Eyeglass Reflection Removal With Joint Learning of Reflection Elimination and Content Inpainting

Wentao Zou, Xiao Lu, Zhilv Yi, Ling Zhang, Gang Fu, Ping Li, *Member, IEEE*, and Chunxia Xiao, *Senior Member, IEEE*

***Abstract*— Eyeglass reflection removal is of great importance to the portrait image processing. However, it remains a challenge to eliminate the reflections on the glass and restore the textual contents of eyes without introducing visual artifacts. Addressing this problem, in this paper, we propose an Eyeglass Reflection Removal Network (ER^{2}Net) by learning reflection elimination and content inpainting jointly. The reflection elimination branch is effective in weak reflection regions, and the content inpainting branch is dedicated to content reasoning in strong reflection regions. We then propose a result fusion module (RFM), which adaptively fuses the elimination result and the inpainting result according to the reflection intensity of each pixel, to produce high-quality result. We also design a memory module for improving the content inpainting result, and propose an eye-symmetry loss to avoid visual artifacts. Additionally, we construct the first Real-world eyeglass Reflection (ReyeR) dataset for eyeglass reflection removal. Extensive quantitative and qualitative experiments demonstrate the superiority of the ER^{2}Net over state-of-the-art methods for eyeglass reflection removal.**

***Index Terms*— Eyeglass reflection removal, dataset, content reasoning, eye symmetry, memory augmentation.**

I. Introduction

REFLECTION is a common optical phenomenon in nature. It can degrade the quality of an image, especially for portrait images with eyeglass reflection. The reflection layer in eyeglass obscures the original eye detail, which poses great challenge for some practical application tasks, such as image-based verification [1], face recognition [2], [3], face keypoint detection [4], [5] and *etc.*. Thus, eyeglass reflection removal is a meaningful and necessary image processing task.

Manuscript received 14 December 2023; revised 19 March 2024 and 24 April 2024; accepted 14 May 2024. Date of publication 27 May 2024; date of current version 30 October 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 61972298 and Grant 62372336. This article was recommended by Associate Editor Q. Xu. *(Wentao Zou and Xiao Lu contributed equally to this work.) (Corresponding author: Chunxia Xiao.)*

Wentao Zou, Zhilv Yi, and Chunxia Xiao are with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: wentaozou@whu.edu.cn; yizhilv@whu.edu.cn; cxxiao@whu.edu.cn).

Xiao Lu is with the College of Engineering and Design, Hunan Normal University, Changsha 410081, China (e-mail: luxiao@hunnu.edu.cn).

Ling Zhang is with the School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430081, China (e-mail: zhling@wust.edu.cn).

Gang Fu and Ping Li are with the Department of Computing and the School of Design, The Hong Kong Polytechnic University, Hong Kong (e-mail: xyzgfu@gmail.com; p.li@polyu.edu.hk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2024.3405576.

Digital Object Identifier 10.1109/TCSVT.2024.3405576

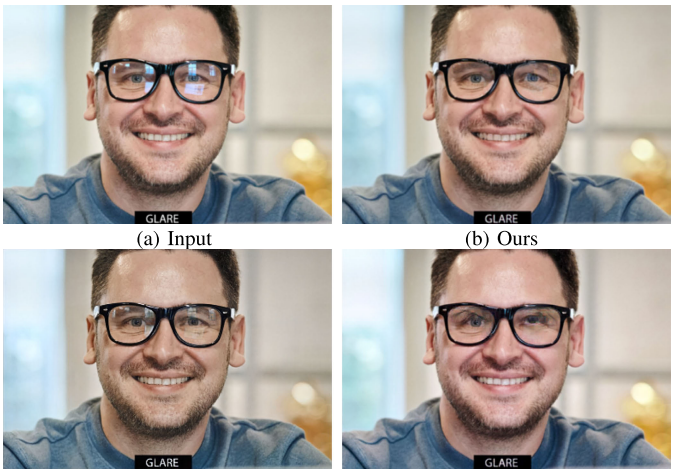
The materials of the eyeglass, the types and intensity of the light source largely affect the intensity of the reflection, which then determines the degradation degree of an image. In the weak reflection regions, only the color and brightness information are changed, while the local texture details are preserved. However, in the strong reflection regions, not only the color and brightness information are changed, but the texture details are completely missing, which makes restoration more difficult. Therefore, it is challenging to reason about the accurate contents without introducing visual artifacts in strong reflection regions and restore the lighting information while preserving the local texture details in the weak reflection regions.

