scientific reports

Check for updates

OPEN

High-resolution image reflection removal by Laplacian-based component-aware transformer

Songnan Chen[1,2] & Zhaoxu Feng[3]✉

Recent data-driven deep learning methods for image reflection removal have made impressive progress, promoting the quality of photo capturing and scene understanding. Due to the massive consumption of computational complexity and memory usage, the performance of these methods degrades significantly while dealing with high-resolution images. Besides, most existing methods for reflection removal can only remove reflection patterns by downsampling the input image into a much lower resolution, resulting in the loss of plentiful information. In this paper, we propose a novel transformer-based framework for high-resolution image reflection removal, termed as the Laplacian pyramid-based component-aware transformer (*LapCAT*). *LapCAT* leverages a Laplacian pyramid network to remove high-frequency reflection patterns and reconstruct the high-resolution background image guided by the clean low-frequency background components. Guided by the reflection mask through pixel-wise contrastive learning, *LapCAT* designs a component-separable transformer block (*CSTB*) which removes reflection patterns from the background constituents through a reflection-aware multi-head self-attention mechanism. Extensive experiments on several benchmark datasets for reflection removal demonstrate the superiority of our *LapCAT*, especially the excellent performance and high efficiency in removing reflection from high-resolution images than state-of-the-art methods.

Image reflection removal, which aims to reconstruct a clean background image from its reflection-contaminated observation, is a crucial yet challenging research topic in computer vision. Present data-driven deep learning methods based on generative models[1–3] have produced impressive results on regular-resolution images. The key challenge is that higher resolution (e.g. 4K) involves quadratically more computational cost and more intricate reflection patterns including diverse variants of reflection regions and scales.

Most conventional methods[4–9] for image reflection removal are built upon the physical model of reflection priors, which has validated its effectiveness in removing simple reflection patterns. However, one crucial limitation is the lack of powerful modeling for variable reflection patterns, which results in the inability to remove reflection components thoroughly in real-life scenes. This limitation is then mitigated by the generative models[10–12] based on convolutional neural networks (CNNs), which enjoy excellent capability of feature representation and select background components from an input image to reconstruct a reflection-free image via pixel-wise supervised learning. The progressive methods[13–15] for image reflection removal are further proposed to address the issue of reflection modeling. However, with the gradual increase in image resolution, a potential drawback of these methods is that they only work well on reflection patterns in regular-resolution images. This drawback leads to two negative consequences: 1) the huge computational cost and memory usage hinder the performance of CNN-based methods for reflection removal on high-resolution images; 2) higher resolution significantly increases the difficulty of reflection removal, since images with megapixels enjoy a greater diversity of reflection patterns.

A straightforward way to let CNN-based methods[10–15] focus on complete reflection patterns in ultra-high resolution images is to first downscale images into regular resolution by the interpolation operations, and then upscale it after removing reflection in a global perspective. The drawback is clear that the scale change of images results in undesired information loss and significantly reduces the image quality. To enlarge the receptive field of deep networks while processing high-resolution images, some methods[16,17] iteratively stack the convolution layers. However, they still suffer from the burden of computational cost in real-life application scenes for the increased image scale.

Inspired by the typical two-stream modeling methods[18,19], some ad hoc solutions[20–23] for ultra-high resolution image processing are proposed. The key idea of these methods is to process images from global and

[1]School of Mathematics and Computer Science, Wuhan Polytechnic University, Wuhan, Hubei, China. [2]Foshan Zhongke Innovation Research Institute of Intelligent Agriculture and Robotics, Foshan, Guangdong, China. [3]China United Network Communications Co., Ltd. Henan Branch, Zhengzhou, Henan, China. ✉email: iefengzhaoxu@163.com

local perspectives respectively. The global branch is designed to coarsely process scale-down but intact images to capture global information and high-level semantics, while the local branch focuses on conducting exquisitely image processing over local but high-resolution patches. In order to further model multi-scale information in the local branch, some multi-stage models[24–28], which progressively enlarge perceptive fields by changing the resolution of local patches in each stage, are proposed. Whilst such learning paradigm shows great potential to deal with reflection removal in ultra-high resolution images, due to the complexity of reflection patterns than other degradations, such as rain streak or haze. For reflection removal task, the multi-stage framework is hard to separate highly coupled components thoroughly, and yet often works awkwardly with low efficiency on ultra-high resolution images.

Achieving a trade-off between both the performance and efficiency within one model is a non-trivial task, especially separating reflection constituents from high-resolution images. To this end, we establish the Laplacian pyramid to separate the high-resolution image into high-frequency information with large image scales and low-frequency information with a lower scale. To thoroughly decouple the reflection constituents from the reflection-contaminated image, we further design a specialized component-aware transformer framework that leverages the self-attention mechanism to model the difference between the background and reflection constituents. As a result, the purified background constituents are able to reconstruct high-quality reflection-free images.

Compared to typical methods for image reflection removal, especially convolution-based generative models, our *LapCAT* framework benefits from the following advantages:

- We introduce a novel Laplacian pyramid-based transformer framework for high-resolution image reflection removal, which preserves high-frequency information by the Laplacian pyramid and removes reflection constituents in a low-resolution image. Thus, it can remove reflections thoroughly and reconstruct higher-quality background images than existing methods while processing high-resolution images.
- We design a Component-Separable Transformer Block (*CSTB*), which is the core unit of proposed transformer and follows a collaborative two-stream structure. Under the guidance of reflection masks through dense contrastive learning, *CSTB* is able to separate background content and reflections, and capture pixel-wise long-range interactions for completely modeling reflection patterns.
- We evaluate our model on both regular- and high-resolution datasets, including Real20, SIR^2, Nature and UHR4K, for image reflection removal respectively. Experimental results comprehensively demonstrate the superiority of our *LapCAT* over the existing state-of-the-art methods, especially removing reflection for high-resolution images.

Related work

Image reflection removal

Optimization-based method Many optimization-based traditional methods for reflection removal rely on handcrafted priors to remove reflection components from the background content. A prominent assumption is that the reflection layer is more blurring than the background content[29–32]. Inspired by this assumption, Wan et al.[7] leveraged Depth-of-Field (DoF) information and predicted the DoF confidence map to reconstruct more precise background edge maps. Besides blurry prior, Shih et al.[33] further discovered the ghosting effect of reflections and leveraged it to thoroughly remove reflection constituents. Further, Levin et al.[4] considered the guidance by user-interaction scheme to accurately locate the reflections regions. Based on such scheme, Heydecker et al.[34] proposed to preserve the veins and structures during reflection removal. Arvanitopoulos et al.[5] optimized the gradient intensity in the Laplacian domain for effective reflection suppression. Aiming to convert the reflection removal problem to a convex optimization problem, Yang et al.[35] adopted the discrete cosine transform and significantly improved the computation efficiency. In addition to above works, some methods attempted to capture multi-view images[36–39] to remove reflections. However, it is still challenging for these methods to cover all reflection patterns and often leads to unpleasant restoration results in real-life applications.

Deep learning-based method Leaning upon its powerful capability of feature representation, deep learning has led to significant improvements for image restoration[10–12,40–43]. Based on the prior knowledge of reflections, Fan et al.[10] introduced a two-stage framework that first estimates the edge map of target background image and further leverages it to reconstruct high-quality background image. Later, Wan et al.[41] designed a two-stream model to collaboratively predict the background as well as the edge map. Dong et al.[44] first located the reflection-contaminated regions with the input image to further iteratively refine the background image. To obtain implicit relation between the reflection and background layers, Yang et al.[13] first predicted the reflection layer and further input it as prior knowledge for thorough reflection removal. Motivated by the effectiveness of perceptual loss in other image restoration tasks[45], Zhang et al.[11] dilated the perceptive fields and performed perceptual loss on the reconstructed image to enhance the image quality. Additionally, Wei et al.[2] introduced an alignment-invariant loss to solve pixel misalignment issues in reflection removal datasets. Later, Li et al.[42] proposed to learn reflection-aware guidance from the contaminated images and then strengthen the reflection-free process. Recently, there are some methods[12,43] adopted pre-trained generative models to enlarge the training data to learn more robust patterns of reflection constituents. Nevertheless, the above methods are able to achieve promising performance when processing regular-resolution images, while significant performance degradation occurs for high-resolution image reflection removal.

Specialized methods for high-resolution image restoration

High-resolution image restoration is a challenging problem due to the explosion of computational cost and the increased diversity of modeling degradation patterns. Though the exploration of high-resolution image reflection removal is scarce, many image restoration tasks have involved this challenge. As an early inspiring work, Chen et al.[21] detach a full high-resolution image into multiple image patches for fidelity details and

an interpolated low-resolution for complete context. It further processes them respectively and aggregates the detached global-local information. However, it often suffers from complex model structure and unstable semantic modeling. Aiming to effectively decrease the computational cost from the degraded high-resolution images, Yi et al.[46] propose an efficient high-resolution image inpainting framework, which first restores the degraded image in low-resolution in a coarse-to-fine manner and then learns a contextual residual for degraded regions to counteract the pixel-wise errors caused by interpolation. Nevertheless, since image inpainting task often provides the mask of degraded regions as a prior, it is unable to cope with diverse reflection patterns, as well as precisely localizing reflection regions. To further strengthen the capability of degradation modeling, Zheng et al.[22] carefully designed a lightweight model that learns a low-resolution affine bilateral grid to directly model haze patterns in each RGB channel for 4K level image dehazing. Though it achieves significant performance for high-resolution image restoration, it shows powerless results of dealing with challenging reflection degradation patterns. As a result, high-resolution image reflection removal is not fully exploited and thus in this work we aim to design a transformer-based framework as the first specialized solution to cope with current challenges for high-resolution image reflection removal.

