# Language-guided Image Reflection Separation

Haofeng Zhong[1,2,3 #] Yuchen Hong[1,2 #] Shuchen Weng[1,2] Jinxiu Liang[1,2] Boxin Shi[1,2,3 *]

[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3] AI Innovation Center, School of Computer Science, Peking University
{hfzhong, shuchenweng, cssherryliang, shiboxin}@pku.edu.cn, yuchenhong.cn@gmail.com

## Abstract

*This paper studies the problem of language-guided reflection separation, which aims at addressing the ill-posed reflection separation problem by introducing language descriptions to provide layer content. We propose a unified framework to solve this problem, which leverages the cross-attention mechanism with contrastive learning strategies to construct the correspondence between language descriptions and image layers. A gated network design and a randomized training strategy are employed to tackle the recognizable layer ambiguity. The effectiveness of the proposed method is validated by the significant performance advantage over existing reflection separation methods on both quantitative and qualitative comparisons.*

## 1. Introduction

When photographing through transparent materials like glass windows or showcases, the presence of reflections can significantly degrade the image quality of captured images and disrupt downstream computer vision tasks like face recognition [53] or depth estimation [2]. As an attractive topic in computational photography, reflection separation aims at decomposing the contaminated mixture image (denoted as $\mathbf{M}$) into two components that correspond to scenes located at different sides of the glass, *i.e.*, the reflection layer (denoted as $\mathbf{R}$) and the transmission layer (denoted as $\mathbf{T}$). Since reflection separation is a severely ill-posed problem, it is imperative to exploit effective priors or auxiliary information for distinguishing the two components.

The primary challenge of solving the reflection separation problem lies in the exploration of distinct clues for distinguishing transmission and reflection layers. Multi-image methods handle this problem by introducing additional constraints. Some of them acquire a series of mixture images in different viewpoints [32, 34, 45] to harness the distinct motions of the two layers, while others adopt specialized capturing setups to obtain complementary scene information [19, 27, 28, 36]. However, the specialized data capture requirements limit the application scope of these methods, especially for images downloaded from the Internet. Single-image methods attempt to tackle reflection separation by utilizing handcrafted priors derived from natural image statistics [30, 33, 44] or leveraging the modeling capacity of neural networks to learn content priors about reflections from a large scale of training data [10, 22, 31]. However, they are prone to fail due to the lack of auxiliary content information about transmission or reflection scenes for solving such a highly ill-posed problem. Recently, language descriptions have shown their effectiveness in providing content information for various vision tasks such as image editing [46, 47, 60, 61], semantic segmentation [57, 65], and image colorization [4–6, 62], which inspires us to think about: *Can we leverage the auxiliary content information brought by language descriptions to facilitate the reflection separation problem?*

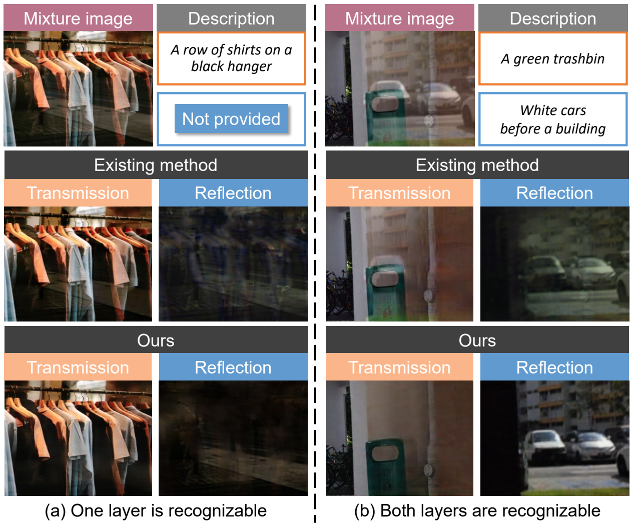

Figure 1. The recognizable layer ambiguity problem causes uncertain quantities of input language descriptions for language-guided image reflection separation. Given language descriptions of either (a) one layer or (b) both two layers, the proposed method achieves robust reflection separation compared with an existing reflection separation method [10].

[#] Equal contributions. [*]Corresponding author.

Since language can effectively convey humans' prior knowledge about the real world [8] and provide auxiliary information of image semantics [65], introducing language descriptions to guide the separation of reflection and transmission layers from mixture images merits exploration. However, leveraging language descriptions for reflection separation is non-trivial in three aspects: **1) Language-image modality inconsistency**. Language and images belong to different modalities, thus it is challenging to establish a cross-modality correspondence between the scene content information provided in language descriptions and the complex blended content present in mixture images. **2) Recognizable layer ambiguity**. Since the image content and brightness of reflection and transmission layers are different, the recognizable extents of them in mixture images are also uncertain. Specifically, as shown in Figure 1, it is possible that only one layer's content is recognizable (clothes in Figure 1(a)), or both two layers exhibit recognizable content (the trashbin and cars in Figure 1(b)), which leads to the difficulty of using uncertain quantities of language descriptions for separating mixture images in practice. **3) Language annotation deficiency**. All existing datasets for the reflection separation task only contain image data but no correlated language description is provided, raising the challenge for network training and evaluation.

In this paper, we introduce the concept of *language-guided image reflection separation* for the *first* time, which leverages flexible natural language to specify the content of one or two layers within a mixture image, relieving the ill-posedness of the reflection separation problem and maintaining a wide applicability for both live captured or online downloaded mixture images. We propose an end-to-end framework that employs adaptive global interaction modules to explore holistic language-image content coherence and utilizes specifically designed loss functions to constrain the correspondence between language descriptions and recovered image layers. A language gate mechanism and a randomized training strategy are designed to deal with the recognizable layer ambiguity problem. To address the language annotation deficiency, we synthesize the training dataset from paired image-language datasets [7, 66] and expand prevailing real reflection separation datasets [31, 49, 67] by manually adding language descriptions. Besides, we further construct a new dataset for visual quality evaluation by collecting mixture images from the Internet and captioning them with language descriptions for recognizable layers. Our contributions are summarized as follows:

- We present the first work that introduces language descriptions to guide the reflection separation task.
- We propose adaptive global interaction modules and language-image loss functions to tackle modality inconsistency.
- We design a language gate mechanism and a randomized training strategy to handle recognizable layer ambiguity.
- We build a dataset with language descriptions to facilitate language-guided image reflection separation.

## 2. Related work

**Single-image reflection separation** methods try to distinguish reflection and transmission layers using a single mixture image, which mainly relies on the assumption that the two layers have different distributions, *i.e.*, reflection layers are more likely to be blurry and appear with lower intensity compared with transmission layers. Conventional methods adopt handcrafted priors derived from natural image statistics in their optimization process, *e.g.*, the gradient sparsity [30], relative smoothness [33], ghosting cues [44], content prior [50], and penalty on the gradient of restored transmission layers [64].

Due to the tremendous progress in the field of deep learning, a series of single-image reflection separation methods concentrate on the improvement of learning strategies or network design, *e.g.*, predicting edges and images with a two-stage [11] or concurrent framework [51, 52], training with the perceptual loss [67], employing generative adversarial network [12] based models [38, 58], adopting iterative refinement strategies [10, 31, 63, 68], and leveraging the complementary two-stream architecture [22]. Meanwhile, research on data synthesis and image models is also ongoing to satisfy the data-driven needs of learning-based methods. Ma *et al.* [38] utilize generative adversarial networks for data generation while Wen *et al.* [59] synthesize mixture images with learned non-linear blending masks. Hu *et al.* [23] introduce a learnable residue term in the mixture image formation model to mitigate the non-linearity caused by the complex camera pipeline. Zheng *et al.* consider physical factors such as reflective amplitude coefficient maps [69] and the absorption effect [70] in the image formation process of mixture images. To facilitate network training and evaluation, researchers [31, 49, 67] also collect real data by using portable glass. Moreover, as a special form of images, panoramic images are introduced to relieve the content ambiguity in mixture images [13, 19, 21, 40]. We refer readers to [54] for a comprehensive and up-to-date survey on single-image reflection separation.

**Multi-image reflection separation** methods usually leverage the auxiliary information introduced by additional images and achieve more robust performance than single-image methods. Polarization-based methods [9, 26, 28, 36, 37, 39, 43] distinguish reflection and transmission layers by using images captured with different angles of polarizers or special polarization cameras. Flash-based methods [3, 18, 20, 27, 29] adopt active light sources to illuminate transmission scenes for obtaining reflection-free guidance. Motion-based methods [32, 34, 35, 45] utilize multiple images captured from different viewpoints to harness

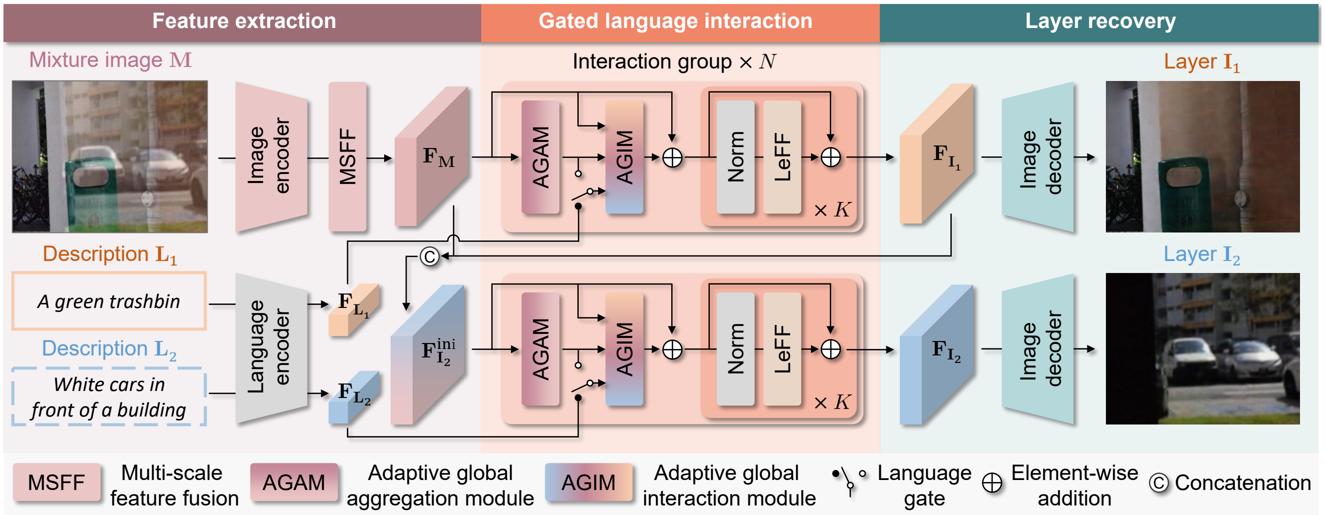
Feature extraction
Gated language interaction
Layer recovery
Mixture image M
Interaction group × N
Layer $\mathbf{I}_1$
Image encoder
MSFF
$\mathbf{F}_\mathrm{M}$
AGAM
AGIM
Norm
LeFF
× K
$\mathbf{F}_{\mathbf{I}_1}$
Image decoder
Description $\mathbf{L}_1$
A green trashbin
Description $\mathbf{L}_2$
White cars in front of a building
Language encoder
$\mathbf{F}_{\mathbf{L}_1}$
$\mathbf{F}_{\mathbf{L}_2}$
©
$\mathbf{F}^\mathrm{ini}_{\mathbf{I}_2}$
AGAM
AGIM
Norm
LeFF
× K
Layer $\mathbf{I}_2$
$\mathbf{F}_{\mathbf{I}_2}$
Image decoder
MSFF Multi-scale feature fusion
AGAM Adaptive global aggregation module
AGIM Adaptive global interaction module
Language gate
⊕ Element-wise addition
© Concatenation

