reflection-distorted image in both RGB and Hue channels.

[19], the captured image is a combination of light and object reflectance, which also determines the color information in the image. Due to the reflection effect of the glass surface, the light reflected by the background object and the light reflected by the glass surface will overlap and occlude each other during the shooting. This mixing process significantly affects the color information of both reflected objects and the background scene, which often leads to obvious color contrast between the reflection-distorted regions and the surrounding regions. Fig. 2 shows a visual example of real-world reflected scenes, and it is clear that reflection-distorted regions are distinctive in the corresponding hue channel map. Moreover, the hue channel contains fewer object details and textures, and it is more likely to help locate the reflection regions. This observation inspired us to utilize color information, which has not been noticed by previous methods, as new useful information to solve SIRR.

In this article, we make an attempt to use the hue map in the HSV color space as a superior constraint and auxiliary information to improve the performance of reflection removal. We found that the hue map of the reflected scene can highlight reflection disturbance, which helps locate the salient reflection regions. In addition, the hue map can be used directly as a color constraint to restore the color of the transmission layer. Accordingly, we design a new hue guidance network (HGNet) with two branches based on the hue information. One branch aims to extract effective hue-related features, while the other one focuses on utilizing these features for transmission restoration. Moreover, we further delicately design a cyclic hue loss to provide a more accurate optimization direction for our network training. Additionally, we propose a lightweight global feature module (LGFM) to improve reflection removal performance, which could capture the long-range dependencies in the feature space at a relatively low computation cost. Without bells and whistles, our method can achieve good de-reflection results. Our contributions are summarized as follows.

1) We propose a novel HGNet architecture for SIRR. The network is composed of two branches to fully explore and exploit the complementarity between image information and hue information for the SIRR process.
2) We propose a new cyclic hue loss that emphasizes the saliency of the reflections to be estimated in the hue map. In addition, this carefully designed loss can be directly applied to previous methods to help them further improve de-reflection.
3) We further propose a LGFM. It is capable of indirectly capturing long-range dependencies of features at a relatively low computation and memory cost to promote our network performance.

II. Related Works

Depending on the number of input images, the existing reflection removal methods can be categorized into two categories: multiple image-based and single image-based ones. For multiple image-based methods, reflections can be more easily removed using inter-image information [6], [20], [21], [22], [23]. For instance, Schechner et al. [24], Farid and Adelson [25] and Kong et al. [26] utilize the physical polarization to obtain multiple images with different polarization angles, which can offer independent information for reflection removal. In this article, we instead focus on removing reflection from a single image, which is significantly more challenging since much less information is available for detecting and removing reflection.

Early methods focus on exploring prior knowledge to constrain this ill-posed problem. The most widely used strategy is to utilize the gradient sparsity prior [1]. For instance, in method [27], Arvanitopoulos et al. impose a Laplacian data fidelity term and ℓ_0 gradient sparsity to suppress the reflection effect. Yang et al. [28] propose a convex model to remove reflection based on a partial differential equation with a gradient threshold. Li and Brown [29] extract two layers from one image by assuming that the reflection layer is smoother than the transmission layer. Wan et al. [30] observe that strong reflections typically only dominate a limited portion of the entire image and develop the region-aware solution to heterogeneously tackle regions with and without reflections. Albeit achieving good performance in certain scenarios, these prior-based methods rely on handcrafted prior assumptions. When the reflected scene violates the assumptions, these methods fail to correctly estimate and remove the reflection.

Recently, deep learning-based SIRR methods were proposed by designing different network architectures and loss functions. CEILNet [15] utilizes edge maps to guide the transmission layer inference network. BDN [16] firstly estimates the reflection layer and then utilizes it to infer background information. Wan et al. [31] design a multiscale-guided learning network and apply gradient features to guide the transmission inference process. In addition to using only low-level auxiliary information, many researchers [18] also introduce high-level semantic information, i.e., perceptual features [32] and object segmentation maps to help the network training. Gandelsman et al. [33] combine multiple DIPs [34] to decompose images into different components in an unsupervised manner. Wieschollek et al. [35] generate a polarization-based dataset and leverage the properties of polarized light to separate layers. Lei et al. [36] build a perfectly aligned polarized-based reflection dataset and propose a two-stage solution to perform polarized reflection removal. Hong et al. [37] focus on solving the problem of panoramic image reflection removal, which

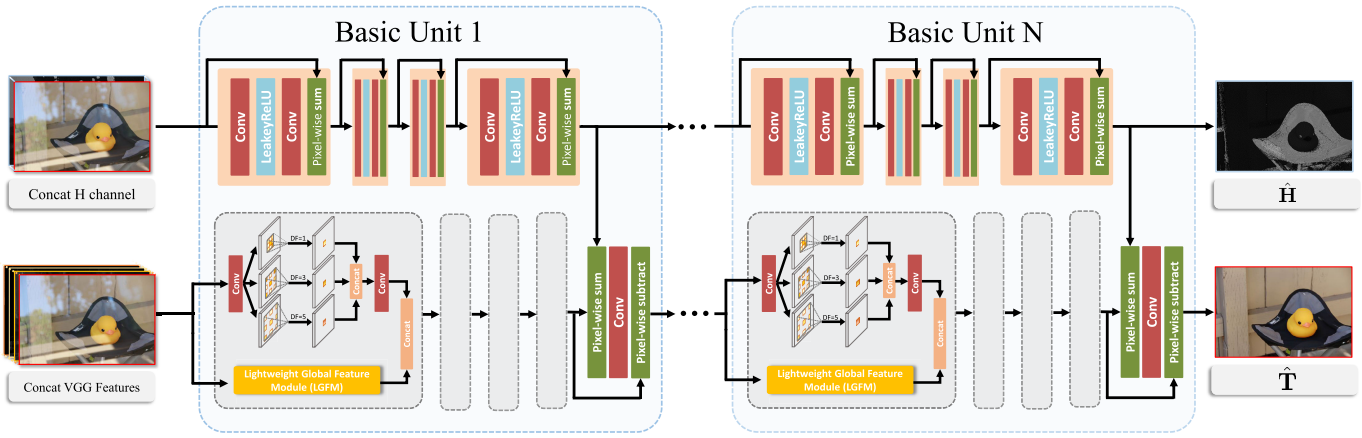


Fig. 3. Our proposed HGNet consists of two branches to remove reflection by integrating image information and corresponding hue information. The image inference branch is closely guided by the associated hue features from the hue inference branch.

exploits the geometric alignment to utilize the reflection cues contained in the panoramic images. IBCLN [17] proposes an iterative strategy to refine the predicted transmission and reflection in a cascaded fashion. Shi et al. [38] adopt the LSTM unit to facilitate the network training. Your trash is my treasure (YTMT) [39] exploits negative activation function techniques to utilize mutual information from the reflection layer and transmission layer. Dong et al. [40] leverage multiscale Laplacian intermediate features to regress the reflection location to achieve reflection removal.

To narrow the gap between synthetic data and real-world scenarios, many researchers focus on how to obtain realistic data. Based on the entanglement and disentanglement mechanisms, Ma et al. [41] design independent reflection generation and separation stages in a weakly-supervised learning framework. Wen et al. [42] predict a nonlinear alpha blending mask to obtain controllable and diverse reflection scenes. Wei et al. [43] collect a large number of easily accessible nonaligned reflection image pairs and define invariant loss to train the deep network. Kim et al. [44] employ physically-based rendering techniques to obtain realistic reflection image pairs.

