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 ∇ represents the gradient operator, and λ 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 ξ_1 and ξ_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 γ_1 and γ_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 γ_1 and γ_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

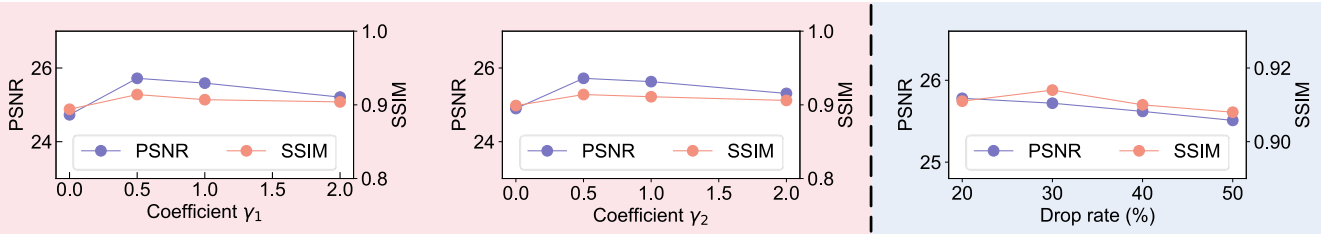

Figure 6. Ablation studies on coefficients of γ_1 and γ_2 (left part) and drop rates of language descriptions (right part).

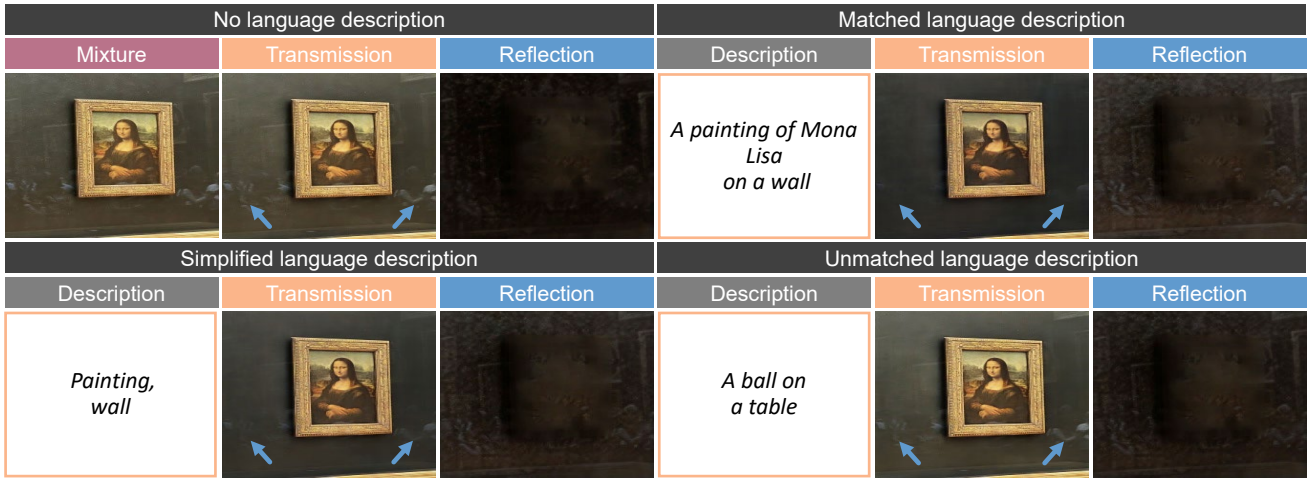

Figure 7. Ablation studies on different types of language descriptions.

