Single Image Reflection Removal through Cascaded Refinement

Chao Li[1,*] Yixiao Yang[1,*] Kun He[1,†] Stephen Lin[2] John E. Hopcroft[3]
[1]School of Computer Science and Technology, Huazhong University of Science and Technology
[2]Microsoft Research Asia, [3]Computer Science Department, Cornell University
brooklet60@hust.edu.cn

Abstract

We address the problem of removing undesirable reflections from a single image captured through a glass surface, which is an ill-posed, challenging but practically important problem for photo enhancement. Inspired by iterative structure reduction for hidden community detection in social networks, we propose an Iterative Boost Convolutional LSTM Network (IBCLN) that enables cascaded prediction for reflection removal. IBCLN is a cascaded network that iteratively refines the estimates of transmission and reflection layers in a manner that they can boost the prediction quality to each other, and information across steps of the cascade is transferred using an LSTM. The intuition is that the transmission is the strong, dominant structure while the reflection is the weak, hidden structure. They are complementary to each other in a single image and thus a better estimate and reduction on one side from the original image leads to a more accurate estimate on the other side. To facilitate training over multiple cascade steps, we employ LSTM to address the vanishing gradient problem, and propose residual reconstruction loss as further training guidance. Besides, we create a dataset of real-world images with reflection and ground-truth transmission layers to mitigate the problem of insufficient data. Comprehensive experiments demonstrate that the proposed method can effectively remove reflections in real and synthetic images compared with state-of-the-art reflection removal methods.

1. Introduction

Undesirable reflections from glass occur frequently in real-world photos. It not only significantly degrades the image quality, but also affects the performance of downstream computer vision tasks like object detection and semantic segmentation. As the reflection removal problem is ill-posed, early works primarily tackle it with multiple input images [24, 19, 16, 32, 6, 23, 5, 7]. More recently, researchers attempt to address the more common and practically significant scenario of a single input image [14, 15, 16, 17, 28, 22, 1, 25].

For single-image reflection removal (SIRR), researchers have observed that some handcrafted priors may help for distinguishing the transmission layer from the reflection layer in a single image. But these priors often do not generalize well to different types of reflections and scenes owing to disparate imaging conditions. In recent years, researchers apply data-driven learning to replace handcrafted priors via deep convolutional neural networks. With abundant labeled data, a network can be trained to perform effectively over a broad range of scenes. However, learning-based single-image methods still have much room for improvement due to complications such as limited training data, disparate imaging conditions, varying scene content, limited physical understanding of this problem, and the performance limitation of various models.

In this work, inspired by the iterative structure reduction approach for hidden community detection in social networks [8, 9], we introduce a cascaded neural network model for transmission and reflection decomposition. Figure 1 illustrates the cascade results in our model, where the transmission and reflection are progressively refined during the iterations. To the best of our knowledge, previous works on reflection removal did not utilize a cascaded refinement approach. Though some methods such as BDN [33] obtain predictions over a sequence of a few sub-networks, they do not iteratively refine the estimates, but rather they conduct a short alternating optimization, e.g., by estimating the reflection from the input image and the initial transmission layer, and then estimating the transmission from the input image and the estimated reflection layer.

For a cascade model on SIRR, a simple approach is to employ one network to generate a predicted transmission that serves as the auxiliary information of the next network, and continue such process with subsequent networks to iteratively improve the prediction quality. With a long cascade, however, the training becomes difficult due to the

*The first two authors contribute equally.
†Corresponding author.

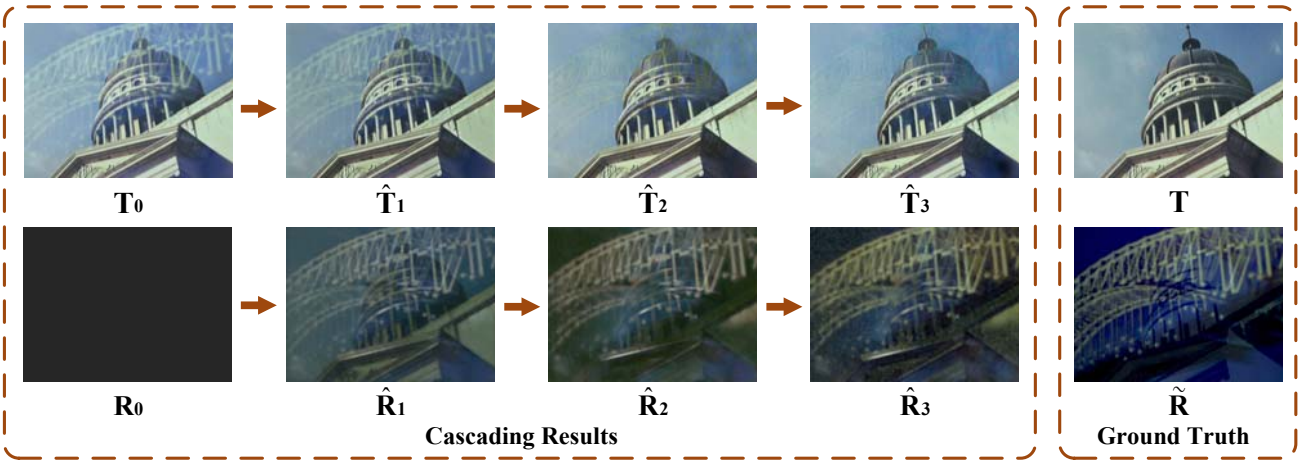

Figure 1. Visualization of results at different cascade steps of the two sub-networks in the proposed model. The estimates of transmissions and residual reflections become increasingly more accurate as they progress through the cascade. More results are in the *suppl. material*.

vanishing gradient problem and limited training guidance at each step. To address this issue, we design a convolutional LSTM (Long Short-Term Memory) network, which saves information from the previous iteration (i.e. time step) and allows gradients to flow unchanged.

