Single Image Reflection Separation with Perceptual Losses

Xuaner Zhang
UC Berkeley

Ren Ng
UC Berkeley

Qifeng Chen
Intel Labs

Abstract

We present an approach to separating reflection from a single image. The approach uses a fully convolutional network trained end-to-end with losses that exploit low-level and high-level image information. Our loss function includes two perceptual losses: a feature loss from a visual perception network, and an adversarial loss that encodes characteristics of images in the transmission layers. We also propose a novel exclusion loss that enforces pixel-level layer separation. We create a dataset of real-world images with reflection and corresponding ground-truth transmission layers for quantitative evaluation and model training. We validate our method through comprehensive quantitative experiments and show that our approach outperforms state-of-the-art reflection removal methods in PSNR, SSIM, and perceptual user study. We also extend our method to two other image enhancement tasks to demonstrate the generality of our approach.

1. Introduction

Reflection from windows and glasses is ubiquitous in the real world, but it is usually undesirable in photographs. Users often want to extract the hidden clean transmission image by removing reflection from an image. For example, we may have been tempted to take photos through an aquarium glass or skyscraper windows, but reflection can often damage the image quality. Removing reflection from a single image allows us to recover visual content with better perceptibility. Thus, separating the reflection layer and transmission layer from an image — the *reflection separation problem* — is an active research area in computer vision.

Let $I \in \mathbb{R}^{m \times n \times 3}$ be the input image with reflection. I can be approximately modeled as the sum of the transmission layer T and the reflection layer R: $I = T + R$. Our goal is to recover the transmission layer T given I, which is an ill-posed problem without additional constraints or priors.

As the reflection separation problem is ill-posed, prior works often require additional input images and hand-crafted priors. A line of previous research uses multiple images as input or requires explicit user guidance [9, 27, 32]. Multiple images, however, are not always available in practice, and user guidance is inconvenient and error-prone. Recent researchers proposed methods for reflection removal from a single image [25, 21], but these approaches rely on hand-crafted priors such as ghost cues and relative smoothness which may not generalize to all images with reflection. More recently, CEILNet [5] uses a deep neural network to train a model with low-level losses on color and edges, but this approach does not directly enable the model to learn high-level semantics which can be highly useful for reflection removal. Low-level information is insufficient for reflection separation when there is color ambiguity or the model needs to ”recognize” objects in the image. For example, in Figure 1, our model trained with perceptual losses may have learned the representations of lamps and faces, and thus correctly removes them from the input image, while CEILNet fails to do so.

In this paper, we present a fully convolutional network with perceptual losses that encode both low-level and high-level image information. Our network takes a single image as input and directly synthesizes two images: the reflection layer and the transmission layer. We further propose a novel exclusion loss that effectively enforces the separation of transmission and reflection at pixel level. To thoroughly evaluate and train different approaches, we build a dataset that contains real-world images and the ground-truth transmission images. Our dataset covers diverse natural environments including indoor and outdoor scenes. We also use this real-world dataset to compare our approach quantitatively to previous methods. In summary, our main contributions are:

- We propose to use a deep neural network with perceptual losses for single image reflection separation. We impose perceptual supervision through two losses with different levels of image information: a feature loss from a visual perception network, and an adversarial loss to refine the output transmission layer.

- We propose a carefully designed exclusion loss that emphasizes independence of the layers to be separated in the gradient domain.

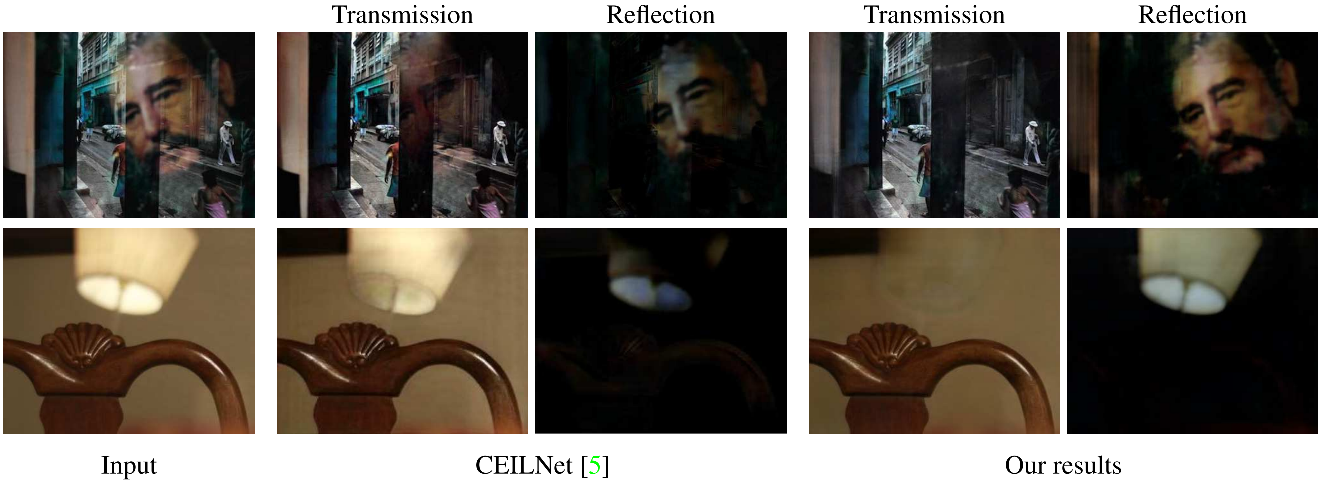


Figure 1: Results by CEILNet [5] and our approach on real-world images. The top row shows a real image from the CEILNet dataset with a window reflecting a poster of a human face; the bottom row shows an image taken by ourselves, with a lamp as the reflection. From left to right: the input images, CEILNet results and our results. Note that our approach trained to learn both low-level and high-level image statistics successfully removes the reflection layers of the face and lamp, while CEILNet does not.

- We build a dataset of real-world images for reflection removal with corresponding ground-truth transmission layers. This new dataset enables quantitative evaluation and comparisons among our approach and existing algorithms.

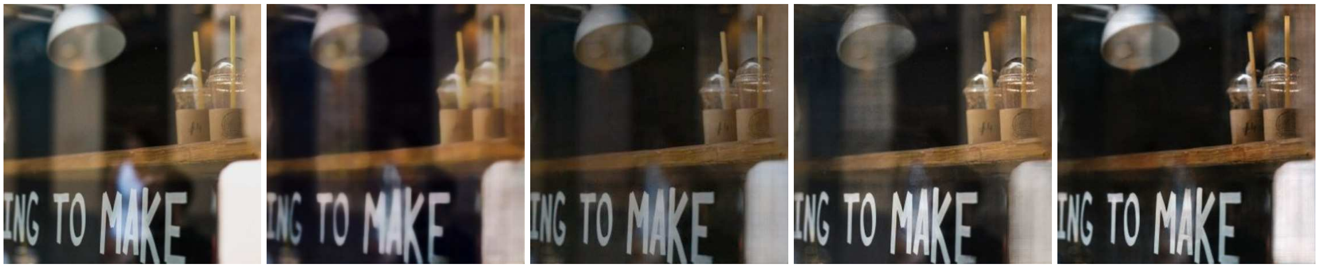
- Our extensive experiments on real data and synthetic data indicate that our method outperforms state-of-the-art methods in SSIM, PSNR, and a perceptual user study on Amazon Mechanical Turk. Our trained model on reflection separation can be directly applied to two other image enhancement tasks, flare removal and dehazing.

2. Related Work

Multiple-image methods. As the reflection separation problem is ill-posed, most previous work tackles this problem with multiple input images. These multi-image approaches often use motion cues to separate the transmission and reflection layers [32, 9, 20, 28, 23, 6, 29, 10]. The motion cues are either inferred from calibrated cameras, or motion parallax that assumes the background and reflection objects have greatly different motion fields. Some other multi-image approaches include the use of flash and no-flash image pairs to improve the flash image with reflection removed [1]. Schechner et al. [24] use a sequence of images with different focus settings to separate layers with depth estimation. Kong et al. [15] exploit physical properties of polarization and use multiple polarized images taken with angular filters to find the optimal separation. More recently, Han and Sim [10] tackle the glass reflection removal problem with multiple glass images, assuming that the gradient field in background image is almost constant while the gradient field in reflection varies much more. Although multiple-image methods have shown promising performance in removing reflection, capturing multiple images is sometimes impossible, for example, these methods can not be applied to existing or legacy photographs.

Single-image methods. Another line of work considers using a single image with predefined priors. A widely used prior is the natural image gradient sparsity [19, 18] to find minimum edges and corners for layer decomposition. The gradient sparsity prior is also explored together with optimal and minimum user assistance to better guide the ill-posed separation problem [17, 27]. A recent work by Arvanitopoulos et al. [2] uses the gradient sparsity constraint, combined with a data fidelity term in the Laplacian space to suppress reflection. However, all these approaches rely on low-level heuristics and are limited in cases where a high-level understanding of the image is needed.

Another prior for reflection separation is that the reflection layer is often out of focus and appears smooth. This is explicitly formulated into an optimization objective by Li and Brown [21], in which they penalize large reflection gradients. Although the assumption of relative smoothness is valid, their formulation can break down when the reflection layer has high contrast. Wan et al. [31] propose a variation of this smoothness prior where depth of field is used as



(a) Input (b) Without L_{feat} (c) Without L_{adv} (d) Without L_{excl} (e) Complete model

Figure 2: Visual comparisons on the three perceptual loss functions, evaluated on a real-world image. In (b), we replace L_{feat} with image space L^1 loss and observed overly-smooth output. (c) shows artifacts of color degradation and noticeable residuals without L_{adv}. In (d), the lack of L_{excl} makes the predicted transmission have undesired reflection residuals. Our complete model in (e) is able to produce better and cleaner prediction.

guidance for edge labeling and layer separation. Additionally, Shih et al. [25] focus on a subset of the problem where reflection has ghost effects, and use estimated convolution kernel to optimize for reflection removal.

Fan et al. [5] recently propose a deep learning network, the Cascaded Edge and Image Learning Network (CEILNet), for reflection removal. They formulate reflection removal as an edge simplification task and learn an intermediate edge map to guide layer separation. CEILNet is trained purely with a low-level loss that combines the differences in color space and gradient domain. The main difference between CEILNet and ours is that they did not explicitly utilize perceptual information during training.