Figure 2. The pipeline of the proposed language-guided image reflection separation framework, which extracts features from mixture images and available language descriptions (the description $L_2$ with dashed lines is possible to be set to null due to the recognizable layer ambiguity) by image and language encoders (in Sec. 3.1), aggregates global visual information by adaptive global aggregation modules (AGAM) and conducts progressive interactions to exploit distinctive image features with gated language guidance by adaptive global interaction modules (AGIM) (in Sec. 3.2), and recovers image layers by image decoders (in Sec. 3.3).

distinct motions of reflection and transmission layers. However, special data capture requirements significantly limit the application scope of these methods, especially for mobile devices or images downloaded from the Internet.

## 3. Proposed method

The pipeline of the proposed framework is illustrated in Figure 2. In this section, we start by introducing the feature extraction stage (in Sec. 3.1) that obtains multi-scale image features and global language features, the gated language interaction stage (in Sec. 3.2) that conducts progressive image-language global interactions to exploit distinctive image features and prevents unavailable language interactions with switchable gates, and the layer recovery stage (in Sec. 3.3) that reprojects features into the image space with a light-weight image decoder. Then we explain loss functions (in Sec. 3.4) employed for network optimization, especially a contrastive correspondence loss and a layer correspondence loss that constrains the network to construct correspondences between the language description and the corresponding layer under the disturbance of the other layer in a superimposed mixture image. Finally, we present our training strategy (in Sec. 3.5) which enables the network to be applicable for varying quantities of input language descriptions and jointly tackles the recognizable layer ambiguity problem with the gated network design.

### 3.1. Image and language feature extraction

The inputs of the proposed method consist of a mixture image $\mathbf{M}$ with two language descriptions $\{\mathbf{L}_i | i = 1, 2\}$ which corresponds to the two image layers. We specify that layer $\mathbf{I}_i$ corresponds to the description $\mathbf{L}_i$. However, due to the recognizable layer ambiguity that in certain cases only one layer of the mixture image is recognizable (usually the transmission layer), for such cases, we set $\mathbf{L}_1$ to be the available language description (for the recognizable transmission layer) and $\mathbf{L}_2$ to null (for the unrecognizable reflection layer) to ensure a unified input setting. Then given the input image and language descriptions, the feature extraction stage aims at obtaining the image feature $\mathbf{F}_\mathbf{M}$ with the image encoder and the multi-scale feature fusion process and extracting the global language feature $\mathbf{F}_{\mathbf{L}_i}$ for each description $\mathbf{L}_i$ via the language encoder for the subsequent interaction procedure.

**Image encoder.** Given a mixture image $\mathbf{M} \in \mathbb{R}^{H \times W \times 3}$, we employ a commonly-used vision backbone ResNet-50 [16] as our image encoder, whose last three layers (*i.e.*, an average pooling layer, a fully connected layer, and a softmax layer) are removed to fit our task. We utilize image features from the first five blocks of the image encoder to form a multi-scale feature pyramid $\{\mathbf{F}_{\mathbf{M}_i}\}_{i=1}^{5}$, where $\mathbf{F}_{\mathbf{M}_i} \in \mathbb{R}^{h_i \times w_i \times C_i}$, $h_i = H/2^i$ and $w_i = W/2^i$, $H$ and $W$ is the height and width of the mixture image, respectively, and $C_i$ is the dimension of the $i$-th extracted feature.

**Multi-scale feature fusion.** Obtaining the extracted feature pyramid, we first transform it into a hypercolumn feature $\mathbf{F}_\mathbf{M}^\mathrm{hyp} \in \mathbb{R}^{h \times w \times C^\mathrm{hyp}}$ [14] (where $h = H/2$, $w = W/2$, and $C^\mathrm{hyp} = \sum_{i=1}^{5} C_i$), which has been proved to be effective in fusing multi-scale contextual information for reflection separation [58, 67]. Considering the computational cost,

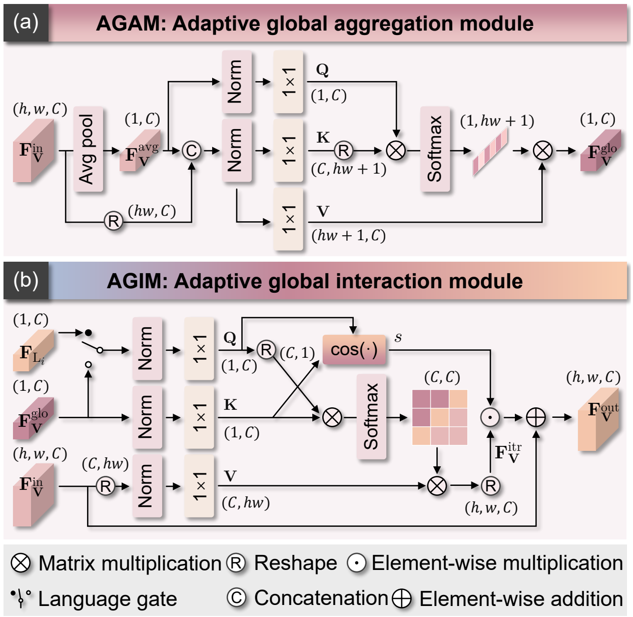
(a)
AGAM: Adaptive global aggregation module
$(h, w, C)$
$\mathbf{F}_{\mathbf{V}}^{\text{in}}$
Avg pool
$(1, C)$
$\mathbf{F}_{\mathbf{V}}^{\text{avg}}$
Norm
$1\times1$
Q
$(1, C)$
K
$(C, hw+1)$
V
$(hw+1, C)$
$(hw, C)$
Softmax
$(1, hw+1)$
$(1, C)$
$\mathbf{F}_{\mathbf{V}}^{\text{glo}}$
(b)
AGIM: Adaptive global interaction module
$(1, C)$
$\mathbf{F}_{\mathbf{L}_i}$
$\mathbf{F}_{\mathbf{V}}^{\text{glo}}$
$(h, w, C)$
$(C, hw)$
Q
$(C, 1)$
$\cos(\cdot)$
$s$
K
Softmax
$(C, C)$
V
$\mathbf{F}_{\mathbf{V}}^{\text{itr}}$
$(h, w, C)$
$\mathbf{F}_{\mathbf{V}}^{\text{out}}$
⊗ Matrix multiplication Ⓡ Reshape ⊙ Element-wise multiplication
Language gate Ⓒ Concatenation ⊕ Element-wise addition

Figure 3. The architecture of the (a) adaptive global aggregation module (AGAM) and (b) adaptive global interaction module (AGIM), which aggregates global contextual information of visual features and achieves feature channel rearrangement with gated language guidance, respectively.

we condense and refine the hypercolumn feature by a $1 \times 1$ convolutional layer with a GELU activation [17] followed by a locally-enhanced feed-forward (LeFF) block [56]. The final fused feature of the mixture image is denoted as $\mathbf{F}_{\mathbf{M}} \in \mathbb{R}^{h \times w \times C}$, which serves as the basis for the subsequent interaction and separation process.

**Language encoder.** Motivated by the rapid development of pre-trained large-scale vision-language models, we employ the language encoder from CLIP [42], which adopts a Transformer architecture [48] to extract language features and obtains a global contextual feature in the multi-modal embedding space by using layer normalization and linear projection layers. Given a language description $\mathbf{L}_i \in \mathbb{R}^L$, we obtain its corresponding global feature $\mathbf{F}_{\mathbf{L}_i} \in \mathbb{R}^C$ by leveraging the modeling capacity of the language encoder to encode the description, thus extracting the holistic contextual information of the corresponding image layer. Here $L$ denotes the length of the language description and $C$ is the feature dimension as the image feature.

## 3.2. Gated language interaction

The gated language interaction stage aims at leveraging the contextual information from available language descriptions to guide the separation of the corresponding layer feature $\mathbf{F}_{\mathbf{I}_i}$, which is composed of $2N$ cascaded interaction groups to separate image layer features successively. As shown in Figure 2, each group consists of an adaptive global aggregation module (AGAM) to gather global information of input visual features, a language gate to prevent detrimental guidance from unavailable descriptions, an adaptive global interaction module (AGIM) to conduct interactions using global features for exploring holistic image-language content correspondence, and $K$ normalization layers with LeFF blocks [56] for feature refinement. The former $N$ interaction groups are utilized for separating $\mathbf{F}_{\mathbf{I}_1}$ from $\mathbf{F}_{\mathbf{M}}$ and $\mathbf{F}_{\mathbf{L}_1}$, and the latter $N$ groups for separating $\mathbf{F}_{\mathbf{I}_2}$ from $\mathbf{F}_{\mathbf{I}_2}^{\text{ini}}$ (obtained by feeding the concatenation of $\mathbf{F}_{\mathbf{M}}$ and $\mathbf{F}_{\mathbf{I}_1}$ into a $1\times1$ convolutional layer) and $\mathbf{F}_{\mathbf{L}_2}$ (if $\mathbf{L}_2$ is available). We set $N = 4$ and $K = 2$ in practice. Details of the gated language interaction are described as follows.

