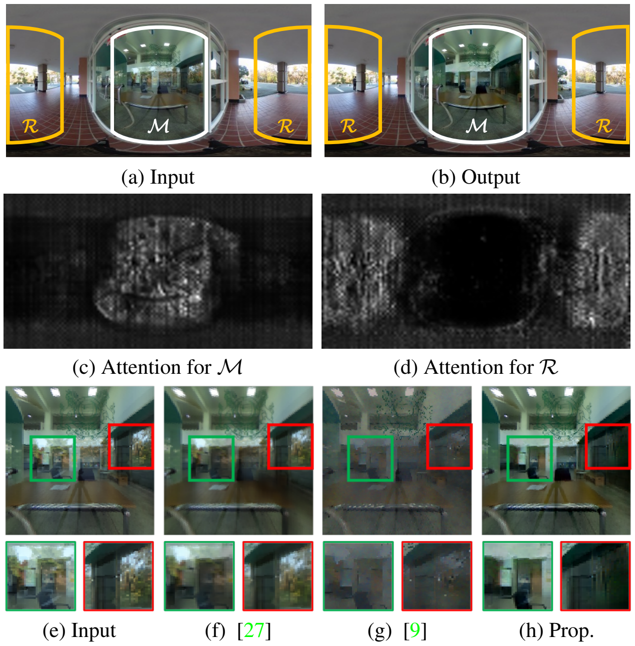
Fully-Automatic Reflection Removal for 360-Degree Images

Jonghyuk Park, Hyeona Kim, Eunpil Park, and Jae-Young Sim*
Ulsan National Institute of Science and Technology, Ulsan, Republic of Korea
{jonghyukpark,gusdk304,cosmos,jysim}@unist.ac.kr

Abstract

Reflection removal (RR) is a technique to reconstruct the transmitted scene behind the glass from a mixed image taken through glass. In 360-degree images, the mixed image region and the reference image region capturing the reflected scene exist together, and the mixed image is often restored by using the information of reference image. In this paper, we first propose a fully-automatic end-to-end RR framework for 360-degree images which automatically detects the mixed and reference image regions and removes the reflection artifacts in the mixed image by using the reference information simultaneously. We devise a transformer based U-Net architecture with horizontal windowing scheme to capture the long-range dependencies between the mixed and reference images via the self-attention mechanism and suppress the reflection artifacts by using the reference information. We also construct a training dataset of 360-degree images by synthesizing realistic reflection artifacts considering diverse geometric relation and photometric variation between the mixed and reference images. The experimental results show that the proposed method detects the mixed and reference image regions reliably without user-annotation and achieves better performance of RR compared with the state-of-the-art methods.



(a) Input (b) Output

(c) Attention for $\mathcal{M}$ (d) Attention for $\mathcal{R}$

(e) Input (f) [27] (g) [9] (h) Prop.

Figure 1. RR for 360-degree images. (a) An input 360-degree image with reflection artifact, and (b) the RR result of the proposed method. The attention maps highlighting (c) the mixed region $\mathcal{M}$ and (d) the reference region $\mathcal{R}$. (e) The input mixed image. The reconstructed transmission images obtained by using (f) [27], (g) [9], and (h) the proposed method, respectively.

1. Introduction

When we take a picture through reflective material (e.g. glass), we obtain a mixed image $\mathbf{M}$ composed of two images: the transmission image $\mathbf{T}$ capturing the scene through the glass and the undesired reflection image $\mathbf{R}$ capturing the scene reflected on the glass. This is formulated as [27]

$$\mathbf{M}(\mathbf{x}) = \Omega(\mathbf{x})\mathbf{T}(\mathbf{x}) + \Phi(\mathbf{x})\mathbf{R}(\mathbf{x}), \tag{1}$$

where $\mathbf{M}(\mathbf{x})$, $\mathbf{T}(\mathbf{x})$, and $\mathbf{R}(\mathbf{x})$ represent the intensities at pixel $\mathbf{x}$ on the mixed, transmission, and reflection images, respectively. Ω and Φ denote the refractive and reflective amplitude coefficient maps [27] associated with the characteristics of the glass.

Reflection removal (RR) [6, 12, 13, 20, 22, 25] is a task to reconstruct $\mathbf{T}$ from $\mathbf{M}$, which is an ill-posed problem. Since both of $\mathbf{T}$ and $\mathbf{R}$ capture natural scenes, it is quite ambiguous to distinguish $\mathbf{T}$ from $\mathbf{R}$. To overcome such *content ambiguity*, RR usually employs the statistics of $\mathbf{R}$, such as the smoothness prior [15] and ghosting cue [18]. Recently, RR for 360-degree images has been introduced [9–11], where a 360-degree image contains both of the mixed region $\mathcal{M}$ with reflection artifact and the reference region $\mathcal{R}$ capturing the original reflected scene, as shown in Figure 1(a). With the help of such reference information, we can effectively remove the reflection artifacts from the mixed image region in a single 360-degree image.

However, there are two major challenges in RR of 360-degree images. First, different from the ordinary images where the reflection artifacts occur over an entire image area in general, a 360-degree image includes both of the mixed and reference regions together. Therefore, we should first find the region $\mathcal{M}$ as well as the reference region $\mathcal{R}$ corresponding to $\mathcal{M}$ within a 360-degree image. Second, we should reconstruct $\mathbf{T}$ from $\mathcal{M}$ by exploiting the reference information in $\mathcal{R}$. Note that the pair of the region detection and the transmission reconstruction is a chicken-and-egg problem. As we detect more accurate region of $\mathcal{R}$, better reconstruction result of $\mathbf{T}$ is yielded. At the same time, more faithful reconstruction result of $\mathbf{T}$ (and $\mathbf{R}$ accordingly) provides more informative clue to find $\mathcal{R}$. Due to these challenges, in the previous methods [9–11], $\mathcal{M}$ was user-provided or assumed to lie in the center area of a 360-degree image, and $\mathcal{R}$ was estimated around the opposite direction of $\mathcal{M}$ in the 360-degree image. Also, they employed the features of $\mathbf{R}$ by matching the pixels of $\mathcal{R}$ and $\mathcal{M}$ in heuristic manners, degrading the performances of RR. In addition, it is impractical to construct a paired dataset of 360-degree images with and without reflection artifacts due to the difficulty of annotation for $\mathcal{M}$ and $\mathcal{R}$.