In our model, two sub-networks use identical convolutional LSTM architecture, one for transmission prediction and the other for reflection prediction. They share input information using the outputs of the previous time step to boost each other's effectiveness. Here we propose a residual reconstruction loss as further training supervision at each cascade step. To simplify the reconstruction loss, we define a new concept of residual reflection, which will be described in Sec. 3.4.

Though a few real-world datasets with ground-truth have been presented [26, 34], the real-world data for SIRR is still insufficient due to the tremendously labor-intensive work. To help resolve the insufficiency of the real-world training data, we also collect a real dataset with densely-labeled ground truth in disparate imaging conditions and varying scenes.

Our main contributions are as follows:

- We propose a new network architecture, a cascaded network, with loss components that achieves state-of-the-art quantitative results on real-world benchmarks for the single image reflection removal problem.
- We design a residual reconstruction loss, which can form a closed loop with the linear method for synthesizing images with reflections, to expand the influence of the synthesis method across the whole network.
- We collect a new real-world dataset containing images with densely-labeled ground-truth, which can serve as baseline data in future research.

2. Related Work

Mathematically speaking, SIRR operates on a captured image $\mathbf{I}$, which is generally assumed to be a linear combination of a transmission layer $\mathbf{T}$ and a reflection layer $\mathbf{R}$. The goal is to infer a transmission layer $\mathbf{T}$ that is free of reflections. In this work, we focus on deep learning-based SIRR, which has produced state-of-the-art results. Previous multiple-image methods [32, 6, 16, 23, 19, 5, 24, 7] and single-image-priors based methods [15, 17, 14, 22, 1, 28, 16, 25] are not considered here.

Due to the advantages in robustness and performance, there is an emerging interest in applying neural networks to SIRR. Fan *et al.* [4] provide the first neural network model to solve this ill-posed problem. They propose a linear method for synthesizing images with reflection for training, and use an edge map as auxiliary information to guide the reflection removal. Wan *et al*. [27] develop two cooperative sub-networks, which predict the transmission layer intensity and gradients concurrently. Both of these works [4, 27] utilize edge or gradient information of the captured layer $\mathbf{I}$, motivated by the idea that the reflection layers are usually not in focus and thus blurry as compared to the transmission layers. From the edge information of the captured image $\mathbf{I}$, the edge map of the transmission image $\mathbf{T}$ is predicted and used in estimating the transmission result. Instead, BDN [33] predicts reflection layers which are then used as auxiliary information in a subsequent network to estimate the transmission.

In several recent methods, improved formulations of the objective function are presented. These include the adoption of perceptual losses [11] to account for both low-level and high-level image information [3, 10, 34]. In these works, images are fed to a deep network pre-trained on ImageNet, and comparisons are made based on extracted multi-

stage features. Adversarial losses have also been applied, specifically to improve the realism of predicted transmission layers [34, 13, 31, 30].

Another direction of study focuses on datasets for training. Moving beyond improvements for the linear synthesis method in [4] and [34], Wen *et al*. [31] synthesize training data with learned non-linear alpha blending masks that better model the real-world imaging conditions. These masks are also used in forming a reconstruction loss that guides the prediction of transmission layers. To deal with the insufficiency of densely-labeled training data, Wei *et al*. [30] present a technique for utilizing misaligned real-world images as the training data, as they are less burdensome to acquire than aligned images and are more realistic than synthetic images.

3. Proposed Method

3.1. Motivation

This work is motivated by research on hidden structures in social networks. He *et al*. [8, 9] define a set of communities as hidden structure if most of the members also belong to other stronger communities. They propose an iterative boost approach to separate a set of strong, dominant communities and another set of weak, hidden communities, and boost the detection accuracy on both sides. The key idea is that, when they detect an approximate set of dominant communities using a base algorithm, and weaken their internal connection to the average connection of the overall graph, the dominant structure is reduced to boost the detection on the set of hidden communities, and vice versa.

Under the scenario of SIRR, a useful trick is to employ sub-networks to learn auxiliary information that can facilitate transmission layer prediction. The types of auxiliary information utilized in existing works include edge information [4, 27] and predicted reflections [33]. The ideal auxiliary information would be the ground truth reflection-free version of the transmission layer, which is what we seek to predict. As this is not available at inference time, we instead use approximations to the ground-truth transmission in the form of predicted transmissions as the auxiliary information. Though certainly not as useful as the ground truth, it nevertheless provides strong guidance, especially as the transmission predictions improve. The key issue then becomes how to drive the transmission estimations closer and closer to the ground truth. Referring to the work of He *et al*. [8, 9], we regard the transmission layer as the strong, dominant structure, and the reflection layer as the weak, hidden structure. By iteratively reducing the more accurate version of the counterpart, we could extract more accurate approximations on the two layers of images.

Our model contains two sub-networks that can collaborate and boost each other's output by reducing the output of one side from the original image as effective auxiliary information for the other complementary side. Such collaborative cascaded refinement of the dominant image (transmission) and the weak image (reflection) is novel for the training of a neural network.