Benchmark datasets. A benchmark dataset by Wan et al. [30] was proposed recently for reflection removal. The authors collected 1500 real images of 40 scenes in a controlled lab environment by imaging pairs of daily objects and postcards, as well as 100 scenes in natural outdoor environments with three different pieces of glasses. However, the dataset has not been released publicly yet at the time of submission. In order to evaluate among different models quantitatively on real-world images, we collect a dataset of 110 real images with ground truth in natural scene environments.

3. Overview

Given an image $I \in [0,1]^{m \times n \times 3}$ with reflection, our approach decomposes I into a transmission layer $f_T(I;\theta)$ and a reflection layer $f_R(I;\theta)$ using a single network $f(I;\theta) = (f_T(I;\theta), f_R(I;\theta))$, where θ is the network weights. We train the network f on a dataset $\mathcal{D} = \{(I,T,R)\}$ where I is the input image, T is the transmission layer of I, and R is the reflection layer of I.

Our loss function contains three terms: a feature loss L_{feat} by comparing the images in feature space, and an adversarial loss L_{adv} for realistic image refinement, an exclusion loss L_{excl} that enforces separation of the transmission and reflection layers in the gradient domain. Our overall loss function is

$$L(\theta) = w_1 L_{\text{feat}}(\theta) + w_2 L_{\text{adv}}(\theta) + w_3 L_{\text{excl}}(\theta), \quad (1)$$

where we set $w_1 = 0.1$, $w_2 = 0.01$ and $w_3 = 1$ to balance the weight of each term.

An ideal model for reflection separation should be able to understand contents in an image. To train our network f with semantic understanding of the input image, we form hypercolumn features [11] by extracting features from a VGG-19 [26] network pre-trained on the ImageNet dataset [22]. The benefit of using hypercolumn features is that the input is augmented with useful features that abstract visual perception of a large dataset such as ImageNet. The hypercolumn feature at a given pixel location is a stack of activation units across selected layers of a network at that location. Here, we sampled the layers 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', and 'conv5_2' in the pre-trained VGG-19 network. The hypercolumn feature has 1472 dimensions in total. We concatenate the input image I with its hypercolumn features as the augmented input for f.

Our network f is a fully convolutional network that has a similar network architecture to the context aggregation network [33, 4]. Our network has a large receptive field of 513 × 513 to effectively aggregate global image information. The first layer of f is a 1 × 1 convolution to reduce feature dimension (1472+3) to 64. The following 8 layers are 3 × 3 dilated convolutions. The dilation rate varies from 1 to 128. All the intermediate layers have 64 feature channels. For the last layer we use a linear transformation to synthesize 2 images in the RGB color space.

We evaluate different methods on the publicly available synthetic and real images from the CEILNet dataset[5] and the real-world dataset we collected. We compare our method to the state-of-the-art reflection removal approach

CEILNet [5], an optimization based approach [21], and Pix2pix [12], a general framework for image translation.

4. Training

4.1. Feature loss

We use a feature loss to measure the difference between our predicted transmission layer and the ground-truth transmission in feature space. As the aforementioned observation in Figure 1 shows, semantic reasoning about the scene would benefit the task of reflection removal. A feature loss that combines low-level and high-level features from a perception network would serve our purpose. Feature loss has also been successfully applied to other tasks such as image synthesis and style transfer [3, 7, 16, 13].

Here, we compute the feature loss by feeding the predicted image layer and the ground truth through a pre-trained VGG-19 network Φ. We compute the L^1 difference between $\Phi(f_T(I;\theta)$ and $\Phi(T)$ in selected feature layers:

$$L_{\text{feat}}(\theta) = \sum_{(I,T)\in\mathcal{D}} \sum_{l} \lambda_l \|\Phi_l(T) - \Phi_l(f_T(I;\theta))\|_1, \quad (2)$$

where Φ_l indicates the layer l in the VGG-19 network. The weights $\{\lambda_l\}$ are used to balance different terms in the loss function. We select the layers 'conv1_2', 'conv2_2', 'conv3_2', 'conv4_2', and 'conv5_2' in the VGG-19 network.

4.2. Adversarial loss

During the course of our research, we find that transmission image can suffer from unrealistic color degradation and undesirable subtle residuals without an adversarial loss. We adopted the conditional GAN [12] for our model. Our generator would be $f_T(I;\theta)$. The architecture of our discriminator, denoted as D, has 4 layers and 64 feature channels wide. The discriminator tries to discriminate between patches in the real transmission images and patches given by $f_T(I;\theta)$ conditioned on I. The goal is to let the network D learn a suitable loss function for further refining layer separation, and to push the predicted transmission layers toward the domain of real reflection-free images.

Loss for the discriminator D is:

$$\sum_{(I,T)\in\mathcal{D}} \log D(I, f_T(I;\theta)) - \log D(I,T), \quad (3)$$

where $D(I,x)$ outputs the probability that x is a natural transmission image given the input image I. Then our adversarial loss is:

$$L_{\text{adv}}(\theta) = \sum_{I\in\mathcal{D}} -\log D(I, f_T(I;\theta)). \quad (4)$$

We optimize over $-\log D(I, f_T(I;\theta))$ instead of $\log(1 - D(I, f_T(I;\theta)))$ for better gradient performance [8].

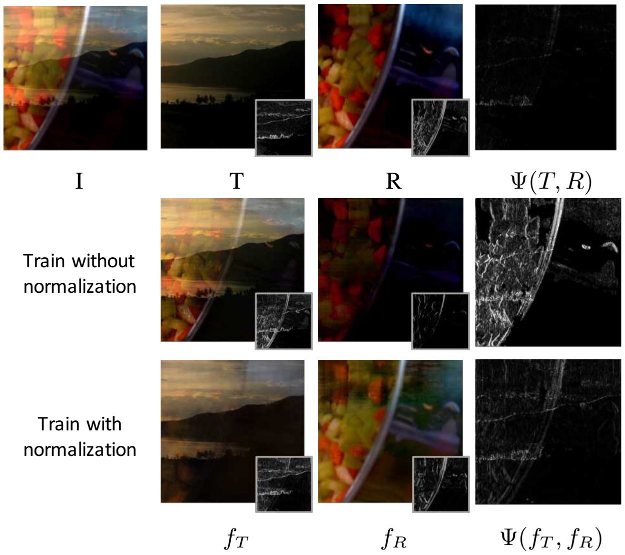


Figure 3: Visual comparisons of training with and without gradient normalization. In the middle two columns, the small window at the right bottom corner of each image shows the gradient magnitude of each image. In the rightmost column, Ψ denotes the normalized gradient product formulated in Equation 6. The first row left to right shows: input, ground truth transmission T, ground truth reflection R, and Ψ. $\Psi(T,R)$ is close to zeros indicating that the gradient fields of T and R are not correlated. The middle row shows results trained with no normalization in the gradient fields. We observe that the reflection prediction trained without normalization is heavily suppressed. Bottom row shows results trained with gradient normalization with better reflection separation.

4.3. Exclusion loss

We further propose an exclusion loss in the gradient domain to better separate the reflection and transmission layers. We explore the relationship between the two layers through analysis of the edges in the two layers. Our key observation is that the edges of the transmission and the reflection layers are unlikely to overlap. An edge in I should be caused by either T or R, but not both. Thus we minimize the correlation between the predicted transmission and reflection layers in the gradient domain. We formulate the exclusion loss as the product of normalized gradient fields of the two layers at multiple spatial resolutions :

$$\begin{aligned} L_{\text{excl}}(\theta) &= \sum_{I\in\mathcal{D}} \sum_{n=1}^{N} \|\Psi(f_T^{\downarrow n}(I;\theta), f_R^{\downarrow n}(I;\theta))\|_F, && (5) \\ \Psi(T,R) &= \tanh(\lambda_T|\nabla T|) \odot \tanh(\lambda_R|\nabla R|), && (6) \end{aligned}$$

where λ_T and λ_R are normalization factors, $\|\cdot\|_F$ is the Frobenius norm, $\odot$ denotes element-wise multiplication, and n is the image downsampling factor: the images f_T and f_R are downsampled by a factor of 2^{n-1} with bilinear interpolation. We set $N = 3$, $\lambda_T = \sqrt{\frac{\|\nabla R\|_F}{\|\nabla T\|_F}}$, and $\lambda_R = \sqrt{\frac{\|\nabla T\|_F}{\|\nabla R\|_F}}$ in our experiments.

Note that the normalization factors λ_T and λ_R are critical in Equation 6, since the transmission and reflection layers may contain unbalanced gradient magnitudes. The reflection layer can be either blurred with low intensity and thus consists of small gradients, or it could reflect very bright light and composes brightest spots in the image, which produces high contrast reflection and thus large gradients. A scale discrepancy between $|\nabla T|$ and $|\nabla R|$ would cause unbalanced updates to the two layer predictions. We observe that without proper normalization factors, the network would suppress the layer with a smaller gradient update rate to close to zero. A visual comparison of results with and without normalization is shown in Figure 3.

L_{excl} is effective in separating the transmission and reflection layers at the pixel level. If we disable L_{excl} in our model, some residual reflection may remain visible in the output transmission image, as shown in Figure 2 (d).

4.4. Implementation

Given the ground-truth reflection layer R, we can further constrain $f_R(I;\theta)$ with R. Reflection layer is usually not in focus and thus blurry. We simply add a L^1 loss in color space to constrain $f_R(I;\theta)$:

$$L_R(\theta) = \sum_{(I,R)\in\mathcal{D}} \|f_R(I;\theta) - R\|_1. \quad (7)$$

We train the network f by minimizing $(L + L_{\text{R}})$ on synthetic and real data jointly. Note that we disable L_{R} when training on a real-world image as it is difficult to estimate R precisely. We tried computing $R = I - T$ but R sometimes contains significant artifacts because $I = R + T$ may not hold when I is overexposed.

For the training data, we use 5000 synthetic images and extract 500 image patches from 90 real-world training images with random resolutions between 256p and 480p. To further augment the data, we randomly resize image patches while keeping the original aspect ratio. We train for 250 epochs with batch size 1 on an Nvidia Titan X GPU and weights are updated using the Adam optimizer [14] with a fixed learning rate of 10^{-4}.