Laplacian pyramid-based component-aware transformer

Given a high-resolution image with reflection I, reflection removal aims to remove the reflection constituents *R* occluded before the background content and reconstruct a clear background image *B*. In this work, we propose a Laplacian pyramid-based component-aware transformer (*LapCAT*) for ultra-high resolution image reflection removal, which first leverages the Laplacian pyramid to downsample input high-resolution image and decouple it into multi-scale high-frequency maps and a low-resolution low-frequency map. The component-aware transformer (*CAT*) follows an encoder-decoder framework and is stacked by a designed Component-Separable Transformer Block (*CSTB*), which separates the reflection and background components from the input feature maps under the modeling of multi-head self-attention mechanism. By performing contrastive learning on sampled pixels according to reflection intensity, *LapCAT* is able to obtain the reflection mask as guidance for each token of input image while coping with different tokens.

In this section, we will first present an overview of the whole framework of our *LapCAT*. Then we will describe how it leverages the Laplacian pyramid to perform high-resolution background image reconstruction. Next, we elaborate on the structure of component-aware transformer and how to obtain the reflection mask through pixel-wise contrastive learning. Lastly, we show how to perform supervision on our *LapCAT* in an end-to-end manner.

Overall framework of LapCAT

As shown in Fig. 1, our proposed *LapCAT* architecture is composed of a Laplacian pyramid network and a component-aware transformer. Specifically, the Laplacian pyramid is employed to downsample the input image to a processable resolution and reconstruct the full-resolution background image referring to the restored low-resolution image and corresponding reflection mask. Then the downscale image is cropped into tokens, which are further performed component separation by designed component-aware transformer (*CAT*), which models various reflection patterns via proposed component-aware multi-head self-attention (*CA-MSA*) mechanism. Moreover, *LapCAT* performs pixel-wise contrastive learning to learn high-level consistency between background and reflection constituents, and estimate a binary reflection mask to locate the reflection regions. The overall framework is detailed below.

Reflection detection by contrastive learning

To locate the reflection-contaminated regions, in the training stage we train a reflection detector that performs pixel-wise contrastive learning in the latent feature space. The reflection detector first obtains a binary reflection distribution mask and then employs it as prior guidance to strengthen the modeling of reflection patterns by component-aware transformer. Even if contrastive learning[47,48] has demonstrated its effectiveness on pixel-wise semantic understanding, locating reflection pixels by contrastive learning is still highly challenging. The crux lies in the lack of pixel-wise labels for contrastive modeling. Therefore, we estimate the probabilistic distribution of reflection pixels by selecting the most representative pixels to perform pixel-wise contrastive learning.

Formulation Given a reflection-contaminated image $\mathrm{I} \in \mathbb{R}^{C\times H\times W}$, our reflection detector aims to learn a latent space $\mathscr{W}$ to capture the intrinsic semantic distinctions between the reflection and background pixels. Then it separates pixels into two clusters by performing clustering analysis and finally obtains a reflection distribution mask. Herein, $\mathrm{RM} \in \mathbb{R}^{1\times H\times W}$ is a binary map where the values of reflection-contaminated regions are 1 and clean background regions are filled with 0. To this end, our *LapCAT* introduces a probability-based training strategy of sample selection for contrastive learning tailored for reflection detection. Specifically, it first estimates the reflection intensity as sampling probability (SP) by calculating the difference between the input image I and corresponding groundtruth I_{gt} during training:

$$\mathrm{SP} = \mathrm{Softmax}(|\mathrm{I} - \mathrm{I}_{\mathrm{gt}}|). \tag{1}$$

Note that a larger SP implies stronger likelihood of a reflection pixel and a lower implies stronger likelihood of a background pixel. Instead of sampling from all pixels, we perform probabilistic sampling according to the value of SP to respectively collect positive and negative sets with high-confidence pixel candidates of reflection or background. Concretely, we first establish two candidate sets for further collection of positive samples $S_c^+ = \left\{r_i \mid SP_i <= t^+\right\}$ and negative samples $S_c^- = \left\{r_i \mid SP_i >= t^-.\right\}$ through thresholds $t^+ = 0.2$ and

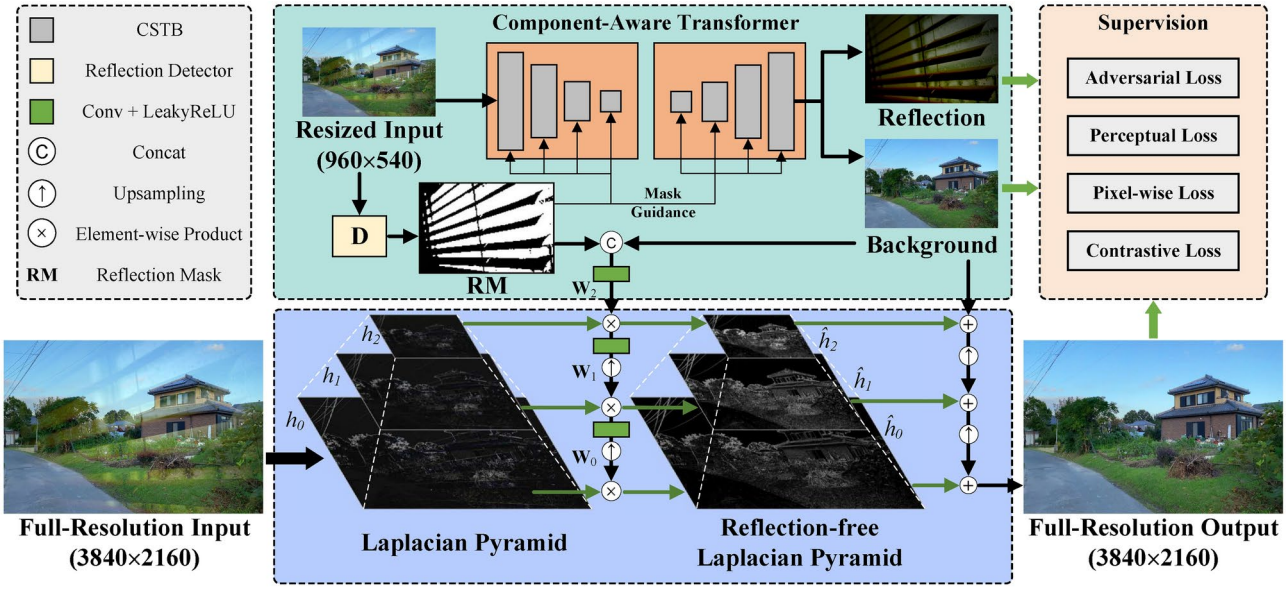

Figure 1. Architecture of the proposed Laplacian pyramid-based component-aware transformer (*LapCAT*), which mainly consists of two phases: reflection-free Laplacian pyramid and component-aware transformer. The reflection mask is obtained from reflection detector through pixel-wise contrastive learning and provides locations of reflection constituents with both phases. Note that the test sample is from the UHR4K-Syn[27] dataset.

$t^- = 0.8$, where r_i is the representation of the i-th pixel obtained by $\mathscr{F}_{\text{proj}}$. Then, by probabilistic sampling, we finally obtain the positive set and negative set for pixel-wise contrastive learning, which follows the formulation:

$$\begin{aligned} S^+ &= \left\{ s_j^+ \mid s_j^+ \in S_c^+, \, \mu_j \leq SP_j \right\}, \\ S^- &= \left\{ s_k^- \mid s_k^- \in S_c^-, \, \mu_k \geq SP_k \right\}, \end{aligned} \tag{2}$$

where S^+ and S^- are the positive and negative sets for contrastive learning, s^{c+} and s^{c-} are the items in the candidate sets, μ is the sampling probability from the normal distribution, and j and k are indexes of candidate samples. In this way, the pixel-wise contrastive learning loss is defined as:

$$\mathscr{L}_c = \frac{1}{N} \sum^{i} -\log \frac{\exp(q_i \cdot t_i^+)/\tau}{\sum \exp(q_i \cdot t_i^+) + \sum \exp(q_i \cdot t_i^-)/\tau}, \tag{3}$$

where q_i denotes the i-th extracted query, and τ is a temperature hyper-parameter. Under the supervision of pixel-wise contrastive learning in the latent space, pixels from the same constituents can be clustered together, and pixels from different constituents are pushed away. Afterward, by conducting clustering analysis to divide them into two clusters, *LapCAT* is able to obtain the reflection mask **RM**, which is defined by:

$$\text{RM} = \text{K-means}(\mathscr{F}_{\text{proj}}(\text{I})), \tag{4}$$

where $\mathscr{F}_{\text{proj}}$ is two fully convolution layers with a GELU[49] layer in-between for non-linear mapping.