III. Methodology

A. Motivation

Since our method is mainly derived from observing existing reflection images, here we further analyze the effect of distortion caused by the reflection phenomenon. First, the reflective layer usually causes regional occlusion or overlap, which leads to obvious **color contrast** between the reflection-distorted regions and the surrounding reflection-free regions. Second, when shooting the object behind glass, we find that the object of interest contains **color degradation**. This is caused by the aforementioned occlusion or overlap, including the loss and scattering of light generated by the glass itself. Hence we argue that reflection significantly affects the color information of the transmission layer. The above analysis of reflection phenomena inspired us to utilize color cues to guide the reflection removal and transmission layer restoration.

To fully utilize the color cue in the specific reflection removal task, here we choose the HSV color space. As a widely used color space, the hue, saturation, and value components are orthogonal to each other, with low correlations in the HSV color space. While in the red green blue (RGB) color space, only three channels work together to determine a certain color, which is far less convenient for representing color information than in HSV color space. Furthermore, we compare the direct mean difference between the RGB channels and hue channels on three public real reflection datasets containing reflection/clear image pairs. As shown in Fig. 2, through calculation and statistics of the difference between the reflection image and the clear image in different color spaces, we find that the average difference values obtained in the hue channel are consistently larger than that in the RGB channels. A larger difference value can better emphasize and highlight the reflection areas, which helps the network to focus on these areas during the training process. Therefore, we intentionally utilize the hue information to guide the construction and training of our network.

The advantages we choose the hue map are threefold: First, the hue is orthogonal to other components so that the color information related to reflection regions can be well separated into the hue map. Second, since the hue map contains fewer object details and textures, it is more suitable for locating the reflection regions. While other color channels, as shown in Fig. 2, contain both reflection and background with rich details, which will increase the difficulty of distinguishing between reflection and object. Last but not least, utilizing hue information can also help restore color distortion, which is a vital aspect that affects human visual perception.

B. Hue Guidance Network

For SIRR, given a reflection-distorted image $\mathbf{I} \in [0, 1]^{3\times W\times H}$, we aim at suppressing and removing unwanted glass effects $\mathbf{R}$, i.e., reflection artifacts and color degradation, to obtain a clean transmission layer $\mathbf{T}$. To achieve this goal, we design two network branches to separately predict the hue map and transmission layer, respectively. The network architecture is presented in Fig. 3.

Specifically, the first branch takes the reflected RGB image $\mathbf{I}$ and the corresponding hue map as inputs and aims to predict the clean hue maps. The second branch receives the concatenation of $\mathbf{I}$ with its hypercolumn features [45] and [18] extracted

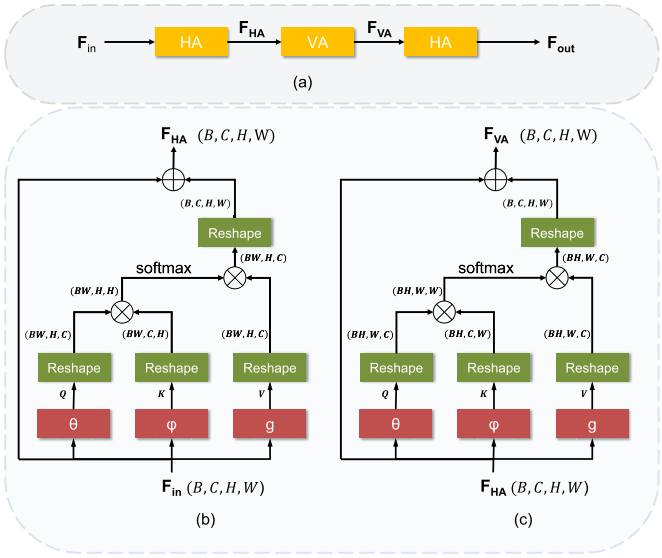


Fig. 4. Visual illustration of the LGFM. (a) LGFM. (b) HA. (c) VA.

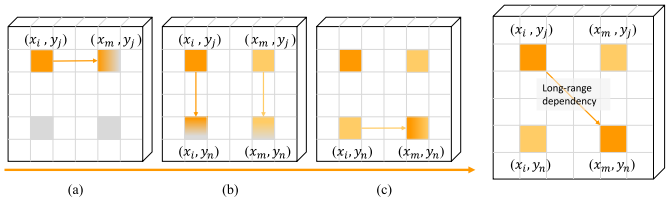


Fig. 5. Simplified illustration of the indirect long-range dependencies construction process between two different position pixels (not in the same horizontal row and vertical column). (a)–(c) Processes of passing point (x_i, y_j) information to point (x_m, y_n) via HA, VA, and HA respectively.

from the pretrained VGG-19 [32] model, which proved to be a very effective augmentation strategy. As described in [43] and [18], delivering the high-level vision features to the de-reflection network could help to boost the semantic understanding capacity of our network.

To achieve a better reflection removal performance, we further introduce dilated convolutions [46] to obtain multiscale representations, which are essential for various vision tasks. Specifically, for the first hue inference branch, we construct a basic feature extraction unit by sequentially deploying dilated Resblocks [47], in which the vanilla convolution operations are replaced with dilated convolutions. Since the hue map contains salient reflection features with fewer image details, using this sequential dilated structure can quickly improve multiscale representation capacity in a layer-wise manner, which is beneficial for locating large reflection regions in space. For the second transmission branch, the basic unit contains two parts: the multiscale local feature module (MSLFM) and the LGFM. MSLFM is derived from [48], which is composed of parallel dilated convolution layers and aims to improve the multiscale representation ability at a more granular level. While MSLFM mainly focuses on the locality contextual information due to the limitations of convolution operations, we further introduce nonlocal feature augmentation to boost the de-reflection performance of our network. However, the pioneering nonlocal feature aggregation manner [49] usually consumes expensive computation and memory, especially for the input of large-size images.

In this way, we propose the LGFM to capture the long-range dependencies in the feature space. Inspired by [50] and [51], LGFM achieves the global relations among pixels by stacking the self-attention operation on horizontal rows and vertical columns pixels, which could progressively capture the long-range dependencies. As shown in Fig. 4, in the horizontal attention (HA), the specific position response is determined by the weighted sum of features in the same horizontal row. Given the input features $\mathbf{X} \in \mathbf{R}^{C\times W\times H}$, C, H, and W present the channel, height, and width of feature maps, respectively. We first employ two vanilla convolution operations $[\theta(\cdot), \varphi(\cdot)]$ with 1×1 kernel size to obtain the query features and key features. Then the similarity computation operation can be defined as

$$f(X_{i,j}, X_{i,k}) = \theta(X_{i,j})^T \varphi(X_{i,k}). \tag{1}$$

Similar to [50], we define horizontal row attention in LGFM as

$$X_{i,j} = \frac{1}{C(X)} \sum_{k=1}^{W} f(X_{i,j}, X_{i,k}) g(X_{i,j}) \tag{2}$$

where $g(X_{i,j})$ computes the representation of feature $X_{i,j}$ at the position (i, j); $C(X)$ means the normalization factor.