In this paper, we propose an end-to-end reflection removal network for 360-degree images, which reconstructs $\mathbf{T}$ by detecting $\mathcal{M}$ and $\mathcal{R}$ in a fully-automatic manner. We first devise the horizontal windowing scheme for the U-Net structure with transformer blocks to capture the relationship between $\mathcal{M}$ and $\mathcal{R}$ in 360-degree images via the self-attention mechanism of transformer, as shown in Figures 1(c) and (d). We also synthesize realistic reflection artifacts for 360-degree images to obtain a paired dataset for training the network by simulating diverse geometric relation and photometric variation between $\mathcal{M}$ and $\mathcal{R}$. The experimental results in Figure 1 demonstrate that the proposed method automatically highlights the attention for $\mathcal{M}$ and $\mathcal{R}$, and successfully removes the reflection artifacts from $\mathcal{M}$ outperforming the state-of-the art methods.

The contributions of this paper are as follows:

- We first propose a fully-automatic RR framework for 360-degree images, that automatically detects the mixed and reference regions without any user-interaction and removes the reflection artifacts in the mixed region by utilizing the information of the reference image.
- We devise a transformer-based U-Net architecture to capture the long-range dependency between the mixed and reference regions.
- We collect real 360-degree images for qualitative and quantitative experiments and demonstrate that the proposed method achieves better performance over state-of-the-art reflection removal methods.

2. Related Works

Single image reflection removal. In general, the reflection images have different characteristics from that of the transmission images that provide useful clues to resolve the content ambiguity in RR. For example, the reflection images are more blurred than the transmission images [15]. The reflection images often have ghost contours due to thick glasses [18]. Traditional methods exploited such characteristics of the reflection images for single image RR. Recently, data-driven priors are being popularly used with deep-learning frameworks [6,12,13,23,25,27]. Fan et al. [6] first applied the convolutional neural network to estimate the edges of the reflection image to reconstruct the transmission image. Yang et al. [23] proposed a cascade network to predict the transmission and reflection images simultaneously. Kim et al. [12] synthesized realistic paired data considering the lens effects and the angle between the light direction and the glass plane to trace the reflection. Zheng et al. [27] proposed a new image formation model considering the absorption effect depending on the incident angle of rays on the glass plane.

Multiple images reflection removal. Several attempts have been made to employ multiple images for RR [7,8,14]. Li et al. [14] utilized the SIFT-flow to align a set of multiple images and classified the features of the transmission and reflection images based on the magnitude of the disparity. Guo et al. [7] exploited structural priors in multiple mixed images to decompose the reflection and transmission images based on the augmented Lagrangian multipliers. Han and Sim [8] decomposed the mixed image into the transmission and reflection layers based on low-rank matrix completion.

360-degree image reflection removal. Whereas the reflection artifacts occur on the entire image area for ordinary images, the artifacts usually lie on local regions only in real-world 360-degree images. Therefore, we need to first find the mixed image region in a 360-degree image, however this is not a trivial task. The existing methods [9–11] assume that the mixed and/or reference image regions are given by user-interaction, and employ the reference information by using heuristic matching schemes. Hong et al. [10] cropped the mixed and reference image regions by multi-step user interaction which are then served as input to single-image RR network. The subsequent method [11] still requires one-time user interaction to acquire the mixed region. Han and Sim [9] proposed a zero-shot learning based RR to overcome the difficulty to collect 360-degree images with ground truth data. In this paper, we propose a fully-automatic end-to-end network that performs RR for 360-degree images based on transformer architecture without any user-interaction.

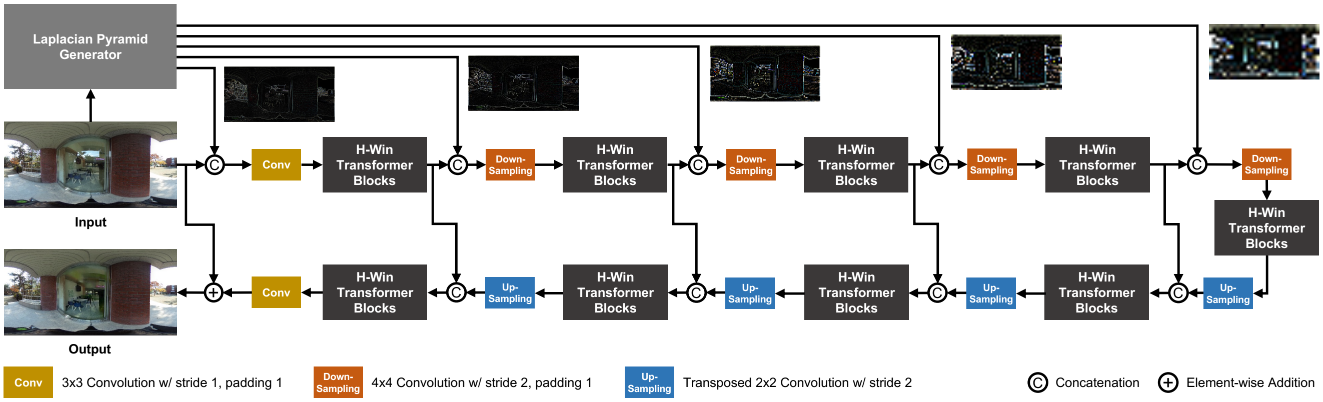

Laplacian Pyramid Generator
Input
Output
Conv
H-Win Transformer Blocks
Down-Sampling
Up-Sampling
C
+
3x3 Convolution w/ stride 1, padding 1
4x4 Convolution w/ stride 2, padding 1
Transposed 2x2 Convolution w/ stride 2
Concatenation
Element-wise Addition

Figure 2. The overall architecture of the proposed end-to-end RR network based on the U-Net with transformer blocks. The horizontal windowing scheme and the laplacian pyramids are used.