Require:
High-resolution reflection image $I \in \mathbb{R}^{H\times W}$;
Reflection mask $RM \in \{0,1\}^{H\times W}$;
Restored low-resolution background $\hat{I}_B \in \mathbb{R}^{\frac{H}{2^k}\times\frac{W}{2^k}}$ by CAT;
Number of pyramid levels $k \geq 2$.
Ensure: Reconstructed high-resolution background $B_{\text{out}} \in \mathbb{R}^{H\times W}$
Initialize operators:
d$(\cdot)$: downsampling by factor of 2;
u$(\cdot)$: upsampling by factor of 2;
$\odot$: element-wise multiplication.
// 1. Construct Laplacian Pyramid:
$I_0 \leftarrow I$
for $\ell = 0$ **to** $k-1$ **do**
  $h_\ell \leftarrow I_\ell - \text{u}(I_{\ell+1})$ // High-freq residual;
  $I_{\ell+1} \leftarrow \text{d}(I_\ell)$ // Downsampling;
end for
// 2. Refine High-frequency Components:
$W_0 \leftarrow \text{Conv}\big(\text{LeakyReLU}([RM, \hat{I}_B])\big)$
for $\ell = 0$ **to** $k-1$ **do**
  $W_{\ell+1} \leftarrow \text{Conv}\big((\text{LeakyReLU}(W_\ell)\big)$ // Learnable weight map;
  $\hat{h}_\ell \leftarrow h_\ell \odot W_\ell$
end for
// 3. Reconstruct Full-resolution Background:
$B \leftarrow \hat{I}_k$ // Start from the smallest scale;
for $\ell = k-1$ **to** 1 **do**
  $B \leftarrow \text{u}(\hat{h}_\ell + B)$ // Iteratively add refined high-freq maps;
end for
$B_{\text{out}} \leftarrow \hat{h}_0 + B$ // High-resolution background.
return B_{out}

Algorithm 1. Laplacian Pyramid-based High-Resolution Background Reconstruction

Laplacian pyramid-based high-resolution background reconstruction

Laplacian pyramid (LP)[50] is widely applied in many vision tasks, such as image super-resolution[51] and image blending[50]. The LP aims to linearly decouple an image into high-frequency and low-frequency parts, and in this way the restored low-resolution reflection-free image can be reconstructed invertibly and efficiently to a high-resolution image. As shown in Fig. 1, the high-frequency maps are multi-scale and the low-frequency map enjoys lower resolution. Specifically, we denote the downsample operation as $d(\cdot)$ and upsample operation as $u(\cdot)$, respectively. Given a reflection-contaminated image $\text{I} \in \mathbb{R}^{H\times W}$, it first goes through a low-pass filter and then is downsampled into a low-frequency map $\mathbf{I_1} \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}}$, as $\mathbf{I_1} = d(\text{I})$. To reversely reconstruct the high-resolution background image, the LP records the residual map $h_0 = \text{I} - u(\mathbf{I_1})$. Besides, the LP iteratively performs such operation to reduce image resolution and obtains a sequence of low-frequency and high-frequency maps. In the reconstruction phase, the LP conducts the backward recurrence: $\mathbf{I_k} = u(\mathbf{I_{k+1}}) + h_k$, where k is the number of levels in the pyramid.

By performing the above operations, we establish the LP in our *LapCAT* framework based on multi-scale input images through bilinear interpolation. To be specific, we first extract their high-frequency residual maps $[h_0, h_1, h_2]$ to construct the Laplacian pyramid. Next, guided by a binary Reflection Mask **RM** (see "Component-aware transformer for reflection removal" section) and the restored low-resolution background image $\hat{\mathbf{I}}_\mathbf{B}$ (see "Laplacian pyramid-based high-resolution background reconstruction" section) by *CAT*, LP learns an updating weight $\mathbf{W_i} = \text{Conv}([\text{RM}, \hat{\mathbf{I}}_\mathbf{B}])$ to refine the multi-scale high-frequency maps. Thus, it enables to obtain reflection-free high-frequency maps:

$$\hat{h}_i = h_i \odot \text{W}_i, \tag{5}$$

where $\odot$ is the Hadamard product, and $\hat{h}_i$ is the high-frequency components in the i-th level of Laplacian pyramid. Lastly, *LapCAT* reversibly reconstructs the full-resolution background image from the restored low-resolution image B_l and iteratively purifies high-frequency maps $[\hat{h}_0, \hat{h}_1, \hat{h}_2]$:

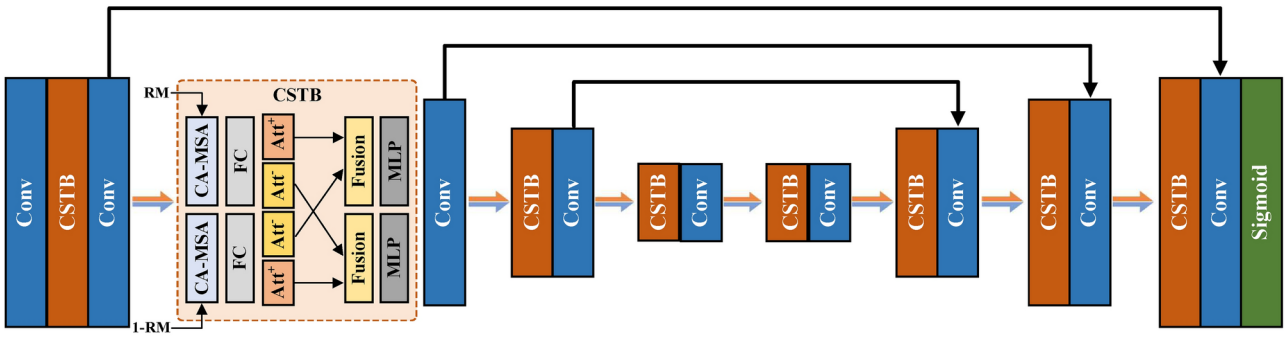

Figure 2. Structure of the proposed U-shaped component-aware transformer.

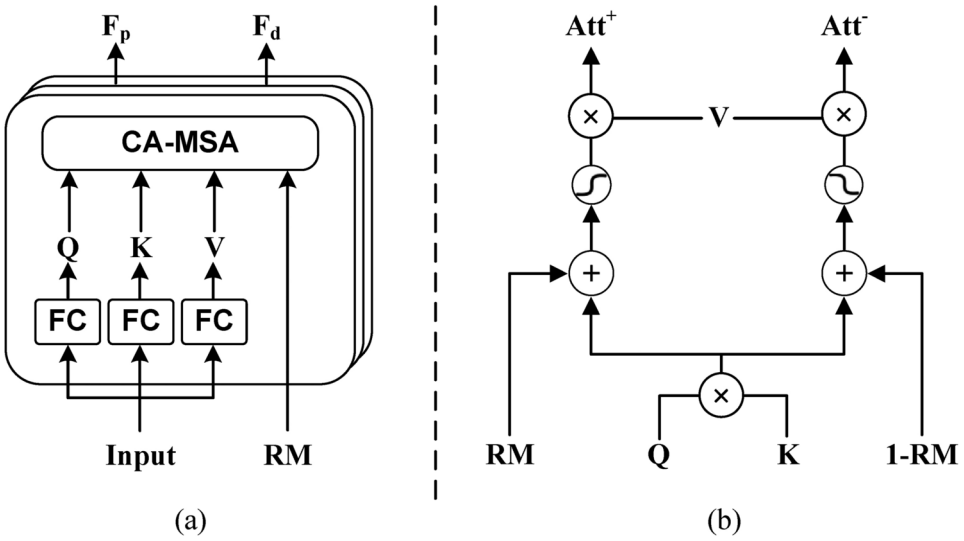

Figure 3. (**a**) Component-separable transformer block. (**b**) Component-aware self-attention mechanism.

$$\hat{B} = \hat{h}_0 + u(\hat{h}_1 + u(\hat{h}_2 + B_l)), \tag{6}$$

where $\hat{B}$ is the reconstructed reflection-free image.

Component-aware transformer for reflection removal

Next we elaborate on the structure of component-aware transformer and how it removes reflection constituents. The goal of component-aware transformer is to thoroughly separate the reflection layer from the background image by modeling the semantic patterns of reflection constituents in different levels of feature spaces. Specifically, we employ the U-shaped vision transformer that exploits shifted windows[52] to save calculation cost and develop the typical transformer block in two perspectives: 1) it introduces a component-aware multi-head self-attention (CA-MSA) mechanism to perform component separation between the background and reflection constituents; 2) under the guidance of reflection mask obtained by pixel-wise contrastive learning, both the transformer block and Laplacian pyramid are able to focus on restoring regions with obvious reflection contamination.

Component-separable transformer block Our component-aware transformer is stacked by a novel transformer block variant in different stages to thoroughly separate complex reflection patterns from the background content. As shown in Fig. 2, the information flow follows a collaboratively dual-branch manner in each component-separable transformer block. Particularly, we leverage the proposed component-aware multi-head self-attention mechanism to separate input features F into the preserved features $\mathbf{F_p}$ and deserted features $\mathbf{F_d}$, and deliver them into the background branch and reflection branch, respectively:

$$\begin{aligned} \mathbf{F_p^+}, \mathbf{F_d^+} &= \mathrm{FC}(\text{CA-MSA}(\mathbf{F}^+)), \\ \mathbf{F_p^-}, \mathbf{F_d^-} &= \mathrm{FC}(\text{CA-MSA}(\mathbf{F}^-)), \end{aligned} \tag{7}$$

where $\mathbf{F}^+$ and $\mathbf{F}^-$ denote the features of background and reflection branch, respectively, and FC is the fully connected layer. Next, we concatenate the preserved features in current branch and deserted features $\mathbf{F_d}$ in the other branch and obtain aggregated features $\mathbf{F_a}$ by leveraging an MLP:

$$\begin{aligned} \mathbf{F}_\mathbf{a}^+ &= \mathrm{MLP}(\mathrm{Concat}[\mathbf{F}_\mathbf{p}^+, \mathbf{F}_\mathbf{d}^-]), \\ \mathbf{F}_\mathbf{a}^- &= \mathrm{MLP}(\mathrm{Concat}[\mathbf{F}_\mathbf{p}^-, \mathbf{F}_\mathbf{d}^+]), \end{aligned} \tag{8}$$

where Concat denotes the concatenation operation. As illustrated in Fig. 2, we employ a convolution layer after each transformer block to reconstruct spatial structure of the intact image. Note that we also adopt the skip-connection operation across stages, which provides the alternative path for the gradient with backpropagation and addresses the vanishing gradient problem during training.