5. Dataset

5.1. Synthetic data

To create synthetic images with reflection, we choose 5000 random pairs of images from Flickr: one outdoor image and one indoor image for each pair. We use an image (either indoor or outdoor) as the transmission layer and the other image as the reflection layer. We assume the transmission and reflection layers locate on different focal planes so that the two layers exhibit noticeable different blurriness. This is a valid assumption in real-life photography, where the object of interest (e.g. artwork through museum windows) is often in the transmission layer and is set to be in focus. In addition, reflection could be intentionally blurred by shooting with a wide aperture. We use this assumption to create a synthetic dataset, by applying a Gaussian smoothing kernel with a random kernel size in the range of 3 to 17 pixels to the reflection image.

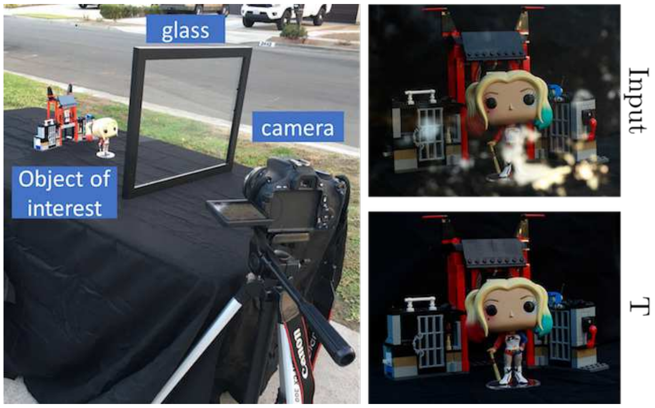


Figure 4: Real data collection setup and captured images. We capture two images with and without the glass with same camera settings in a static scene. Right column from top to bottom: captured image with reflection and the ground-truth transmission image T.

Our image composition approach is similar to the one proposed by Fan et al. [5], but our forward model has the following differences. We remove gamma correction from the images and operate in linear space to better approximate the physical formation of images. Instead of fixing the intensity decay on R, we apply variation to the intensity decay since we observe that reflection in real images could have comparable or higher intensity level than the transmission layer. We apply slight vignette centered at random position in the reflection layer, which simulates the scenario when camera views the reflection from oblique angles.

5.2. Real data

At the time of developing this work, there is no publicly available benchmark with ground-truth transmission to evaluate different reflection removal approaches on real data. We collected a dataset of 110 real image pairs: image with reflection and its corresponding ground-truth transmission image. The images with reflection were taken with a

	Synthetic		Real	
Method	SSIM	PSNR	SSIM	PSNR
Input	0.689	15.09	0.697	17.66
Pix2pix [12]	0.583	14.47	0.648	16.92
Li and Brown [21]	0.742	15.30	0.750	18.29
CEILNet [5]	0.826	20.47	0.762	19.04
Ours	**0.853**	**22.63**	**0.821**	**21.30**

Table 1: Quantitative comparison results among our method and 3 other previous methods. We evaluated on synthetic data provided by CEILNet [5], and our real image test set. We also provide a trivial baseline that takes the input image as the result transmission image.

Canon 600D camera on a tripod with a portable glass in front of the camera. The ground-truth transmission layer was captured when the portable glass was removed. Each image pair was taken with the same exposure setting. Our setup for data capture is shown in Figure 4. We captured the dataset with the following considerations:

- environments: indoor and outdoor;
- lighting conditions: skylight, sunlight, and incandescent;
- camera viewing angles: front view and oblique view;
- and camera apertures (affecting the reflection blurriness): *f*/2.0 — *f*/16.

We split the dataset randomly into a training set and a test set. We extract 500 patches from 90 training images for training and use 20 images for quantitative evaluation.

6. Experiments

6.1. Comparison to prior work

We compare our model to CEILNet [5], the layer separation method by Li and Brown [21], and Pix2pix [12]. We evaluated different methods on the publicly available synthetic images from the CEILNet dataset [5] and the real images from the test set of our real-world dataset.

Our model is only trained on our generated synthetic dataset and the training set of our real-world dataset. For CEILNet, we evaluate its pre-trained model on the CEILNet synthetic images. To evaluate CEILNet on our real data, we fine-tune its model with our real training images (otherwise it performs poorly). We evaluate the approach of Li and Brown [21] with the provided default parameters. Pix2pix is a general image translation model, we train its model on our generated synthetic dataset and the training set of our collected real dataset.

The quantitative results are shown in Table 1. We compute the PSNR and SSIM between the result transmission images of different methods and ground-truth transmission layer. We demonstrate strong quantitative performance over previous works on both synthetic and real data.

	Preference rate
Ours$>$CEILNet [5]	84.2%
Ours$>$Li and Brown [21]	87.8%

Table 2: User study results. The preference rate shows the percentage of comparisons in which users prefer our results.

We also conduct a user study on Amazon Mechanical Turk, following the protocol by Chen and Koltun [3]. During the user study, each user is presented with a input real-world image with reflection, our predicted transmission image, and the predicted transmission image by a baseline in the same row. Then the user needs to choose an output image that is closer to the reflection-free version of the input image between the two predicted transmission images. There are 80 real-world images for comparisons from our dataset and the CEILNet dataset. The results are reported in Table 2. 84.2% of the comparisons to CEILNet and 87.8% of the comparisons to Li and Brown have our results rated to contain less reflection. The results are statistically significant with $p < 10^{-3}$ and 20 users participate in the user study.

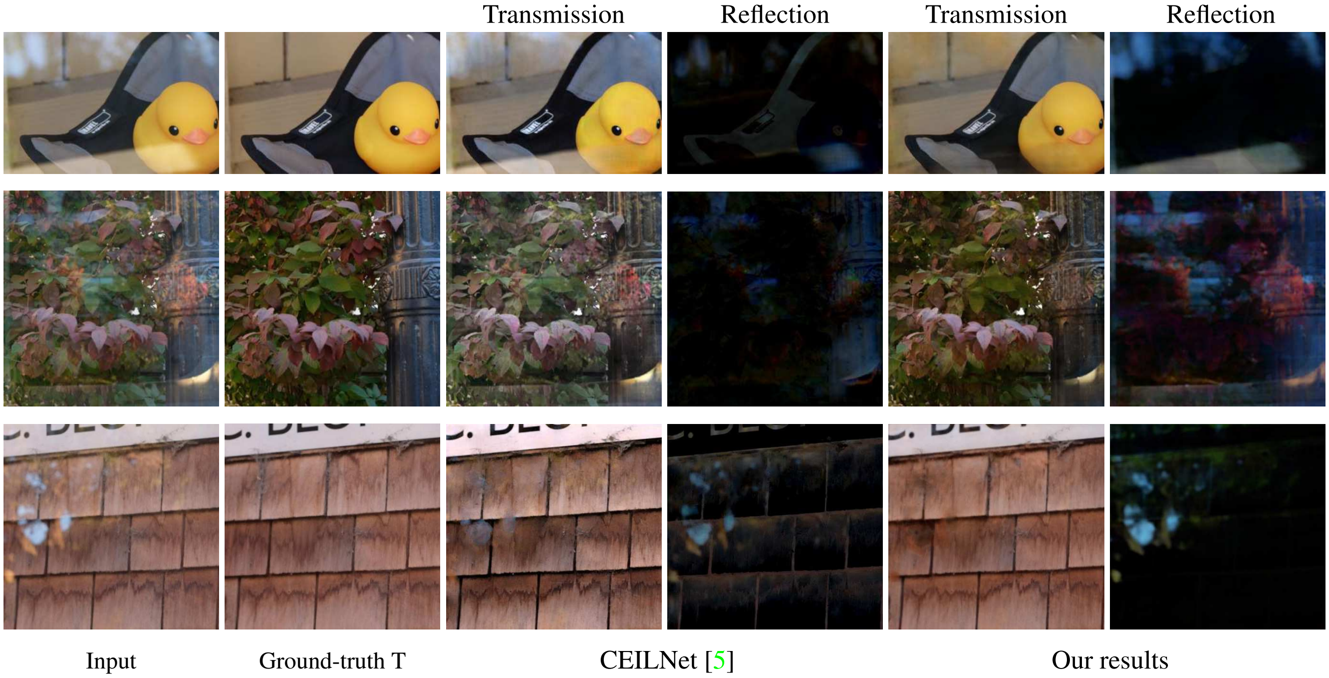
More experimental details and results are reported in the supplement.

	Synthetic		Real	
Method	SSIM	PSNR	SSIM	PSNR
Ours w/o L_{feat}	0.683	18.24	0.743	19.07
Ours w/o L_{adv}	0.818	20.80	0.793	21.12
Ours w/o L_{excl}	0.796	19.58	0.802	20.22
Ours L_{adv}-only	0.765	18.05	0.782	19.52
Ours complete	**0.853**	**22.63**	**0.821**	**21.30**

Table 3: Quantitative comparisons on synthetic and real images among multiple ablated models of our method. We remove each of the three losses and evaluate on the re-trained models. 'Ours L_{adv}-only' denotes our method trained with only an adversarial loss. Our complete model shows better performance on both synthetic and real data. We evaluate on synthetic data provided by CEILNet [5], and our real test images described in Section 5.2.