Similar to HA, vertical attention (VA) can be defined as

$$X_{i,j} = \frac{1}{C(X)} \sum_{k=1}^{H} f(X_{i,j}, X_{i,k}) g(X_{i,j}). \tag{3}$$

In the VA module, the specific position response is only determined by the weighted sum of features at the same vertical column. Hence, for the pixels of the same horizontal row or the same vertical column, the long-range dependencies could be built via just one HA or VA operation. For the other condition (not in the same horizontal row or the same vertical column), the long-range dependency relationship of two different pixels is indirectly constructed via stacking the HA and VA modules. To make this condition easier and clear to understand, we provide the simplified illustration of information propagation in Fig. 5.

Given an input feature $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ (C, H and W are the channel, height, and width, respectively). The primary difference between the original nonlocal module and our proposed LGFM is the calculation process of the similarity attention map. The original nonlocal module [49] directly calculates the interaction among all positions along the channel dimension. The computational complexity is $O(C \times HW \times HW)$. In contrast, LGFM achieves global relations among different positions by stacking the HA and VA operations, which could progressively capture the long-range dependencies. The computational complexity of LGFM is $O(C \times H^2 + C \times W^2)$, which largely reduces both computation and memory costs.

Moreover, as shown in Fig. 3, the transmission network branch is further refined by borrowing complementary information from intermediate features of the hue network branch. To achieve feature integration, we have considered the concatenating fusion operation [52] and boosting fusion

operation [53]. Since the boosting manner can more effectively exploit two different features of information [53], we use it as the fusion operation in our framework.

C. Training Loss

Our training loss consists of three terms: one content loss $\mathcal{L}_c$, one cyclic hue loss $\mathcal{L}_{\text{hue}}$ and one adversarial loss $\mathcal{L}_{\text{adv}}$.

1) Content Loss: A pretrained VGG-19 model from ImageNet [54] can be directly applied as an image feature extractor, which could provide supervised constraints related to high-level semantic information [55]. This content loss is defined as

$$\mathcal{L}_c = \sum_l \lambda_l \left\| \Phi_l(\mathbf{T}) - \Phi_l(\hat{\mathbf{T}}) \right\|_1 \tag{4}$$

where $\mathbf{T}$ and $\hat{\mathbf{T}}$ are the ground truth (GT) transmission layer and the estimated transmission layer, respectively. $\{\lambda_l\}$ denotes the balancing weights, same as the values of [18]. $\{\Phi_l\}$ indicates the "conv1_2," "conv2_2," "conv3_2," "conv4_2," and "conv5_2" layers of VGG-19 [32] network.

2) Cyclic Hue Loss: We further design a new cyclic hue loss in the hue domain to provide a more credible optimization direction for our network training. Since the hue distribution is cyclic in the HSV color space, the distance between two different hue maps ($\mathbf{H^a}, \mathbf{H^b} \in [0, 1]^{W \times H}$) measurement needs to be adjusted accordingly. Based on the cyclic distribution property, the maximum distance between two hue values should be 0.5. In the general case, the distance between $H^a_{i,j}$ and $H^b_{i,j}$ can be defined as

$$\mathcal{D}_{\text{hue}}\left(H^a_{i,j}, H^b_{i,j}\right) = \begin{cases} \left|H^a_{i,j} - H^b_{i,j}\right|, & \left|H^a{}_{i,j} - H^b{}_{i,j}\right| \leqslant 0.5 \\ 1 - \left|H^a_{i,j} - H^b_{i,j}\right|, & \left|H^a{}_{i,j} - H^b{}_{i,j}\right| > 0.5. \end{cases} \tag{5}$$

Therefore, two cases should be considered for the hue loss calculation: the case where the distance is less than 0.5, and the opposite. Based on the above analysis, the cyclic hue loss is defined as

$$\mathcal{D}_{\text{hue}}(\mathbf{H^a}, \mathbf{H^b}) = \left\| \left|\mathbf{H^a} - \mathbf{H^b}\right| \bigodot \mathcal{M} \right\|_1 + \left\| \left|\mathbf{H^a} - \mathbf{H^b} - \mathbb{1}\right| \bigodot (\mathbb{1} - \mathcal{M}) \right\|_1 \tag{6}$$

where $\mathcal{M}_{i,j} = 1$ when $|\mathbf{H^a}_{i,j} - \mathbf{H^b}_{i,j}| < 0.5$ and $\mathcal{M}_{i,j} = 0$ in the opposite case. $\mathbb{1}$ denotes a matrix that all values equal 1. $\bigodot$ denotes element-wise multiplication. According to our two branch network, we define entire cyclic hue loss as

$$\mathcal{L}^1_{\text{hue}} = \mathcal{D}_{\text{hue}}(\hat{\mathbf{H}}, \mathbf{H_T}) \tag{7}$$

$$\mathcal{L}^2_{\text{hue}} = \mathcal{D}_{\text{hue}}(\mathbf{H}_{\hat{\mathbf{T}}}, \mathbf{H_T}) \tag{8}$$

$$\mathcal{L}_{\text{hue}} = \mathcal{L}^1_{\text{hue}} + \beta * \mathcal{L}^2_{\text{hue}} \tag{9}$$

where $\mathbf{H_T}$, $\hat{\mathbf{H}}$ and $\mathbf{H}_{\hat{\mathbf{T}}}$ are the GT hue map, the estimated hue map and the estimated background's hue map respectively and β belongs to hyperparameter.

3) Adversarial Loss: Similar to previous methods [17], [18], we also introduce the adversarial loss to enhance realism of the the generated transmission results. Following the Conditional GAN [56], adversarial loss $\mathcal{L}_{\text{adv}}$ consists of $\mathcal{L}^{\mathcal{G}}_{\text{adv}}$ and $\mathcal{L}^{\mathcal{D}}_{\text{adv}}$, which are defined as (refer to [18] for details)

$$\mathcal{L}^{\mathcal{G}}_{\text{adv}} = -\log \mathcal{D}(\mathbf{I}, \hat{\mathbf{T}}) \tag{10}$$

$$\mathcal{L}^{\mathcal{D}}_{\text{adv}} = -\log \mathcal{D}(\mathbf{I}, \mathbf{T}) - \log(1 - \mathcal{D}(\mathbf{I}, \hat{\mathbf{T}})) \tag{11}$$

where $\mathcal{D}$ denotes the discriminator.

4) Overall Loss: Finally, the overall loss function is

$$\mathcal{L} = \mathcal{L}_c + \alpha_1 * \mathcal{L}_{\text{hue}} + \alpha_1 * \mathcal{L}_{\text{adv}} \tag{12}$$

where α_1 and α_2 are the balancing weights.

IV. Experiments

A. Implementation Details

We implement our network in the Pytorch framework on a PC with an NVIDIA Geforce GTX 1080 Ti GPU. Our model is trained for 70 epochs and adopts the Adam optimizer [57] with an initial learning rate of 1×10^{-4}. The learning rate change strategy is halved at epoch 30 and reduced to 1×10^{-5} at epoch 50. Random flipping and rotation are performed as data augmentation in the training phase. It takes 0.13 seconds to process an image with the resolution of 256×256 on an NVIDIA GTX 1080 Ti GPU. We empirically set the hyperparameters as: $\alpha_1 = 100$, $\alpha_2 = 0.5$ and $\beta = 2$. The number N of the basic unit is set as three as default.