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 remark-

able 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

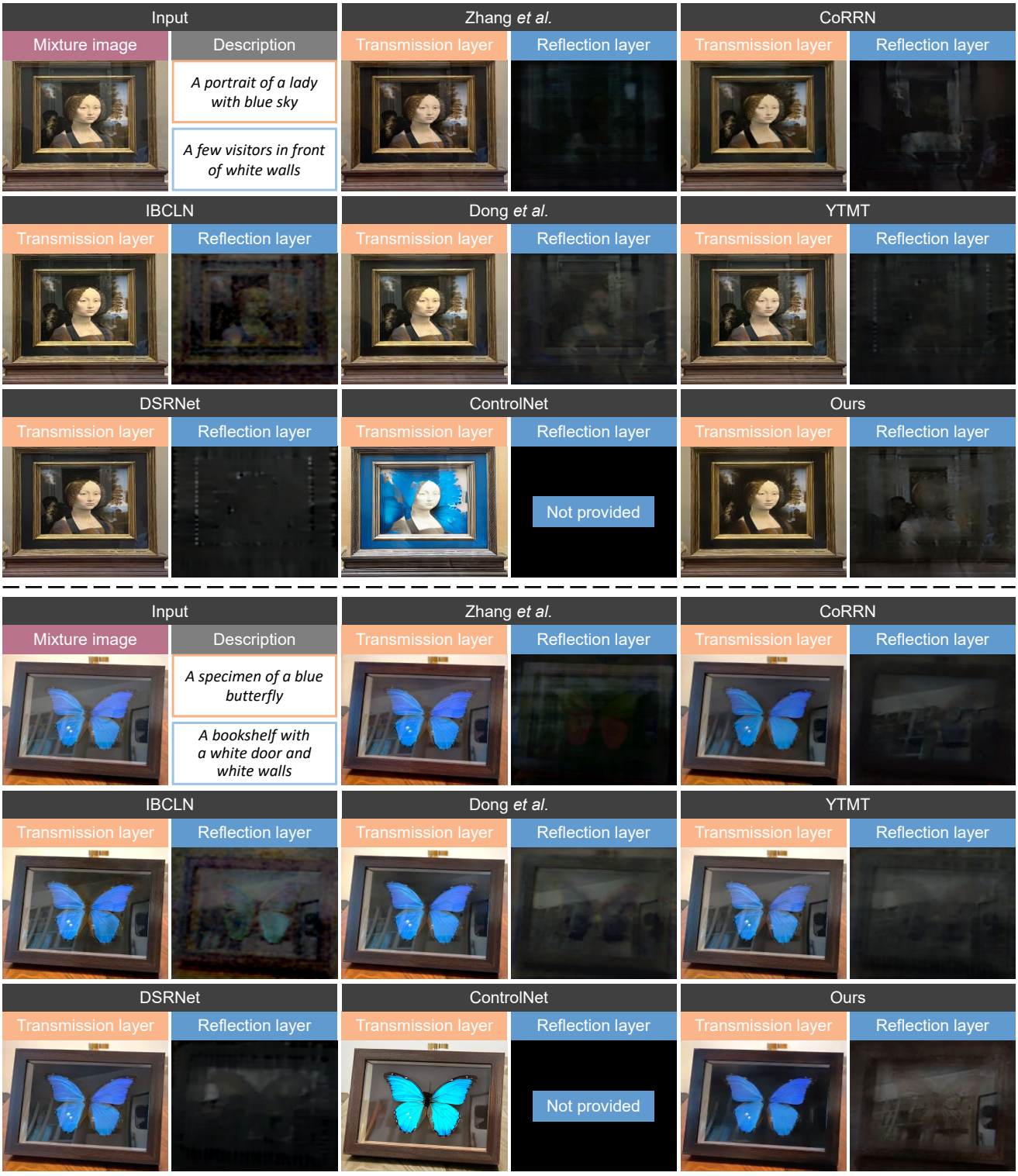

Figure 9. Qualitative comparison of estimated transmission and reflection layers on the proposed REFOL dataset, compared with several state-of-the-art single-image methods [1, 5–7, 12, 16] and a diffusion-based method ControlNet [15]. Please zoom in for details.