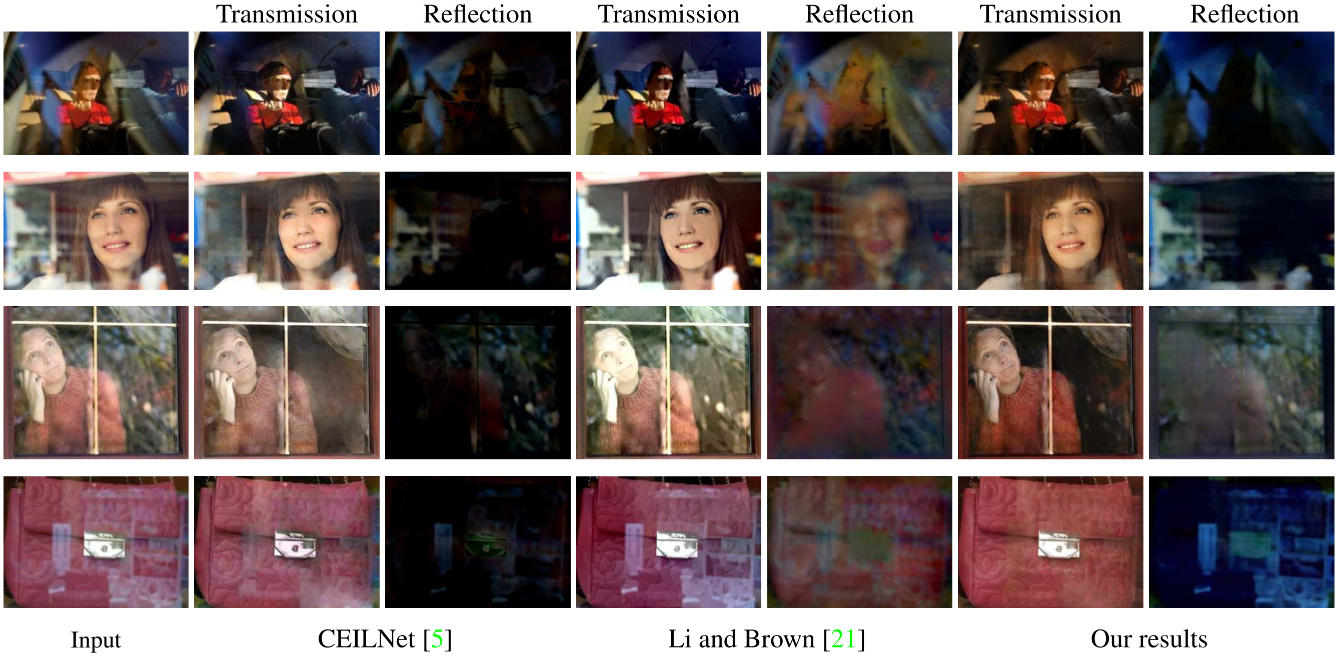
6.2. Qualitative results

We present qualitative results of different methods in Figure 5 and Figure 6, evaluated on real-world images from our dataset (with ground truth) and from CEILNet [5] (with-



Input Ground-truth T CEILNet [5] Our results

Figure 5: Visual results comparison between CEILNet [5] and our method, evaluated on real images from our dataset described in Section 5.2. From left to right: input, ground truth transmission layer, CEILNet [5] predictions and our predictions. Notice that our method produces better and cleaner predictions in both the transmission and reflection layers. Additional results are provided in the supplement.



Input CEILNet [5] Li and Brown [21] Our results

Figure 6: Qualitative comparisons among CEILNet [5], Li and Brown [21] and our method, evaluated on real images in the CEILNET dataset. Note that even though we have no supervision on the reflection layer for real data, but our method predicts cleaner reflection layer as well. Additional results are provided in the supplement.

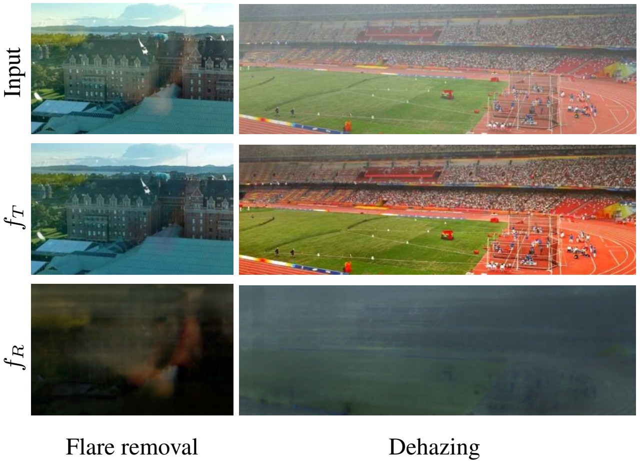


Figure 7: Extension applications on camera flare removal and image dehazing. For each column, from top to bottom: input, our predicted enhanced layer, our predicted removed layer.

out ground truth), respectively.

6.3. Controlled experiments

To analyze how each loss contributes to the final performance of our network, we remove or replace each loss in the combined objective and re-train the network. A visual comparison is shown in Figure 2. When we replace the feature loss L_{feat} with a L^1 loss in color space, the output images tend to be overly-smooth; similar observation is also discussed in [34, 12]. Without L_{excl}, we notice that visible contents of the reflection layer may appear in the transmission prediction. The adversarial refinement loss L_{adv} helps recover cleaner and more natural results, as shown in (e).

The quantitative results are shown in Table 3. We also analyze the performance of the model with only an adversarial loss, which is similar to a conditional GAN [12].

7. Extensions

We demonstrate two additional image enhancement applications, flare removal and dehazing, using our trained model to remove an undesired layer. Note that we directly apply our trained reflection removal model without training or fine-tuning on any flare removal or dehazing dataset. These two tasks can be treated as layer separation problems, similar to reflection separation. For flare removal, we aim to remove the optical artifacts of lens flare, which is caused by light reflection and scattering inside the lens. For dehazing, we target at removing the hazy layer. The hazy images suffer from contrast loss caused by light scattering, reflection and attenuation of particles in the air. We show the extension results in Figure 7. Our trained model can achieve image enhancement by removing undesirable layers from the input images for flare removal and dehazing. More extension results are provided in the supplement.

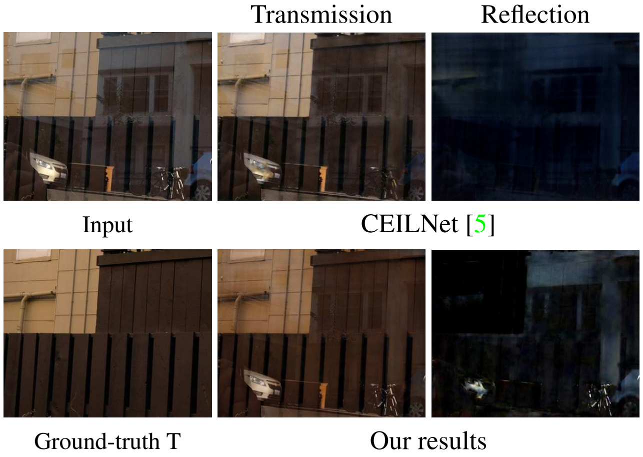


Figure 8: A challenging case with sharp reflection. Our method produces better reflection separation results than CEILNet, but is not able to remove reflection completely.

8. Discussion

We presented an end-to-end learning approach for single image reflection separation with perceptual losses and a customized exclusion loss. To decompose an image into the transmission and reflection layers, we found it effective to train a network with combined low-level and high-level image features. In order to evaluate different methods on real data, we collected a new dataset of real-world images for reflection removal that contains ground-truth transmission layers. We additionally extend our approach to two other photo enhancement applications to show generality of our approach for layer separation problems.

Although our reflection separation model outperforms state-of-the-art approaches on both synthetic and real images, we believe the performance can be further improved in the future. Figure 8 illustrates one challenging scenario where the reflection layer is almost as sharp as the transmission layer in a real-world image. We hope our model and dataset will inspire subsequent work on reflection separation and the challenging scenarios. Our dataset and code will be made publicly to facilitate future research.

9. Acknowledgement

We thank You Zhang for great help collecting the reflection dataset. We also thank Yichao Zhou and Daniel Seita for constructive writing feedback. This work is supported by UC Berkeley EECS departmental fellowship and hardware donations from NVIDIA.

References

[1] A. Agrawal, R. Raskar, S. K. Nayar, and Y. Li. Removing photography artifacts using gradient projection and flash-exposure sampling. *TOG*, 2005. 2

[2] N. Arvanitopoulos, R. Achanta, and S. Süsstrunk. Single image reflection suppression. In *CVPR*, 2017. 2

[3] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 4, 6

[4] Q. Chen, J. Xu, and V. Koltun. Fast image processing with fully-convolutional networks. In *ICCV*, 2017. 3

[5] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *ICCV*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[6] K. Gai, Z. Shi, and C. Zhang. Blind separation of superimposed moving images using image statistics. *IEEE PAMI*, 34, 2012. 2

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

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 4

[9] X. Guo, X. Cao, and Y. Ma. Robust separation of reflection from multiple images. In *CVPR*, 2014. 1, 2

[10] B.-J. Han and J.-Y. Sim. Reflection removal using low-rank matrix completion. In *CVPR*, 2017. 2

[11] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 3

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 4, 6, 8

[13] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4

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

[15] N. Kong, Y.-W. Tai, and J. S. Shin. A physically-based approach to reflection separation: from physical modeling to constrained optimization. *PAMI*, 2014. 2

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 4

[17] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE PAMI*, 2007. 2

[18] A. Levin, A. Zomet, and Y. Weiss. Learning to perceive transparency from the statistics of natural scenes. In *NIPS*, 2003. 2

[19] A. Levin, A. Zomet, and Y. Weiss. Separating reflections from a single image using local features. In *CVPR*, 2004. 2

[20] Y. Li and M. S. Brown. Exploiting reflection change for automatic reflection removal. In *CVPR*, 2013. 2

[21] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014. 1, 2, 4, 6, 7

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[23] B. Sarel and M. Irani. Separating transparent layers through layer information exchange. *ECCV*, 2004. 2

[24] Y. Y. Schechner, N. Kiryati, and R. Basri. Separation of transparent layers using focus. *IJCV*, 2000. 2

[25] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman. Reflection removal using ghosting cues. In *CVPR*, 2015. 1, 3

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 3

[27] O. Springer and Y. Weiss. Reflection separation using guided annotation. *arXiv preprint arXiv:1702.05958*, 2017. 1, 2

[28] C. Sun, S. Liu, T. Yang, B. Zeng, Z. Wang, and G. Liu. Automatic reflection removal using gradient intensity and motion cues. In *Proceedings of the 2016 ACM on Multimedia Conference*, 2016. 2

[29] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *CVPR*, 2000. 2

[30] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot. Benchmarking single-image reflection removal algorithms. In *CVPR*, 2017. 3

[31] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot. Depth of field guided reflection removal. In *ICIP*, 2016. 2

[32] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman. A computational approach for obstruction-free photography. *ACM Trans. Graph.*, 34(4), 2015. 1, 2

[33] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3

[34] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for neural networks for image processing. *IEEE Trans. Computational Imaging*, 2017. 8

CVPR2018 Paper Translation

姓名：　　廖梓涵

学号：　　2015302350

班号：　　10011502

具有感知损失的单图像反射分离

Xuaner Zhang
UC Berkeley

Ren Ng
UC Berkeley

Qifeng Chen
Intel Labs

摘要

我们提出了一种将光反射从单个图像中分离出来的方法。该方法使用一个完全卷积的网络训练端到端的损失，利用低水平和高水平的图像信息。我们的损失函数包括两个感知损失：视觉感知网络的特征损失和编码传输层图像特征的对抗性损失。我们还提出了一种新的排斥损失，它强制像素级层分离.我们创建了一个真实世界图像的数据集，其中包含了光反射和相应的地面真相传输层，用于定量评估和模型训练。通过综合定量实验验证了该方法的有效性，结果表明，该方法在PSNR、Ssim和感知用户研究等方面均优于现有的光反射方法。我们还将我们的方法扩展到另外两个图像增强任务，以演示我们的方法的通用性。