3.2. General Design Principles

We use two convolutional LSTM networks to separately generate the predicted transmission layers and the predicted reflection layers. The input of each sub-network includes the outputs of both the transmission and reflection sub-networks. Besides, the outputs of the two sub-networks are combined within a reconstruction loss to supervise the whole model at each time step. The synergy between the two sub-networks leads to a mutual boost in their predictions, resulting in progressive improvements of the auxiliary information and finally accurate estimates of the transmission.

To ensure that the transmission sub-network and the reflection sub-network generate complementary outputs, we enforce a reconstruction loss where the image $\hat{\mathbf{I}}$ synthesized from the estimated transmission and reflection is expected to match the input image $\mathbf{I}$.

A related constraint is employed in RmNet [31], which synthesizes an image $\mathbf{I}$ from the ground-truth transmission layer with no reflection, the reflection layer used to produce reflections off the glass, and an alpha blending mask $\mathbf{W}$. Thus, $\mathbf{I} = \mathbf{W} \circ \mathbf{T} + (\mathbf{1} - \mathbf{W}) \circ \mathbf{R}$, where $\circ$ denotes element-wise multiplication. The reconstructed image $\hat{\mathbf{I}}$ is then compared to the synthetic input image $\mathbf{I}$. However, their alpha blending model only approximates the complex physical mechanisms involved in forming an actual input image with reflections, as it does not model effects such as spatially varying blurs and Gamma correction [2], which is used to correct for the differences between the way a camera captures content and the way our visual system processes light. This will limit reconstruction quality on real-world input images and consequently degrade prediction results as we found from experiments reported in Table 1.

To avoid the problem that RmNet encounters, we use a scale parameter α instead of the element-wise mask matrix $\mathbf{W}$, and we directly calculate the *residual reflection* $\widetilde{\mathbf{R}}$ by $\mathbf{I} - \alpha \cdot \mathbf{T}$. In this way, we do not require modeling the complicated physical process involved in the formation of images with reflection, and our performance does not suffer from deficiencies in such a synthesis model. The benefit of predicting residual reflection instead of the reflection layer used to produce reflections off the glass is that image reconstruction becomes simplified as just the sum of the predicted transmission and the predicted residual reflection. Also, different from RmNet, all our linear operations are done in the linear color space, removing Gamma correction [2].

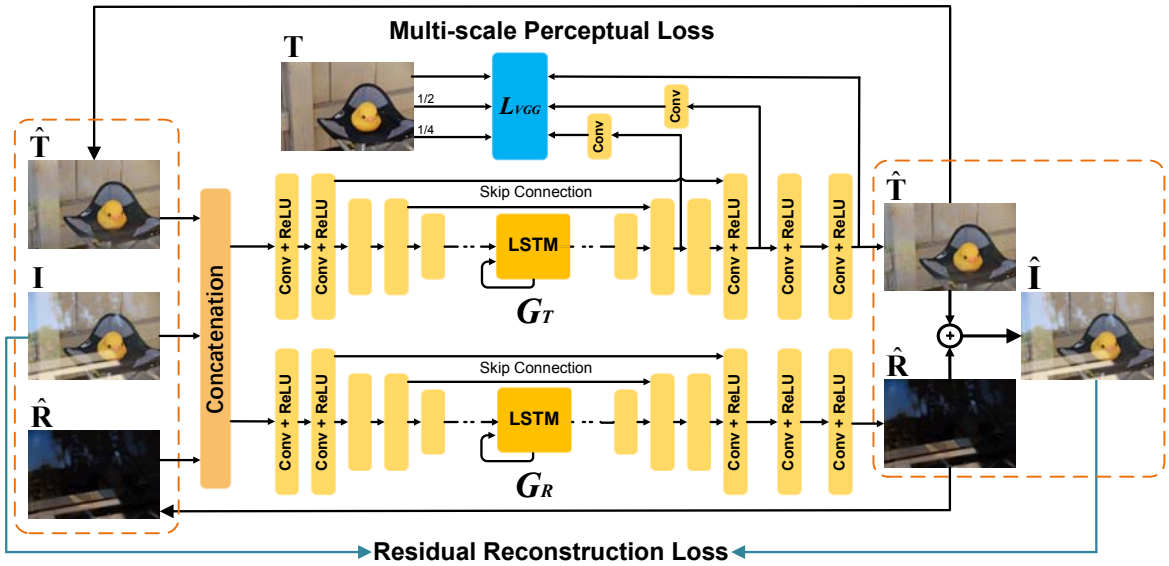

Figure 2. The architecture of IBCLN. The cascaded network consists of a transmission generative sub-network $\boldsymbol{G_T}$ and a reflection generative sub-network $\boldsymbol{G_R}$ with skip connections, both of which are convolutional LSTM networks. The images generated at each time step by the two sub-networks will be fed back at the next time step. The overall network is trained in an end-to-end manner.

3.3. Network Architecture

The architecture of the proposed network is illustrated in Figure 2[1]. IBCLN consists of two sub-networks: a transmission-prediction network $\boldsymbol{G_T}$ and a reflection-prediction network $\boldsymbol{G_R}$. The two sub-networks are both convolutional LSTM networks with the same architecture but different goals. The former aims to learn the transmission $\mathbf{T}$ while the latter aims to learn the residual reflection $\widetilde{\mathbf{R}}$, so they learn completely different weight parameters. Each sub-network consists of an encoder with 11 Conv-relu blocks that extract the features from the input image, a convolutional LSTM unit [20] and a decoder with 8 convolutional layers for generating the predicted transmission layer or the predicted residual reflection layer. Each convolutional layer is followed by a ReLU activation, except for the LSTM layers which are followed by a Sigmoid activation or a Tanh activation. In each sub-network, there are two skip connections between the encoder and the decoder to prevent blurred outputs. The convolutional layers and skip connections are similar to those of a contextual autoencoder [18]. Different from previous works, our objective function includes the proposed residual reconstruction loss and a multi-scale perceptual loss.