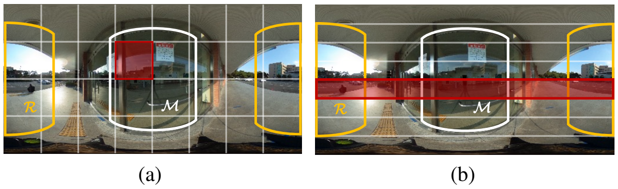


Figure 3. Windowing schemes in transformer. (a) The conventional square-shaped windowing scheme [21] and (b) the proposed horizontal windowing scheme.

3. Proposed Method

We explain the transformer-based end-to-end RR network for 360-degree images. The overall architecture of the proposed network is shown in Figure 2.

3.1. Transformer-Based End-to-End Network

The goals of the proposed end-to-end 360-degree image RR network are to detect the regions of $\mathcal{M}$ and $\mathcal{R}$ as well as to exploit the information from $\mathbf{R}$ to reconstruct $\mathbf{T}$ in a fully-automatic manner. As the multi-head self-attention mechanism (a.k.a. transformer) effectively captures the long-range dependencies between the tokens [19], we expect that this capability is suited to capture the dependencies between $\mathcal{M}$ and $\mathcal{R}$ within a single 360-degree image where the reflection image $\mathbf{R}$ contains the relevant information to the mixed image $\mathbf{M}$. However, the mixed image $\mathbf{M}$ also yields different characteristics from natural single layer images. Based on the multi-head self-attention mechanism of transformer, we implicitly estimate the regions of $\mathcal{M}$ and $\mathcal{R}$ by calculating the dot-product similarity between the tokens. Then the feed-forward network reconstructs the transmission image $\mathbf{T}$ from $\mathcal{M}$ by exploiting the information from the estimated $\mathbf{R}$. We design the encoder and decoder based on the U-shaped hierarchical network [21] using transformer blocks, since the hierarchical structure helps to increase the receptive field.

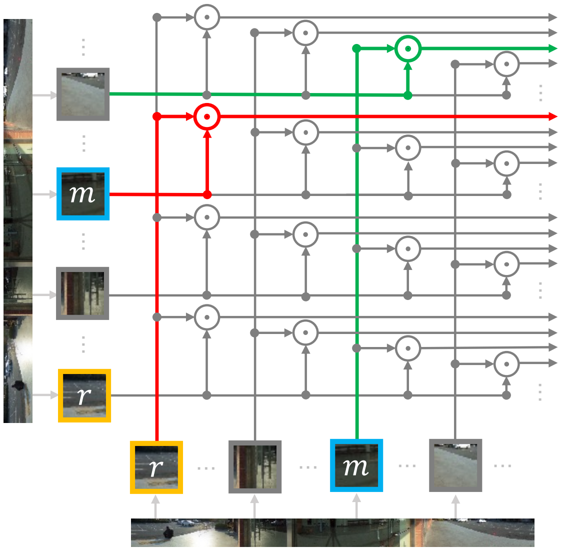


Figure 4. Illustration of the self-attention mechanism of transformer. Whereas the red colored dot product denotes the high attention between $m \in \mathcal{M}$ and $r \in \mathcal{R}$, the green colored dot product denotes low attention between $m \in \mathcal{M}$ and a certain token in single layer image regions outside of $\mathcal{R}$.

Horizontal windowing scheme. The square-shaped window based transformer is popularly employed for image restoration [16, 21], however it is not suitable for RR of 360-degree images due to the large field-of-view and the geometric distortion between $\mathcal{M}$ and $\mathcal{R}$. As illustrated in Figure 3 (a), $\mathcal{M}$ and $\mathcal{R}$ are usually located far from each other in a typical 360-degree image, and thus the tokens from $\mathcal{M}$ and $\mathcal{R}$ may not usually contained to a same window due to the limited size of window. Enlarging the window size to include both of $\mathcal{M}$ and $\mathcal{R}$ also brings the extra quadratic computational complexity.

To overcome this drawback, we redesign the window to have horizontally elongated shapes by setting the window size as $n \times W$, where W is the width of 360-degree image and n is a hyper-parameter. Figure 3 (b) shows the proposed windowing scheme where we see that the tokens from both of $\mathcal{M}$ and $\mathcal{R}$ are contained to a same window, which are then considered to measure the similarity via the self-attention mechanism of transformer.

Figure 4 illustrates the self-attention based region estimation for $\mathcal{M}$ and $\mathcal{R}$ by using the horizontal windowing scheme. Since the mixed image region exhibits different characteristics from the regions of single layer image, the attention between the tokens from inside and outside of $\mathcal{M}$, as depicted by the green colored dot product, would be highlighted and helps to estimate the region of $\mathcal{M}$. On the other hand, the features of $\mathbf{R}$ can be also observed in $\mathbf{M}$, and thus the attention between a certain token $r \in \mathcal{R}$ and the tokens in $\mathcal{M}$, as depicted by the red colored dot product, is boosted by relatively high similarity and eventually detects $\mathcal{R}$.

Laplacian pyramid. We also apply the Laplacian pyramid to use the multi-scale high-frequency features. Due to the photometric and geometric misalignment between the reflection image $\mathbf{R}$ embedded in the mixed image and the reference image extracted from $\mathcal{R}$, it is challenging to match $\mathcal{M}$ and $\mathcal{R}$. Since the high-frequency components, e.g., edges and contours, are useful for alignment, [9, 10] used the image gradient to directly match the images. In contrary, we implicitly align them based on deep semantic features encoded by the U-Net structure. We use the output of trainable convolutional neural network (CNN) initialized by the Laplacian filter [5] to extract the high frequency components.

3.2. Reflection Synthesis for Training

We generate a 360-degree image $\mathbf{I}$ containing synthetic reflection artifacts by using a pair of real 360-degree outdoor image $\hat{\mathbf{J}_\mathbf{o}}$ and indoor image $\hat{\mathbf{J}_\mathbf{i}}$ without reflection artifacts considering diverse geometric relation and photometric variation between the mixed and reference images within a 360-degree image.