Component-aware multi-head self-attention Our component-aware transformer captures long-range interactions between pixels to perform component separation by component-aware multi-head self-attention according to the mathematical definition[13,53] of image reflection removal:

$$\mathrm{I} = \alpha f_1(\mathbf{I_B}) + \beta f_2(\mathbf{I_R}), \tag{9}$$

where α and β are the coefficients, and *f* denotes the nonlinear mappings for background and reflection image.

Our key design in CA-MSA aims to concentrate on precisely separating the reflection component from the background content under the guidance of predicted Reflection Mask (see "Reflection detection by contrastive learning" section). Specifically, the input features of CA-MSA are first projected into Q, K, and V through fully connection (FC) layers. Then, we further introduce a reflection mask (RM) that employs contrastive learning to provide pixel-wise prior of reflection intensities within the input degraded image. Mathematically, each CA-MSA produces two separate attention maps:

$$\begin{aligned} \mathrm{Att}^+ &= \mathrm{Softmax}(\frac{\mathrm{QK}^T + \mathrm{RM}}{\sqrt{d_k}})\mathrm{V}, \\ \mathrm{Att}^- &= \mathrm{Softmin}(\frac{\mathrm{QK}^T + (1 - \mathrm{RM})}{\sqrt{d_k}})\mathrm{V}, \end{aligned} \tag{10}$$

where Att^+ and Att^- are the attention maps targeting the background and reflection content respectively, and $\sqrt{d_k}$ is the scaling factor. This process facilitates to precise component separation for the following two reasons: (i) Since contrastive learning in our method effectively models the distributions of reflection patterns and background content in pixel level, the reflection mask is able to provide strong prior while separating different components through typical self-attention. (ii) The attention maps Att^+ and Att^- from our CA-MSA are complement, and our Transformer model is composed of stacked Component-Separable Transformer blocks with CA-MSA operations. Thus, it is able to progressively separate the reflection component under effective supervision and finally restore a clean background image.

Intuition The rationale behind our design is that our CAT model serves as a collaborative information distiller between background and reflection constituents guided by an estimated reflection mask. Each Component-Separable Transformer Block can use the preserved feature $\mathbf{F_p}$ and deserted feature $\mathbf{F_d}$ for background and reflection reconstruction. As described in Equation 7, the total information of $\mathbf{F}^+$ and $\mathbf{F}^-$ is equivalent to that in $\mathbf{F}_\mathbf{a}^+$ and $\mathbf{F}_\mathbf{a}^-$. This property guarantees no information flowing away from the interactive process, which substantially avoids the problems of vanishing/exploding gradients and redundant features. Besides, Equation 8 shows that $\mathbf{F}_\mathbf{p}^+$ is complementary to $\mathbf{F}_\mathbf{d}^-$, and $\mathbf{F}_\mathbf{d}^+$ is complementary to the $\mathbf{F}_\mathbf{p}^-$. By merging the complementary counterparts, there is no wasted information in our framework and the repeated use of CA-MSA leads to gradually enhancing the background stream while attenuating the reflection constituents, thus it is able to show promising performance while coping with challenging reflection patterns for high-resolution images.

End-to-end supervised parameter learning

We optimize the parameters of our *LapCAT* in an end-to-end manner. Besides pixel-wise contrastive learning loss $\mathscr{L}_c$, there are three loss functions, multi-scale pixel reconstruction loss $\mathscr{L}_{\mathrm{pixel}}$, multi-scale perceptual loss $\mathscr{L}_{\mathrm{perc}}$, and adversarial loss $\mathscr{L}_{\mathrm{adv}}$, to train our *LapCAT*.

Multi-scale pixel reconstruction loss We employ the L_1 loss to learn the pixel-level reconstruction for each resolution of background image and reflection image:

$$\mathscr{L}_{\mathrm{pixel}} = \alpha_1 \sum_{i=1}^{N} \mathscr{L}_1 \left(\mathrm{I}_{\mathrm{gt}}^i, \hat{\mathrm{I}}^i \right) + \beta_1 \mathscr{L}_1 \left(\mathbf{I}_{\mathrm{gt}}, \hat{\mathrm{I}}_{\mathrm{B}}^{\mathrm{out}} \right), \tag{11}$$

where $\hat{\mathrm{I}}^i$ denotes the separated image in the *i*-th stage, and $\hat{\mathrm{I}}_{\mathrm{B}}^{\mathrm{out}}$ is the final result by Laplacian Pyramid. Besides, we obtain various scales of groundtruth $\mathrm{I}_{\mathrm{gt}}^i$ by bilinear interpolation. Empirically, we set $\alpha_1 = 0.1$ and $\beta_1 = 0.5$.

Multi-scale perceptual loss To learn the consistency of semantic information, we perform supervision of the perceptual loss[45] in each resolution:

$$\mathscr{L}_{\mathrm{perc}} = \alpha_2 \sum_{i=1}^{N} \mathscr{L}_{\mathrm{VGG}} \left(\mathrm{I}_{\mathrm{gt}}^i, \hat{\mathrm{I}}^i \right) + \beta_2 \mathscr{L}_{\mathrm{VGG}} \left(\mathrm{I}_{\mathrm{gt}}, \hat{\mathrm{I}}_{\mathrm{B}}^{\mathrm{out}} \right), \tag{12}$$

where $\mathscr{L}_{\mathrm{VGG}}$ denotes perceptual distance between two images measured by a pre-trained VGG-19[54]. Empirically, we set $\alpha_2 = 0.1$ and $\beta_2 = 0.5$.

Conditional adversarial loss, which encourages the final separated image $\hat{\mathrm{I}}^{\mathrm{out}}$ to be as realistic as the ground-truth image I_{gt}. We employ the spectral normalization[55] in our discriminator to stabilize the adversarial learning process:

$$\mathscr{L}_{\mathrm{adv}} = -\mathbb{E}_{\mathrm{I}_{\mathrm{gt}} \backsim \mathbb{P}_{\mathrm{LapCAT}}}[D^{sn}(G(\mathrm{I}))], \tag{13}$$

where D^{sn} is the discriminator with the spectral normalization operation after each convolution layer. In sum, the total loss function is defined as follows:

$$\mathscr{L} = \lambda_1 \mathscr{L}_{\mathrm{pixel}} + \lambda_2 \mathscr{L}_{\mathrm{perc}} + \lambda_3 \mathscr{L}_{\mathrm{adv}} + \lambda_4 \mathscr{L}_{\mathrm{c}}. \tag{14}$$

By tuning on a held-out validation set, the hyper-parameters, we empirically set $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_2 = 0.01$, and $\lambda_4 = 0.01$.

Experiments

In this section, we perform experiments to demonstrate the effectiveness of our proposed *LapCAT*. We first elaborate the experimental settings, including datasets, evaluation metrics, and implementation details. Next, we compare the experimental results with state-of-the-art methods for reflection removal and further analyze how LapCAT dominates in both performance and model efficiency while coping with high-resolution images. Finally, we conduct ablation study to investigate the effectiveness of proposed techniques and further discuss the problems to be solved in future work.

Experimental settings

Datasets We evaluate our method for image reflection removal using four real-world benchmark datasets: SIR^2[53], Nature[14], Real20[11], and UHR4K[27]. According to the resolution of images, we roughly divide them into regular-resolution datasets (SIR^2[53] and Nature[14]) and high-resolution datasets (Real20[11] and UHR4K[27]).

SIR^2 dataset The SIR^2 dataset[53] is a real-world benchmark dataset to evaluate the performance of reflection removal, and all images have a resolution 540×400. It consists of three sub-datasets, Solid, Postcard, and Wild. 1) Solid dataset, composed of 200 triplets of images describing indoor solid object scenes; 2) Postcard dataset, which is a dataset containing 199 triplets of images obtained from postcards; 3) Wild dataset which contains 55 triplets of images about wild scenes.

Nature The Nature dataset[14] contains 220 real-world image pairs, and 200 images are used for training and the rest for evaluation. The images are captured under seven considerations to simulate diverse image conditions and the resolution of images in the nature dataset is 600×400.

Real20 The Real20 dataset[11] includes 90 images for training and uses 20 images for evaluation, which are captured with a portable glass in front of the camera under four conditions. The average resolution of images in Real20 dataset is about 1106×902 and altogether 83 different scenes are included.

UHR4K The UHR4K dataset[27] is a 4K (3840×2160) level dataset that includes both synthetic and real-world paired images for ultra high-resolution image reflection removal:

1) UHR4K-Syn. The UHR4K-Syn dataset extracts 2,167 frames from numerous 4K videos to simulate real-world scenes. It synthesizes the reflection image by applying the Gaussian kernel to blur the images of DIV2K dataset[56]. Finally, the degraded images are synthesized according to the physical principle of reflection formation. Consequently, UHR4K-Syn has 2,167 image pairs, and divides 2,117 image pairs for training and 50 image pairs for evaluation.
2) UHR4K-Real. The UHR4K-Real dataset collects real-world 4K images using two cameras, Nikon D300 and Huawei mobile phone. Altogether 336 image pairs(about 116 real-world scenes) are captured and it divides them into 316 image pairs for training and 20 image pairs for evaluation. Thus, the UHR4K-Real data is at present one of the most challenging datasets that has both high-resolution and diverse scenes.*Evaluation metrics* We adopt two commonly used evaluation metrics in low-level vision to measure the quality of the generated background images in our experiments quantitatively: *PSNR* and *SSIM*. *PSNR* computes the peak signal-to-noise ratio in decibels between two images while *SSIM* measures the perceptual similarity between two images. Higher value of *PSNR* or *SSIM* denotes higher quality of the restored background image.