B. Dataset

1) *SIR-454 Dataset [58]:* **SIR-454** is the first public real-world reflection scenes dataset, which includes three scenarios: *solid*, *postcard*, and *wild*. As described in [58] and [44], *solid* and *postcards* belong to the controlled indoor scenes, which use different blur levels and glass thickness to purposely explore the impact of varying parameters. While the *wild* subset, which contains 55 images and is denoted as *wild-55*, covers general daily conditions scenes captured in the natural environments.
2) *Real-110 Dataset [18]:* **Real-110** contains 110 real-world reflected/GT image pairs with multiple natural scenes. This dataset has been divided into 90 pairs for training, denoted as *real-90*, and 20 pairs for testing, denoted as *real-20*. Note that *real-20* consists of many high-resolution images (e.g., $2942 \times 1935 \times 3$). Due to the limitation of our hardware, we resize the images to avoid out-of-memory. Specifically, referring to the previous method (ERRNet [43]), retaining the aspect ratio of the original images, we resize the images on the *real-20* dataset [18] to keep the short side size at 512. Finally, we adopt resized images for all comparison methods for fair evaluations.
3) *Nature-220 Dataset [17]:* **Nature-220** composed of 220 real-world reflected/GT image pairs. This dataset was randomly separated into 200 images (*nature-200*) for training and 20 images (*nature-20*) for testing.

TABLE I

QUANTITATIVE COMPARISONS ON BENCHMARK DATABASES. WE COMPARE PSNR AND SSIM VALUES CALCULATED IN THE Y CHANNEL OF THE YCRCB COLOR SPACE AND IN THE RGB COLOR SPACE, RESPECTIVELY. THE LATTER IS RECORDED AS THE AVERAGE OF THE RESULTS OF THE THREE CHANNELS. THE BEST RESULTS ARE IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED. THE METHODS MARKED ‡ REFER TO THE FINE-TUNED VERSIONS

Dataset (size)	Metrics	Methods						
		Input Images	FRS [29]	BDN [16]	Zhang *et al.*‡ [18]	CoRRN ‡ [32]	RmNet [43]	R2Net [8]
real (20)	PSNR ↑	18.88	18.24	18.57	21.76	20.39	18.99	19.93
	SSIM ↑	0.796	0.670	0.726	0.785	0.743	0.676	0.718
nature (20)	PSNR ↑	20.32	19.27	18.83	23.27	21.17	19.36	21.08
	SSIM ↑	0.766	0.731	0.737	0.795	0.743	0.725	0.730
wild (55)	PSNR ↑	25.91	22.75	22.01	23.96	24.66	21.98	21.92
	SSIM ↑	0.892	0.844	0.823	0.878	0.879	0.821	0.830
postcard (199)	PSNR ↑	20.92	20.45	21.01	17.37	20.41	19.71	21.22
	SSIM ↑	0.866	0.822	0.873	0.804	0.831	0.808	0.847
solid (200)	PSNR ↑	23.65	21.67	23.45	22.64	22.78	20.33	22.03
	SSIM ↑	0.883	0.837	0.880	0.883	0.891	0.793	0.843
Trainable Parameters (M: 10^6)		-	-	58.49M	0.39M	59.51M	65.43M	14.02M

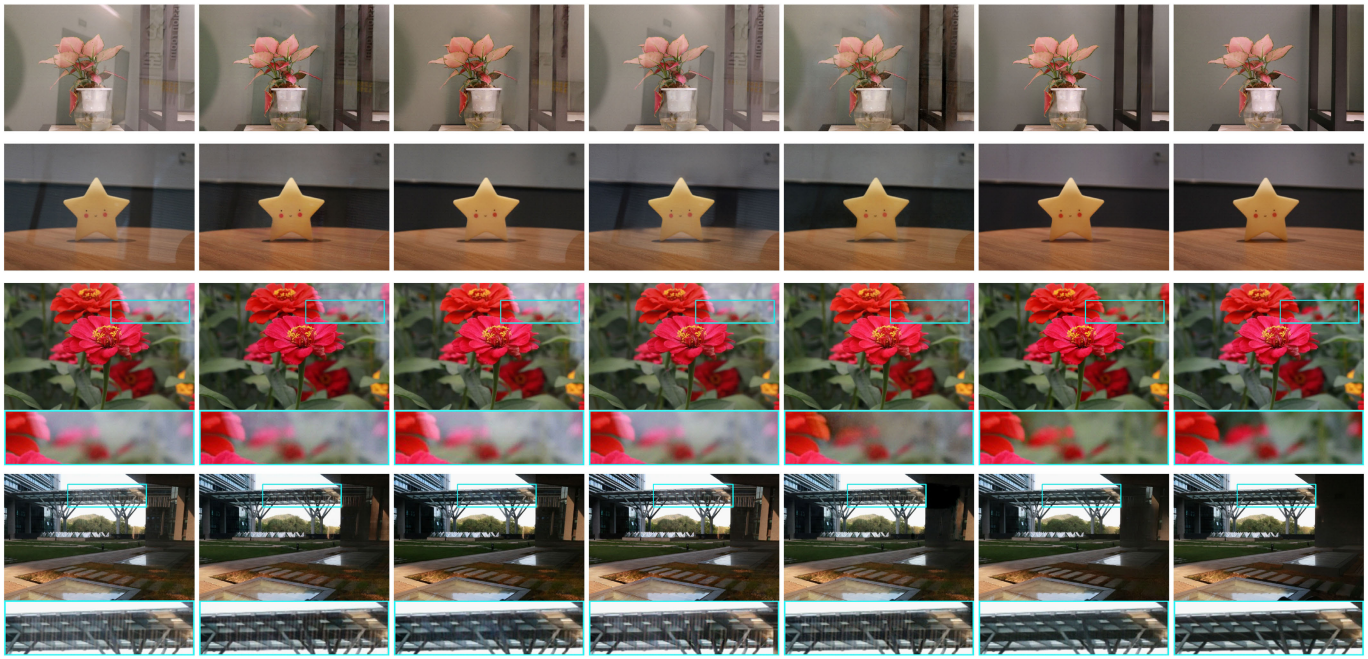
Dataset (size)	Metrics	Methods						
		MLEFGN [7]	ERRNet ‡ [44]	ICBLN [17]	GR-Net [19]	YTMT [40]	LANet [41]	Ours
real (20)	PSNR ↑	19.48	22.93	21.74	19.64	23.13	**23.44**	23.15
	SSIM ↑	0.747	0.807	0.780	0.729	0.813	0.825	**0.861**
nature (20)	PSNR ↑	21.38	22.32	23.84	16.73	20.87	23.52	**25.52**
	SSIM ↑	0.756	0.795	0.783	0.600	0.767	0.810	**0.891**
wild (55)	PSNR ↑	22.34	25.71	24.47	25.58	25.38	25.89	**26.16**
	SSIM ↑	0.831	0.894	0.887	0.888	0.894	0.904	**0.923**
postcard (199)	PSNR ↑	20.28	22.25	23.35	23.16	22.74	21.28	**23.52**
	SSIM ↑	0.807	0.868	0.871	0.876	0.869	0.891	**0.893**
solid (200)	PSNR ↑	22.68	24.76	24.75	19.73	24.77	23.99	**25.16**
	SSIM ↑	0.872	0.899	0.897	0.727	0.900	0.901	**0.915**
Trainable Parameters (M: 10^6)		15.42M	18.58M	21.61M	13.72M	38.32M	10.93M	14.51M

For the training dataset, we totally employ 5000 reflection image pairs, including 3500 synthetic and 1500 real-scene images. To be specific, we first adopt the method [18] to synthesize reflected/GT image pairs from the PASCAL VOC dataset [59]. We then utilize the real-world benchmark dataset from *real-90* and *nature-200* to generate reflected/GT image pairs through the random cropping operation. The patch size is set as 256×256 during the training phase. For testing, we validate the performance of our method with the five test datasets (*nature-20*, *real-20*, *wild-55*, *postcard-199*, and *solid-200*) which have been partitioned by previous methods [17], [18], [60]. Following the previous setting, we conduct fair comparisons throughout the experiment.