Geometric relation. The reflection artifacts locally occur in a 360-degree image, and hence we first select the regions of $\mathcal{M}_o$ and $\mathcal{M}_i$ at the same locations in $\hat{\mathbf{J}_\mathbf{o}}$ and $\hat{\mathbf{J}_\mathbf{i}}$, respectively. We also select the region $\mathcal{R}_o$ corresponding to $\mathcal{M}_o$ at the opposite location to $\mathcal{M}_o$ with respect to the image center of $\hat{\mathbf{J}_\mathbf{o}}$. Then we take the transmission image $\hat{\mathbf{T}}$ from $\mathcal{M}_i$ and the reflection image $\hat{\mathbf{R}}$ from $\mathcal{R}_o$, respectively, following [9], and synthesize a mixed image $\hat{\mathbf{M}}$. However, [9] puts a strong assumption that the direction from the camera center to the glass center is orthogonal to the glass plane such that $\mathcal{R}$ lies at the opposite direction to $\mathcal{M}$ with respect to the image center. We mitigate this assumption to consider more generalized geometric relations between $\mathcal{M}$ and $\mathcal{R}$ by randomly trimming away the vertical sides of $\mathcal{M}$ and the corresponding areas of $\mathcal{R}$ accordingly.

Photometric adjustment. In typical Low-Dynamic Range (LDR) 360-degree images, the reference region usually exhibits brighter pixel intensities than that of the mixed region or even saturated intensity values due to dynamic range clipping. We consider this characteristics to synthesize the mixed image $\hat{\mathbf{M}}$ in the linear image space following [11]. Using the transmission image $\hat{\mathbf{T}}$ and the reflection image $\hat{\mathbf{R}}$ in the linear image space, we revise the image formulation model of $\hat{\mathbf{M}}$ as

$$\hat{\mathbf{M}}(\mathbf{x}) = \Omega(\mathbf{x})\hat{\mathbf{T}}(\mathbf{x}) + \Phi(\mathbf{x})\hat{\mathbf{R}}(\mathbf{x}). \tag{2}$$

We basically adopt the results of [27] to set the values of Ω and Φ in (2) which were exhaustively surveyed using the Monte Carlo simulation.

We then have a pair of the resulting 360-degree image $\hat{\mathbf{I}}$ such that

$$\hat{\mathbf{I}}(\mathbf{x}) = \begin{cases} \hat{\mathbf{M}}(\mathbf{x}), & \text{for } \mathbf{x} \in \mathcal{M}_o, \\ \hat{\mathbf{J}_\mathbf{o}}(\mathbf{x}), & \text{for } \mathbf{x} \notin \mathcal{M}_o, \end{cases} \tag{3}$$

and its ground truth image and $\hat{\mathbf{J}}$ without reflection artifacts as

$$\hat{\mathbf{J}}(\mathbf{x}) = \begin{cases} \hat{\mathbf{T}}(\mathbf{x}), & \text{for } \mathbf{x} \in \mathcal{M}_o, \\ \hat{\mathbf{J}_\mathbf{o}}(\mathbf{x}), & \text{for } \mathbf{x} \notin \mathcal{M}_o. \end{cases} \tag{4}$$

We apply the dynamic range clipping and the gamma correction to $\hat{\mathbf{I}}$ and $\hat{\mathbf{J}}$ to generate the pair of training images $\mathbf{I}$ and $\mathbf{J}$ in non-linear space, which are then used to train the network in an end-to-end manner.

4. Experimental Results

4.1. Implementation Details

We construct about $28K$ pairs of 360-degree images by using HDR outdoor 360-degree images [1] and LDR indoor 360-degree images without reflection artifacts [26]. Note that, before synthesis, we transform LDR images into linear space by applying inverse gamma correction, following [5, 25]. Among them, we set aside 121 images as the validation dataset. We down-sample the images to 256×512 size due to memory consumption and computational complexity. For inference, we use 60 real 360-degree images with reflection artifacts captured in various scenarios including the test images from [2, 9]. For the quantitative evaluation, we additionally captured 19 pairs of real 360-degree images and the corresponding ground truth images by using a portable glass pane.

We trained the network for 20 epochs using AdamW optimizer [17] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and set the weight decay as 0.02. We used the cosine decay strategy to decrease the learning rate from 2×10^{-4} to 1×10^{-6}, and