Figure 3 illustrates IBCLN from a different perspective. All $\boldsymbol{G_T}$ illustrated in this figure is exactly the same network with the same parameters, but at different time steps in the cascade. We connect $\boldsymbol{G_T}$ at adjacent time steps with convolutional LSTM units that save information from the previous time step. In the actual model, the convolutional LSTM unit is in the middle of the sub-network and connected with convolutional layers. The convolutional LSTM unit has four gates, including an input gate, a forget gate, an output gate, as well as a cell state. The cell state encodes the state information that will be fed to the next LSTM. The LSTM's output feature is fed into the next convolutional layer. More details can be found in ConvLSTM [20]. At time step t, both of the sub-networks take nine channels of input, specifically a concatenation of the synthetic image $\mathbf{I}$, the predicted transmission $\hat{\mathbf{T}}_{t-1}$ and residual reflection $\hat{\mathbf{R}}_{t-1}$ predicted at time step $t-1$ ($1 < t \leq N$). $\mathbf{T}_0$ is set to be the synthetic image $\mathbf{I}$ and $\mathbf{R}_0$ is set to 0.1 for all entries. The output of the transmission prediction network $\boldsymbol{G_T}$ at the final time step N serves as the final result.

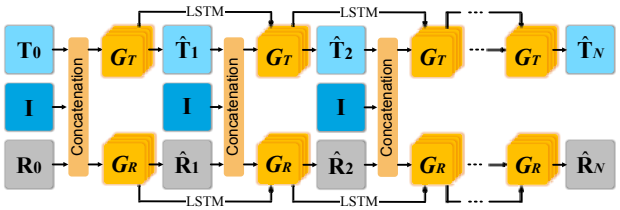

Figure 3. Characterizing IBCLN with increasing number of time steps. All blocks labeled as $\boldsymbol{G_T}$ indicate one sub-network and all blocks labeled as $\boldsymbol{G_R}$ indicate another sub-network. The output at time step $t-1$ serves as the input at time step t. $\hat{\mathbf{T}}_1$, $\hat{\mathbf{T}}_2$, ..., $\hat{\mathbf{T}}_N$ are the predicted transmission. $\hat{\mathbf{R}}_1$, $\hat{\mathbf{R}}_2$, ..., $\hat{\mathbf{R}}_N$ are the predicted residual reflection.

Many previous works consider auxiliary information to

[1]Code and model: https://github.com/JHL-HUST/IBCLN/.

be important for predicting reflection-free transmission layers [4, 33, 27, 31], since it indicates to the network where the removal should be focused on. In our work, $\hat{\mathbf{T}}_{t-1}$ and $\hat{\mathbf{R}}_{t-1}$ are saved to serve as the auxiliary information of step t ($1 < t \leq N$). The auxiliary information will improve with increasing numbers of time steps (see Figure 1). Since the predicted transmissions represent what the network can infer at a given time step, using them as auxiliary information is effective. Additionally, the predicted residual reflection is complementary to the predicted transmission in an image, so it also contains meaningful information.

Considering that the iterative process may require a long cascade, using conventional convolutional networks as the sub-networks would make the full model hard to train. This motivates our use of two convolutional LSTM networks, each with a convolutional LSTM unit. The continuity among time steps makes the model easy to train. Additionally, a cascaded architecture has fewer parameters to learn, as both of the sub-networks are iterated multiple times and each instance of a sub-network shares the same weights. Moreover, a convolutional LSTM network has more complete information exchange either within itself or between the two sub-networks, which is more in line with our iterative boost idea.

3.4. Objective Function

Residual Reconstruction Loss. For the existing linear models [4, 34] for generating synthetic images, the general steps are to perform a series of complex operations on a reflection image to produce a reflection layer $\mathbf{R}$, then to generate a synthetic image $\mathbf{I}$ by a linear operation: $\mathbf{I} = \text{clip}(\alpha \cdot \mathbf{T} + \mathbf{R})$. Usually $\alpha \in [0.8, 1]$ due to the slight attenuation of light as it passes through a glass plane. The weight of the reflection layer $\mathbf{R}$ is 1 as the original reflection image has been subtracted adaptively by the synthesis method. The clipping operation forces all values of the synthetic image to be in [0,1].

We introduce a new loss to the proposed network, called the *residual reconstruction loss*. We adopt the above synthesis model, but replace $\mathbf{R}$ with $\widetilde{\mathbf{R}}$, where $\widetilde{\mathbf{R}}$ is determined from $\mathbf{I}$ and $\mathbf{T}$. $\widetilde{\mathbf{R}}$ offers more effective auxiliary information for transmission prediction, and a more convincing ground truth, as compared to the artificially constructed $\mathbf{R}$. $\widetilde{\mathbf{R}}$ is obtained by reverting the linear synthesis model, as

$$\widetilde{\mathbf{R}} = \mathbf{I} - \alpha \cdot \mathbf{T}. \tag{1}$$

With this definition of $\widetilde{\mathbf{R}}$, the clipping operation is not needed and we avoid its loss of information. After $\widetilde{\mathbf{R}}$ is calculated, it can be used as the ground truth of $\boldsymbol{G_R}$ to guide the generation of the predicted residual reflection $\hat{\mathbf{R}}$. Then, we can simply revert Eq. (1) in the objective function, as

$$\hat{\mathbf{I}} = \alpha \cdot \hat{\mathbf{T}} + \hat{\mathbf{R}}, \tag{2}$$

where $\hat{\mathbf{T}}$, $\hat{\mathbf{R}}$ and $\hat{\mathbf{I}}$ are the predicted transmission, predicted residual reflection and the reconstructed image, respectively. α is the same as in the synthesis model.