**Adaptive global aggregation module (AGAM).** As the network structure shown in Figure 3(a), given an input visual feature $\mathbf{F}_{\mathbf{V}}^{\text{in}} \in \mathbb{R}^{h \times w \times C}$ with the spatial resolution of $h \times w$, AGAM is designed for adaptively obtaining a global feature $\mathbf{F}_{\mathbf{V}}^{\text{glo}} \in \mathbb{R}^C$ that aggregates the contextual information for the subsequent interaction. The visual feature $\mathbf{F}_{\mathbf{V}}^{\text{in}}$ is firstly averaged by an average pooling layer to obtain $\mathbf{F}_{\mathbf{V}}^{\text{avg}} \in \mathbb{R}^C$. Then the aggregation process is accomplished via a cross-attention mechanism which contains three linear projection layers with layer normalization [1] to conduct query, key, and value projections: $\mathbf{F}_{\mathbf{V}}^{\text{glo}} = \text{Softmax}(\mathbf{Q}\mathbf{K}^\top/\tau)\mathbf{V}$, where the query $\mathbf{Q} \in \mathbb{R}^C$ is projected from $\mathbf{F}_{\mathbf{V}}^{\text{avg}}$, and the key $\mathbf{K} \in \mathbb{R}^{(hw+1)\times C}$ and value $\mathbf{V} \in \mathbb{R}^{(hw+1)\times C}$ are from the concatenation of $\mathbf{F}_{\mathbf{V}}^{\text{in}}$ and $\mathbf{F}_{\mathbf{V}}^{\text{avg}}$, and $\tau$ is a learnable scaling factor to control the magnitude of the dot product of $\mathbf{Q}$ and $\mathbf{K}$ before applying the softmax function.

**Adaptive global interaction module (AGIM).** Inspired by existing reflection separation approaches [23, 58] which attempt to distinguish transmission and reflection layers in the feature space by feature channel rearrangement (*i.e.*, allocating distinct channels from mixture image features to the two layers), we propose to integrate language interactions at the feature channel level. To achieve this, as illustrated in Figure 3(b), we propose the adaptive global interaction module (AGIM), which employs the channel-wise cross-attention mechanism to conduct interactions between global language features and image features for channel rearrangement. The query, key, and value projections using linear projection layers with layer normalization [1] are conducted on the global language feature $\mathbf{F}_{\mathbf{L}_i}$, the global visual feature $\mathbf{F}_{\mathbf{V}}^{\text{glo}}$, and the visual feature $\mathbf{F}_{\mathbf{V}}^{\text{in}}$ to generate $\mathbf{Q} \in \mathbb{R}^C$, $\mathbf{K} \in \mathbb{R}^C$, and $\mathbf{V} \in \mathbb{R}^{C \times hw}$, respectively. Before the interaction, a language gate is designed to prevent impacts of unavailable language descriptions caused by the recognizable layer ambiguity problem that users may only input one description for the layer recognizable in the mixture image. Specifically, if the description $\mathbf{L}_i$ corresponding to the current global language feature $\mathbf{F}_{\mathbf{L}_i}$ is available (not set to null), the gate will feed $\mathbf{F}_{\mathbf{L}_i}$ into the following interaction process, otherwise the gate will feed $\mathbf{F}^{\text{glo}}_{\mathbf{V}}$, which turns the interaction process to be a channel-wise self-attention.

After filtering by the language gate, an interacted feature can be obtained through the channel-wise cross-attention: $\mathbf{F}^{\text{itr}} = \text{Softmax}(\mathbf{Q}^\top\mathbf{K}/\eta)\mathbf{V}$, where $\eta$ is a learnable scaling factor to control the magnitude of the attention map. Besides, to adaptively adjust the influence of language guidance based on the correspondence between image and language features, we define another scaling factor $s$ valued as the cosine distance between $\mathbf{Q}$ and $\mathbf{K}$ to multiply with the interacted feature $\mathbf{F}^{\text{itr}}$. Finally, the output of AGIM is obtained by a residual structure: $\mathbf{F}^{\text{out}}_{\mathbf{V}} = \mathbf{F}^{\text{in}}_{\mathbf{V}} + s\mathbf{F}^{\text{itr}}$, which integrates the contextual information from the language description and adjusts features for layer separation.

### 3.3. Layer recovery

After global language-image interaction and progressive refinement by the gated language interaction stage, we obtain layer features $\mathbf{F}_{\mathbf{I}_i}$ which integrates holistic contextual information from the corresponding language description $\mathbf{L}_i$ (if available). The layer recovery stage is dedicated to reconstructing each image layer $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ from its corresponding layer features $\mathbf{F}_{\mathbf{I}_i}$, which is achieved by using individual image decoders. In the image decoder, the layer feature is firstly upsampled by a transposed convolutional layer with a GELU activation [17], then refined by a residual block, and finally projected from the feature domain back into the image domain by a $1 \times 1$ convolutional layer with a sigmoid activation, thus accomplishing a precise layer recovery.

### 3.4. Loss functions

In this section, we introduce the contrastive correspondence loss $\mathcal{L}_{\text{ctr}}$ and the layer correspondence loss $\mathcal{L}_{\text{lcr}}$ for constraining the proposed method to establish the cross-modality correspondence between language descriptions $\mathbf{L}_i$ and corresponding image layers under the content disturbance from counter layers in mixture images. We also briefly describe the image layer loss $\mathcal{L}_{\text{img}}$ which consists of image- and feature-level loss functions for layer recovery. We denote the estimated image layers as $\tilde{\mathbf{I}}_i$ and ground truths as $\mathbf{I}_i$. Details of the loss functions are as follows.

**Contrastive correspondence loss.** CLIP [42] conducts contrastive language-image pre-training within batches which successes in establishing cross-modality correspondence between language descriptions and clean images, while our language-guided image reflection separation task requires finding the correspondence between the given language description and a certain layer in the mixture image with the interference of extraneous contents. To tackle the above issue, we propose a contrastive correspondence loss that conducts contrastive learning between image layers to establish correct cross-modality correspondence. Specifically, given a language description $\mathbf{L}_i$, the goal of the contrastive correspondence loss is to force the network to learn the relation that the corresponding image layer $\mathbf{I}_i$ is more relevant to $\mathbf{L}_i$ than the counter layer $\mathbf{I}_j$ ($j \neq i$), thus constraining the estimated image layer $\tilde{\mathbf{I}}_i$ to conform to the associated language description $\mathbf{L}_i$.

To measure the relevance between a language description $\mathbf{L}$ and an image layer $\mathbf{I}$, we define a feature-level similarity function $\mathcal{D}(\cdot)$ as:

$$\mathcal{D}(\mathbf{L}, \mathbf{I}) = \sigma(\Psi(\mathbf{F}^{\text{glo}}_{\mathbf{L}}, \mathbf{F}^{\text{glo}}_{\mathbf{I}})), \tag{1}$$

where $\sigma$ represents the sigmoid function, $\Psi$ represents the cosine distance, $\mathbf{F}^{\text{glo}}_{\mathbf{L}}$ is the global language feature obtained by the language encoder (in Sec. 3.1), and $\mathbf{F}^{\text{glo}}_{\mathbf{I}}$ is the global image feature produced by the AGAM (in Sec. 3.2). For an available language description $\mathbf{L}_i$, we calculate its contrastive correspondence loss as:

$$\mathcal{L}_{\text{ctr}}(\mathbf{L}_i, \tilde{\mathbf{I}}_i, \mathbf{I}_j) = -\log\left(\frac{\mathcal{D}(\mathbf{L}_i, \tilde{\mathbf{I}}_i)}{\mathcal{D}(\mathbf{L}_i, \tilde{\mathbf{I}}_i) + \mathcal{D}(\mathbf{L}_i, \mathbf{I}_j)}\right), \tag{2}$$

and we sum up the contrastive correspondence loss of each available language description as the final one.

**Layer correspondence loss.** We further define a layer correspondence loss to encourage the relevance between the language description $\mathbf{L}_i$ and the estimated image layer $\tilde{\mathbf{I}}_i$ to approach the relevance between $\mathbf{L}_i$ and the ground truth image layer $\mathbf{I}_i$:

$$\mathcal{L}_{\text{lcr}}(\mathbf{L}_i, \tilde{\mathbf{I}}_i, \mathbf{I}_i) = \left\|\mathcal{D}(\mathbf{L}_i, \tilde{\mathbf{I}}_i) - \mathcal{D}(\mathbf{L}_i, \mathbf{I}_i)\right\|_1, \tag{3}$$

where $\mathcal{D}(\cdot)$ is the same feature-level similarity function as in the contrastive correspondence loss. We also sum up the layer correspondence loss of each available language description for final supervision.

**Image layer loss.** To achieve high-fidelity recovery of image layers (*i.e.*, transmission and reflection layers), the proposed method is also optimized with loss functions following previous reflection separation methods [23, 52, 67]. Specifically, we utilize loss functions that conduct constraints on the visual quality of estimated images (*i.e.*, the pixel $\mathcal{L}_{\text{pix}}$, structural similarity $\mathcal{L}_{\text{ssim}}$, and perceptual loss $\mathcal{L}_{\text{per}}$) or exploit the inherent relationship between two layers (*i.e.*, the exclusion $\mathcal{L}_{\text{exc}}$ and reconstruction loss $\mathcal{L}_{\text{rec}}$), and we denote the combination of the above image- or feature-level loss functions as the image layer loss $\mathcal{L}_{\text{img}}$[1].

Overall, the total loss function is then formulated as:

$$\mathcal{L}_{\text{total}} = \gamma_1\mathcal{L}_{\text{ctr}} + \gamma_2\mathcal{L}_{\text{lcr}} + \mathcal{L}_{\text{img}}, \tag{4}$$

where coefficients are set as $\gamma_1 = \gamma_2 = 0.5$.

[1] Details of $\mathcal{L}_{\text{img}}$ are provided in the supplementary material.

### 3.5. Training strategy

Due to the recognizable layer ambiguity that sometimes only one layer in the mixture image is recognizable, we propose a randomized training strategy to synergize with the gated language interaction mechanism (in Sec. 3.2). Since our training data simulates the recognizable layer ambiguity that some image layers do not have corresponding language descriptions (introduced in the next section), we only feed the available language description corresponding to the other image layer to guide the separation. For data with descriptions of both layers available, which indicates that both layers are recognizable, we also randomly drop one language description and feed the remaining one into the network to improve the generalization capacity for the proposed method. In practice, we set the ratio of dropping language descriptions to $30\%$.

We implement the proposed method with PyTorch [41] with a batch size of 16 on two Nvidia GeForce RTX 3090 GPUs. The model is trained for 40 epochs with Adam optimizer [25] to update learnable parameters. Weights are initialized as in [15]. The learning rate is set to $10^{-4}$ initially and decreases to $10^{-5}$ at epoch 30.

## 4. Data preparation

Though existing works have constructed several datasets for single-image reflection separation [31, 49, 67], they are unavailable for the proposed language-guided reflection separation framework due to the lack of corresponding language descriptions. Therefore, we build a dataset containing both synthetic and real data to overcome the data deficiency and facilitate network training and evaluation. Each group of data is composed of a mixture image, a transmission layer, a reflection layer, and two language descriptions. Details of synthetic and real data are as follows.