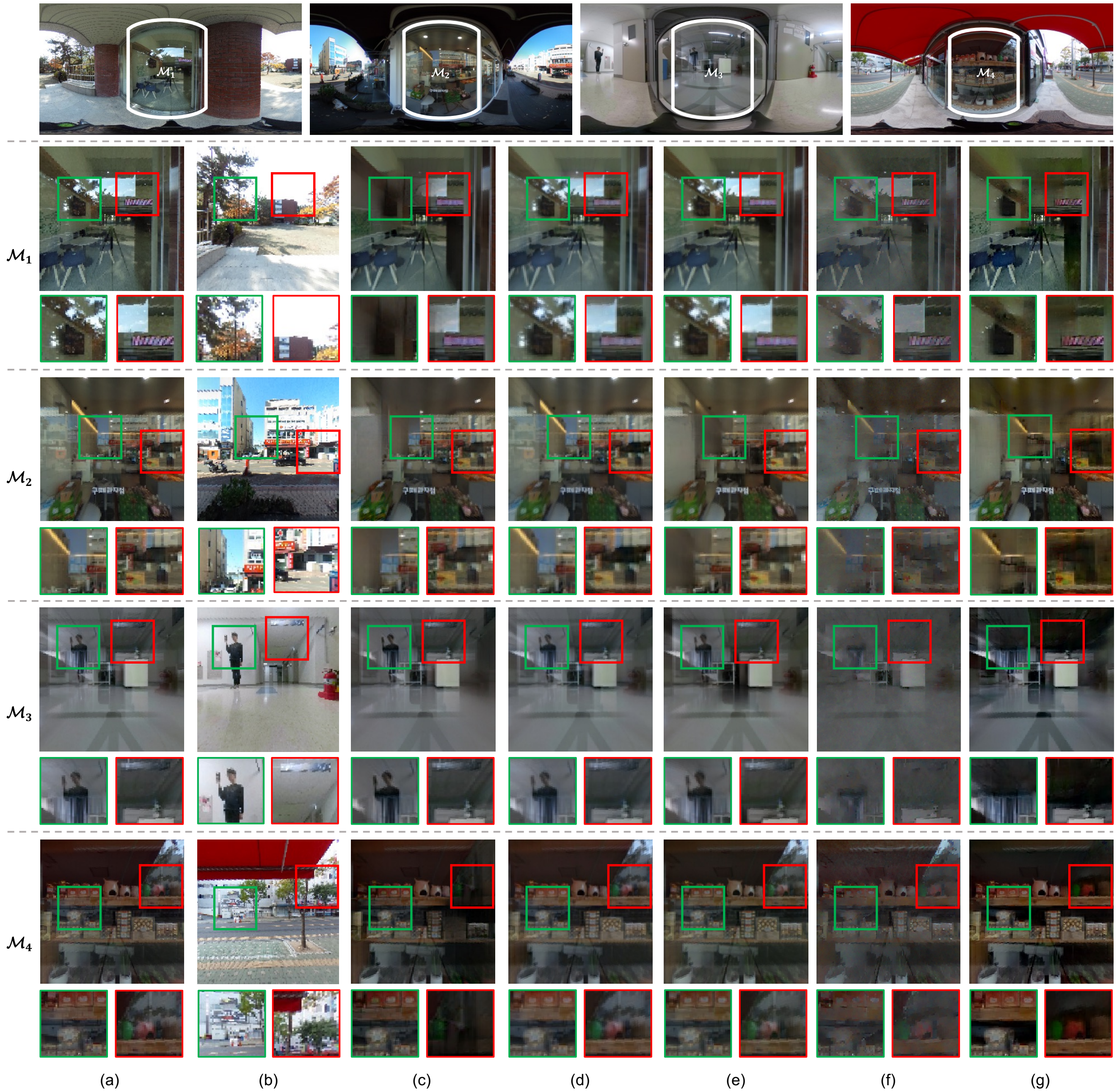


Figure 5. Qualitative results of the proposed method compared with that of the state-of-the art RR methods. (a) Input 360-degree images with reflection artifacts and (b) the reference image in $\mathcal{R}$. The RR results of (c) ERRNet [22], (d) IBCLN [13], (e) ABS [27], (f) ZS360 [9] and (g) the proposed method, respectively. We manually selected the reference images from $\mathcal{R}$ for visualization purpose

used the Charbonnier loss [3] with $\epsilon = 10^{-3}$. We set $n = 1$ and batch size as 4. We used the horizontal flip and shift for data augmentation.

4.2. Comparison With Existing Methods

In Figure 5, we first qualitatively compare the results of the proposed method with that of the three state-of-the art single-image RR methods (ERRNet [22], IBCLN [13], and ABS [27]) and the existing 360-degree image RR method (ZS360 [9]). For fair comparison, we re-train the compared methods by using our synthesized training dataset.

The reflection artifacts are bright and strong in $\mathcal{M}_1$ and $\mathcal{M}_2$ increasing the content ambiguity. The strong edges are observed as reflection artifacts in $\mathcal{M}_3$. $\mathcal{M}_4$ has a rel-

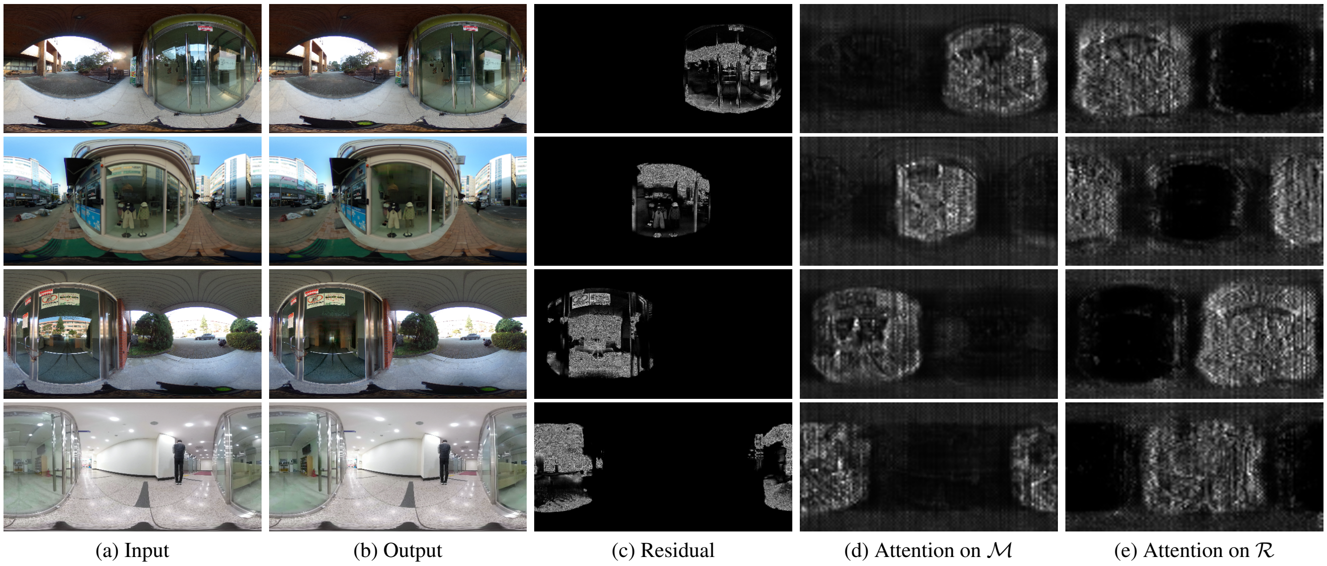

(a) Input
(b) Output
(c) Residual
(d) Attention on $\mathcal{M}$
(e) Attention on $\mathcal{R}$

Figure 6. Region detection performance of the proposed 360-degree image RR. (a) Input 360-degree images with real reflection artifacts. (b) The RR results of the proposed method. (c) The residual maps between (a) and (b). The intermediate attention maps highlighting (d) $\mathcal{M}$ and (e) $\mathcal{R}$, respectively.