Note that all the above linear operations are done in the linear color space, so the Gamma correction [2] on each image is removed before inclusion in linear operations.

It is intuitive that the reconstructed image $\hat{\mathbf{I}}$ should be similar to the original input through a well-trained network. The residual reconstruction loss is defined as:

$$\mathcal{L}_{residual} = \sum_{I \in \mathcal{D}} \sum_{t=1}^{N} \mathcal{L}_{MSE}(\mathbf{I}, \hat{\mathbf{I}}_t). \tag{3}$$

$\mathcal{L}_{MSE}$ indicates the mean squared error. t denotes the time step of the two sub-networks. N represents the final time step when $\hat{\mathbf{T}}$ converges.

The residual reconstruction loss works well experimentally. One potential reason is that the two sub-networks have the same architecture but complementary objectives. With the same architecture, they may be under-trained or over-trained concurrently. The complementary objectives within the residual reconstruction loss can balance the error from the two sub-networks. If both of the two sub-networks are either under-trained or over-trained, the error will be doubled in the residual reconstruction loss.

Multi-scale Perceptual Loss. Multi-scale losses are effective in image decomposition tasks such as raindrop removal [18]. A multi-scale loss extracts the features from different decoder layers and feeds them into a convolutional layer to form outputs at different resolutions. The outputs are then compared to those of real images by their $\mathcal{L}_{MSE}$ distance. By adopting such a loss in our task, we can capture more contextual information from various scales. We change the $\mathcal{L}_{MSE}$ distance to the perceptual distance between the predicted image and the real image over different scales. This loss thus considers different scales of both low-level and high-level information. We define the loss function as:

$$\begin{aligned}\mathcal{L}_{MP} = \sum_{T, T^3, T^5 \in \mathcal{D}} & (\mathcal{L}_{VGG}(\mathbf{T}, \hat{\mathbf{T}}) + \gamma_3 \mathcal{L}_{VGG}(\mathbf{T^3}, \hat{\mathbf{T}}^{\mathbf{3}}) \\ & + \gamma_5 \mathcal{L}_{VGG}(\mathbf{T^5}, \hat{\mathbf{T}}^{\mathbf{5}})),\end{aligned} \tag{4}$$

where $\hat{\mathbf{T}}$, $\hat{\mathbf{T}}^{\mathbf{3}}$, $\hat{\mathbf{T}}^{\mathbf{5}}$ indicate the outputs of the last, 3^{rd} last and 5^{th} last layers at time step N, whose sizes are 1, $\frac{1}{2}$ and $\frac{1}{4}$ of the original size, respectively. $\mathbf{T}$, $\mathbf{T^3}$ and $\mathbf{T^5}$ indicate the ground truth that has the same scale as that of the outputs, respectively. Layers with smaller size are not considered since their information is relatively insignificant. We set $\gamma_3 = 0.8$ and $\gamma_5 = 0.6$. All the images are fed into the VGG19 network [21]. We compare the outputs of the layers 'conv1_2' and 'conv2_2' in the VGG19 network.

Pixel Loss. To ensure that the outputs become as close to the ground truth as possible, we utilize the $\mathcal{L}_{MSE}$ loss to

measure the pixel-wise distance between them. Our pixel loss is defined as follows:

$$\mathcal{L}_{pixel} = \sum_{T\in\mathcal{D}}\sum_{t=1}^{N}[\mathcal{L}_{MSE}(\mathbf{T},\hat{\mathbf{T}}_t) + \mathcal{L}_{MSE}(\widetilde{\mathbf{R}},\hat{\mathbf{R}}_t)], \quad (5)$$

where $\widetilde{\mathbf{R}}$ is the residual reflection. $\hat{\mathbf{T}}_t$ and $\hat{\mathbf{R}}_t$ are the outputs at time step t.

Adversarial Loss. To improve the realism of the generated transmission layers, we further add an adversarial loss. We define an opponent discriminator network $\boldsymbol{D}$. The adversarial loss is defined as (refer to [34] for details):

$$\mathcal{L}_{adv} = \sum_{T\in\mathcal{D}} -\log \boldsymbol{D}(\mathbf{T},\hat{\mathbf{T}}). \quad (6)$$

Overall Loss. Overall, our objective function of IBCLN is defined as:

$$L = \lambda_1\mathcal{L}_{residual} + \lambda_2\mathcal{L}_{MP} + \lambda_3\mathcal{L}_{pixel} + \lambda_4\mathcal{L}_{adv}, \quad (7)$$

where we empirically set the weights as $\lambda_1 = 2, \lambda_2 = 1, \lambda_3 = 2, \lambda_4 = 0.01$ throughout our experiments.

3.5. Implementation Details

We implement the proposed IBCLN in Pytorch on a PC with an Nvidia Geforce GTX 2080 Ti GPU. The overall model is trained for 80 epochs with a batch size of 2, using the Adam optimizer [12]. The learning rate for the overall network training is set to 0.0002. For the training data, we use 4000 images containing 2800 synthetic images and 1200 image patches of size 256×256 from 290 real-world training images, containing 200 images from our created dataset and 90 images from Zhang et al. [34].