### 4.1. Synthetic data

The synthetic dataset is generated for network training to satisfy the data-driven need of the proposed method. Due to the demand for paired image-language data, we utilize two prevailing image captioning datasets (*i.e.*, Flickr30k [66] and COCO Captions [7]) for data generation, which contain 31,000 and 330,000 images respectively, and each image has 5 independent human-generated language descriptions. We randomly select images from the above two datasets as transmission $\mathbf{T}_\mathrm{S}$ and reflection scene images $\mathbf{R}_\mathrm{S}$ and conduct an image synthesis process with linear blending [23]:

$$\hat{\mathbf{M}} = \hat{\mathbf{T}} + \hat{\mathbf{R}} = \alpha\hat{\mathbf{T}}_\mathrm{S} + \beta\hat{\mathbf{R}}_\mathrm{S}, \tag{5}$$

where $\hat{\mathbf{T}}_\mathrm{S} = g_\mathrm{inv}(\mathbf{T}_\mathrm{S})$ and $\hat{\mathbf{R}}_\mathrm{S} = g_\mathrm{inv}(\mathbf{R}_\mathrm{S})$, $g_\mathrm{inv}$ represents the inverse gamma correction, and $\alpha \in [0.8, 1]$ and $\beta \in [0.4, 1]$ are the blending attenuation coefficients as in [23].

Considering the recognizable layer ambiguity that sometimes only one layer is recognizable in a mixture image, we assign a language description only when the corresponding layer is obvious enough in the synthesized mixture image, *i.e.*, $\frac{\mathrm{mean}(\hat{\mathbf{V}}_l)}{\mathrm{mean}(\hat{\mathbf{V}}_\mathbf{M})} \geqslant \mu$, where $\hat{\mathbf{V}}_l$ represents the brightness image of $\hat{\mathbf{T}}$ or $\hat{\mathbf{R}}$ in the HSV color space and $\hat{\mathbf{V}}_\mathbf{M}$ represents the brightness image of $\hat{\mathbf{M}}$, and we set $\mu = 0.3$ in the data generation process. Finally, gamma correction is applied to image triplets $\{\hat{\mathbf{T}}, \hat{\mathbf{R}}, \hat{\mathbf{M}}\}$ to obtain $\{\mathbf{T}, \mathbf{R}, \mathbf{M}\}$, and we generate 50000 triplets of data in total for network training.

### 4.2. Real data

Existing real datasets collected for the single-image reflection separation task usually contain mixture images with ground truth of transmission layers (*e.g.*, Zhang *et al.* [67] and Nature [31]), and SIR$^2$ [49, 54] further captures ground truths of reflection layers. For these off-the-shelf real datasets, we augment them by manually adding language descriptions for each group of data to satisfy the input setting of the proposed language-guided reflection separation task. Specifically, following the annotation principle of COCO Captions [7], we first describe the content of transmission layers in mixture images with entities, attributes (*e.g.*, colors or materials), and relative positions between different entities. If the content of reflection layers is recognizable in mixture images, we also give language descriptions for reflection layers in the same way. To further evaluate the generalization capacity of the proposed method, we collect a real dataset (denoted as REFOL dataset) containing 100 mixture images from the Internet that are captured in different scenes and with different cameras. We annotate these mixture images with language descriptions in the same manner as mentioned above. Following the training strategy of previous methods [10, 31], we utilize 200 image pairs from Nature dataset [31] and 90 pairs from Zhang *et al.* [67] for training, and the rest of real data are used for quantitative and qualitative evaluation.

## 5. Experiments

### 5.1. Comparison with state-of-the-art methods

To evaluate the performance of the proposed method, we conduct quantitative and qualitative experiments on existing real datasets [31, 49, 67] (with our manually annotated language descriptions) and our newly collected dataset REFOL. We compare with state-of-the-art single-image reflection separation methods, including DSRNet [23], YTMT [22], Dong *et al.* [10], IBCLN [31], CoRRN [52], and Zhang *et al.* [67]. For fair comparisons, we finetune the above methods on our training data if their training codes are provided. We report better results between the original pre-trained model and the finetuned version.

Table 1. Comparison of quantitative results in terms of PSNR [24] and SSIM [55] on real datasets for evaluating the recovery of transmission layers. ↑ (↓) indicates larger (smaller) values are better. **Bold** numbers indicate the best-performing results.

| Dataset (size) | Metrics | Methods | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Zhang *et al.* [67] | CoRRN [52] | IBCLN [31] | Dong *et al.* [10] | YTMT [22] | DSRNet [23] | Ours |
| Postcard (199) | PSNR↑<br>SSIM↑ | 20.85<br>0.872 | 22.04<br>0.870 | 23.41<br>0.872 | 23.72<br>0.903 | 22.82<br>0.885 | 24.88<br>0.910 | **25.02**<br>**0.915** |
| Object (200) | PSNR↑<br>SSIM↑ | 23.84<br>0.872 | 25.13<br>0.912 | 24.52<br>0.891 | 24.36<br>0.898 | 24.68<br>0.892 | 26.44<br>0.921 | **26.51**<br>**0.927** |
| Wild (101) | PSNR↑<br>SSIM↑ | 24.97<br>0.875 | 25.17<br>0.889 | 24.78<br>0.884 | 25.75<br>0.903 | 25.70<br>0.897 | 25.86<br>0.908 | **26.23**<br>**0.925** |
| Real20 (20) | PSNR↑<br>SSIM↑ | 22.34<br>0.795 | 21.43<br>0.801 | 21.47<br>0.762 | 23.34<br>0.812 | 23.23<br>0.802 | 23.88<br>0.816 | **24.05**<br>**0.824** |
| Nature (20) | PSNR↑<br>SSIM↑ | 20.62<br>0.753 | 20.75<br>0.783 | 23.72<br>0.806 | 23.45<br>0.808 | 21.53<br>0.778 | 22.26<br>0.801 | **23.87**<br>**0.812** |
| Average (540) | PSNR↑<br>SSIM↑ | 22.77<br>0.865 | 23.70<br>0.883 | 24.02<br>0.875 | 24.31<br>0.894 | 24.01<br>0.883 | 25.51<br>0.906 | **25.72**<br>**0.914** |

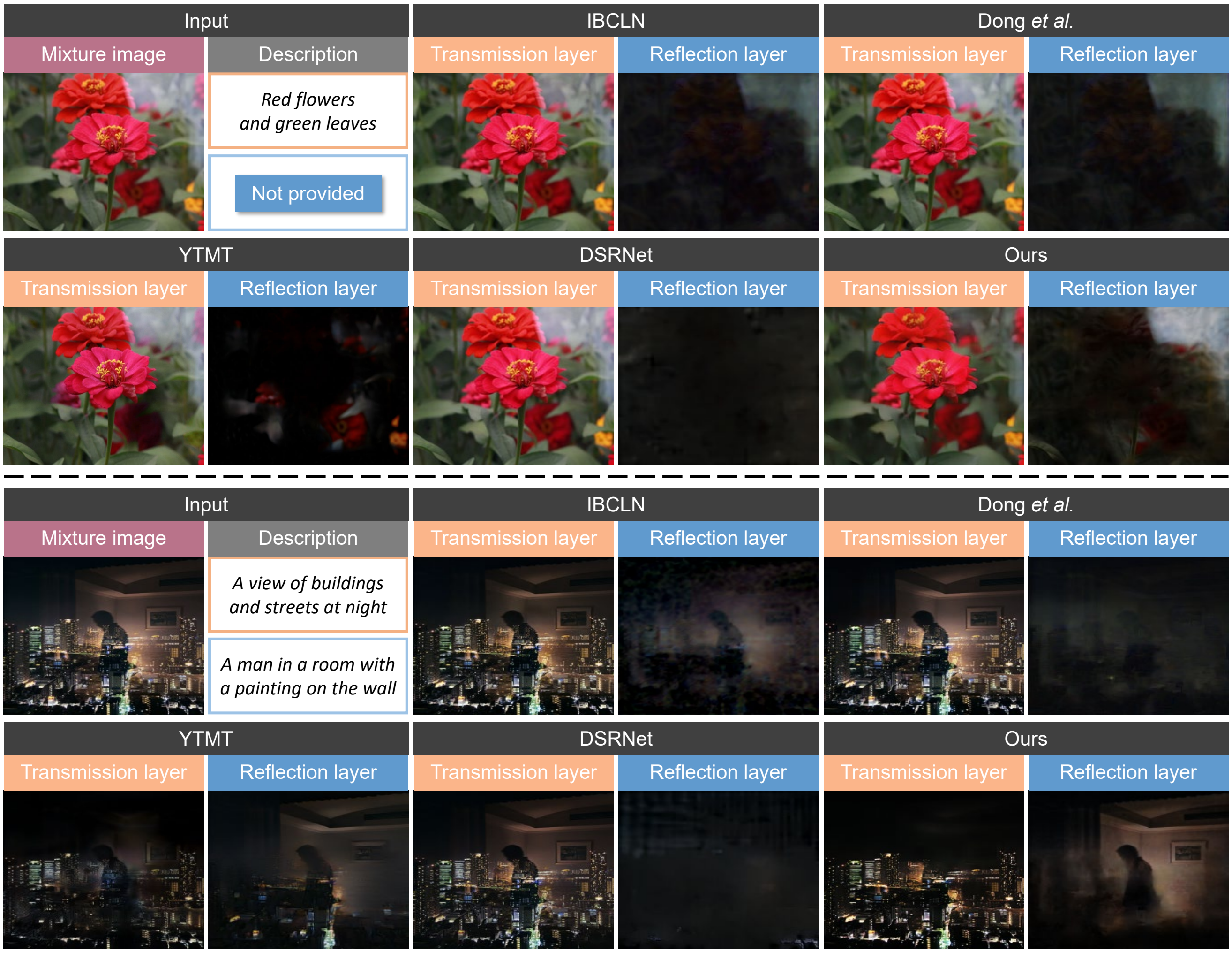

Figure 4. Qualitative comparison of estimated transmission and reflection layers on real data, compared with the state-of-the-art methods including DSRNet [23], YTMT [22], Dong *et al.* [10], and IBCLN [31]. Please zoom in for details.

**Quantitative comparison.** Quantitative experiments are conducted on three real datasets for reflection separation, *i.e.*, Nature [31], Real20 [67], and three subsets of SIR$^2$ [49] dataset. Following the setting of existing reflection separa-

Table 2. Quantitative results of ablation studies.

| Metrics | w/o language | w/o AGIM | $\mathcal{L}_{\text{img}}$ only | Ours |
|---|---|---|---|---|
| PSNR↑ | 24.31 | 24.69 | 24.52 | **25.72** |
| SSIM↑ | 0.885 | 0.901 | 0.893 | **0.914** |

tion methods [10, 36], we utilize PSNR [24] and SSIM [55] as error metrics for evaluating the recovery of transmission layers[2]. As quantitative results shown in Table 1, the proposed method achieves the best performance of both PSNR and SSIM, which validates its generalization capacity and the effectiveness of language descriptions.