Implementation details We implement our *LapCAT* in distribution mode with 4 RTX 3090 GPUs under Pytorch framework. Adam[57] is employed for gradient descent optimization with batchsize set to 8. The initial learning rate is set to be 1×10^{-4} and the training process takes maximally 100 epochs. Random flipping, random cropping and resizing are used for data augmentation. To have fair comparisons between different methods for reflection removal, we ensure that all methods are optimized on the same training set and evaluated on the same test set following the same optimization and evaluation protocols.

Results analysis

Baselines (1) *RmNet*[12], which specially trains a generative model to obtain more training samples, and thus the model for reflection removal is able to become more robust; (2) *ERRNet*[2], which focuses on exploiting the misaligned training data and multi-scale features fusion to strengthen the performance of reflection removal; (3) *IBCLN*[14], which is a cascaded network that can iteratively refine the quality of synthesized background images; (4) *Zou* ***et al.***[3] introduces an adversarial learning-based model to improve the performance of image separation;

(5) *CFDNet* [58], which performs contrastive supervision in the latent feature space to improve the performance of feature decoupling for reflection removal; (6) *DMGN*[15], which designs a residual deep-masking cell to filter out undesired information and restore a clean background image in a coarse-to-fine manner; (7) *MPRNet*[26] serves as a general framework for image restoration, which adopts a progressive strategy to model noise patterns; 98) *LAS* [44], which proposes to detect the guidance map of reflection regions to locate the reflection patterns and further leverages it to iteratively remove reflection components; (9) *YTMT* [59] is a two-stream framework for reflection removal that aims to construct block-wisely communication with each branch; (10) *Restormer*[60], which is an efficient Transformer model by developing the multi-head attention mechanism to capture long-range pixel interactions; (11) *Zheng **et al.***[22], which is a specialized method for ultra-high resolution image dehazing and uses the bilateral filtering to reconstruct full-resolution images efficiently; (12) *GLSGN* [27] is a global-local stepwise generative network that progressively learns multiple consistencies between different pathways to efficiently restore ultra-high resolution images;

Quantitative comparison with state-of-the-art methods

We conduct experiments to compare our *LapCAT* with twelve state-of-the-art methods[2,3,12,14,15,22,26,27,44,58–60] for image reflection removal. Table 1 lists the quantitative results of different methods for image reflection removal on five real-world benchmark datasets in terms of PSNR and SSIM. In particular, according to the resolution of benchmark datasets, we denote Nature[14] and SIR^2[53] as regular-resolution datasets and Real20[11] and UHR4K[27] as high-resolution datasets.

Results on regular-resolution datasets As listed in Table 1, our *CAT* model achieves the best performance on regular-resolution datasets, including SIR^2[53] and Nature[14]. While our model outperforms other competing methods, methods based on reflection location can boost the performance over other priors based state-of-the-art methods. For instance, LAS[44] predicts a confidence map to locate reflection constituents and further employs it to model reflection patterns by convolution-based models. In contrast, our *CAT* framework synthesizes a precise binary mask to locate the reflection pixels within the input image through pixel-wise contrastive learning. Furthermore, compared to the state-of-the-art Transformer model for image restoration Restormer[60], our *CAT* develops a component-aware self-attention mechanism that captures long-range interactions between pixels and explicitly injects the predicted mask into it to refine the details of reflection-contaminated regions. Additionally, though Softmax/Softmin activation functions are first proposed by CFDNet[58], our *LapCAT* first introduces an iterative scheme to gradually remove reflection constituents with the guidance of reflection mask in the latent feature space and finally restores a cleaner background image. As a result, the comparisons on regular-resolution datasets demonstrate the superiority of our Transformer framework while modeling various reflection patterns.

Results on high-resolution datasets In order to investigate the performance of our method on high-resolution images, we conduct experiments to compare our *LapCAT* with other state-of-the-art methods on UHR4K[27] (3840×2160) and Real20[11] (avg. 1106×902) datasets.

We can make the following observations from the experimental results presented in Table 1. First, our *LapCAT* outperforms other state-of-the-art methods for image reflection removal by a large margin on real-world high-resolution datasets, which demonstrates the superiority of our proposed Laplacian pyramid stage for high-resolution image reconstruction. Besides, benefiting from the powerful capability of feature decoupling

Method	Real20[11] (1106×902)		Nature[14] (600×400)		SIR^2[53] (540×400)		UHR4K-Syn[27] (3860×2140)		UHR4K-Real[27] (3860×2140)	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
RmNet[12]	19.47	0.748	19.07	0.755	20.74	0.836	17.05	0.829	20.86	0.808
ERRNet[2]	18.87	0.735	20.79	0.796	22.60	0.856	17.54	0.818	21.69	0.819
IBCLN[14]	19.99	0.759	23.57	0.783	22.68	0.860	20.66	0.879	22.59	0.824
CFDNet[58]	18.90	0.706	23.79	0.811	23.89	0.884	16.99	0.785	20.84	0.773
Zou *et al.*[3]	19.76	0.752	22.34	0.803	23.52	0.877	18.89	0.841	20.57	0.799
DMGN[15]	19.26	0.745	23.23	0.839	23.41	0.875	20.10	0.865	22.91	0.833
MPRNet[26]	21.37	0.781	23.42	0.848	23.82	0.880	21.26	0.888	22.71	0.821
LAS[44]	22.09	0.786	23.45	0.810	24.25	0.900	22.70	0.885	21.94	0.828
YTMT[59]	20.30	0.735	23.85	0.810	24.08	0.894	19.23	0.850	20.22	0.733
Zheng *et al.*[22]	20.61	0.756	22.59	0.782	22.46	0.868	23.57	0.886	20.08	0.712
Restormer[60]	21.89	0.778	23.98	<u>0.852</u>	24.25	0.889	23.07	0.882	22.77	0.822
V-DESIRR[61]	<u>22.94</u>	0.801	23.83	0.808	**26.49**	0.902	23.82	0.890	22.65	0.825
GLSGN[27]	22.20	0.790	<u>24.27</u>	**0.856**	24.11	0.903	<u>25.96</u>	<u>0.911</u>	<u>24.35</u>	<u>0.841</u>
DSRNet[62]	22.32	<u>0.806</u>	22.26	0.801	25.70	<u>0.919</u>	22.96	0.871	23.23	0.816
LGIRS[63]	22.47	**0.809**	23.87	0.812	<u>25.86</u>	**0.921**	23.87	0.889	22.57	0.804
CAT (ours)	22.32	0.794	**24.31**	**0.856**	24.29	0.899	23.69	0.891	22.86	0.818
LapCAT (ours)	**23.01**	**0.809**	23.95	0.849	24.16	0.897	**26.32**	**0.924**	**25.18**	**0.866**

Table 1. Quantitative results of different models for image reflection removal on five datasets in terms of PSNR and SSIM. The best results are in bold and the second best results are underlined.

by the developed component-separable transformer block, *CAT* without Laplacian pyramid can also obtain comparable performance with these ad hoc methods for high-resolution image restoration. It is reasonable since effective long-range interactions by self-attention mechanism contribute to thoroughly modeling reflection patterns within the input image. Second, compared to specialized methods for high-resolution image restoration, such as Zheng *et al.*[22] and GLSGN[27], our method still boosts the performance over them by 0.81 dB on Real20 and 0.83 dB PSNR on UHR4K-Real using the proposed Laplacian pyramid-based reconstruction strategy. Zheng *et al.*[22] leverages the bilateral filtering to efficiently restore high-resolution images. However, due to the diversity of reflection patterns, such model is not satisfied to adequately model them. GLSGN[27] is a multi-pathway framework that learns the global-local consistency between pathways. Although it achieves the second best performance, it is still challenging for it to remove complex reflections thoroughly owing to the lack of capturing long-range interactions between pixels. While both methods are designed specifically for high-resolution image restoration, our *LapCAT* enjoys higher performance due to both effective constituent separation and high-resolution detail reconstruction.

Qualitative comparison with state-of-the-art methods

Visual comparison on regular-resolution images To evaluate the visual performance for reflection removal by our proposed model on regular-resolution images, we perform a qualitative comparison with other state-of-the-art methods on real-world test images from SIR^2[53] and Nature[14]. Specifically, we visualize the restored results by different methods in Fig. 4. The results show that our model is able to restore cleaner background images than other competing methods on regular-resolution images and the obvious superiority of our method mainly benefits from two factors: 1) our reflection detection mechanism can provide the location of reflection constituents so that the details of reconstructed images can be obviously improved; 2) the developed component-aware self-attention mechanism captures long-range interactions between pixels, which models reflection patterns more precisely. Besides, its two-stream collaborative structure leads to effective feature decoupling. Thus, almost no obvious reflection constituents remain in the second sample in Fig. 4 by our method.