1. 导言

在现实世界中，窗户和眼镜上的光反射随处可见，但在照片中通常是不受欢迎的。用户通常希望通过删除图像中的光反射来提取隐藏的干净传输图像。例如，我们可能被诱惑通过水族馆玻璃或摩天大楼窗户拍照，但光反射往往会损害图像质量。从单个图像中删除光反射可以使我们以更好的感知能力恢复视觉内容。因此，将光反射层和传输层从图像中分离-光反射分离问题-是计算机视觉中一个活跃的研究领域。

设I ∈ RM×n×3为输入图像，再进行光反射扫描。我可以近似地建模为传输层T和光反射层R：i=TR之和，我们的目标是恢复给定的传输层T，这是一个不需要附加约束或先验的不适定问题。

由于光反射选择分离问题不合适，以前的工作通常需要额外的输入图像和精心制作的优先级。先前的一系列研究使用多幅图像作为输入，或者需要明确的用户指导[9，27，32]。然而，在实践中并不总是可以获得多种图像，而且用户指导也不方便，而且容易出错。研究人员提出了从单个图像中删除光反射的方法[25，21]，但这些方法依赖于手工制作的先验信息，如鬼提示和相对平滑-这可能并不能概括到所有具有光反射选择的图像。最近，CEILNet[5]使用一个深层神经网络来训练一个颜色和边缘损失较低的模型，但是这种方法并不能直接使模型学习高级语义，这对于去除光反射选择是非常有用的。当图像中存在颜色模糊或模型需要“识别”对象时，低层信息是用于光反射分离的。例如，在图1中，我们经过感知损失训练的模型可能已经学会了灯和脸的表示，从而正确地从输入图像中删除了它们，而CEILNet却没有这样做。

在本文中，我们提出了一个具有感知损失的全卷积网络，它既编码低层次图像信息，又编码高层次图像信息。我们的网络以一幅图像作为输入，直接合成两幅图像：光反射层和传输层。我们进一步提出了一种新的排斥损失，在像素级有效地加强了传输分离和光反射。为了深入评估和训练不同的方法，我们建立了一个包含真实世界图像和地面真相传输图像的数据集。我们的数据集涵盖各种自然环境，包括室内和室外场景。我们还使用这个真实的数据集来定量地比较我们的方法和以前的方法。总之，我们的主要贡献是：

- 我们建议使用具有感知损失的深度神经网络来进行单个图像反射分离。 我们通过两个具有不同级别的图像信息的损失来实施感知监督：来自视觉感知网络的特征丢失，以及用于重新定义输出传输层的对抗性损失。

- 我们提出了一个精心设计的排除损失算法，强调在梯度域中要分离的层的独立性

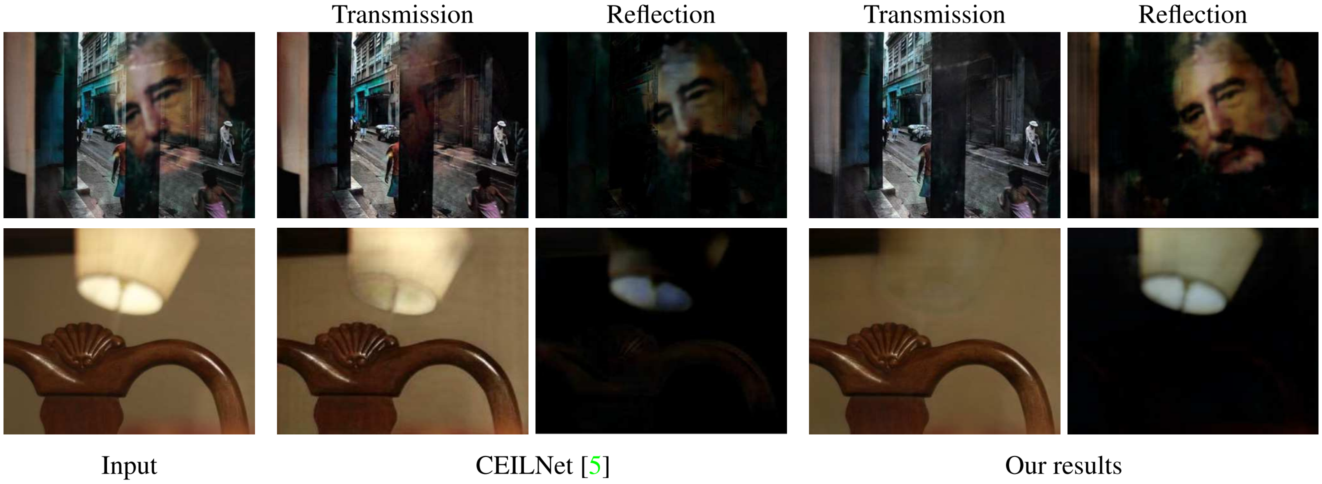


图1：CEILNet [5]的结果和我们对真实世界图像的处理方法。 顶行显示来自CEILNet数据集的真实图像，其中一个窗口反映了人脸的海报；底行显示我们自己拍摄的图像，其中一盏灯作为反射。 从左到右：输入图像，CEILNet结果和我们的结果。 请注意，我们培训的方法是学习低级和高级图像统计数据，成功地消除了面部和灯泡的反射层，而CEILNet却没有。

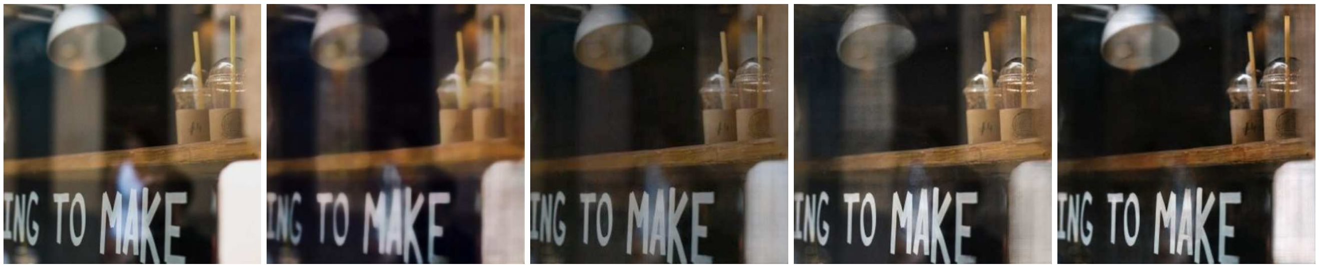
- 我们使用相应的地面实况传输层构建真实世界图像的数据集，以便反射消除。 这个新数据集可以对我们的方法和现有算法进行定量评估和比较。
- 我们对实际数据和合成数据的广泛实验表明，我们的方法在SSIM，PSNR和亚马逊机械土耳其人的感知用户研究方面优于最先进的方法。 我们训练的反射分离模型可以直接应用于另外两个图像增强任务，即移除和去雾。

2. 相关工作

多图像方法。由于光反射选择分离问题是不适定的，以往的大部分工作都是用多个输入图像来解决这个问题。这些多图像处理方法经常使用运动提示来分离传输层和重光反射层[32、9、20、28、23、6、29、10]。运动线索要么是从标定的相机中推断出来的，要么是假设背景和光反射对象具有很大不同的运动光反射屏蔽的运动视差。其他一些多图像方法包括使用光反射灰分和无光反射灰分图像对来改进光反射灰分图像，并删除了光反射图像[1]。谢克纳等人。[24]使用具有不同焦点设置的图像序列，将分层和深度估计分开。Kong等人[15]利用极化的物理性质，利用角fi透镜拍摄的多幅极化图像进行fi和最佳分离。最近，HANN和Sim[10]解决了多幅玻璃图像的玻璃再扫描光反射去除问题，假设背景图像中的梯度fiLD几乎是恒定的，而光反射成像中的梯度fiLD变化更大。虽然多图像方法在去除光反射图像方面表现出了良好的性能，但捕捉多幅图像有时是不可能的，例如，这些方法不能应用于现有的或遗留的照片。

单图像方法。另一项工作是考虑使用带有前置fiNed优先级的单个图像。一个广泛应用的先验是自然图像梯度稀疏度[19，18]到fi和最小边缘和角落进行层分解。梯度稀疏先验还与最优和最小用户辅助一起探索，以更好地指导不适定分离问题
[17，27]。Arvanitopoulos等人最近的工作。[2]利用梯度稀疏约束，结合Laplacian空间中的数据fiDelity项来抑制光反射选择。然而，所有这些方法都依赖于低层次的启发式方法，并且在需要对图像进行高层次理解的情况下受到限制。

光反射选择分离的另一个前提是，光反射层经常脱离焦点，并呈现平滑的状态。这是由Li和Brown[21]明确提出的一个优化目标，在这个目标中，他们惩罚了大的光反射梯度。虽然相对平滑的假设是有效的，但当光反射层具有高对比度时，它们的公式就会破裂。万等人[31]提出了一种改变这种平滑性的方法，在此之前，使用fiLD的深度作为边缘标记和层分离的指导。此外，Shih等人。[25]集中讨论了光反射选择有鬼效应的问题的一个子集，并使用估计的卷积核来优化去除光反射。



(a) Input　(b) Without L_{feat}　(c) Without L_{adv}　(d) Without L_{excl}　(e) Complete model

图2：在真实世界图像上评估的三种感知损失函数的视觉比较。 在（b）中，我们将Lfeat替换为图像空间L1损失并观察到过度平滑的输出。 （c）显示没有Ladv的颜色退化和明显残差的伪影。 在（d）中，Lexcl的缺乏使预测的传输具有不希望的反射残差。 我们在（e）中的完整模型能够产生更好和更清晰的预测。

FAN等人[5]最近提出了一种深度学习网络，即级联边缘和图像学习网络(CEIL-net)，用于去除光反射。他们将光反射去除描述为一个边缘单纯fi阳离子任务，并学习一个中间边缘映射来引导层分离。CEILNet是纯粹与低水平损失的训练结合在颜色空间和梯度域的差异。CEILNet和我们的网络的主要区别在于它们在训练中没有明确地利用知觉信息。