**Qualitative comparison.** To evaluate the visual quality of reflection separation results, we compare the proposed method with four single-image reflection separation methods, including DSRNet [23], YTMT [22], Dong *et al.* [10], and IBCLN [31]. Qualitative results on recovering both transmission and reflection layers are shown in Figure 4. As can be observed in Figure 4, IBCLN [31] can only remove parts of reflections, while Dong *et al.* [10] have trouble in dealing with complex semantic images like the reflections of a man in a room (the second example). YTMT [22] separates layers incorrectly and thus brings the content of transmission layers into reflections. DSRNet [23] fails to recover transmission and reflection layers for these challenging cases. Contributing to the language guidance, the proposed method generates better visual results and recovers both transmission and reflection layers neatly.

### 5.2. Ablation study

In this section, we conduct several ablation studies with quantitative results shown in Table 2 to investigate the influence of the additional input of language descriptions (denoted as 'w/o language'), the network design of AGIM (denoted as 'w/o AGIM'), and the loss functions for cross-modality correspondence (denoted as '$\mathcal{L}_{\text{img}}$ only'). The performance of the variant 'w/o language' suffers from an obvious degradation as we remove language descriptions, which shows the effectiveness of the additional contextual information for layer separation. The variant 'w/o AGIM' replaces AGIMs with simple feature fusion blocks which directly concatenate language and image features and feed them into self-attention blocks, and the decline in performance validates the necessity of our gated interaction mechanism. The variant '$\mathcal{L}_{\text{img}}$ only' is trained with $\mathcal{L}_{\text{img}}$, which obtains results slightly better than the variant 'w/o language', indicating the significance of establishing the cross-modality correspondence.

To further verify the effectiveness of the language interaction, we conduct an ablation study by gradually increasing the number of input descriptions. As shown in yellow boxes of Figure 5, by utilizing more input language descriptions, the separation of transmission (clothes) and reflection layers (a man with a camera) becomes more thorough, which also demonstrates the efficacy and robustness of the proposed method.

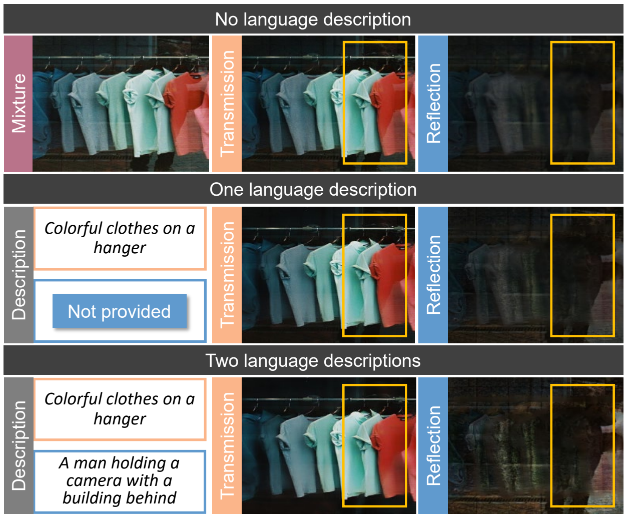

Figure 5. Results with different numbers of input language descriptions.

## 6. Conclusion

This paper introduces natural language to provide contextual information about image layers for relieving the ill-posed reflection separation problem. We develop an end-to-end framework with adaptive global interaction modules and language-image loss functions to effectively manage the modality inconsistency, and we adopt a language gate mechanism with randomized training strategies to handle the recognizable layer ambiguity. To address data deficiency, a specially built dataset with language annotations significantly aids in training and evaluating the proposed language-guided image reflection separation framework. Quantitative and qualitative experiments on real data demonstrate the effectiveness of introducing language descriptions for reflection separation.

**Limitations**. The proposed method may not distinguish transmission and reflection layers accurately when their contents are similar. For such ambiguous cases, a more flexible language-guided mechanism is needed. This might be solved by exploring a better interaction approach, which is left as our future work.

## Acknowledgement

This work is supported by National Natural Science Foundation of China under Grant No. 62136001, 62088102.

[2] Quantitative evaluations on the recovery of reflection layers are provided in the supplementary material.

## References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[2] Yakun Chang, Cheolkon Jung, and Jun Sun. Joint reflection removal and depth estimation from a single image. *IEEE TCYB*, 2020. 1

[3] Yakun Chang, Cheolkon Jung, Jun Sun, and Fengqiao Wang. Siamese dense network for reflection removal with flash and no-flash image pairs. *IJCV*, 2020. 2

[4] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. L-CoDer: Language-based colorization with color-object decoupling transformer. In *Proc. of ECCV*, 2022. 1

[5] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-CoIns: Language-based colorization with instance awareness. In *Proc. of CVPR*, 2023.

[6] Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. L-CAD: Language-based colorization with any-level descriptions using diffusion priors. In *Proc. of NeurIPS*, 2023. 1

[7] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2, 6

[8] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. NeRDi: Single-view NeRF synthesis with language-guided diffusion as general image priors. In *Proc. of CVPR*, 2023. 2

[9] Yaron Diamant and Yoav Y Schechner. Overcoming visual reverberations. In *Proc. of CVPR*, 2008. 2

[10] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *Proc. of ICCV*, 2021. 1, 2, 6, 7, 8

[11] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *Proc. of ICCV*, 2017. 2

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proc. of NeurIPS*, 2014. 2

[13] Byeong-Ju Han and Jae-Young Sim. Zero-shot learning for reflection removal of single 360-degree image. In *Proc. of ECCV*, 2022. 2

[14] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proc. of CVPR*, 2015. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. of ICCV*, 2015. 6

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, 2016. 3

[17] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GeLUs). *arXiv preprint arXiv:1606.08415*, 2016. 4, 5

[18] Yuchen Hong, Youwei Lyu, Si Li, and Boxin Shi. Near-infrared image guided reflection removal. In *Proc. of ICME*, 2020. 2

[19] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C. Kot, and Boxin Shi. Panoramic image reflection removal. In *Proc. of CVPR*, 2021. 1, 2

[20] Yuchen Hong, Youwei Lyu, Si Li, Gang Cao, and Boxin Shi. Reflection removal with NIR and RGB image feature fusion. *IEEE TMM*, 2022. 2

[21] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C. Kot, and Boxin Shi. PAR$^2$Net: End-to-end panoramic image reflection removal. *IEEE TPAMI*, 2023. 2

[22] Qiming Hu and Xiaojie Guo. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. In *Proc. of NeurIPS*, 2021. 1, 2, 6, 7, 8

[23] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proc. of ICCV*, 2023. 2, 4, 5, 6, 7, 8

[24] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics letters*, 2008. 7, 8

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[26] Naejin Kong, Yu-Wing Tai, and Sung Yong Shin. A physically-based approach to reflection separation. In *Proc. of CVPR*, 2012. 2

[27] Chenyang Lei and Qifeng Chen. Robust reflection removal with reflection-free flash-only cues. In *Proc. of CVPR*, 2021. 1, 2

[28] Chenyang Lei, Xuhua Huang, Mengdi Zhang, Qiong Yan, Wenxiu Sun, and Qifeng Chen. Polarized reflection removal with perfect alignment in the wild. In *Proc. of CVPR*, 2020. 1, 2

[29] Chenyang Lei, Xudong Jiang, and Qifeng Chen. Robust reflection removal with flash-only cues in the wild. *IEEE TPAMI*, 2023. 2

[30] Anat Levin and Yair Weiss. User assisted separation of reflections from a single image using a sparsity prior. *IEEE TPAMI*, 2007. 1, 2

[31] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *Proc. of CVPR*, 2020. 1, 2, 6, 7, 8

[32] Yu Li and Michael S Brown. Exploiting reflection change for automatic reflection removal. In *Proc. of ICCV*, 2013. 1, 2

[33] Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *Proc. of CVPR*, 2014. 1, 2

[34] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *Proc. of CVPR*, 2020. 1, 2

[35] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to see through obstructions with layered decomposition. *IEEE TPAMI*, 2021. 2

[36] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Reflection separation using a pair of unpolar-

ized and polarized images. In *Proc. of NeurIPS*, 2019. 1, 2, 8

[37] Youwei Lyu, Zhaopeng Cui, Si Li, Marc Pollefeys, and Boxin Shi. Physics-guided reflection separation from a pair of unpolarized and polarized images. *IEEE TPAMI*, 2022. 2

[38] Daiqian Ma, Renjie Wan, Boxin Shi, Alex C Kot, and Ling-Yu Duan. Learning to jointly generate and separate reflections. In *Proc. of ICCV*, 2019. 2

[39] Shree K Nayar, Xi-Sheng Fang, and Terrance Boult. Separation of reflection components using color and polarization. *IJCV*, 1997. 2

[40] Jonghyuk Park, Hyeona Kim, Eunpil Park, and Jae-Young Sim. Fully-automatic reflection removal for 360-degree images. In *Proc. of WACV*, 2024. 2

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Proc. of NeurIPS*, 2019. 6

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, 2021. 4, 5

[43] Yoav Y Schechner, Nahum Kiryati, and Ronen Basri. Separation of transparent layers using focus. *IJCV*, 2000. 2

[44] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *Proc. of CVPR*, 2015. 1, 2

[45] Christian Simon and In Kyu Park. Reflection removal for in-vehicle black box videos. In *Proc. of CVPR*, 2015. 1, 2

[46] Jimeng Sun, Shuchen Weng, Zheng Chang, Si Li, and Boxin Shi. UniCoRN: A unified conditional image repainting network. In *Proc. of CVPR*, 2022. 1

[47] Jiajun Tang, Haofeng Zhong, Shuchen Weng, and Boxin Shi. LuminAIRe: Illumination-aware conditional image repainting for lighting-realistic generation. In *Proc. of NeurIPS*, 2023. 1

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. of NeurIPS*, 2017. 4

[49] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *Proc. of ICCV*, 2017. 2, 6, 7

[50] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, Wen Gao, and Alex C. Kot. Region-aware reflection removal with unified content and gradient priors. *IEEE TIP*, 2018. 2

[51] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. CRRN: Multi-scale guided concurrent reflection removal network. In *Proc. of CVPR*, 2018. 2

[52] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. CoRRN: Cooperative reflection removal network. *IEEE TPAMI*, 2019. 2, 5, 6, 7

[53] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, and Alex C Kot. Face image reflection removal. *IJCV*, 2021. 1

[54] Renjie Wan, Boxin Shi, Haoliang Li, Yuchen Hong, Ling-Yu Duan, and Alex C Kot. Benchmarking single-image reflection removal algorithms. In *IEEE TPAMI*, 2022. 2, 6

[55] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Proc. of ACSSC*, 2003. 7, 8

[56] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proc. of CVPR*, 2022. 4

[57] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-driven referring image segmentation. In *Proc. of CVPR*, 2022. 1