C. Comparison Results

We compare our model against several approaches including two model-driven methods: Li and Brown [29], and FRS [28], and several deep learning-based methods: BDN [16], Zhang et al. [18], CoRRN [31], RmNet [42], ERRNet [43], IBCLN [17], GR-nets [44], YTMT [39], and LANet [40]. We also test the networks R2Net [8] and MLEFGN [7], which are specially designed for the classical super-resolution and denoising tasks. We further retrained R2Net and MLEFGN using our training data to compare the de-reflection performance, in which we increase the number of channels to be similar to the number of parameters of our method. All experiments are performed under the same experimental settings (e.g., the same inputs and evaluation codes). For fair comparisons, we choose the best of two models: the provided pretrained model and the fine-tuned version by our training data. For example, we directly employ the publicly pretrained model provided by ICBLN [17] to conduct comparison experiments other than the reported results in their original paper. For the evaluation metrics, we use the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) [61], which are widely used to measure image restoration performance for quantitative evaluation. Higher PSNR and SSIM values generally indicate the generated image is closer to the GT. Comparison results are shown in Table I. Evaluation results demonstrate that our method has the best overall performance compared to the recent state-of-the-art methods. The limited performances of the retrained R2Net and MLEFGN show that the networks for classical image restoration tasks are not suitable for de-reflection. Our method obtains 1.86 dB improvement in PSNR on *nature-20* dataset compared with ICBLN [17], and even reaches 25.52 dB in PSNR on *nature-20* dataset. Because the reflection scenes in *wild-55*, *postcard-199*, and *solid-200* are quite different from our training data, our model achieves the best performance on these three benchmarks. This also demonstrates the powerful generalization ability of our approach.

We also show two visual comparisons of real-world reflection scenes in Figs. 6 and 7. The images in Fig. 6 are obtained from *nature-20*, including indoor and outdoor scenes. Notice that our method produces better and cleaner transmission images and removes most reflection. For example, as shown



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

Fig. 6. Visual comparison with state-of-the-art methods. The examples above are obtained from *nature-20*. **(Best viewed on-screen.)** (a) Input. (b) Zhang et al. (c) ERRNet. (d) YTMT. (e) ICBLN. (f) Ours. (g) GT.

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

Fig. 7. Visual comparison with state-of-the-art methods. The examples above are obtained from ERRNet [43], which acquired many real-world reflection images without the corresponding aligned GT. The above results prove that our method can be well generalize to various reflection scenes. **(Best viewed on-screen.)** (a) Input. (b) Zhang et al. (c) ERRNet. (d) YTMT. (e) ICBLN. (f) Ours.

in the enlarged part in the second row, only our network successfully removes the complex reflection textures, while other results have obvious reflection phenomena. The test images in Fig. 7 are obtained from [43], which contains plenty of misaligned real-world natural reflection images. Since these reflection-distorted scenes are far different from

TABLE II
ABLATION STUDY ON THE EFFECT OF EACH COMPONENT IN PROPOSED DE-REFLECTION NETWORK ON THE THREE REAL-WORLD BENCHMARK DATASETS

Models	Transmission branch		Hue branch	Guidance mannner		PSNR/SSIM		
	MSLFM	LGFM		Concat	Boosting	*real* (20)	*nature* (20)	*wild* (55)
Model-1	✓		✓		✓	22.14/0.820	24.23/0.819	25.46/0.887
Model-2		✓	✓		✓	22.30/0.826	24.66/0.847	26.02/0.912
Model-3	✓	✓				22.89/0.855	24.93/0.860	25.71/0.890
Model-4	✓	✓	✓	✓		23.01/0.857	25.23/0.887	25.93/0.906
Ours	✓	✓	✓		✓	23.15/0.861	25.52/0.891	26.16/0.923

the aforementioned real benchmark dataset, the results can demonstrate the generalization ability of our method. For example, as shown in the third row, our method can simultaneously suppress strong reflection (bright regions) and remove relatively weak reflection (handrail). In the last row, ERRNet [43] fails to handle such strong reflection scenes. While methods [17] generate obvious color degradation in the restored image. Compared with other methods, our method successfully alleviates strong reflections and restores the color information of the transmission layer well. In conclusion, our method achieves a better trade-off between refection removal and color restoration.

D. Ablation Study

In this section, we first discuss the effect of each network component. Then, we compare our hue information with the widely used side information in SIRR, i.e., the edge information. Last, we test the effectiveness of our cyclic hue loss by directly applying it to the previous methods.

1) Effects of Network Modules: To verify the effectiveness of the proposed network modules, we compare our default network with other variants, as shown in Table II. Model-1 and Model-2 verify the effects of the MSLFM and the LGFM. It is clear that the LGFM module brings a significant improvement in PSNR values and captures global information features, indeed boosting the de-reflection performance. By adding the hue branch to the network, the performance can be further boosted in Model-3. This indicates that the complementary features borrowed from the hue network branch indeed provide useful information. To verify our conjecture, we further visualize the intermediate feature maps extracted from the hue branch, which are sent to the transmission branch to provide complementary information in Fig. 8. As can be seen, some features contain obvious reflection-related information, as shown in Fig. 8(c). While others can hardly observe reflection-related content, as shown in Fig. 8(d). This demonstrates that using the hue information can help locate and isolate reflection areas and improve the network capability of reflection removal. Model-4 verify the boosting module performs better than the simple concatenation operation in terms of the evaluation results.

Additionally, we further conduct experiments to verify the effectiveness of the LGFM from two perspectives. Firstly, we replace LGFM with the original nonlocal module. However, our hardware cannot afford the memory burden brought by the original nonlocal module. In order to successfully evaluate the nonlocal module on the reflection benchmarks, we have to reduce the number of the original nonlocal module. The specific experimental results can be found in Table V. Compared to the original nonlocal module, our proposed lightweight module can get better results under similar memory usage conditions. Secondly, as aforementioned, we adopt the HA-VA-HA order of LGFM as default. We adjust the order of HA and VA to verify the robustness of LGFM. The other order is HA-VA-HA. As shown in Table V, the performances of the model before and after the order adjustment are Obviously comparable. This is because both two different orders are capable of capturing global long-range dependencies from different directions in the feature space.