基准数据集最近，WAN等人提出了一个基准数据集，用于删除光反射。作者通过对日常物品和明信片的成像，收集了1500个受控实验室环境中40个场景的真实图像，以及3个不同眼镜在自然环境中的100个场景。然而，数据集在提交时尚未公开发布。为了对真实场景中的不同模型进行定量评价，我们在自然场景环境中采集了110个真实图像的数据集。

3. 概述

给出了图像I∈[0，1]m×n×3的光反射，利用单个网络f(i； θ)=(FT(I； θ)，Fr(I； θ)将I分解为传输层FT(I； θ)和光反射层Fr(I； θ)，其中θ是网络权值.我们在数据集D={(i，T，R)}上训练网络f，其中I是输入图像，T是I的传输层，R是I的光反射层。

我们的损失函数包含三个项：特征空间中比较图像的特征丢失Lfeat和用于真实图像refinement的对抗性损失LADV，强制梯度域中传输层和光反射层分离的排除损失Lexl。我们的总损失函数是

$$L(\theta) = w_1 L_{\text{feat}}(\theta) + w_2 L_{\text{adv}}(\theta) + w_3 L_{\text{excl}}(\theta), \qquad (1)$$

其中我们设置W1=0.1，w2=0.01和w3=1来平衡每个项的权重。

理想的光反射分离模型应该能够理解图像中的内容。为了训练我们的网络对输入图像的语义理解，我们从VGG-19[26]网络中提取在ImageNet数据集[22]上预先训练过的特征，从而形成超列特征[11]。使用超列特性的Benefit是通过抽象大型数据集(如ImageNet)的视觉感知的有用特性来增强输入。给定像素位置上的超列功能是跨该位置的网络的选定层的激活单元的堆栈。在此，我们对预先训练的VGG-19网络中的“卷积1 2”、“卷积2”、“conc 3 2”、“conf 4 2”和“conf 5 2”进行了采样。超列功能共有1472个维度。我们将输入图像i与它的超列特性连接起来，作为f的增广输入。

我们的网络f是一个完全卷积的网络，其网络结构与上下文聚合网络[33，4]相似。我们的网络具有513×513的光反射接收能力，可以有效地聚合全球图像信息。f的first层为1×1卷积，将特征维数(1472 3)降至64。以下8层为3×3膨胀卷积。扩张率为1~128。所有中间层都有64个特征通道。对于最后一层，我们使用线性变换在RGB颜色空间中合成2幅图像。

我们从CEILNet数据集[5]和我们收集的真实世界数据集中，对可公开使用的、合成的和真实的图像进行了不同的评价。我们将我们的方法与最先进的光反射去除方法CEILNet[5]、基于优化的方法[21]和图像翻译通用框架Pix2pix[12]进行了比较。

4. 训练

4.1. 特征损失

我们使用一个特征损失来测量我们预测的传输层和特征空间中的地面真相传输之间的差异。如图1中的上述观察所示，关于场景的语义推理将有利于光反射执行重新删除光反射的任务。将感知网络中的低层次和高层次特征结合在一起的特征丢失将为我们的目标服务。特征丢失也已成功地应用于其他任务，如图像合成和风格转换[3，7，16，13]。

在这里，我们通过预先训练的VGG-19网络Φ，通过提供预测的图像层和地面真相来计算特征损失。我们计算了Φ(FT(I；θ)和Φ(T)在选定的特征层中的L1差：其中Φl表示VGG-19网络中的层l。

$$L_{\text{feat}}(\theta) = \sum_{(I,T)\in\mathcal{D}} \sum_{l} \lambda_l \|\Phi_l(T) - \Phi_l(f_T(I;\theta))\|_1, \quad (2)$$

权值{λ1}用于平衡损失函数中的不同项。在VGG-19网络中，我们选择了“卷积1 2”、“卷积2”、“conc 3 2”、“conf 4 2”和“conf 5 2”。

4.2. 对抗损失

在我们的研究过程中，我们光反射和传输任务图像可以遭受不现实的颜色退化和不可接受的微妙残差，而不会造成对抗性损失。我们采用条件GaN[12]作为我们的模型。我们的发电机将是FT(I；θ)。我们的鉴别器的结构，表示为D，有4层，64个特征通道宽。鉴别器试图区分真实传输图像中的斑块和以I为条件的FT(I；θ)给出的块，目的是让网络D学习一个适当的损失函数，以进一步实现光反射层的分离，并将预测的传输层推向真实无光反射图像的区域。

鉴别器D的损失是：

$$\sum_{(I,T)\in\mathcal{D}} \log D(I, f_T(I;\theta)) - \log D(I,T), \quad (3)$$

其中D（I，x）输出x是给定输入图像I的自然透射图像的概率。然后我们的经验损失是：

$$L_{\text{adv}}(\theta) = \sum_{I\in\mathcal{D}} -\log D(I, f_T(I;\theta)). \quad (4)$$

我们优化了 $-\log D(I, f_T(I;\theta))$ 代替 $\log(1 - D(I, f_T(I;\theta)))$ 以得到更好的梯度性能 [8].

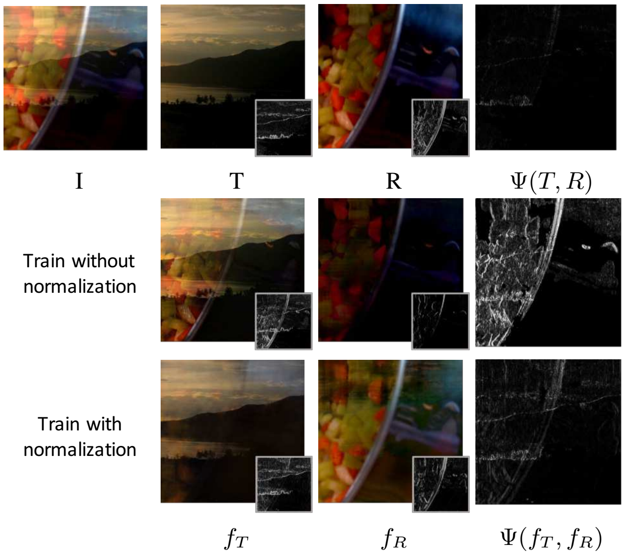


Figure 3: 训练与外梯度归一化的视觉比较.在中间两列中，每幅图像右下角的小窗口显示每幅图像的梯度大小。在最右边的列中，Ψ表示方程6中的归一化梯度积.从左到右的first行显示：输入、地面真实传输T、地面真相光反射选择R和Ψ.Ψ(T，R)接近于零，表明T和R的梯度光反射屏蔽不相关。中间一行显示在梯度光反射屏蔽中没有规范化的训练结果。我们观察到，没有归一化训练的光反射预测被严重抑制。底部行显示经过梯度归一化训练的结果，并能较好地重新进行光反射分离。

4.3. 排除损失

为了更好地分离光反射传输层和传输层，我们进一步提出了梯度中的排斥损耗。通过对两层边界的分析，探讨了这两层之间的关系。我们的主要观察是，传输的边缘和光反射层是不可能重叠的。I中的边缘应该是由T或R引起的，而不是两者兼而有之。因此，在梯度域中，我们将预测的传输层和光反射层之间的相关性最小化。我们将排除损失作为两层在多个空间分辨率下的归一化梯度场的乘积：

$$\begin{aligned} L_{\text{excl}}(\theta) &= \sum_{I\in\mathcal{D}} \sum_{n=1}^{N} \|\Psi(f_T^{\downarrow n}(I;\theta), f_R^{\downarrow n}(I;\theta))\|_F, \quad (5) \\ \Psi(T,R) &= \tanh(\lambda_T|\nabla T|) \odot \tanh(\lambda_R|\nabla R|), \quad (6) \end{aligned}$$

请注意，归一化因子λT和λR在方程6中是关键的，因为传输层和光反射层可能包含不平衡梯度大小。

在实验中我们设：$N=3, \lambda_T = \sqrt{\frac{\|\nabla R\|_F}{\|\nabla T\|_F}}$, and $\lambda_R = \sqrt{\frac{\|\nabla T\|_F}{\|\nabla R\|_F}}$

光反射层既可以是低强度模糊的，也可以是由小梯度组成的，也可以是光反射等非常明亮的光，并在图像中形成最亮的斑点，从而产生高对比度的光反射，从而产生较大的梯度。对两个层的预测进行不平衡的更新，将导致∇、T_x和∇_R_x之间的比例差异。我们观察到，如果没有适当的归一化因子，网络会抑制梯度上升速率较小的层，使其接近于零。图3显示了有标准化和没有规范化的结果的可视化比较。

Lexl在像素级有效地分离传输层和光反射层。如果我们在我们的模型中禁用LESL，那么在输出传输映像中可能仍然可以看到一些剩余的光反射，如图2(D)所示。

4.4. 实施

给出了地面真实光反射层R，我们可以用R. 光反射层对Fr(I; θ)进行进一步的约束，使得光反射层通常不聚焦，因而模糊。我们只需在颜色空间中添加L1损失来约束Fr(i; θ):

$$L_R(\theta) = \sum_{(I,R)\in\mathcal{D}} \|f_R(I;\theta) - R\|_1. \quad (7)$$

我们通过对合成数据和真实数据的最小化(LLR)来训练网络f。请注意，当我们在现实世界的图像上进行训练时，我们禁用LR，因为精确估计R是很难的。我们试着计算R=i T，但是R有时包含伪影，因为当我暴露过高时，I=RT可能不起作用。

训练数据采用5000幅合成图像，从90幅随机分辨率为256~480 p的真实训练图像中提取500个图像补丁。为了进一步增加数据，我们随机调整图像块的大小，同时保持原来的高宽比。我们在NvidiaTitan X GPU上训练250个批次大小为1，并使用ADAM优化器[14]更新权重，其fi学习速率为10 4。

5. 数据集

5.1. 综合数据

为了用光反射方法生成合成图像，我们从Flickr中随机选取了5000对图像：一幅室外图像，一幅室内图像。

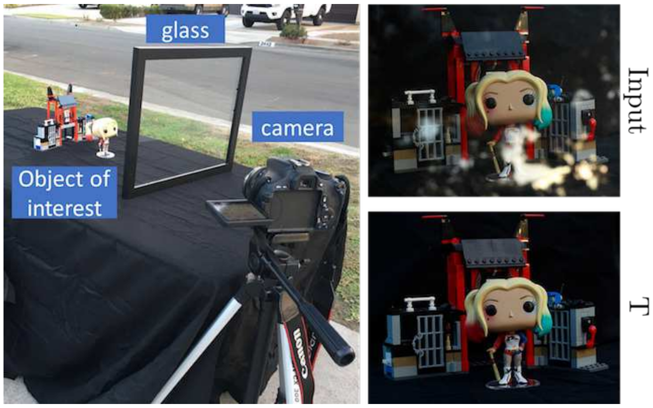


图4：真实数据收集设置和捕获的图像。 我们在静态场景中使用相同的相机设置捕获带有和不带有玻璃的两个图像。 从上到下的右列：具有反射的捕获图像和地面实况透射图像T.