[58] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proc. of CVPR*, 2019. 2, 3, 4

[59] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *Proc. of CVPR*, 2019. 2

[60] Shuchen Weng and Boxin Shi. Conditional image repainting. *IEEE TPAMI*, 2023. 1

[61] Shuchen Weng, Wenbo Li, Dawei Li, Hongxia Jin, and Boxin Shi. MISC: Multi-condition injection and spatially-adaptive compositing for conditional person image synthesis. In *Proc. of CVPR*, 2020. 1

[62] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. L-CoDe: Language-based colorization using color-object decoupled conditions. In *Proc. of AAAI*, 2022. 1

[63] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proc. of ECCV*, 2018. 2

[64] Yang Yang, Wenye Ma, Yin Zheng, Jian-Feng Cai, and Weiyu Xu. Fast single image reflection suppression via convex optimization. In *Proc. of CVPR*, 2019. 2

[65] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *Proc. of CVPR*, 2022. 1, 2

[66] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 2, 6

[67] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proc. of CVPR*, 2018. 2, 3, 5, 6, 7

[68] Ya-Nan Zhang, Linlin Shen, and Qiufu Li. Content and gradient model-driven deep network for single image reflection removal. In *Proc. of ACM MM*, 2022. 2

[69] Qian Zheng, Jinnan Chen, Zhan Lu, Boxin Shi, Xudong Jiang, Kim-Hui Yap, Ling-Yu Duan, and Alex C Kot. What does plate glass reveal about camera calibration? In *Proc. of CVPR*, 2020. 2

[70] Qian Zheng, Boxin Shi, Jinnan Chen, Xudong Jiang, Ling-Yu Duan, and Alex C. Kot. Single image reflection removal with absorption effect. In *Proc. of CVPR*, 2021. 2

# Language-guided Image Reflection Separation (Supplementary Material)

Haofeng Zhong[1,2,3#] Yuchen Hong[1,2#] Shuchen Weng[1,2] Jinxiu Liang[1,2] Boxin Shi[1,2,3*]

[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3] AI Innovation Center, School of Computer Science, Peking University

{hfzhong, shuchenweng, cssherryliang, shiboxin}@pku.edu.cn, yuchenhong.cn@gmail.com

In the supplementary material, we provide details about the image layer loss $\mathcal{L}_{\text{img}}$, report quantitative results on reflection recovery, conduct additional ablation studies, provide a comparison on model size and efficiency, and conduct additional qualitative comparisons with state-of-the-art reflection separation methods.

## 7. Details of the image layer loss

In this section, we provide details of the image layer loss $\mathcal{L}_{\text{img}}$ (corresponding to footnote 1 in the main paper), which consists of several image- or feature-level loss functions following previous reflection separation methods [2, 3, 6, 12, 16] to impose constraints on the visual quality of estimated transmission and reflection layers or to exploit the inherent relationship between the two layers. We denote the estimated transmission and reflection layers as $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{R}}$ and their ground truths as $\mathbf{T}$ and $\mathbf{R}$, respectively, and mixture images are denoted as $\mathbf{M}$.

**Pixel loss** $\mathcal{L}_{\text{pix}}$. We apply the $l_1$ distance to penalize the pixel-wise discrepancy on estimated images and gradients with their ground truths, which is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{pix}} =& \|\mathbf{T}-\tilde{\mathbf{T}}\|_1 + \|\mathbf{R}-\tilde{\mathbf{R}}\|_1 \\ &+\lambda(\|\nabla\mathbf{T}-\nabla\tilde{\mathbf{T}}\|_1 + \|\nabla\mathbf{R}-\nabla\tilde{\mathbf{R}}\|_1),\end{aligned} \tag{6}$$

where $\nabla$ represents the gradient operator, and $\lambda$ is set to 1.

**Structural similarity loss** $\mathcal{L}_{\text{ssim}}$. We incorporate the structural similarity index (SSIM) to form a loss function, which conforms to human perception and evaluates the similarity in luminance, contrast, and structure between image pairs. The structural similarity loss $\mathcal{L}_{\text{ssim}}$ [12] is defined as:

$$\mathcal{L}_{\text{ssim}} = 2 - (\text{SSIM}(\mathbf{T},\tilde{\mathbf{T}}) + \text{SSIM}(\mathbf{R},\tilde{\mathbf{R}})). \tag{7}$$

**Perceptual loss** $\mathcal{L}_{\text{per}}$. To measure the multi-scale discrepancy between estimated images layers and their ground truths in the feature domain, we utilize the VGG-19 model to extract low-level and high-level image features and calculate the perceptual loss [16] as:

$$\mathcal{L}_{\text{per}} = \sum_k \phi_k(\mathcal{D}_k^{\text{vgg}}(\mathbf{T},\tilde{\mathbf{T}}) + \mathcal{D}_k^{\text{vgg}}(\mathbf{R},\tilde{\mathbf{R}})), \tag{8}$$

where $\{\phi_k\}$ are the weights for balancing multi-scale feature discrepancies, and $\mathcal{D}_k^{\text{vgg}}$ represents the $l_1$ distance between features extracted from the $k$-th convolutional layer in the VGG-19 model. We adopt the same selection of convolutional layers and the setting of $\{\phi_k\}$ as [16].

**Exclusion loss** $\mathcal{L}_{\text{exc}}$. To ensure the gradient irrelevance between estimated transmission and reflection layers for diminishing content residues from each other, we employ the exclusion loss [16] as:

$$\mathcal{L}_{\text{exc}} = \frac{1}{M}\sum_{m=0}^{M-1}\|\Theta(\tilde{\mathbf{T}}^{\downarrow m},\tilde{\mathbf{R}}^{\downarrow m})\|_{\text{F}}, \tag{9}$$

$$\Theta(\tilde{\mathbf{T}},\tilde{\mathbf{R}}) = \tanh(\xi_1|\nabla\tilde{\mathbf{T}}|)\odot\tanh(\xi_2|\nabla\tilde{\mathbf{R}}|), \tag{10}$$

where $\|\cdot\|_{\text{F}}$ denotes the Frobenius norm, $\tilde{\mathbf{T}}^{\downarrow m}$ and $\tilde{\mathbf{R}}^{\downarrow m}$ represent down-sampling $\tilde{\mathbf{T}}$ and $\tilde{\mathbf{R}}$ by a factor of $2^m$ with bilinear interpolation ($2^M$ at most where $M=3$ as in [16]), $\odot$ is the element-wise multiplication, and $\xi_1$ and $\xi_2$ are the normalization factors as in [16].

**Reconstruction loss** $\mathcal{L}_{\text{rec}}$. To constrain the relation between transmission layers, reflection layers, and mixture images, we employ a reconstruction loss following [6]:

$$\mathcal{L}_{\text{rec}} = \|\tilde{\mathbf{T}}+\tilde{\mathbf{R}}+\Omega(\tilde{\mathbf{T}},\tilde{\mathbf{R}})-\mathbf{M}\|)_1, \tag{11}$$

where $\Omega(\tilde{\mathbf{T}},\tilde{\mathbf{R}})$ is a residue term estimated from an additional learnable residue module $\Omega(\cdot)$ [6], which is designed for handling the non-linearity in the mixture image formation process caused by the non-linear mapping and dynamic range clipping [4] in the camera pipeline.

Overall, the image layer loss is formulated as:

$$\mathcal{L}_{\text{img}} = \omega_1\mathcal{L}_{\text{pix}}+\omega_2\mathcal{L}_{\text{ssim}}+\omega_3\mathcal{L}_{\text{per}}+\omega_4\mathcal{L}_{\text{exc}}+\omega_5\mathcal{L}_{\text{rec}}. \tag{12}$$

Following previous methods [1, 6, 12, 16], the weights are set as $\omega_1=1$, $\omega_2=1$, $\omega_3=0.01$, $\omega_4=1$, and $\omega_5=0.2$.

## 8. Quantitative results on reflection recovery

In this section, we conduct quantitative experiments on three subsets (*i.e.*, Postcard, Object, and Wild) of a real reflection separation dataset SIR$^2$ [13] with our manually

[#] Equal contributions. [*] Corresponding author.

Table 3. Quantitative results in terms of PSNR and SSIM on three subsets of the SIR$^2$ dataset [13] for evaluating the recovery of reflection layers, compared with state-of-the-art single-image reflection separation methods [1, 5–7, 12, 16]. Averaged results are shown at the bottom. ↑ indicates larger values are better. **Bold** numbers indicate the best-performing results.

| Dataset (size) | Metrics | Methods | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Zhang *et al.* [16] | CoRRN [12] | IBCLN [7] | Dong *et al.* [1] | YTMT [5] | DSRNet [6] | Ours |
| Postcard (199) | PSNR↑ | 17.02 | 17.68 | 17.95 | 18.13 | 17.53 | 17.66 | **18.37** |
| | SSIM↑ | 0.519 | 0.574 | 0.528 | 0.592 | 0.557 | 0.566 | **0.611** |
| Object (200) | PSNR↑ | 21.87 | 22.52 | 22.08 | 23.62 | 22.91 | 23.56 | **23.88** |
| | SSIM↑ | 0.531 | 0.561 | 0.524 | 0.688 | 0.605 | 0.669 | **0.699** |
| Wild (101) | PSNR↑ | 20.33 | 20.93 | 20.82 | 21.53 | 21.22 | 21.64 | **21.94** |
| | SSIM↑ | 0.544 | 0.568 | 0.554 | 0.606 | 0.581 | 0.613 | **0.627** |
| Average (500) | PSNR↑ | 19.63 | 20.27 | 20.18 | 21.01 | 20.43 | 20.82 | **21.30** |
| | SSIM↑ | 0.529 | 0.568 | 0.532 | 0.633 | 0.581 | 0.617 | **0.649** |

Table 4. Ablation studies on the network structure and the size of training dataset.

| CLIP-L-encoder | Llama2-L-encoder | AGAM | CLIP-I-encoder | AGIM | Cross att | 50K data | 13K data | PSNR↑ | SSIM↑ |
|---|---|---|---|---|---|---|---|---|---|
| ✓ | | ✓ | | ✓ | | ✓ | | 25.72 | 0.914 |
| | ✓ | ✓ | | ✓ | | ✓ | | 25.68 | 0.917 |
| ✓ | | | ✓ | ✓ | | ✓ | | 24.80 | 0.891 |
| ✓ | | ✓ | | | ✓ | ✓ | | 24.92 | 0.903 |
| ✓ | | ✓ | | ✓ | | | ✓ | 25.55 | 0.909 |

annotated language descriptions (as mentioned in Sec. 4 of the main paper) to evaluate the recovery of reflection layers (corresponding to footnote 2 in the main paper), since other datasets such as Real20 [16] and Nature [7] do not provide ground truths of reflection layers. We compare the proposed method with state-of-the-art single-image reflection separation methods [1, 5–7, 12, 16]. PSNR and SSIM are selected as error metrics. As shown in Table 3, the proposed method achieves the best performance, which indicates the efficacy of introducing language descriptions for relieving the ambiguity in separating strong reflections from mixture images.