TABLE III
ABLATION STUDY ON THE EFFECT OF THE DIFFERENT LOSS FUNCTIONS ON THE THREE REAL-WORLD BENCHMARK DATASETS

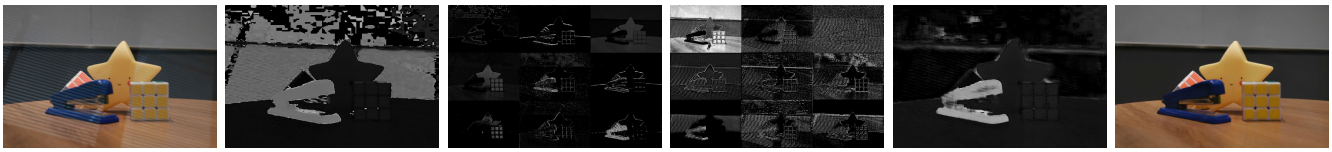
Models	*real* (20)	*nature* (20)	*wild* (55)
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
w/o $\mathcal{L}_{hue}^{1}$	22.89/0.855	24.93/0.860	25.71/0.890
w/o $\mathcal{L}_{hue}^{2}$	22.39/0.805	24.73/0.853	25.06/0.879
w/o $\mathcal{L}_{adv}$	22.93/0.830	25.13/0.865	25.86/0.901
Ours	23.15/0.861	25.52/0.891	26.16/0.923

TABLE IV
COMPARISONS ON EDGE INFORMATION AND HUE INFORMATION

	real (20)	*nature* (20)	*wild* (55)
	PSNR/SSIM	PSNR/SSIM	PSNR/SSIM
Edge guidance network	21.96/0.803	24.37/0.823	25.32/0.864
Hue guidance network	**23.15/0.861**	**25.52/0.891**	**26.16/0.923**

2) Effects of Loss Functions: We further investigate the de-reflection performance of our network architecture when removing partial loss constraints. Note that without $\mathcal{L}_{\text{hue}}^{1}$ means our network only contains the transmission inference branch. It can be clearly seen that each loss function contributes to our method's performance. Comparison results on real-world datasets are shown in Table III and Fig. 9. We found that without $\mathcal{L}_{\text{hue}}^{2}$ largely affects the de-reflection performance and brings obvious color bias in the estimated result. This indicates that adding the cyclic hue loss can better guide the network training and increase the accuracy of reflection removal.

In addition, we further verify the model performance when ignoring the cyclic distribution property of the HSV color

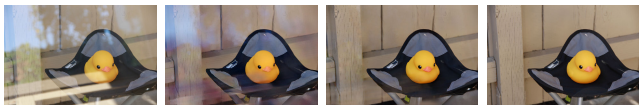


(a) (b) (c) (d) (e) (f)

Fig. 8. Visual results about the hue information. Obviously, the intermediate features contain reflection-free information, which is useful for reflection removal. **(Best viewed on-screen with zoom.)** (a) Input. (b) Hue map of (a). (c) Reflection-free features. (d) Reflection-related features. (e) Estimated hue map. (f) Final result.

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

Fig. 9. Visual comparisons on the three loss items, which are evaluated on a real-world reflection scene image. (a) Input. (b) without $\mathcal{L}^1_{\text{hue}}$. (c) without $\mathcal{L}^2_{\text{hue}}$. (d) without $\mathcal{L}_{\text{adv}}$. (e) Ours.



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

Fig. 10. Visual examples of edge guidance model and our hue guidance model. Both (b) and (c) adopt the same network architecture and parameter setting. (a) Input. (b) Edge guidance. (c) Hue guidance. (d) GT.

TABLE V

ABLATION STUDY ON THE DIFFERENT TEST DATASETS. **MODEL-1**: USING THE PLAIN MAE LOSS TO MEASURE THE DISTANCE BETWEEN DIFFERENT HUE MAPS; **MODEL-2**: REPLACING OUR PROPOSED LGFM WITH THE ORIGINAL NONLOCAL MODULES UNDER THE SIMILAR MEMORY COST CONDITION; **MODEL-3**: ADJUSTING THE ORDER OF ROW ATTENTION AND COLUMN ATTENTION IN LGFM. BOTH THE LARGER VALUES OF PSNR AND SSIM MEAN BETTER RESULTS

Dataset (size)	Metrics	**Model-1**	**Model-2**	**Model-3**	Default Model
real (20)	PSNR	22.70	22.93	23.11	23.15
	SSIM	0.795	0.831	0.853	0.861
nature (20)	PSNR	25.24	25.36	25.43	25.52
	SSIM	0.836	0.890	0.887	0.891
wild (55)	PSNR	26.01	26.02	26.21	26.16
	SSIM	0.908	0.913	0.928	0.923
postcard (199)	PSNR	22.87	22.93	23.60	23.52
	SSIM	0.874	0.881	0.895	0.893
soild (200)	PSNR	24.62	24.90	25.09	25.16
	SSIM	0.891	0.904	0.913	0.915

TABLE VI

EVALUATION OF THE EFFECTIVENESS OF THE CYCLIC HUE LOSS. $\mathcal{L}^2_{\text{hue}}$ APPLIED TO THE PREVIOUS METHODS HELPS TO IMPROVE THEIR DE-REFLECTION PERFORMANCE

		PSNR		
		nature (20)	*real* (20)	*wild* (55)
Zhang *et al.*‡ [18]	w/o $\mathcal{L}^2_{hue}$	23.27	21.76	23.96
	w/ $\mathcal{L}^2_{hue}$	23.48	22.03	24.64
ERRNet ‡ [44]	w/o $\mathcal{L}^2_{hue}$	22.32	22.93	25.71
	w/ $\mathcal{L}^2_{hue}$	22.67	23.10	26.04

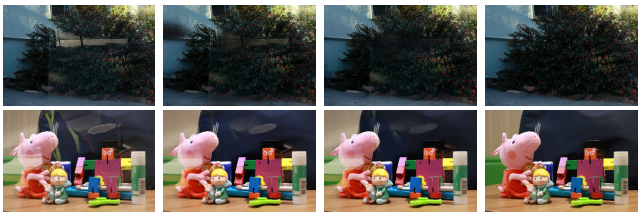
space in Table V. To be specific, we replace the loss calculation of two hue maps with the plain MAE loss. And we find that directly applying MAE loss between two different hue maps (**Model-1**) would lead to a certain degree of de-reflection performance decline. The decline in model performance is due to the inconsistent calculation of the MAE loss in the HSV space [62], which would inaccurately measure the distances along the hue channel of the HSV color space.

3) Superiority of the Hue Map Information: Due to the massively ill-posed nature of this specific problem, previous methods introduce different auxiliary information conducive to SIRR. In these methods, the widely used auxiliary information is the prior knowledge containing the edge of the image, which assumes that the gradient of the transmission layer is usually stronger than the gradient of the reflection [29]. While in this article, we argue that compared to edge information, using hue information as auxiliary information is more suitable for SIRR. Compared with gradients or edges that usually provide sparse features, using the hue map can further provide dense features and help locate the reflection regions.

To verify the Superiority of the hue information, we replace the hue map with the edge map in the hue branch and, at the same time, replace the cyclic hue loss function with the gradient loss. Unlike previous methods [29], [31] that utilize edge maps to perceive the reflectance areas, we exploit the color information to help our network perceive and recover the reflection areas. The comparison results are shown in Table IV, and it is clear that the results of using the hue guidance model are significantly better than that of the edge guidance model. This is because the effects of reflection are usually regional, and the hue map enables highlighting the reflection-distorted regions. On the contrary, as a sparse feature of an image, edge information only considers the difference between adjacent pixels and cannot describe reflection from a larger spatial range. This makes it difficult to distinguish the difference between the transmission layer and the reflection layer using edge information. Note that our edge guidance network version also performs superior to other methods, e.g., CoRRN [31], which also utilizes the edge information. This also proves the effectiveness of our network architecture.