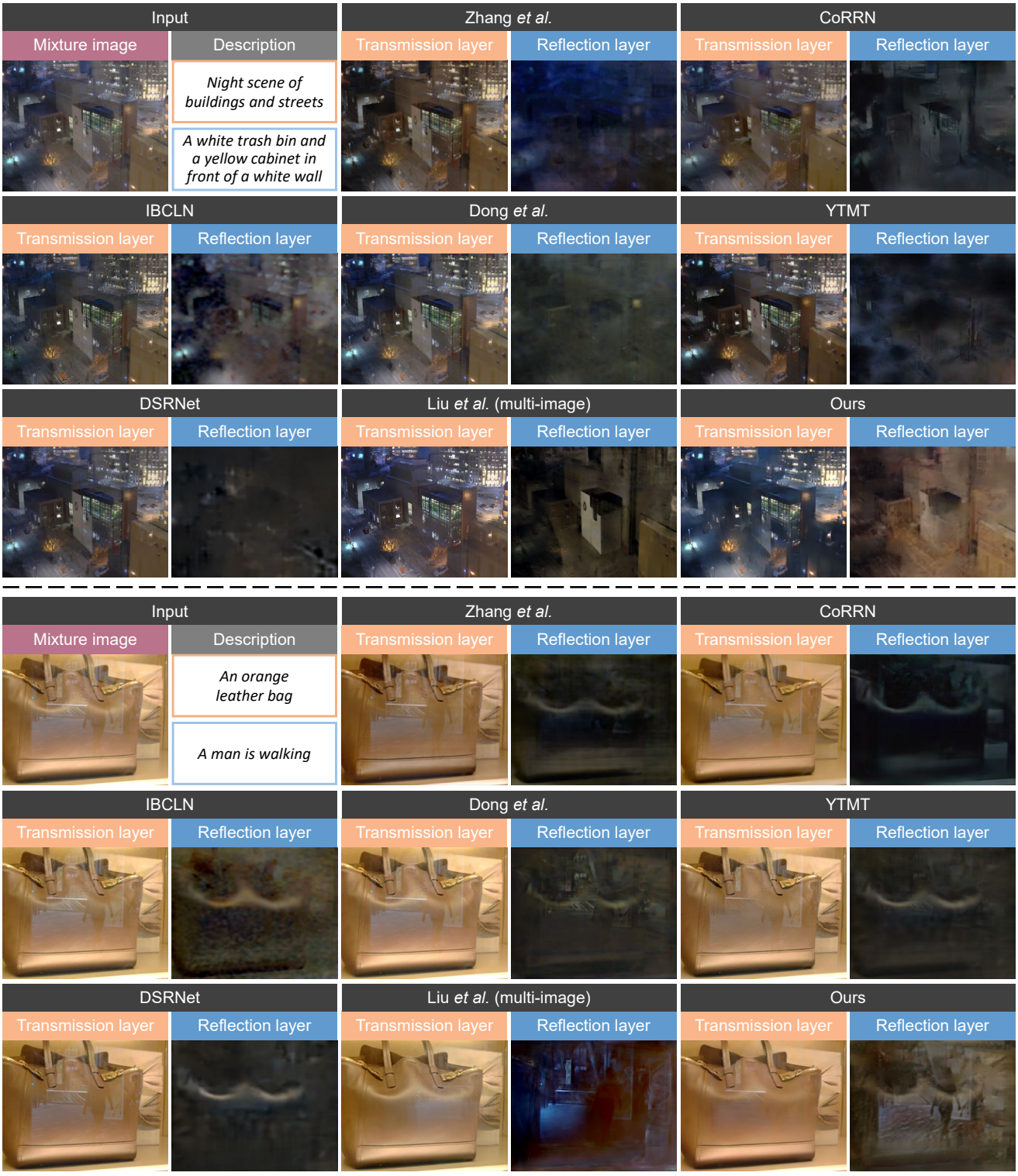

Figure 10. Qualitative comparison of estimated transmission and reflection layers on real data from [8] and [14], compared with several state-of-the-art single-image methods [1, 5–7, 12, 16] and a multi-image method Liu *et al*. [9]. Please zoom in for details.