atively dark transmission image $\mathbf{T}$. The single image RR methods [13, 22, 27] almost fail to remove the strong reflection artifacts due to the lack of the reference information, as shown in $\mathcal{M}_1$ and $\mathcal{M}_2$. In contrary, ZS360 [9] utilizes the reference $\mathcal{R}$ and alleviates the reflection artifacts of $\mathcal{M}_1$ and $\mathcal{M}_2$. However, it tends to blur the details of the transmission image and still remains some amount of the reflection artifacts. On the other hand, the proposed method not only removes the reflection artifacts but enhances the visibility of the transmission image, e.g., the interior walls, bread shelves, and curtain, outperforming the compared methods. Similarly, while the existing methods also fail to work on for $\mathcal{M}_3$ and $\mathcal{M}_4$, the proposed method effectively suppresses the reflection artifacts and clearly reconstructs the transmission images, as shown in the green box in $\mathcal{M}_3$ and the red box in $\mathcal{M}_4$. More experimental results can be found in the supplementary material.

Next, we quantitatively compare the performance of the proposed method with that of the existing methods using the 19 pairs of test images. We measure the PSNR, SSIM, LPIPS [24], and DISTS [4] for evaluation. Note that the proposed method achieves the best performance and surpasses the existing methods, as shown in Table 1.

4.3. Detailed Results of the Proposed Method

Performance of region detection. We feed the entire 360-degree image to our network without any annotation or user interaction to specify $\mathcal{M}$ and $\mathcal{R}$. As shown in Figures 6 (a) and (b), the proposed method successfully alleviates the reflection artifacts, even when the glass regions are not fixed to the image center. Note that only the mixed region of $\mathcal{M}$ is changed when we measure the difference between the input and output, as shown in Figure 6 (c). It means that the proposed method reliably detects the region of $\mathcal{M}$. To further analyze the RR performance of the proposed method, we plot the intermediate attention maps of the transformers. We found that some of the head of the transformer block highlight the mixed region $\mathcal{M}$ (Figure 6 (d)), while the others focus on the reference region $\mathcal{R}$ (Figure 6 (e)). This shows that the self-attention-based detection for $\mathcal{M}$ and $\mathcal{R}$ works as we expected in the proposed network.

Metric	Methods				
	ERRNet [22]	IBCLN [13]	ABS [27]	ZS360 [9]	Ours
PSNR (↑)	24.294	22.524	23.940	16.501	**27.028**
SSIM (↑)	0.9275	0.9384	0.9450	0.7020	**0.9634**
LPIPS (↓)	0.1436	0.1029	0.0905	0.3359	**0.0620**
DISTS (↓)	0.1039	0.0920	0.0788	0.1107	**0.0493**

Table 1. Comparison of quantitative performance of RR in terms of PSNR, SSIM, LPIPS and DISTS evaluated on our quantitative test dataset. The best scores are in bold.

Performance with arbitrary locations of $\mathcal{M}$. We evaluate the performance of RR with different locations of $\mathcal{M}$ in 360-degree images. We see that the sky and the structure of the trees are removed while the indoor structures are more vividly reconstructed, as shown in the first two rows in Figure 7 (d). Next, we also test our method on a more challenging case where the mixed region of $\mathcal{M}$ is separated

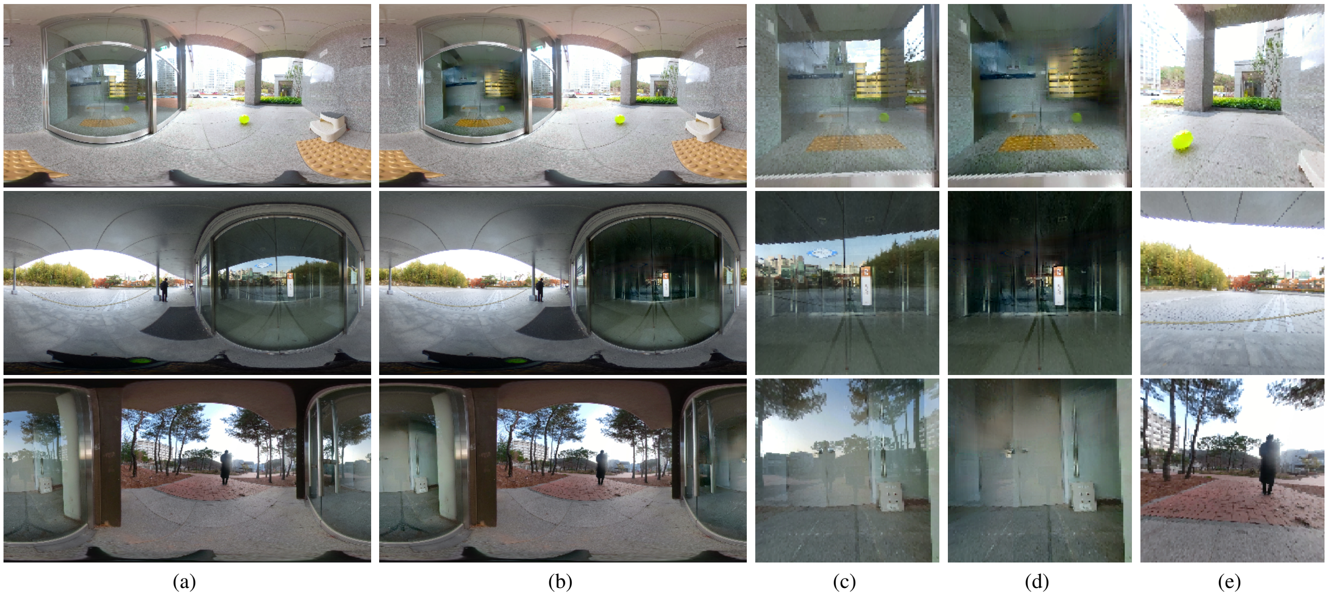


Figure 7. Results of RR with various locations of $\mathcal{M}$ within 360-degree images. (a) Input 360-degree images with real reflection artifacts. (b) The outputs of the proposed method. We show the images cropped from (c) $\mathcal{M}$ of input, (d) $\mathcal{M}$ of output, and (e) the reference image in $\mathcal{R}$, respectively.