我们使用一个图像(室内或室外)作为传输层，另一个图像作为光反射检测层。我们假设透射层和再光反射层位于不同的焦平面上，使这两层呈现出明显不同的模糊度。在现实摄影中，这是一个有效的假设，其中感兴趣的对象(例如，通过博物馆窗口的艺术品)通常位于传输层，并设置为焦点。另外，通过大光圈射击可以故意模糊光反射检测。我们利用这个假设来创建一个合成数据集，将一个具有随机核大小的高斯平滑核在3到17像素范围内应用于光反射检测图像。

我们的图像合成方法类似范等人提出的方法。[5]但我们的前向模型有以下不同之处。我们从图像中去除伽马校正，并在线性空间中操作，以更好地逼近图像的物理形成。我们用变分法代替了R上的光反射星强度衰减，因为我们观察到真实图像中的强度衰减可能比透射层具有相似或更高的强度水平。我们在光反射检测层中采用了以随机位置为中心的微缩体，它模拟了摄像机从斜角度观察光反射横线时的场景。

5.2. 真实数据

在开展这项工作时，没有公开可用的基准与地面真相传输，以评估不同的光反射删除方法对实际数据。我们收集了110个真实图像对的数据集：带重光反射的图像及其相应的地面真实传输图像。

	Synthetic		Real	
Method	SSIM	PSNR	SSIM	PSNR
Input	0.689	15.09	0.697	17.66
Pix2pix [12]	0.583	14.47	0.648	16.92
Li and Brown [21]	0.742	15.30	0.750	18.29
CEILNet [5]	0.826	20.47	0.762	19.04
Ours	**0.853**	**22.63**	**0.821**	**21.30**

光反射扫描图像采用表1：我们的方法与其他3种方法的定量比较结果。我们对CEILNet[5]提供的合成数据和我们的真实图像测试集进行了评估。我们还提供了一个将输入图像作为结果传输图像的微不足道的基线。

佳能600 D相机在三脚架与便携式玻璃在相机前面。当便携式玻璃被移除时，地面真相传输层被捕获。每个图像对都是在相同的曝光设置下拍摄的。我们的数据捕获设置如图4所示。我们捕获数据集时需要注意以下事项：

- 环境：室内和室外
- 照明条件：天窗，阳光和白炽灯
- 摄像机视角：前视图和斜视图
- 相机光圈（影响反射模糊）：$f/2.0$ — $f/16$.

我们将数据集随机分成训练集和测试集。我们从90幅训练图像中提取500个斑块进行训练，并使用20幅图像进行定量评价。

6. 实验

6.1. 与以往工作的比较

我们将我们的模型与CEILNet[5]、Li和Brown[21]和Pix2pix[12]的分层分离方法进行了比较。我们评估了来自CEILNet数据集的公开合成图像[5]和来自我们真实世界数据集测试集的真实图像的不同方法。

我们的模型只在我们生成的合成数据集和我们的真实世界数据集的训练集上进行训练。对于CEILNet，我们在CEILNet合成图像上评估它的预训练模型。为了评估CEILNet对我们的真实数据，我们光反射Ne调优它的模型与我们真正的训练映像(否则它表现不佳)。我们用给定的缺省参数对Li和Brown[21]的方法进行了评价。Pix2pix是一种通用的图像翻译模型，我们在生成的合成数据集和所收集的真实数据集的训练集上对其模型进行训练。

定量结果如表1所示。计算了不同传输方法的结果传输图像与地面真实传输层之间的PSNR和Ssim。与以往的合成和实际数据相比，我们展示了较强的定量性能。

	Preference rate
Ours>CEILNet [5]	84.2%
Ours>Li and Brown [21]	87.8%

表2：用户研究结果。 偏好率显示用户更喜欢我们的结果的比较百分比。

我们还按照Chen和Koltun[3]的协议，对亚马逊机械土耳其语进行了用户研究。在用户研究过程中，每个用户都会得到一幅输入的真实世界图像，其中包含光反射选择，我们的预测传输图像，以及通过同一行基线预测的传输图像。然后，用户需要选择一个输出图像，该输出图像更接近于两个预测的传输图像之间的输入图像的无光反射选择版本。我们的数据集和CEILNet数据集有80幅真实图像可供比较.结果见表2。84.2%的与CEILNet的比较和87.8%的与Li和Brown的比较结果表明，我们的结果包含较少的光反射。结果表明，光反射不可行，P<10 3，20名用户参与了用户研究。

更多的实验细节和结果报告在补编中。

	Synthetic		Real	
Method	SSIM	PSNR	SSIM	PSNR
Ours w/o L_{feat}	0.683	18.24	0.743	19.07
Ours w/o L_{adv}	0.818	20.80	0.793	21.12
Ours w/o L_{excl}	0.796	19.58	0.802	20.22
Ours L_{adv}-only	0.765	18.05	0.782	19.52
Ours complete	**0.853**	**22.63**	**0.821**	**21.30**

表3：我们方法的多个消融模型中合成和实际图像的定量比较。 我们重新移动三个损失中的每一个并评估重新训练的模型。'我们只有Ladv'表示我们的训练方法只有对抗性的损失。 我们的完整模型在合成和实际数据上都表现出更好的性能。 我们评估了CEILNet [5]提供的合成数据，以及5.2节中描述的实际测试图像。

6.2. 定性结果

我们在图5和图6中给出了不同方法的定性结果，分别对来自我们的数据集的真实图像(包含地面真相)和CEILNet[5](没有地面真相)的真实图像进行了评估。

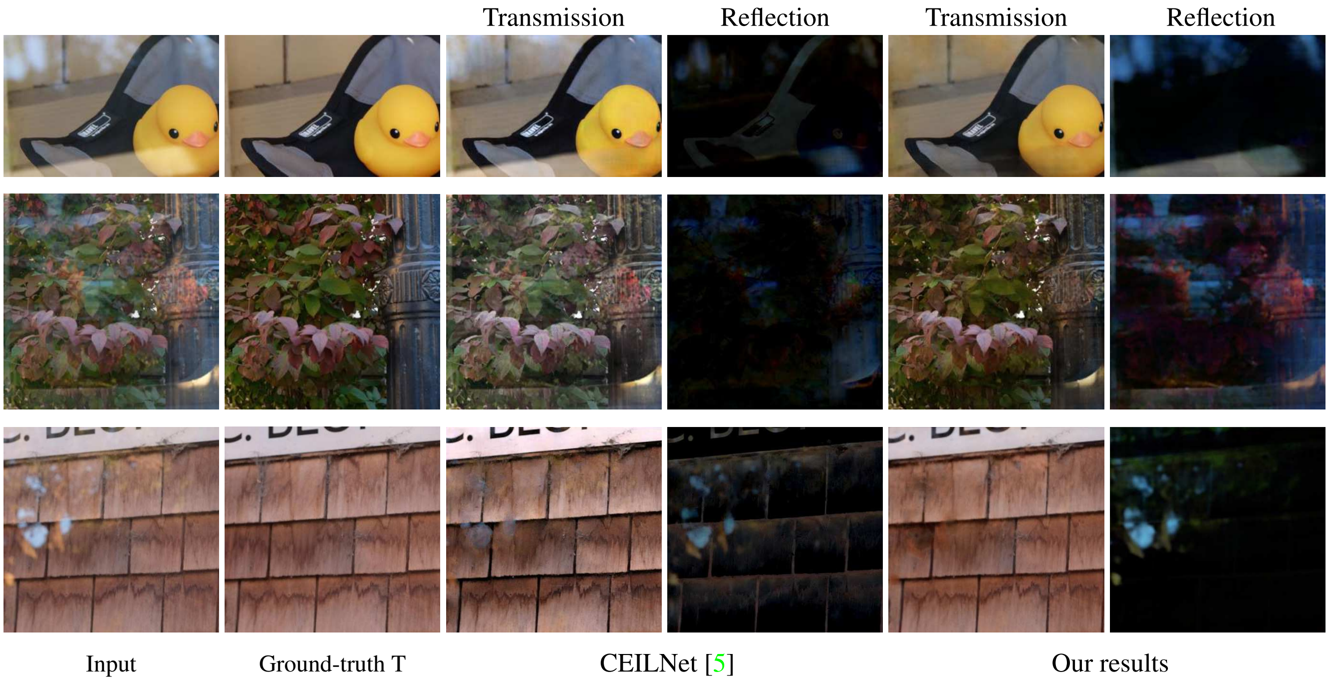


图5：CEILNet [5]和我们的方法之间的视觉结果比较，根据第5.2节中描述的数据集的真实图像进行评估。 从左到右：输入，地面实况传输层，CEILNet [5]预测和我们的预测。 请注意，我们的方法在传输层和反射层中产生更好，更清晰的预测。 补充中提供了额外的结果。