## 9. Additional ablation studies

**Ablation studies on the network structure.** We conduct ablation studies on the network structure to investigate the effectiveness of the language encoder, global image feature, and interaction module by replacing the language encoder of CLIP [10] with the encoder of a large language model Llama2 [11] (with 13B parameters), replacing the AGAM with the global image feature encoder of CLIP [10], and replacing the AGIM with standard cross-attention modules, respectively. As shown in Table 4, the proposed method (the first row) achieves competitive results with the variant (the second row) using the language encoder of Llama2 [11], which indicates our generalizability. Besides, directly using global image features from pretrained CLIP [10] (the third row) leads to performance degradation since they are trained for classification. Using standard cross-attention modules also degrades the performance (the fourth row), indicating the efficacy of AGIM for channel rearrangement.

**Ablation studies on the network training.** We investigate the influence of the training dataset size by training our model with 13,000 images from our dataset following [1]. Table 4 shows a slight performance decrease with fewer training data (the fifth row) while we still outperform baselines (Table 1 of main paper). Besides, we conduct an ablation study by setting loss coefficients $\gamma_1$ and $\gamma_2$ in Eq. (4) of the main paper to 0, 0.5, 1.0, and 2.0, respectively. As shown in the left part of Figure 6, setting both $\gamma_1$ and $\gamma_2$ as 0.5 yields the best results. In addition, we investigate the drop rate of language descriptions mentioned in Sec. 3.5 of the main paper. As depicted in the right part of Figure 6, the drop rate of 30% strikes an optimal balance, which is adopted in the paper.

**Ablation studies on language descriptions.** We investigate different types of language descriptions as shown in Figure 7. Using the simplified description achieves comparable performance to the complete matched description, while using the unmatched description fails in reflection separation, indicating the efficacy of incorporating language modality. Besides, since reflection layers are sometimes too dark and blurry to be recognizable [12] which might make descriptions of reflection layers unobtainable, we empirically set $\mathbf{I}_1$ and $\mathbf{I}_2$ to be transmission and reflection layers, respectively. If exchanging the order of descriptions (shown in Figure 8), though results are degraded due to different statistics of transmission and reflection layers, the contents still conform to descriptions, validating the effectiveness of language guidance.

## 10. Comparison on model size and efficiency

We show the model size (number of parameters), computational cost (FLOPs), and inference time of the proposed method and state-of-the-art single-image methods in Table 5. The input image size is set as 224 × 288, and we run the inference on an Nvidia RTX 2080 Ti GPU. While having the

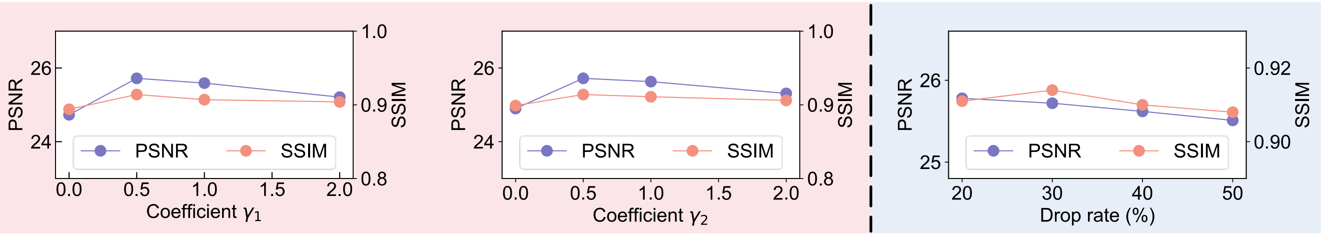

Figure 6. Ablation studies on coefficients of $\gamma_1$ and $\gamma_2$ (left part) and drop rates of language descriptions (right part).

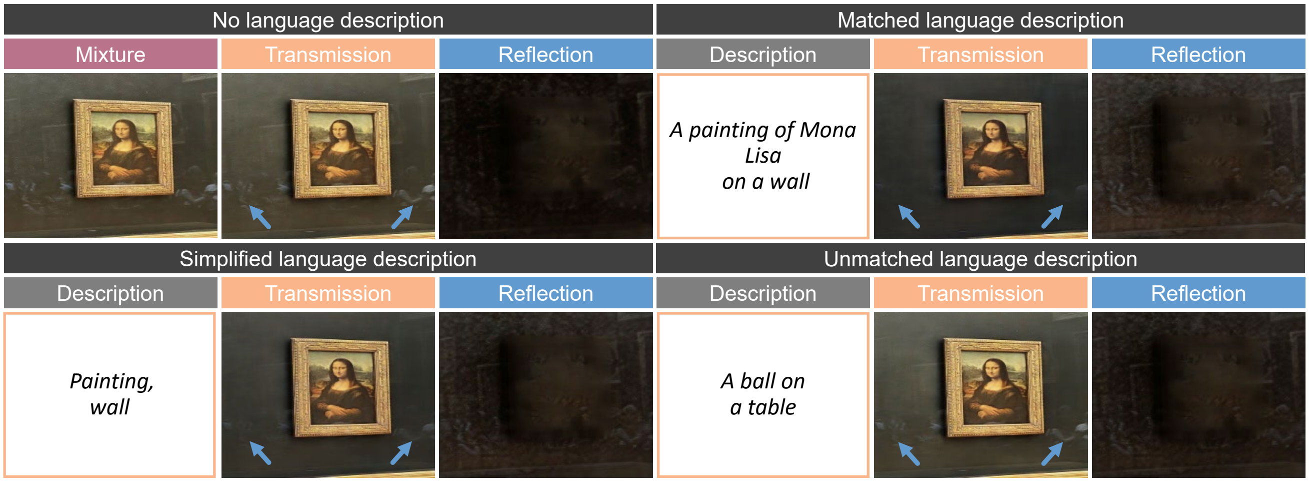

Figure 7. Ablation studies on different types of language descriptions.

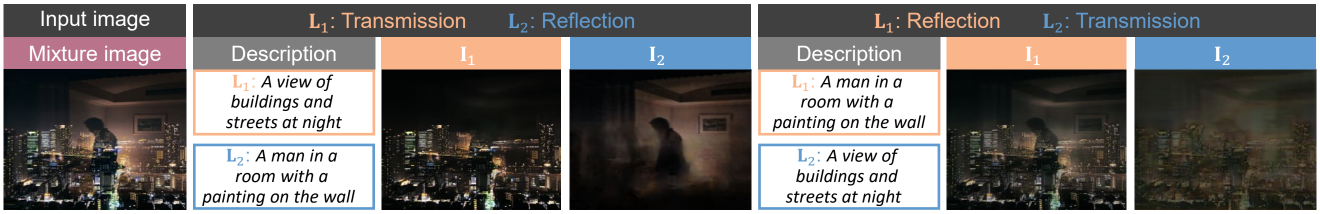
Input image
Mixture image
$L_1$: Transmission
$L_2$: Reflection
Description
$I_1$
$I_2$
$L_1$: A view of buildings and streets at night
$L_2$: A man in a room with a painting on the wall
$L_1$: Reflection
$L_2$: Transmission
Description
$I_1$
$I_2$
$L_1$: A man in a room with a painting on the wall
$L_2$: A view of buildings and streets at night

Figure 8. Results of exchanging the order of language descriptions.

Table 5. Comparisons on the model size, computational cost, and inference time, compared with single-image methods [1, 5–7, 12, 16].

| Metric | Method | | | | | | |
|---|---|---|---|---|---|---|---|
| | Zhang *et al*. [16] | CoRRN [12] | IBCLN [7] | Dong *et al*. [1] | YTMT [5] | DSRNet [6] | Ours |
| Params | 22.06M | 59.51M | 21.61M | 10.93M | 73.43M | 137.63M | 75.54M |
| FLOPs | 99.66G | 75.53G | 386.16G | 329.28G | 437.16G | 406.97G | 320.95G |
| Time (s) | 0.028 | 0.017 | 0.034 | 0.044 | 0.062 | 0.115 | 0.056 |

comparable model size, computational cost, and inference time with recent single-image methods (*e.g*., YTMT [5] and DSRNet [6]), the proposed method outperforms them in reflection separation as shown in Table 1 of the main paper, indicating our trade-off between practicality and efficiency.

## 11. Additional qualitative results

In this section, additional qualitative experiments are conducted on real datasets to show the effectiveness and unique advantages of the proposed language-guided reflection separation method. We compare with several single-image methods including DSRNet [6], YTMT [5], Dong *et al*. [1], IBCLN [7], CoRRN [12], and Zhang *et al*. [16]. Besides, a representative diffusion-based image generation method, *i.e*., ControlNet [15], is selected to show the performance of the prevailing diffusion models on reflection separation. We also compare with a multi-image reflection separation method Liu *et al*. [9] to demonstrate the robustness of the proposed method. Details are as follows.

**Comparison with ControlNet** [15]. ControlNet [15] is a conditional generative model modified from large pre-trained text-to-image diffusion models, achieving remarkable performance in image generation and editing. To make ControlNet [15] fit our input setting, we finetune it following the official instruction[1] by using mixture images as source images (control images), language descriptions of transmission layers as prompts, and transmission layers as target images. Qualitative results on the proposed REFOL dataset are shown in Figure 9. It can be observed that ControlNet [15] performs modifications on mixture images in a generative manner, *e.g.*, the portrait in the first example is infused with the blue hue and the blue butterfly in the second example is transformed into cyan, which leads to a divergence in the content of generated results from original mixture images, indicating that ControlNet [15] cannot be trivially adapted to the task of reflection separation. By utilizing global scene contextual information from language descriptions to interact with visual features for channel rearrangement (mentioned in Sec. 3.2), the proposed method outperforms single-image methods in achieving a more thorough separation of transmission and reflection layers and obtains results whose image content remains faithful to input images. For instance, as shown in Figure 9, the proposed method distinguishes reflections of visitors from the portrait in the first example while other single-image methods fail in recognizing the visitors, and in the second example, the bookshelf and the white door are also correctly separated from the butterfly by the proposed method, indicating the efficacy of language descriptions.