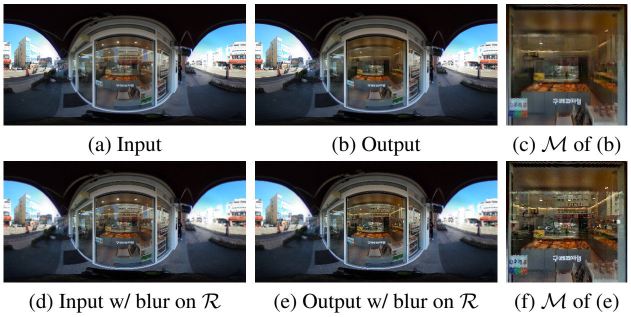

(a) Input
(b) Output
(c) $\mathcal{M}$ of (b)
(d) Input w/ blur on $\mathcal{R}$
(e) Output w/ blur on $\mathcal{R}$
(f) $\mathcal{M}$ of (e)

Figure 8. Analysis on the utilization of $\mathbf{R}$ for RR. (a) An input 360-degree image with reflection artifact and (b) the result of the proposed method. (c) A zoomed-in image on $\mathcal{M}$ of (b). (d) An input 360-degree image with Gaussian blurring on $\mathcal{R}$ and (e) the result of the proposed method. (f) A zoomed-in image on $\mathcal{M}$ of (e).

toward the two ends of the image, as shown in the third row of Figure 7. Surprisingly, we see that the proposed method removes the sky, tree, and building effectively, while reconstructing the transmission images clearly.

Utilization of reference information. We validate whether the proposed method utilizes the reference information in $\mathcal{R}$ to remove the reflection artifacts in the mixed image $\mathbf{M}$, which is the main motivation of 360-degree image RR to resolve the content ambiguity. We applied the Gaussian blurring with 3×3 kernel to the outside of $\mathcal{M}$ to suppress the information in $\mathcal{R}$. The first row in Figure 8 represents the input/output of the proposed method without blurring, while the second row shows the performance degradation due to the Gaussian blurring. As shown in Figures 8 (c) and (f), the proposed method alleviates the reflection artifacts but it fails to reconstruct $\mathbf{T}$ faithfully when $\mathcal{R}$ is degraded. This implies that the proposed RR method actually utilizes the information $\mathcal{R}$ for reflection removal.

4.4. Ablation Study

Transformer. We validate the effect of the transformer module in the proposed design. We test four cases: (1) without the self-attention mechanism (w/o transformer), using the square-shaped window with the size of (2) 8×8, (3) 16×16 and (4) the horizontal-shaped window with $n = 2$, respectively. Note that all the test cases yield low performance qualitatively and quantitatively compared to the proposed method, as shown in Figure 9 and Table 2.

Laplacian pyramid. We validate the effect of the Laplacian pyramid. Although we remove the Laplacian pyramid from the proposed method, it achieves high quantitative scores as shown in Table 2. However, it fails to remove the high-frequency components of reflection, remaining the reflection artifacts as shown in Figure 10 (c). This implies that the Laplacian pyramid helps to employ the high frequency features for 360-degree RR.

5. Conclusion

In this paper, we proposed a fully-automatic 360-degree image reflection removal method without requiring any user-annotation. Considering the mechanism of 360-degree

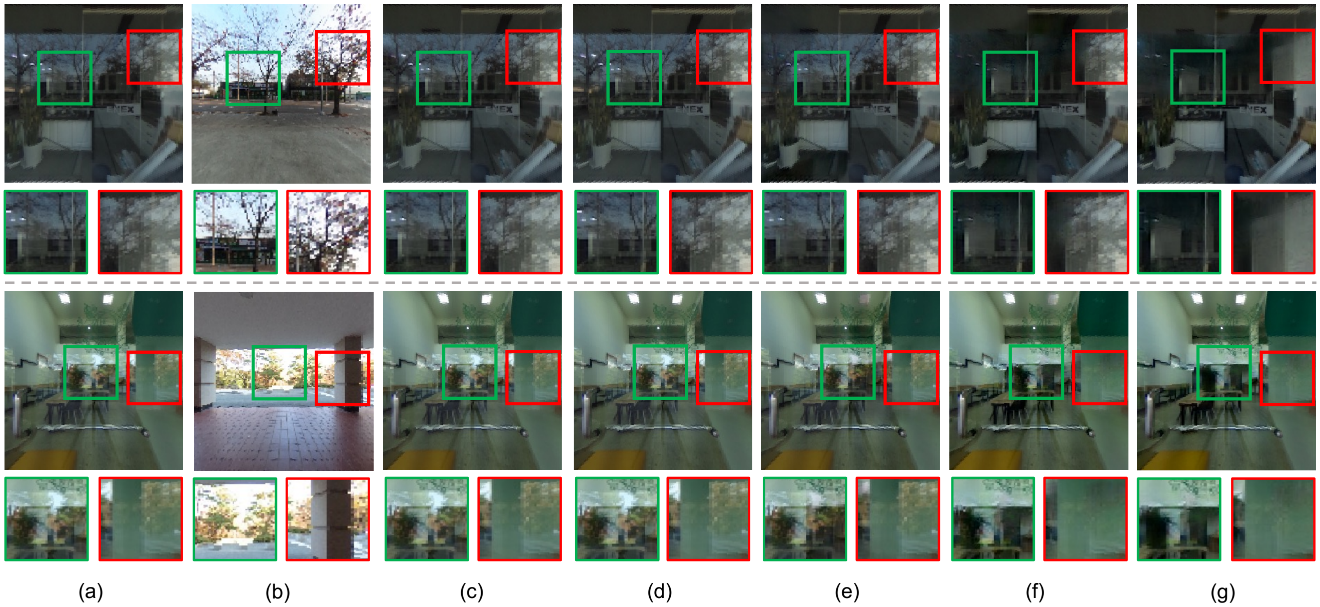


Figure 9. Ablation study on the transformer module. The cropped images of (a) the input mixed image and (b) the reference image in $\mathcal{R}$. The RR results (c) without the self-attention mechanism, (d) with the 8×8 square-shaped windowing scheme [21], (e) with the 16×16 square-shaped windowing scheme [21], (f) with the horizontal-shaped windowing scheme ($n = 2$), and (g) the proposed method.