4. Experiments

4.1. Dataset Preparation

Figure 4. Samples from our real world *Nature* dataset. Top: images with reflection. Bottom: the corresponding ground-truth transmission layers.

Similar to current deep learning methods, our method requires a relatively large amount of data with ground truth for training. Our synthesis model is the same as the recently proposed linear method [34] except for the clipping operation. We utilize their synthetic dataset as well. In our experiments, different methods are evaluated on the publicly available real-world images from the SIR^2 datasets [26], Zhang et al. [34] and the real-world dataset we create.

Our created dataset, called `Nature`, contains 220 real-world image pairs: images with reflection and the corresponding ground-truth transmission layers (see samples in Figure 4). We use a Canon EOS 750D for image acquisition. Each ground-truth transmission layer is captured when the portable glass is removed. The dataset is randomly partitioned into a training set and a testing set. We use 200 images for training and 20 images for quantitative evaluation. Inspired by Zhang et al. [34], we captured the images with the following considerations to simulate diverse imaging conditions: 1) Environments: indoor and outdoor; 2) Lighting conditions: skylight, sunlight, and incandescent; 3) Thickness of the glass slabs: 3 mm and 8 mm; 4) Distance between the glass and the camera: 3 to 15 cm; 5) Camera viewing angles: front view and oblique view; 6) Camera exposure value: 8.0 - 16.0; 7) Camera apertures (affecting the reflection blurriness): f/4.0 — f/16.

4.2. Comparison to State-of-the-art Methods

4.2.1 Quantitative Evaluations

We compare our IBCLN against state-of-the-art methods including CEILNet [4], Zhang et al. [34], BDN [33], RmNet [31] and ERRNet [30]. For an apples-to-apples comparison, we finetune each model (if the model provides training code) on our training dataset and report the best result of the original pre-trained model and finetuned version (denoted with a suffix '-F'). RmNet [31] has three models for different reflection types, and we report the best result from among the three models.

Table 1 summarizes results of all the competing methods on five real-world datasets, including three sub-datasets from SIR^2 [26], Zhang et al. [34] and our dataset. The number of images in each dataset is shown after the name. The quality metrics include PSNR and SSIM [29]. Larger values of PSNR and SSIM indicate better performance. IBCLN achieves the best performance on four of the datasets, but not on 20 images of "Zhang et al.". As ERRNet [30] is developed based on model Zhang et al. [34], EERNet and Zhang et al. both have better performance on the dataset "Zhang et al.". In terms of overall performance over all the test datasets, IBCLN surpasses the other methods.

4.2.2 Qualitative Evaluations

Figure 5 presents visual results and the ground truth on real-world images from SIR^2 [26], Zhang et al. [34] and our dataset. We select two images from each dataset. It can be seen that Zhang et al. [34] tends to over-remove the reflection layer, while the other baseline methods tend to under-remove. Our model is more accurate and removes most of the undesirable reflections.

Table 1. Quantitative comparison of different methods on three real-world benchmark datasets. The best results are in **bold** and orange color, and the second best results are underlined and in blue color. 'Average' is obtained by averaging the metric scores of all images from all the above real-world datasets.

Dataset (size)	Index	Methods					
		CEILNet-F [4]	Zhang et al. [34]	BDN-F [33]	RmNet [31]	ERRNet-F [30]	IBCLN
Object (200)	PSNR	22.81	22.68	23.02	20.33	24.85	**24.87**
	SSIM	0.801	0.874	0.853	0.793	0.889	**0.893**
Postcard (199)	PSNR	20.08	16.81	20.71	19.71	21.99	**23.39**
	SSIM	0.810	0.797	0.857	0.808	0.874	**0.875**
Wild (55)	PSNR	22.14	21.52	22.34	21.98	24.16	**24.71**
	SSIM	0.819	0.829	0.821	0.821	0.847	**0.886**
Zhang et al. (20)	PSNR	18.79	22.42	19.47	18.77	**23.35**	21.86
	SSIM	0.749	0.792	0.720	0.681	**0.811**	0.762
Nature (20)	PSNR	19.33	19.56	18.92	19.36	22.18	**23.57**
	SSIM	0.745	0.736	0.737	0.725	0.756	**0.783**
Average (494)	PSNR	21.31	20.85	21.68	20.19	23.45	**24.08**
	SSIM	0.806	0.829	0.841	0.795	0.870	**0.875**