Visual comparison on high-resolution images We perform a visual comparison on high-resolution datasets for reflection removal to investigate whether our model enjoys obvious advantages of reflection removal on high-resolution images compared to the baseline methods. Figure 5 illustrates four examples of high-resolution image reflection removal from Real20[11] and UHR4K[27] datasets. In these cases, it is quite challenging to thoroughly

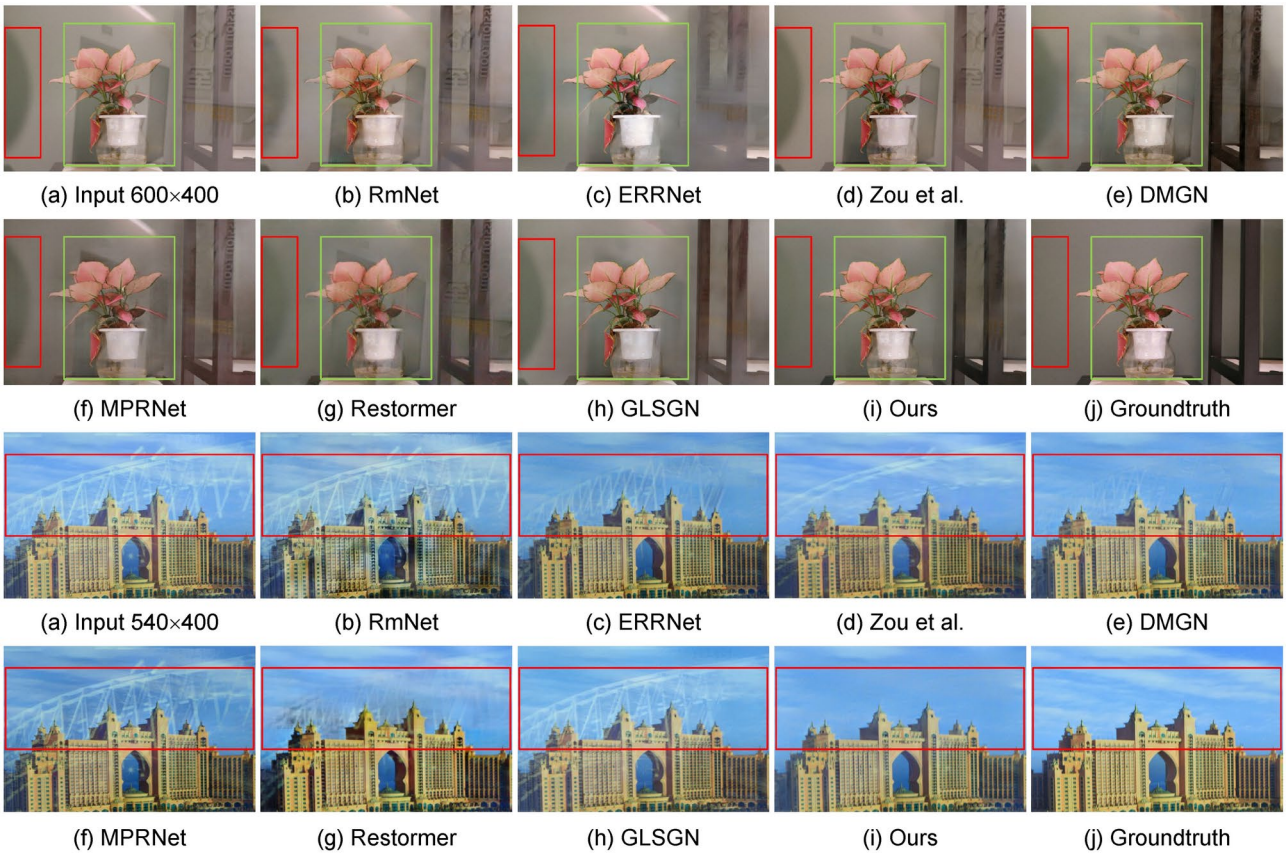

Figure 4. Visual comparison on regular-resolution real-world images from Nature[14] and SIR^2[53] for reflection removal. Our model can recover higher-quality details in the reconstructed images. Best viewed in zoom-in mode.

Figure 5. Visual comparison on high-resolution test images from Real20[11] and UHR4K[27] for reflection removal. Prominent reflection-contaminated regions are highlighted by the bounding boxes.

remove the reflections in the high-resolution input image since the reflection constituents are highly coupled to the background content and become more difficult to be completely recognized.

We can make following observations from the visual comparison with state-of-the-art methods for reflection removal in Fig. 5. First, our *LapCAT* is able to synthesize the cleanest background with higher-quality details on test images of high-resolution datasets over all baseline methods, which demonstrates the superiority of our proposed model. Besides, compared to the convolution-based generative models, our *LapCAT* is able to

model more complex reflection patterns. For example, although the reflection constituents in the first example in Fig. 5 significantly degrade the quality of background content, our *LapCAT* can remove most visible reflections and synthesize the most realistic and exquisite details compared to other methods. It is reasonable since our component-aware multi-head self-attention mechanism first captures pixel-wise long-term dependencies under the guidance of reflection detection. Additionally, our model performs contrastive learning to locate reflection regions in the latent feature space by the designed reflection detection mechanism and further leverages the predicted reflection mask to guide the component-aware transformer to refine the background content. Thus, our *LapCAT* is able to synthesize higher-quality background images compared to the other transformer-based method[60].

Similar to our *LapCAT*, GLSGN[27] also designs a specialized framework that adopts a global-local mechanism to restore high-resolution images and it learns the global-local consistency to capture more effective information. However, since the reflection constituents and background content are highly coupled in the input image, it is still challenging for GLSGN to completely model reflection patterns without long-range interactions between pixels. In contrast, our *LapCAT* introduces two specializedly designed mechanisms for image reflection removal: the reflection detection mechanism which locates the regions of reflection constituents by pixel-wise contrastive learning and the component-aware self-attention mechanism which iteratively filters out reflection constituents and absorbs the background content. As a result, our model is able to enjoy a more powerful capability of recognizing and removing complex reflection patterns in high-resolution image reflection removal.

User study. To further evaluate the visual quality of restored background images, we conduct a user study to compare our model with the top-three most powerful methods for reflection removal, including Restormer[60], GLSGN[27] and LAS[44]. We randomly selected 50 high-resolution test samples from Real20[11] and UHR4K[27] and presented the restoration results by *LapCAT* and other three methods to 50 human subjects for manual ranking of restoring quality. As shown in Table 2, *LapCAT* achieves 65.72% of the 2500 votes, which is much higher than the competing models. Additionally, when we aggregated the evaluation results of all subjects for each sample, our model won on 38 test samples, while the other three models won on a total of 12.

Ablation study

In this section, We conduct experiments to investigate the effect of each proposed functional module, including contrastive learning-based reflection detector, component-separable Transformer block, and Laplacian pyramid-based high-resolution image reconstruction mechanism, on restoration performance. To this end, we conduct ablation experiments on five variants of our *LapCAT*, which incrementally activates these proposed functional modules:

- *Base-Transformer*, which only employs a shifted window-based U-shaped transformer as the base framework without generating the reflection image. Thus, no component-aware transformer block or Laplacian pyramid strategy is used with this model. It serves as the baseline to gauge any improvements contributed by each subsequent module.
- *RUT* builds on the Base-Transformer by adding a separate reflection branch to synthesize the *R*eflection image in *U*-shaped *T*ransformer. Note that RUT only learns to divide reflection-contaminated images into two parts, the background image and the reflection image.
- *RUT-CSTB*, which employs the proposed *C*omponent-*S*eparable *T*ransformer *B*lock to replace typical Transformer blocks, thereby further introducing feature exchange between branches without the guidance of reflection mask and serving as the core unit of information collaboration for separating background and reflection images.
- *CAT*, which augments the model by incorporating a reflection detector and performs pixel-wise contrastive learning to predict a reflection mask and further leverages the reflection mask into the self-attention mechanism of Transformer blocks to improve the performance of background reconstruction.
- *LapCAT* is the complete model and further adopts a Laplacian pyramid mechanism to efficiently reconstruct high-resolution background images.

Table 3 lists the experimental results of five variants of our proposed *LapCAT* on high-resolution datasets Real20[11] and UHR4K[27] in terms of PSNR and SSIM. The performance improvement from *Base Transformer* to *RUT* demonstrates the necessity of the reflection branch. The final performance jump from *CAT* to *LapCAT* confirms the impact of the Laplacian pyramid strategy on high-resolution image reconstruction. The increasingly better performance of five variants shows the effectiveness of proposed technical components in our *LapCAT*.

Model	Share of the vote (%)	Winning samples
Restormer[60]	1.96	1
GLSGN[27]	21.0	8
LAS[44]	11.32	3
Ours	65.72%	38

Table 2. User study on the reflection removal results. 50 human subjects are asked to perform comparison between our *LapCAT* and other three methods on the restoration results of 50 randomly selected high-resolution test samples. Our model obtains 65.72% votes among $50 \times 50 = 2500$ comparisons and wins on 38 samples.

Method	Real20[11]		UHR4K-Real	
	PSNR	SSIM	PSNR	SSIM
Base transformer	19.42	0.712	20.54	0.785
RUT	20.37	0.739	21.88	0.801
CSTB	21.02	0.754	22.48	0.812
CAT	22.32	0.794	22.86	0.818
LapCAT (complete)	**23.01**	**0.809**	**25.18**	**0.866**
w/o $\mathscr{L}_{\text{pixel}}$	22.25	0.791	24.33	0.848
w/o $\mathscr{L}_{\text{per}}$	22.51	0.795	24.71	0.855
w/o $\mathscr{L}_{\text{adv}}$	22.73	0.801	24.86	0.860
w/o $\mathscr{L}_{\text{c}}$	22.42	0.789	24.54	0.852

Table 3. Ablation study on our *LapCAT* in terms of PSNR and SSIM to investigate the effectiveness of each proposed technique in our model. The best results are in bold.

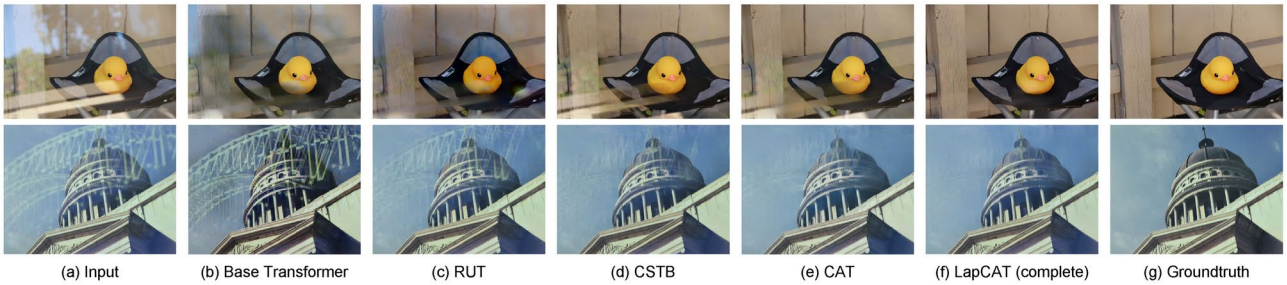

Figure 6. Visualization of reflection removal results by five variants of *LapCAT* on two test images.