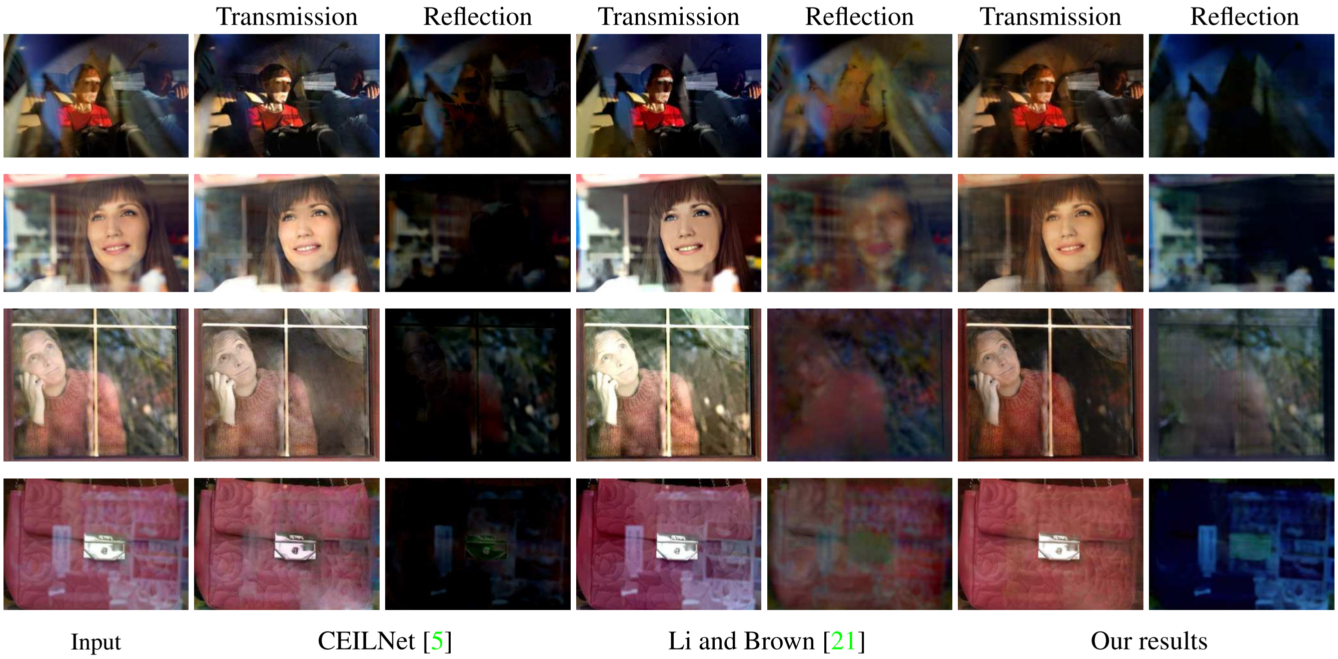


图6：CEILNet [5]，Li和Brown [21]以及我们的方法之间的定性比较，在CEILNET数据集中的真实图像上进行了评估。 请注意，即使我们对实际数据的反射层没有监督，但我们的方法也预测了更清晰的反射层。 补充剂中提供了其他结果。

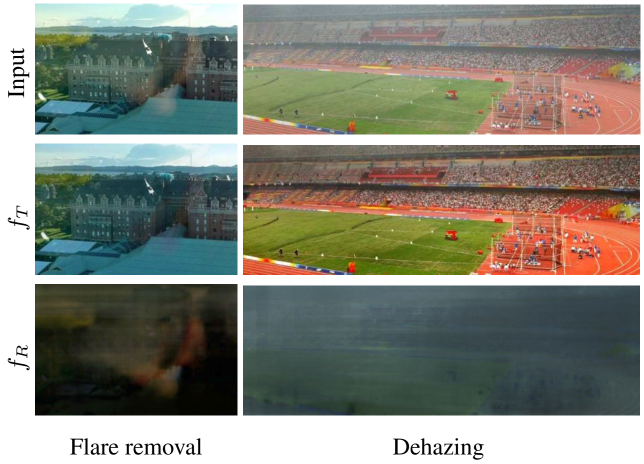


图7：相机flare上的扩展应用程序是移除和图像去雾。对于每列，从上到下：输入，我们预测的增强层，我们预测的删除层。

6.3. 受控实验

为了分析每一种损失对我们网络的光反射性能的贡献，我们在合并的目标中消除或替换了每个损失，并对网络进行了重新训练。图2显示了一个可视化的比较。当我们用颜色空间中的L1损失代替特征损失Lfeat时，输出图像往往过于平滑；类似的观察在[34，12]中也得到了讨论。在没有Lexl的情况下，我们注意到在传输预测中可能会出现光反射检测层的可见内容。如(E)所示，对抗性的光反射损失有助于恢复更清洁和更自然的结果。

定量结果如表3所示。我们还分析了与条件GaN[12]相似的只有对抗性损失的模型的性能。

7. 扩展

我们演示了两个额外的图像增强应用，光反射是去除和解除，使用我们训练的模型删除一个不想要的层。请注意，我们直接应用我们训练过的光反射选择去除模型，没有训练或光反射重新调优的任何光反射是删除或解除数据集。这两个任务可以看作是层分离问题，类似于光反射选择分离。对于光反射是去除的，我们的目的是去除透镜光反射的光学伪影，这是由透镜内部的光重反射和散射引起的。对于去雾，我们的目标是清除迷雾层。模糊图像由于光散射、光反射散射和粒子在空气中的衰减而造成对比度损失。

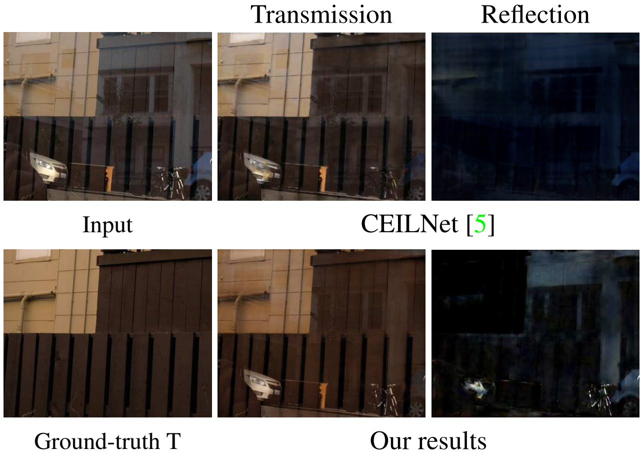


图8：具有急剧反射的具有挑战性的案例。 我们的方法比CEILNet产生更好的反射分离结果，但不能完全消除反射。

我们在图7中显示了扩展结果。我们的训练模型可以通过从输入图像中去除不需要的层来实现图像增强，用于光反射去除和去噪。补充部分提供了更多的扩展结果。

8. 讨论

提出了一种基于感知损失和自定义排除损失的单图像光反射分离的端到端学习方法。为了将图像分解为传输层和光反射检测层，我们发现对低层次和高层次图像特征相结合的网络进行训练是有效的。为了评估对真实数据的不同处理方法，我们收集了一个新的真实世界图像数据集，用于去除包含地面真相传输层的光反射。此外，我们还将我们的方法扩展到另外两个照片增强应用程序，以显示我们的方法对于层分离问题的通用性。

虽然我们的光反射选择分离模型在合成图像和真实图像上都优于现有的方法，但我们相信在未来的性能上还可以进一步提高。图8展示了一个具有挑战性的场景，其中光反射检测层几乎与真实世界图像中的传输层一样尖锐。我们希望我们的模型和数据集能够对光反射分离和具有挑战性的场景的后续工作有所启发。我们的数据集和代码将公开发布，以便于今后的研究。

9. 感谢

我们感谢您对收集光反射数据集的巨大帮助。我们也感谢周宜超和赛塔的建设性写作反馈。这项工作得到加州大学伯克利分校EECS部门奖学金和NVIDIA公司的硬件捐赠的支持.

参考

[1] A. Agrawal, R. Raskar, S. K. Nayar, and Y. Li. Removing photography artifacts using gradient projection and flash-exposure sampling. *TOG*, 2005. 2

[2] N. Arvanitopoulos, R. Achanta, and S. Süsstrunk. Single image reflection suppression. In *CVPR*, 2017. 2

[3] Q. Chen and V. Koltun. Photographic image synthesis with cascaded refinement networks. In *ICCV*, 2017. 4, 6

[4] Q. Chen, J. Xu, and V. Koltun. Fast image processing with fully-convolutional networks. In *ICCV*, 2017. 3

[5] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *ICCV*, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[6] K. Gai, Z. Shi, and C. Zhang. Blind separation of superimposed moving images using image statistics. *IEEE PAMI*, 34, 2012. 2

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

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 4

[9] X. Guo, X. Cao, and Y. Ma. Robust separation of reflection from multiple images. In *CVPR*, 2014. 1, 2

[10] B.-J. Han and J.-Y. Sim. Reflection removal using low-rank matrix completion. In *CVPR*, 2017. 2

[11] B. Hariharan, P. A. Arbeláez, R. B. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In *CVPR*, 2015. 3

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 4, 6, 8

[13] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 4

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

[15] N. Kong, Y.-W. Tai, and J. S. Shin. A physically-based approach to reflection separation: from physical modeling to constrained optimization. *PAMI*, 2014. 2

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, 2016. 4

[17] A. Levin and Y. Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE PAMI*, 2007. 2

[18] A. Levin, A. Zomet, and Y. Weiss. Learning to perceive transparency from the statistics of natural scenes. In *NIPS*, 2003. 2

[19] A. Levin, A. Zomet, and Y. Weiss. Separating reflections from a single image using local features. In *CVPR*, 2004. 2

[20] Y. Li and M. S. Brown. Exploiting reflection change for automatic reflection removal. In *CVPR*, 2013. 2

[21] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014. 1, 2, 4, 6, 7

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015. 3

[23] B. Sarel and M. Irani. Separating transparent layers through layer information exchange. *ECCV*, 2004. 2

[24] Y. Y. Schechner, N. Kiryati, and R. Basri. Separation of transparent layers using focus. *IJCV*, 2000. 2

[25] Y. Shih, D. Krishnan, F. Durand, and W. T. Freeman. Reflection removal using ghosting cues. In *CVPR*, 2015. 1, 3

[26] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015. 3

[27] O. Springer and Y. Weiss. Reflection separation using guided annotation. *arXiv preprint arXiv:1702.05958*, 2017. 1, 2

[28] C. Sun, S. Liu, T. Yang, B. Zeng, Z. Wang, and G. Liu. Automatic reflection removal using gradient intensity and motion cues. In *Proceedings of the 2016 ACM on Multimedia Conference*, 2016. 2

[29] R. Szeliski, S. Avidan, and P. Anandan. Layer extraction from multiple images containing reflections and transparency. In *CVPR*, 2000. 2

[30] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, and A. C. Kot. Benchmarking single-image reflection removal algorithms. In *CVPR*, 2017. 3

[31] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot. Depth of field guided reflection removal. In *ICIP*, 2016. 2

[32] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman. A computational approach for obstruction-free photography. *ACM Trans. Graph.*, 34(4), 2015. 1, 2

[33] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016. 3

[34] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for neural networks for image processing. *IEEE Trans. Computational Imaging*, 2017. 8