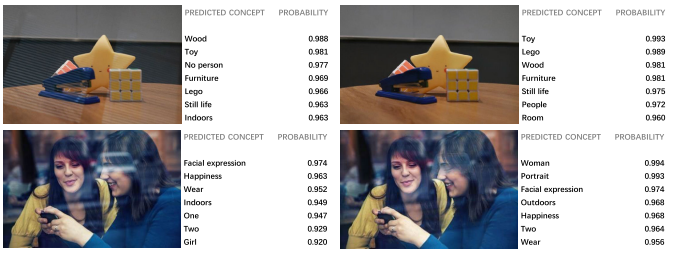
We also provide visual comparisons in Fig. 10. As can be seen, using the edge guidance model fails to handle regional reflection with a bright appearance. While using our hue guidance model can well alleviate the reflection effect and significantly improve the visual quality.

4) Effectiveness of the Cyclic Hue Loss: Since our proposed cyclic hue loss is independent of the network architecture,



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

Fig. 11. Comparison results after adding $\mathcal{L}_{\text{hue}}$. Row 1 from Zhang et al. [18] and row 2 from ERRNet [43]. (a) Input image. (b) without $\mathcal{L}_{\text{hue}}$. (c) w/ $\mathcal{L}_{\text{hue}}$. (d) GT.



(a) (b)

Fig. 12. Real-world examples of object recognition improvements were obtained by our method. Recognition performance is more accurate when our network is used as a preprocessing step. (a) Recognition on input images. (b) Recognition on our results.

it can be directly used to train other SIRR methods to further improve their performance. To verify the effectiveness of our proposed loss, we integrate it into Zhang et al. [18], and ERRNet [43]. Note that ERRNet [43] has two training stages, and the second stage is on unaligned data. Therefore, for a fair comparison, we only train ERRNet with aligned image pairs. We directly add our cyclic hue loss function without changing the default loss functions used by these two methods. Since both methods provide source code, we retrain two models for each network. The first one is trained by using the default loss functions, and the second one is trained by adding our cyclic hue loss.

Table VI shows the quantitative results, and it is clear that adding our cyclic hue loss $\mathcal{L}^2_{\text{hue}}$ can improve the PSNR values of these two methods on testing benchmark datasets. Note that on *wild-55* data set, $\mathcal{L}^2_{\text{hue}}$ helps Zhang et al. [18] almost achieve an average gain of 0.68 dB while 0.33 dB for ERRNet [43]. We also present visual comparisons in Fig. 11. With the help of $\mathcal{L}^2_{\text{hue}}$, not only can quantitative performance be boosted, but also visual quality can be improved.

V. Discussion and Extension

A. Why Hue Information Is Useful?

In most cases, the color difference between the reflection region and its surrounding area is obvious. This indicates that utilizing hue information could guide the network to effectively distinguish between reflection and nonreflection areas and help restore the transmission layer. In addition, compared with gradient-based or edge-based loss that usually provides sparse constraints due to the continuity of the color region, the use of our cyclic hue loss can also provide dense constraints, helping to locate reflection regions.

B. Limitation

As shown in Fig. 1, pixel values of the reflection images are saturated, leading to the loss of content in the brightness reflectance areas. In this case, although our network performs better than previous methods, it is still difficult to completely remove reflections and accurately restore the transmission layer information. More advanced generative models or external information should be introduced for restoring these reflection cases, which will be our future work.

C. Extension

We also test our method as a preprocessing for high-level vision tasks. Specifically, we test our model on Clarifai,[1] which is a widely used commercial image recognition system based on CNNs. We show two examples in Fig. 12. As can be seen, in the first row, the input and our result are recognized as wood and toy with the highest probability, respectively. While in the second row, the system judges that the probability of two people is increased to 0.964. Since reflections often cause undesirable occlusion and color degradation in the captured image, this may affect the performance of downstream computer vision tasks. This test demonstrates that our proposed network has potential value for serving high-level vision systems.

VI. Conclusion

We introduce a new network with two branches to effectively eliminate reflection from a single image. Unlike most traditional methods that utilize the image edge as auxiliary information, we found that the hue map of the captured reflection image can highlight the reflection-distorted regions. This motivates us to introduce hue information as an effective guide for reflection removal. We further design a new cyclic hue loss, which can be directly plugged into previous methods, to help network training and improve SIRR performance. Compared with state-of-the-art methods, our method achieves better performance in both quantitative and qualitative aspects. Moreover, our network has potential value for helping downstream high-level vision tasks.

References

[1] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1647–1654, Sep. 2007.

[2] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–11, Jul. 2015.

[3] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman, "Reflection removal using ghosting cues," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3193–3201.

[4] Q. Zheng, B. Shi, J. Chen, X. Jiang, L.-Y. Duan, and A. C. Kot, "Single image reflection removal with absorption effect," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13395–13404.

[1] https://www.clarifai.com/models/general-image-recognition

[5] C. Lei and Q. Chen, "Robust reflection removal with reflection-free flash-only cues," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14811–14820.

[6] X. Guo, X. Cao, and Y. Ma, "Robust separation of reflection from multiple images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2187–2194.

[7] F. Fang, J. Li, Y. Yuan, T. Zeng, and G. Zhang, "Multilevel edge features guided network for image denoising," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 9, pp. 3956–3970, Sep. 2021.

[8] S. Anwar, N. Barnes, and L. Petersson, "Attention-based real image restoration," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 13, 2021, doi: 10.1109/TNNLS.2021.3131739.

[9] Y. Wang, X. Zhao, T. Jiang, L. Deng, Y. Chang, and T. Huang, "Rain streaks removal for single image via kernel-guided convolutional neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3664–3676, Aug. 2021.

[10] K. Jiang et al., "Multi-scale hybrid fusion network for single image deraining," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Sep. 24, 2021, doi: 10.1109/TNNLS.2021.3112235.

[11] L. Ma, R. Liu, J. Zhang, X. Fan, and Z. Luo, "Learning deep context-sensitive decomposition for low-light image enhancement," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5666–5680, Oct. 2022.

[12] R. Liu, Z. Jiang, X. Fan, and Z. Luo, "Knowledge-driven deep unrolling for robust image layer separation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1653–1666, May 2020.

[13] J.-F. Hu, T.-Z. Huang, L.-J. Deng, T.-X. Jiang, G. Vivone, and J. Chanussot, "Hyperspectral image super-resolution via deep spatiospectral attention convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 7251–7265, Dec. 2022.

[14] C.-M. Feng, Z. Yang, H. Fu, Y. Xu, J. Yang, and L. Shao, "DONet: Dual-octave network for fast MR image reconstruction," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jul. 1, 2021, doi: 10.1109/TNNLS.2021.3090303.

[15] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3238–3247.

[16] J. Yang, D. Gong, L. Liu, and Q. Shi, "Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 654–669.

[17] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3565–3574.

[18] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4786–4794.

[19] D. J. Jobson, Z. Rahman, and G. A. Woodell, "A multiscale retinex for bridging the gap between color images and the human observation of scenes," *IEEE Trans. Image Process.*, vol. 6, no. 7, pp. 965–976, Jul. 1997.

[20] Y. Li and M. S. Brown, "Exploiting reflection change for automatic reflection removal," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2432–2439.