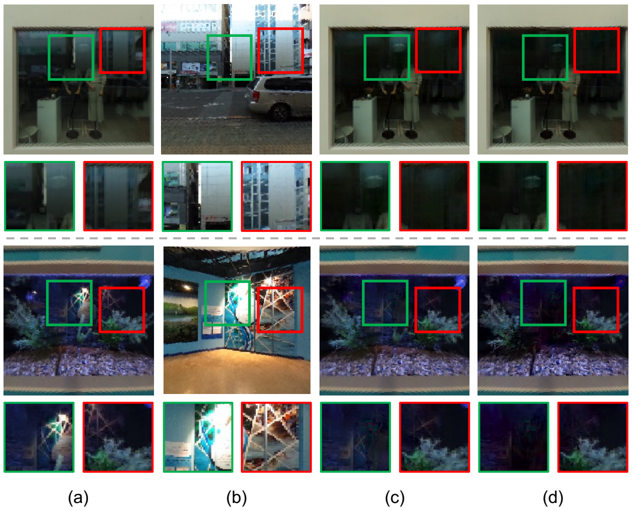


Figure 10. Ablation study on the Laplacian pyramid. The cropped image from (a) the input mixed image and (b) the reference image in $\mathcal{R}$. The RR results (c) without the Laplacian pyramid and (d) with the Laplacian pyramid, respectively.

Metric	Architecture					
	w/o trans.	8×8 win.	16×16 win.	$n = 2$	w/o Lap.	Prop.
PSNR (↑)	24.578	24.160	24.677	26.854	26.963	**27.028**
SSIM (↑)	0.9507	0.9499	0.9508	0.9611	0.9631	**0.9634**
LPIPS (↓)	0.0771	0.0745	0.0739	0.0646	0.0621	**0.0620**
DISTS (↓)	0.0642	0.0627	0.0604	0.0521	0.0501	**0.0493**

Table 2. Ablation study on the architecture design. We evaluate PSNR, SSIM, LPIPS, and DISTS on our quantitative test dataset. The best scores are in bold.

image reflection removal, we adopted the U-Net with transformer employing the horizontal-shaped windows to capture the long-range dependency between the mixed and reference regions. We additionally applied the laplacian pyramids to use multi-scale high frequency features. To obtain the training data, we synthesized the reflection artifacts on 360-degree images considering the geometric relation and photometric variation between the mixed and reference regions. The experimental results demonstrated that the proposed method automatically highlights the mixed region and the reference region, and removes the reflection artifacts in the mixed region faithfully by using the information of the reference region, outperforming the state-of-the-art RR methods qualitatively and quantitatively.

Acknowledgments

This work was supported by the National Research Foundation of Korea within the Ministry of Science and ICT (MSIT) under Grant 2020R1A2B5B01002725, and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub) and (No.2020-0-01336, Artificial Intelligence Graduate School Program(UNIST)).

References

[1] https://polyhaven.com/. 4

[2] https://github.com/bjhan1/zeroshot_reflection_removal. 4

[3] P. Charbonnier, L. Blanc-Feraud, G. Aubert, and M. Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *ICIP*, volume 2, pages 168–172 vol.2, 1994. 5

[4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE TPAMI*, 44(5):2567–2581, 2020. 6

[5] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *ICCV*, pages 5017–5026, 2021. 4

[6] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *ICCV*, pages 3238–3247, 2017. 1, 2

[7] Xiaojie Guo, Xiaochun Cao, and Yi Ma. Robust separation of reflection from multiple images. In *CVPR*, pages 2187–2194, 2014. 2

[8] Byeong-Ju Han and Jae-Young Sim. Reflection removal using low-rank matrix completion. In *CVPR*, pages 5438–5446, 2017. 2

[9] Byeong-Ju Han and Jae-Young Sim. Zero-shot learning for reflection removal of single 360-degree image. In *ECCV*, pages 533–548. Springer, 2022. 1, 2, 4, 5, 6

[10] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C Kot, and Boxin Shi. Panoramic image reflection removal. In *CVPR*, pages 7762–7771, 2021. 1, 2, 4

[11] Yuchen Hong, Qian Zheng, Lingran Zhao, Xudong Jiang, Alex C Kot, and Boxin Shi. Par2net: End-to-end panoramic image reflection removal. *IEEE TPAMI*, 2023. 1, 2, 4

[12] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *CVPR*, pages 5164–5173, 2020. 1, 2

[13] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E Hopcroft. Single image reflection removal through cascaded refinement. In *CVPR*, pages 3565–3574, 2020. 1, 2, 5, 6

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

[15] Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *CVPR*, pages 2752–2759, 2014. 1, 2

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[18] YiChang Shih, Dilip Krishnan, Fredo Durand, and William T Freeman. Reflection removal using ghosting cues. In *CVPR*, pages 3193–3201, 2015. 1, 2

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3

[20] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C Kot. Crrn: Multi-scale guided concurrent reflection removal network. In *CVPR*, pages 4777–4785, 2018. 1

[21] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *CVPR*, pages 17683–17693, 2022. 3, 8

[22] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *CVPR*, pages 8178–8187, 2019. 1, 5, 6

[23] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *ECCV*, pages 654–669, 2018. 2

[24] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 6

[25] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *CVPR*, pages 4786–4794, 2018. 1, 2, 4

[26] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *ECCV*, pages 519–535. Springer, 2020. 4

[27] Qian Zheng, Boxin Shi, Jinnan Chen, Xudong Jiang, Ling-Yu Duan, and Alex C Kot. Single image reflection removal with absorption effect. In *CVPR*, pages 13395–13404, 2021. 1, 2, 4, 5, 6