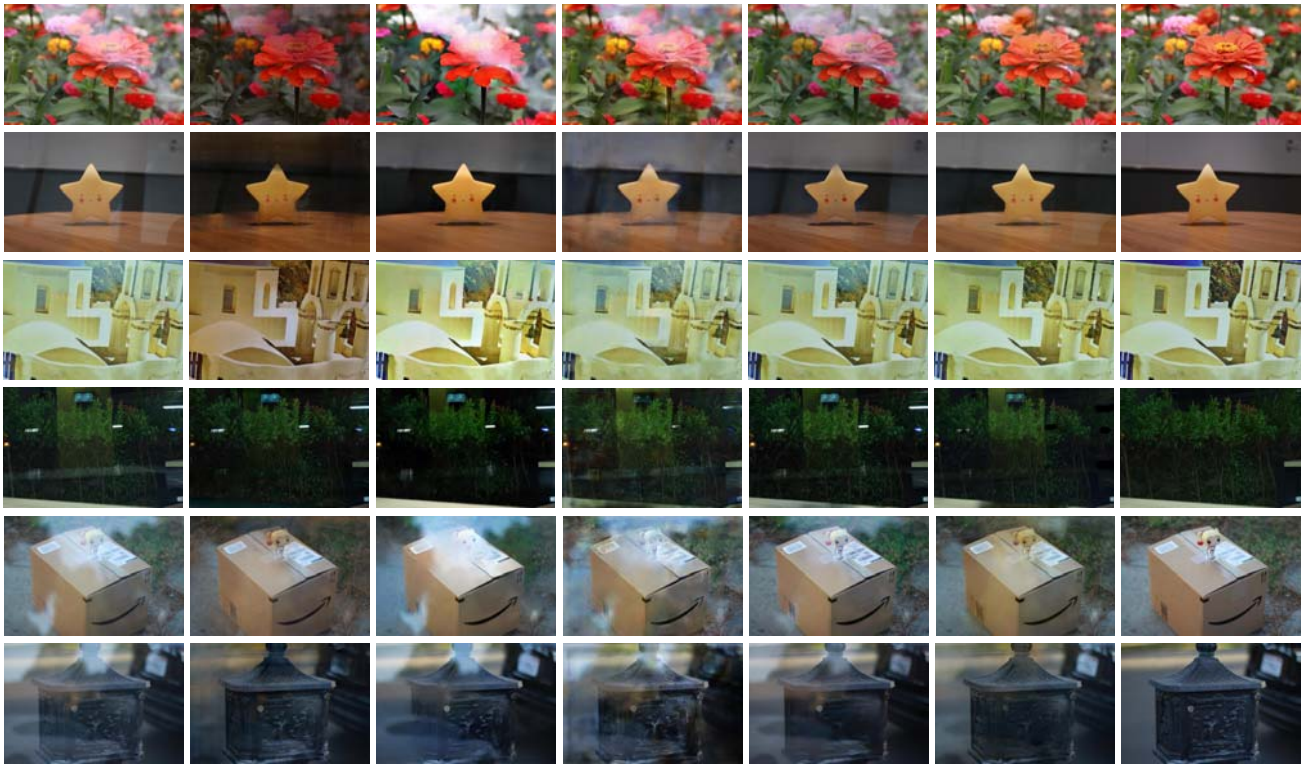

Figure 5. Visual comparison among state-of-the-art approaches and our method on images from three real-world image datasets, namely, *Nature* (Rows 1-2), SIR^2 (Rows 3-4) and Zhang et al. (Rows 5-6). More results can be found in the *suppl. material*.

4.3. Controlled Experiments

For better analyzing our network architecture and the objective function of IBCLN, we separately remove the sub-network $\boldsymbol{G_R}$, the iteration step, and the three-loss terms one by one. Then we train new models with the modified networks. The results from these ablations on the architecture are given in Table 2. The result of a cascade network without LSTM is not shown in the table because it cannot be effectively trained. The ablation study on the loss terms is shown in Table 3. And visual comparisons among all the modified networks and IBCLN are displayed in Figure 6 and Figure 7. We observe that using two iterative sub-networks, time steps, L_{adv}, $L_{residual}$ and L_{MP} all enhance the performance of IBCLN, and all the blocks and the losses

Figure 6. Visual comparison among IBCLN and versions with a modified loss on real-world images. More results are in the *suppl. material*.

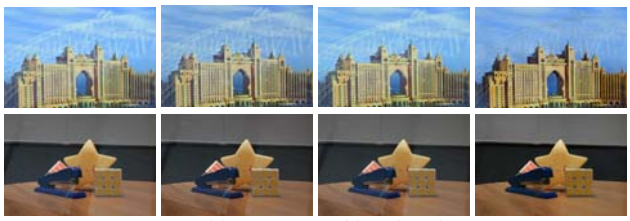

Figure 7. Visual comparison among IBCLN and versions with architecture modifications on real-world images. More results can be found in the *suppl. material*.

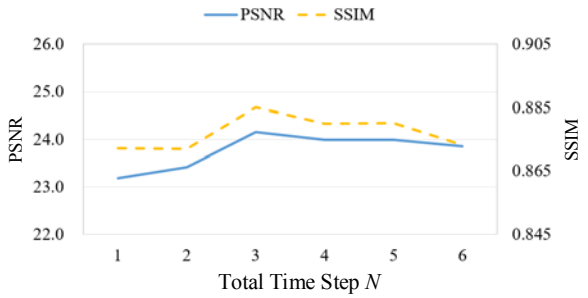

Figure 8. Results using different total time steps N in IBCLN on SIR^2 [26]. Total time steps $N = 3$ yields the best performance.

Table 2. Ablation study of IBCLN for architecture on three testing sets. w/o $\boldsymbol{G_R}$ means training with only one sub-network $\boldsymbol{G_T}$. w/o iteration means the total time steps is 1. Each term contributes to the SIRR performance, and combining all achieves the best results.

Model	Nature		Zhang et al.		SIR^2	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/o G_R	21.79	0.759	20.65	0.742	22.36	0.868
w/o iteration	21.82	0.764	20.49	0.739	23.09	0.872
Complete	**23.57**	**0.783**	**21.86**	**0.762**	**24.20**	**0.884**

yield different contributions to the removal performance. The complete IBCLN with all structures and objective function terms yields the best results.

To explore how many time steps are appropriate for the predicted transmission to converge, we train the model with different total time steps. Figure 8 exhibits the results. We see that the output approximately converges when total time steps are equal to 3. We experimented with having the network learn the total time steps automatically for different images, but we found that providing this much flexibility causes the performance to decay.