[21] B. Sarel and M. Irani, "Separating transparent layers through layer information exchange," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2004, pp. 328–341.

[22] C. Sun, S. Liu, T. Yang, B. Zeng, Z. Wang, and G. Liu, "Automatic reflection removal using gradient intensity and motion cues," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 466–470.

[23] J. Yang, H. Li, Y. Dai, and R. T. Tan, "Robust optical flow estimation of double-layer images under transparency or reflection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1410–1419.

[24] Y. Y. Schechner, J. Shamir, and N. Kiryati, "Polarization and statistical analysis of scenes containing a semireflector," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 17, no. 2, pp. 276–284, 2000.

[25] H. Farid and E. H. Adelson, "Separating reflections and lighting using independent components analysis," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 1999, pp. 262–267.

[26] N. Kong, Y.-W. Tai, and J. S. Shin, "A physically-based approach to reflection separation: From physical modeling to constrained optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 209–221, Feb. 2014.

[27] N. Arvanitopoulos, R. Achanta, and S. Susstrunk, "Single image reflection suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4498–4506.

[28] Y. Yang, W. Ma, Y. Zheng, J.-F. Cai, and W. Xu, "Fast single image reflection suppression via convex optimization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8141–8149.

[29] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2752–2759.

[30] R. Wan, B. Shi, L. Duan, A. Tan, W. Gao, and A. C. Kot, "Region-aware reflection removal with unified content and gradient priors," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2927–2941, Jun. 2018.

[31] R. Wan, B. Shi, H. Li, L. Duan, A. Tan, and A. C. Kot, "CoRRN: Cooperative reflection removal network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 12, pp. 2969–2982, Dec. 2020.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

[33] Y. Gandelsman, A. Shocher, and M. Irani, "'Double-DIP': Unsupervised image decomposition via coupled deep-image-priors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11026–11035.

[34] V. Lempitsky, A. Vedaldi, and D. Ulyanov, "Deep image prior," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9446–9454.

[35] P. Wieschollek, O. Gallo, J. Gu, and J. Kautz, "Separating reflection and transmission images in the wild," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 89–104.

[36] C. Lei, X. Huang, M. Zhang, Q. Yan, W. Sun, and Q. Chen, "Polarized reflection removal with perfect alignment in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1747–1755.

[37] Y. Hong, Q. Zheng, L. Zhao, X. Jiang, A. C. Kot, and B. Shi, "Panoramic image reflection removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7762–7771.

[38] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[39] Q. Hu and X. Guo, "Trash or treasure? An interactive dual-stream strategy for single image reflection separation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24683–24694.

[40] Z. Dong, K. Xu, Y. Yang, H. Bao, W. Xu, and R. W. H. Lau, "Location-aware single image reflection removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5017–5026.

[41] D. Ma, R. Wan, B. Shi, A. Kot, and L. Duan, "Learning to jointly generate and separate reflections," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2444–2452.

[42] Q. Wen, Y. Tan, J. Qin, W. Liu, G. Han, and S. He, "Single image reflection removal beyond linearity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3771–3779.

[43] K. Wei, J. Yang, Y. Fu, D. Wipf, and H. Huang, "Single image reflection removal exploiting misaligned training data and network enhancements," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8178–8187.

[44] S. Kim, Y. Huo, and S.-E. Yoon, "Single image reflection removal with physically-based training images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5164–5173.

[45] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 447–456.

[46] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "ESPNet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 552–568.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[48] M. Wang, X. Fu, Z. Sun, and Z.-J. Zha, "JPEG artifacts removal via compression quality ranker-guided networks," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 566–572.

[49] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[50] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[51] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-deeplab: Stand-alone axial-attention for panoptic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 108–126.

[52] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*. Cham, Switzerland: Springer, 2015, pp. 234–241.

[53] H. Dong et al., "Multi-scale boosted dehazing network with dense feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2157–2167.

[54] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[55] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.

[56] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[57] M. Everingham, S. M. A. Eslami, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, 2015.

[58] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3922–3930.

[59] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[60] R. Wan, B. Shi, L. Duan, A. Tan, and A. C. Kot, "CRRN: Multi-scale guided concurrent reflection removal network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4777–4785.

[61] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[62] K. R. Castleman, *Digital Image Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

Yurui Zhu (Student Member, IEEE) received the B.Eng. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2018, where he is currently pursuing the Ph.D. degree with the School of Cyber Science and Technology.

His research interests focus on computer vision, especially low-level vision tasks.

Xueyang Fu (Member, IEEE) received the Ph.D. degree in signal and information processing from Xiamen University, Xiamen, China, in 2018.

He was a Visiting Scholar with Columbia University, New York, NY, USA, sponsored by the China Scholarship Council, from 2016 to 2017. He is currently an Associate Professor with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. He is also a member of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application (NEL-BITA). His research interests include machine learning and image processing.

Zheyu Zhang received the B.Eng. degree in electronic information engineering from Xidian University, Xi'an, China, in 2019. He is currently pursuing the Ph.D. degree with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China.

His research interests focus on computer vision, especially medical image analysis and low-level vision tasks.

Aiping Liu (Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2009, and the M.S. and Ph.D. degrees in electrical and computer engineering from The University of British Columbia (UBC), Vancouver, BC, Canada, in 2011 and 2016, respectively.

She has been a Post-Doctoral Research Fellow with the Pacific Parkinson's Research Center, UBC. Currently, she is an Associate Professor with the Department of Electrical Engineering and Information Science, University of Science and Technology of China. She is also a member of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application (NEL-BITA). She has published over 80 scientific papers in prestigious journals and conferences. Her research interests include biomedical signal processing and neuroimaging analysis.

Dr. Liu is serving as an Associate Editor for IEEE SIGNAL PROCESSING LETTERS and an Guest Editor for *Journal of Neuroscience Methods* and *Frontiers in Aging Neuroscience*.

Zhiwei Xiong (Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2006 and 2011, respectively.

He has been a Professor with USTC since 2016. Before that, he was a Researcher with Microsoft Research Asia (MSRA), Beijing, China. He has authored or coauthored more than 100 papers in premium journals and conferences. His research interests include computational photography, low-level vision, and biomedical image analysis.

Dr. Xiong received the Best Paper Award of IEEE International Conference on Visual Communications and Image Processing (VCIP) 2016 and the Microsoft Research Asia (MSRA) Fellowship in 2009. He and his students were winners of eight technical challenges held in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), ACM Multimedia (MM), International Conference on Multimedia and Expo (ICME), and International Symposium on Biomedical Imaging (ISBI).

Zheng-Jun Zha (Member, IEEE) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2004 and 2009, respectively.

He is currently a Full Professor with the School of Information Science and Technology, University of Science and Technology of China, and the Executive Director of the National Engineering Laboratory for Brain-Inspired Intelligence Technology and Application (NEL-BITA). He has authored or coauthored more than 200 papers in his research field with a series of publications on top journals and conferences, which include multimedia analysis and understanding, computer vision, pattern recognition, and brain-inspired intelligence.

Dr. Zha was a recipient of multiple paper awards from prestigious multimedia conferences, including the Best Paper Award and the Best Student Paper Award in Association for Computing Machinery (ACM) Multimedia. He is an Associate Editor of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and *ACM Transactions on Multimedia Computing, Communications, and Applications*.