**Comparison with Liu *et al.*** [9]. We further conduct experiments on real datasets collected for multi-image reflection separation [8, 14]. We compare the proposed method with the aforementioned single-image methods [1, 5–7, 12, 16] and a multi-image method Liu *et al.* [9] which leverages different motions of the two layers to guide the separation. Qualitative results are shown in Figure 10. By introducing language descriptions, the proposed method achieves comparable performance with Liu *et al.* [9] in reflection separation, *e.g.*, the trash bin and the cabinet in the first example and the walking man in the second example of Figure 10, where other single-image methods fails in discerning the content of reflection layers. Moreover, multi-image reflection separation methods [8, 9, 14] typically require additional images (with the quantity ranging from one to four) with specialized capture settings compared with single-image methods, while the proposed method only demands a maximum of two additional language descriptions for network inputs, which significantly relieves the burden of data acquisition and storage associated with multi-image methods. Concurrently, the proposed method maintains the broad applicability as single-image methods, indicating its potential for practical applications.

[1] https://github.com/lllyasviel/ControlNet/blob/main/docs/train.md

## References

[1] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *Proc. of ICCV*, 2021. 1, 2, 3, 4, 5, 6

[2] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C. Kot, and Boxin Shi. Panoramic image reflection removal. In *Proc. of CVPR*, 2021. 1

[3] Yuchen Hong, Youwei Lyu, Si Li, Gang Cao, and Boxin Shi. Reflection removal with nir and rgb image feature fusion. *IEEE TMM*, 2022. 1

[4] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C. Kot, and Boxin Shi. PAR$^2$Net: End-to-end panoramic image reflection removal. *IEEE TPAMI*, 2023. 1

[5] Qiming Hu and Xiaojie Guo. Trash or treasure? an interactive dual-stream strategy for single image reflection separation. In *Proc. of NeurIPS*, 2021. 2, 3, 4, 5, 6

[6] Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *Proc. of ICCV*, 2023. 1, 2, 3

[7] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *Proc. of CVPR*, 2020. 2, 3, 4, 5, 6

[8] Yu Li and Michael S Brown. Exploiting reflection change for automatic reflection removal. In *Proc. of ICCV*, 2013. 4, 6

[9] Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Learning to see through obstructions. In *Proc. of CVPR*, 2020. 3, 4, 6

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, 2021. 2

[11] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[12] Renjie Wan, Boxin Shi, Haoliang Li, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. CoRRN: Cooperative reflection removal network. *IEEE TPAMI*, 2019. 1, 2, 3, 4, 5, 6

[13] Renjie Wan, Boxin Shi, Haoliang Li, Yuchen Hong, Ling-Yu Duan, and Alex C. Kot. Benchmarking single-image reflection removal algorithms. In *IEEE TPAMI*, 2022. 1, 2

[14] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T Freeman. A computational approach for obstruction-free photography. *ACM TOG*, 2015. 4, 6

[15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. of ICCV*, 2023. 3, 4, 5

[16] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proc. of CVPR*, 2018. 1, 2, 3, 4, 5, 6

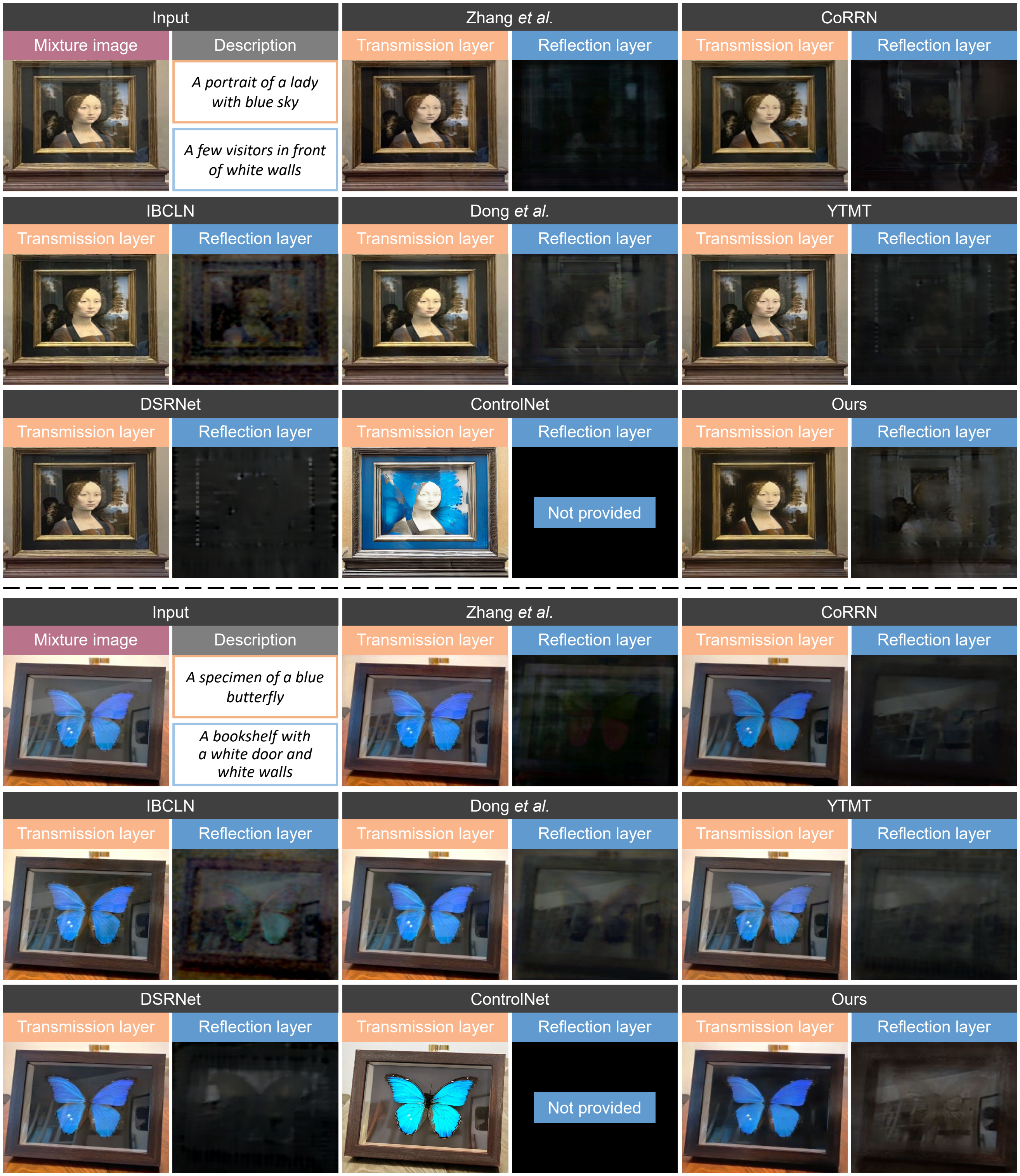

Figure 9. Qualitative comparison of estimated transmission and reflection layers on the proposed REFOL dataset, compared with several state-of-the-art single-image methods [1, 5–7, 12, 16] and a diffusion-based method ControlNet [15]. Please zoom in for details.

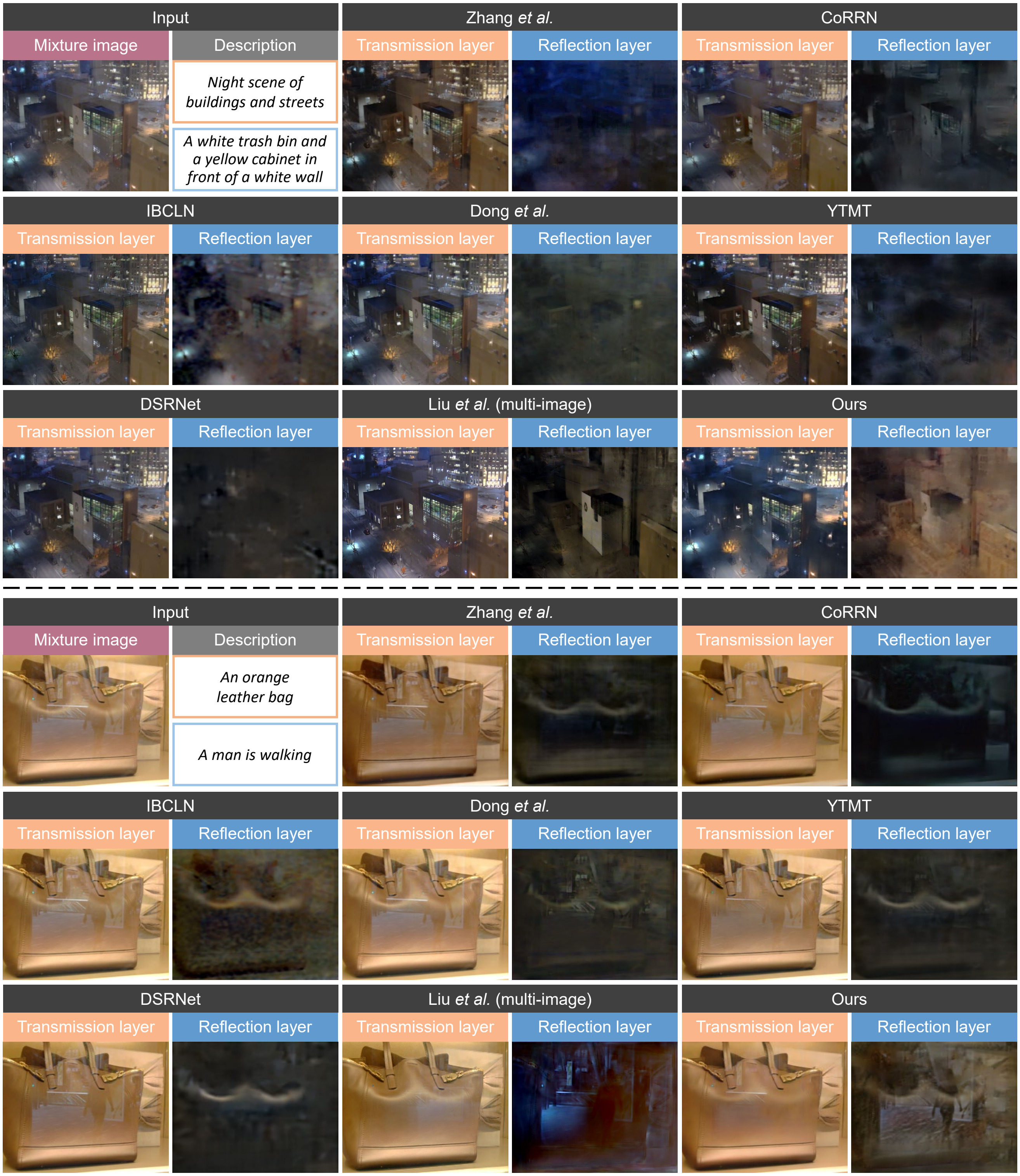

Figure 10. Qualitative comparison of estimated transmission and reflection layers on real data from [8] and [14], compared with several state-of-the-art single-image methods [1, 5–7, 12, 16] and a multi-image method Liu *et al*. [9]. Please zoom in for details.