Table 3. Ablation study of IBCLN for loss terms on three testing sets. Each loss contributes to IBCLN's performance, and combining all achieves the best result.

Model	Nature		Zhang et al.		SIR^2	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
L_{pixel} only	21.98	0.739	19.54	0.722	22.91	0.843
w/o L_{adv}	23.24	0.746	21.74	0.755	23.86	**0.885**
w/o $L_{residual}$	22.54	0.770	20.98	0.755	23.74	0.881
w/o L_{MP}	23.14	0.744	21.47	0.734	22.96	0.863
Complete	**23.57**	**0.783**	**21.86**	**0.762**	**24.20**	0.884

5. Conclusion

We present an Iterative Boost Convolutional LSTM Network (IBCLN) that can effectively remove the reflection from a single image in a cascaded fashion. To formulate an effective cascade network, we propose to iteratively refine the transmission and reflection layers at each step in a manner that they can boost prediction quality for each other, and to employ LSTM to facilitate training over multiple cascade steps. The intuition is that a better estimate of the complementary residual reflection can boost the prediction of the transmission, and vice versa. Besides, we incorporate a residual reconstruction loss as further training guidance at each cascade step. Moreover, we combine a multi-scale loss with the perceptual loss to form a multi-scale perceptual loss. Quantitative and qualitative evaluations on five datasets (including ours) demonstrate that the proposed IBCLN outperforms state-of-the-art methods on the challenging single image reflection removal problem. In future work, we will try our cascaded prediction refinement approach on other image layer decomposition tasks such as raindrop removal, flare removal and dehazing.

Acknowledgments

This work is supported by the Fundamental Research Funds for the Central Universities (2019kfyXKJC021) and Microsoft Research Asia.

References

[1] Nikolaos Arvanitopoulos, Radhakrishna Achanta, and Sabine Susstrunk. Single image reflection suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4498–4506, 2017.

[2] David R. Bull. Chapter 4 - digital picture formats and representations. In David R. Bull, editor, *Communicating Pictures*, pages 99 – 132. Academic Press, Oxford, 2014.

[3] Zhixiang Chi, Xiaolin Wu, Xiao Shu, and Jinjin Gu. Single image reflection removal using deep encoder-decoder network. *arXiv preprint arXiv:1802.00094*, 2018.

[4] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3238–3247, 2017.

[5] Kun Gai, Zhenwei Shi, and Changshui Zhang. Blind separation of superimposed moving images using image statistics. *IEEE transactions on pattern analysis and machine intelligence*, 34(1):19–32, 2011.

[6] Xiaojie Guo, Xiaochun Cao, and Yi Ma. Robust separation of reflection from multiple images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2187–2194, 2014.

[7] Byeong-Ju Han and Jae-Young Sim. Reflection removal using low-rank matrix completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5438–5446, 2017.

[8] Kun He, Yingru Li, Sucheta Soundarajan, and John E Hopcroft. Hidden community detection in social networks. *Information Sciences*, 425:92–106, 2018.

[9] Kun He, Sucheta Soundarajan, Xuezhi Cao, John E. Hopcroft, and Menglong Huang. Revealing multiple layers of hidden community structure in networks. *CoRR*, abs/1501.05700, 2015.

[10] Meiguang Jin, Sabine Süsstrunk, and Paolo Favaro. Learning to see through reflections. In *2018 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12. IEEE, 2018.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

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

[13] Donghoon Lee, Ming-Hsuan Yang, and Songhwai Oh. Generative single image reflection separation. *arXiv preprint arXiv:1801.04102*, 2018.

[14] Anat Levin, Assaf Zomet, and Yair Weiss. Learning to perceive transparency from the statistics of natural scenes. In *Advances in Neural Information Processing Systems*, pages 1271–1278, 2003.

[15] Anat Levin, Assaf Zomet, and Yair Weiss. Separating reflections from a single image using local features. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–VIII. IEEE, 2004.

[16] Yu Li and Michael S Brown. Exploiting reflection change for automatic reflection removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2432–2439, 2013.

[17] Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2759, 2014.

[18] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2482–2491, 2018.

[19] Bernard Sarel and Michal Irani. Separating transparent layers through layer information exchange. In *European Conference on Computer Vision*, pages 328–341. Springer, 2004.

[20] Xingjian SHI, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Ofer Springer and Yair Weiss. Reflection separation using guided annotation. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1192–1196. IEEE, 2017.

[23] Chao Sun, Shuaicheng Liu, Taotao Yang, Bing Zeng, Zhengning Wang, and Guanghui Liu. Automatic reflection removal using gradient intensity and motion cues. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 466–470. ACM, 2016.

[24] Richard Szeliski, Shai Avidan, and P Anandan. Layer extraction from multiple images containing reflections and transparency. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 246–253. IEEE, 2000.

[25] Mehmet Ali Cagri Tuncer and Ali Cafer Gurbuz. Ground reflection removal in compressive sensing ground penetrating radars. *IEEE Geoscience and remote sensing letters*, 9(1):23–27, 2011.

[26] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3922–3930, 2017.

[27] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2018.

[28] Renjie Wan, Boxin Shi, Tan Ah Hwee, and Alex C Kot. Depth of field guided reflection removal. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 21–25. IEEE, 2016.

[29] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to

structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[30] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019.

[31] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3771–3779, 2019.

[32] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T Freeman. A computational approach for obstruction-free photography. *ACM Transactions on Graphics (TOG)*, 34(4):79, 2015.

[33] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 654–669, 2018.

[34] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4786–4794, 2018.