Methods	RmNet[12]	ERRNet[2]	IBCLN[14]	LAS[44]	DMGN[15]	Zheng *et al.*[22]	GLSGN[27]	Restormer[60]	Ours
Params (M)	65.43	18.95	21.61	–	45.49	34.54	15.69	–	**13.22**
Runtime (s)	0.407	3.210	0.682	–	0.767	0.464	0.082	–	**0.063**

Table 4. Model complexity of our *LapCAT* and eight state-of-the-art methods for image reflection on a 4K (3840×2160) test image in terms of trainable parameters and runtime. The best results are in bold. '–' denotes out of memory by a RTX 3090 GPU.

Additionally, we visualize the restored background images by different variants of *LapCAT* in Fig. 6 and it qualitatively shows the quality improvements of restored images by our framework as well.

Effect of each loss. Table 3 also shows the performance of our LapCAT model without pixel loss $\mathscr{L}_{\text{pixel}}$, perceptual loss $\mathscr{L}_{\text{per}}$, contrastive loss $\mathscr{L}_{\text{c}}$ or without the adversarial loss $\mathscr{L}_{adv}$. We observe that both $\mathscr{L}_{\text{pixel}}$, $\mathscr{L}_{\text{per}}$ and $\mathscr{L}_{\text{c}}$ yield notable performance improvement. It is reasonable that $\mathscr{L}_{\text{pixel}}$ provides direct supervision on each pixel of restored images, $\mathscr{L}_{\text{per}}$ improves the correctness of reconstructed semantics, and $\mathscr{L}_{\text{c}}$ provides strong prior of reflection regions for iterative Transformer blocks. Besides, adversarial loss aims to improve the visual quality of restored results, thus the performance drop by $\mathscr{L}_{\text{adv}}$ is not significant compared to other losses.

Investigation on the model efficiency In Table 4, we further compare the computational efficiency of our proposed *LapCAT* with competing methods[2,12,14,15,22,27,44,60] on a 4K (3840×2160) image in terms of trainable parameters and inference time. The results show that our *LapCAT* is able to enjoy high efficiency as well as achieving the best performance on high-resolution image reflection removal. Note that the results of computational efficiency by Restormer[60] and LAS[44] are not shown on account of the out of memory problem when processing a 4K image.

Effect of reflection detection module To explore the superiority of our designed reflection detection module, we visualize the predicted mask of reflection regions by our model and location-aware method LAS[44] in Fig. 7. LAS[44] is a location-aware state-of-the-art method for image reflection removal. Benefiting from the pixel-wise contrastive learning, our reflection detection module enables our framework to model reflection constituents more precisely and synthesize higher-quality reflection image than the competing method.

Effect of component-separable transformer block We conduct experiments on Real20[11] and SIR^2[53] datasets to investigate the effectiveness of reflection separation by our proposed component-separable transformer block. To this end, we compare it with other ways of component separation, including ReLU-E, Feat-E, and w/o E. ReLU-E delivers the deactivated features by ReLU activation to the other branch, Feat-E sends all features, and w/o E denotes no feature exchange in our model. The experimental results are illustrated in Table 5, and the

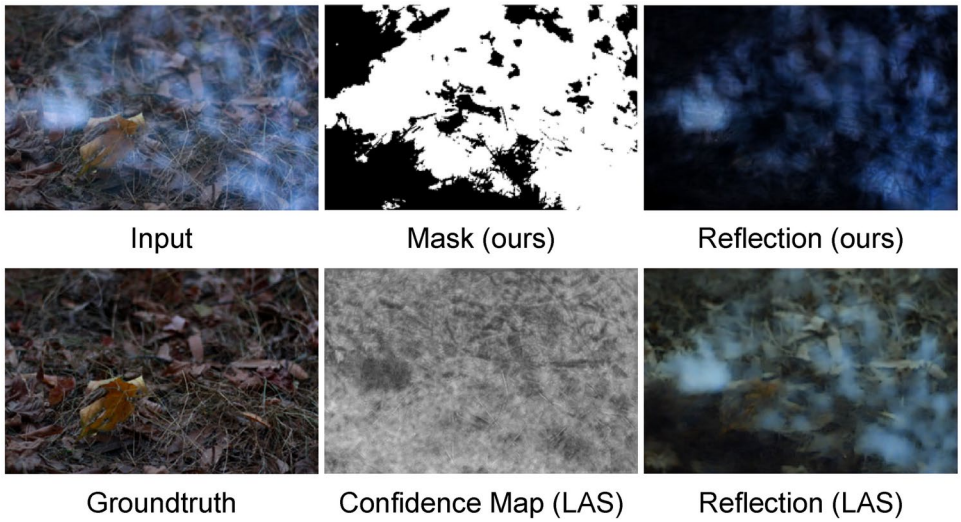

Figure 7. Visualization of reflection perception by our model and LAS[44] on a test image from Real20[11]. Best viewed in zoom-in mode.

Method	Real20[11]		Nature[14]	
	PSNR	SSIM	PSNR	SSIM
w/o E	19.93	0.726	22.09	0.795
Feat-E	20.37	0.742	23.13	0.838
ReLU-E	21.69	0.763	23.76	0.846
LapCAT (ours)	**23.01**	**0.809**	**23.95**	**0.849**

Table 5. Comparison of our model with other ways of feature exchange on Real20[11] and Nature[14] for image reflection removal in terms of PSNR and SSIM. The best results are in bold.

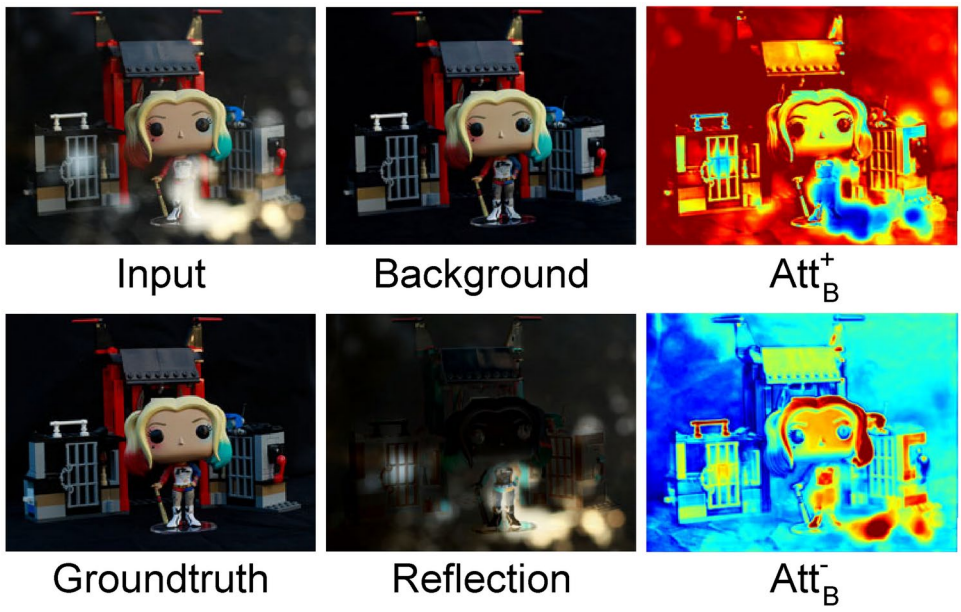

Figure 8. Visualization of attention maps in the component-aware transformer block. The test sample is from the Real20[11] dataset.

Dataset	Metric	Down-2	Down-4	Down-8
UHR4K-Real	PSNR	**25.18**	24.71	24.06
	SSIM	**0.866**	0.847	0.840

Table 6. Performance analysis of different downsampling strides. The best results are in bold.

results show that our designed component-aware transformer block obviously enables our model to reconstruct higher-quality background images than other ways of reflection separation. Although our CAT model is inspired by YTMT , we find that the performance of reflection removal is significantly improved, benefiting from the self-attention exchange than features after ReLU activation. Furthermore, we visualize attention maps of separated features in the background branch in Fig. 8.

Effect of Laplacian pyramid The performance comparisons between *LapCAT* and *CAT* in Table 1 and 3 have demonstrated the effectiveness of our designed Laplacian pyramid for high-resolution image reflection removal. Besides, we conduct experiments to investigate the effect of downsampling stride in Laplacian pyramid by setting it to Down-2, Down-4, and Down-8, and the experimental results are shown in Table 6. The performance comparison reveals that our *LapCAT* achieves the best performance while setting the downsampling stride to two.

Conclusion

In this work, we design a Laplacian pyramid-based component-aware Transformer model *LapCAT* for high-resolution image reflection removal. Our *LapCAT* designs a Laplacian pyramid module to preserve and synthesize high-fidelity details as well as downsampling the high-resolution image into a processable resolution. Additionally, it also introduces a component-aware self-attention mechanism to precisely separate reflection constituents from the background content, and such mechanism enables our model to capture long-range interactions between pixels in a high-resolution image. Benefiting from the reflection mask from reflection detection module through pixel-wise contrastive learning, our *LapCAT* is able to locate reflection constituents in the input image and thus reconstruct cleaner background images.

Limitation Even though our *LapCAT* shows powerful capability of modeling reflection patterns in high-resolution images, it is still challenging for it to simultaneously cope with different types of degradation in the same image. This is mainly due to the fact that our model is under pixel-wise supervised learning, and thus we plan to investigate it in an unsupervised way in our future work.

Data availability

The datasets and codes used and analyzed during the current study are available from the corresponding author upon reasonable request.

Received: 2 December 2024; Accepted: 13 March 2025
Published online: 22 March 2025

References

1. Goodfellow, I. J. *et al.* Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems* 2672–2680 (2014).
2. Wei, K., Yang, J., Fu, Y., Wipf, D. & Huang, H. Single image reflection removal exploiting misaligned training data and network enhancements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8178–8187 (2019).
3. Zou, Z., Lei, S., Shi, T., Shi, Z. & Ye, J. Deep adversarial decomposition: A unified framework for separating superimposed images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 12806–12816 (2020).
4. Levin, A. & Weiss, Y. User assisted separation of reflections from a single image using a sparsity prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 1647–1654 (2007).
5. Arvanitopoulos, N., Achanta, R. & Susstrunk, S. Single image reflection suppression. In *CVPR* 4498–4506 (2017).
6. Yan, Q., Xu, Y., Yang, X. & Nguyen, T. Separation of weak reflection from a single superimposed image. *IEEE Signal Process. Lett.* **21**, 1173–1176 (2014).
7. Wan, R., Shi, B., Hwee, T. A. & Kot, A. C. Depth of field guided reflection removal. In *2016 IEEE International Conference on Image Processing* 21–25 (IEEE, 2016).
8. Wan, R., Shi, B., Tan, A.-H. & Kot, A. C. Sparsity based reflection removal using external patch search. In *ICME* 1500–1505 (2017).
9. Wang, H., Lin, S., Ye, X. & Gu, W. Separating corneal reflections for illumination estimation. *Neurocomputing* **71**, 1788–1797 (2008).
10. Fan, Q., Yang, J., Hua, G., Chen, B. & Wipf, D. A generic deep architecture for single image reflection removal and image smoothing. In *Proceedings of the IEEE International Conference on Computer Vision* 3238–3247 (2017).
11. Zhang, X., Ng, R. & Chen, Q. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4786–4794 (2018).
12. Wen, Q. *et al.* Single image reflection removal beyond linearity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3771–3779 (2019).
13. Yang, J., Gong, D., Liu, L. & Shi, Q. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *Proceedings of the European conference on computer vision* 654–669 (2018).
14. Li, C., Yang, Y., He, K., Lin, S. & Hopcroft, J. E. Single image reflection removal through cascaded refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3565–3574 (2020).
15. Feng, X. et al. Deep-masking generative network: A unified framework for background restoration from superimposed images. *IEEE Trans. Image Process.* **30**, 4867–4882 (2021).
16. Szegedy, C. *et al.* Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1–9 (2015).

17. Xu, R., Xiao, Z., Huang, J., Zhang, Y. & Xiong, Z. Edpn: Enhanced deep pyramid network for blurry image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 414–423 (2021).
18. Iizuka, S., Simo-Serra, E. & Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **36**, 1–14 (2017).
19. Benzenati, T., Kallel, A. & Kessentini, Y. Stf-trans: A two-stream spatiotemporal fusion transformer for very high resolution satellites images. *Neurocomputing* **563**, 126868 (2024).
20. Quan, W. et al. Image inpainting with local and global refinement. *IEEE Trans. Image Process.* **6**, 66 (2022).
21. Chen, W., Jiang, Z., Wang, Z., Cui, K. & Qian, X. Collaborative global-local networks for memory-efficient segmentation of ultra-high resolution images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8924–8933 (2019).
22. Zheng, Z. *et al.* Ultra-high-definition image dehazing via multi-guided bilateral learning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16180–16189 (IEEE, 2021).
23. Chen, Z., Wang, W., Xie, E., Lu, T. & Luo, P. Towards ultra-resolution neural style transfer via thumbnail instance normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 36 393–400 (2022).
24. Zhang, H., Dai, Y., Li, H. & Koniusz, P. Deep stacked hierarchical multi-patch network for image deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 5978–5986 (2019).
25. Suin, M., Purohit, K. & Rajagopalan, A. Spatially-attentive patch-hierarchical network for adaptive motion deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 3606–3615 (2020).
26. Zamir, S. W. *et al.* Multi-stage progressive image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 14821–14831 (2021).
27. Feng, X. *et al.* Global-local stepwise generative network for ultra high-resolution image restoration. arXiv preprint arXiv:2207.08808 (2022).
28. Yi, W. et al. Semi-supervised progressive dehazing network using unlabeled contrastive guidance. *Neurocomputing* **551**, 126494 (2023).
29. Li, Y. & Brown, M. S. Single image layer separation using relative smoothness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2752–2759 (2014).
30. Yan, Q., Xu, Y., Yang, X. & Nguyen, T. Separation of weak reflection from a single superimposed image. *IEEE Signal Process. Lett.* **21**, 1173–1176 (2014).
31. Wan, R. et al. Region-aware reflection removal with unified content and gradient priors. *IEEE Trans. Image Process.* **27**, 2927–2941 (2018).
32. Dong, Z. et al. A polarization-based image restoration method for both haze and underwater scattering environment. *Sci. Rep.* **12**, 1836 (2022).
33. Shih, Y., Krishnan, D., Durand, F. & Freeman, W. T. Reflection removal using ghosting cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3193–3201 (2015).
34. Heydecker, D. et al. Mirror, mirror, on the wall, who's got the clearest image of them all?—A tailored approach to single image reflection removal. *IEEE Trans. Image Process.* **28**, 6185–6197 (2019).
35. Yang, Y., Ma, W., Zheng, Y., Cai, J.-F. & Xu, W. Fast single image reflection suppression via convex optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 8141–8149 (2019).
36. Gai, K., Shi, Z. & Zhang, C. Blind separation of superimposed moving images using image statistics. *IEEE Trans. Pattern Anal. Mach. Intell.* **34**, 19–32 (2011).
37. Yang, J., Li, H., Dai, Y. & Tan, R. T. Robust optical flow estimation of double-layer images under transparency or reflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 1410–1419 (2016).
38. Sun, C. *et al.* Automatic reflection removal using gradient intensity and motion cues. In *Proceedings of the 24th ACM International Conference on Multimedia* 466–470 (2016).
39. Xue, T., Rubinstein, M., Liu, C. & Freeman, W. T. A computational approach for obstruction-free photography. *ACM Trans. Graph.* **34**, 1–11 (2015).
40. Qiu, G., Tao, D., You, D. & Wu, L. Low-illumination and noisy bridge crack image restoration by deep cnn denoiser and normalized flow module. *Sci. Rep.* **14**, 18270 (2024).
41. Wan, R., Shi, B., Duan, L.-Y., Tan, A.-H. & Kot, A. C. Crrn: Multi-scale guided concurrent reflection removal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4777–4785 (2018).
42. Li, Y. et al. Two-stage single image reflection removal with reflection-aware guidance. *Appl. Intell.* **53**, 19433–19448 (2023).
43. Ma, D., Wan, R., Shi, B., Kot, A. C. & Duan, L.-Y. Learning to jointly generate and separate reflections. In *Proceedings of the IEEE International Conference on Computer Vision* 2444–2452 (2019).
44. Dong, Z. *et al.* Location-aware single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 5017–5026 (2021).
45. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision* 694–711 (Springer, 2016).
46. Yi, Z., Tang, Q., Azizi, S., Jang, D. & Xu, Z. Contextual residual aggregation for ultra high-resolution image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 7508–7517 (2020).
47. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9729–9738 (2020).
48. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
49. Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016).
50. Burt, P. J. & Adelson, E. H. The Laplacian pyramid as a compact image code. In *Readings in Computer Vision* 671–679 (Elsevier, 1987).
51. Lai, W.-S., Huang, J.-B., Ahuja, N. & Yang, M.-H. Deep Laplacian pyramid networks for fast and accurate super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 624–632 (2017).
52. Liu, Z. *et al.* Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 10012–10022 (2021).
53. Wan, R., Shi, B., Duan, L.-Y., Tan, A.-H. & Kot, A. C. Benchmarking single-image reflection removal algorithms. In *Proceedings of the IEEE International Conference on Computer Vision* 3922–3930 (2017).
54. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
55. Miyato, T., Kataoka, T., Koyama, M. & Yoshida, Y. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957 (2018).
56. Agustsson, E. & Timofte, R. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (2017).
57. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
58. Feng, X. *et al.* Contrastive feature decomposition for image reflection removal. In *2021 IEEE International Conference on Multimedia and Expo* 1–6 (IEEE, 2021).
59. Hu, Q. & Guo, X. Trash or treasure? An interactive dual-stream strategy for single image reflection separation. *Adv. Neural Inf. Process. Syst.* **34**, 24683–24694 (2021).

60. Zamir, S. W. *et al.* Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).
61. Prasad, B., Boregowda, L. R., Mitra, K., Chowdhury, S. *et al.* V-desirr: Very fast deep embedded single image reflection removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 2390–2399 (2021).
62. Hu, Q. & Guo, X. Single image reflection separation via component synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 13138–13147 (2023).
63. Zhong, H., Hong, Y., Weng, S., Liang, J. & Shi, B. Language-guided image reflection separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* 24913–24922 (2024).

Acknowledgements
This research was supported by the Research and Innovation Initiatives of WHPU (Grant no. 2023Y46), the Research Funding of Wuhan Polytechnic University (Grant no.2024R005), the Basic and Applied Basic Research Foundation of Guangdong Province (Grant no.2023A1515110776). The authors gratefully acknowledge this support.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Z.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.