Recent years, we have witnessed the success of deep learning-based methods in various image restoration problems. Several large-scale real-world datasets [6], [8], [9] and a few of learning-based methods [6], [8], [10], [11], [12], [13] have been proposed to eliminate the reflections in natural images. However, removing reflections in eyeglass images has been rarely addressed in the literature, and one of the important reasons is that there is no large-scale real-world datasets available. Although existing reflection elimination methods can be directly used in eyeglass reflection removal task, they are only effective for the weak reflections and has poor ability to reason about the contents in strong reflection regions, as shown in Fig. 1 (c).

Another straightforward solution is to leverage the inpainting-based methods to recover the missing contents in reflection regions. We have trained the human eye inpainting method ExGANs [7] for our reflection removal task. As shown in Fig. 1 (d), although ExGANs can produce semantically-plausible results, it destroys the textual details and introduce visual artifacts, since it restores the degradation regions using the surroundings without considering the texture details in the weak reflection regions.

Leveraging multiple images of the same scene are helpful to reason about the contextual contents in strong reflection regions [14]. However, existing multiple-images-based methods capture the images with specially designed equipment, such as the polarization camera [15] or camera with different focus lengths [16], *etc.*, which hinders the application of such kind of methods.

In this work, we propose a novel eyeglass reflection removal method to combine the advantages of the texture detail preserved elimination-based method and the contents restored inpainting-based method. To do that, we present an Eyeglass



(a) Input (b) Ours

(c) IBCLN [6] (d) ExGANs [7]

Fig. 1. Visualization results of our ER^{2}Net and other SOTA methods. IBCLN [6] fails to remove the reflection, especially in the strong reflection regions, it also cannot reason about the accurate contents in the strong reflection regions. ExGANs [7] cannot preserve the local texture details in weak reflection regions and introduces obvious visual artifacts. Our ER^{2}Net performs well on both weak and strong reflections and can produce fine-grained results.

Reflection Removal Network (ER^{2}Net) including two main modules, a multi-task network (MTNet) and a result fusion module (RFM). MTNet consists of a reflection detection branch, an elimination branch, and an inpainting branch. The elimination branch learns to eliminate the reflections and preserve the textual details, which would be effective for weak reflections. While the inpainting branch learns to restore the missing contents according to the detection result, which would be more effective for strong reflections. Therefore, we design the RFM module to fuse the elimination result and the inpainting result according to the reflection intensity of each pixel. Also, we design a memory module to record the prototypical feature cross images to reason about the true contents in the inpainting branch. We also propose an eye-symmetry loss to avoid visual artifacts by employing the symmetry prior of human eyes.

Additionally, since there is no publicly available real-world eyeglass reflection dataset, we construct the first **R**eal-word **eye**glass **R**eflection (ReyeR) dataset containing 13,610 pairs of high-quality images. All the captured ground-truth images are performed with pixel offset alignment and color correction to guarantee the pixel and color tone consistency with the input images.

Our main contributions are summarized as follows:

1) We present an Eyeglass Reflection Removal Network (ER^{2}Net) consisting of a multi-task network and a result fusion module. ER^{2}Net learns to combine the advantages of the reflection elimination and detail preservation for weak reflection regions and the contents restoration for strong reflection regions, and thus can deal with both weak and strong reflections.

2) We design a memory module to record the prototypical features cross images for reasoning about the accurate contents, and propose an eye symmetry loss to optimize the final results to avoid introducing visual artifacts.

3) We construct a real-world eyeglass reflection dataset (ReyeR) with 13,610 pairs of reflection- and reflection-free images, which is the first large-scale dataset for training and evaluation of the learning-based eyeglass reflection removal method.

II. Related Work

A. Single-Image-based Reflection Removal

Many methods exploit a single image for reflection removal using prior knowledge, such as the gradient sparsity prior [17], [18], [19], [20] and the smoothness prior [21], [22]. Sandhan and Choi [19] extracted prior gradient information from the reflection layer and the background layer, and transformed the reflection removal problem into a convex optimization problem by maximizing the joint probability. Conte et al. [23] proposed to remove the object reflections in videos by considering the reflection removing as a global optimization problem. However, these priors are often handcrafted, and the results are usually unrealistic. Learning-based methods [11], [12], [13], [24], [25], [26], [27], [28], [29], [30], [31] have been proposed to solve the problem of single image reflection removal. Zhang et al. [8] designed several loss functions utilizing the characteristic of reflections to guide the network training. Watanabe and Hasegawa [24] used auto-encode and U-net to remove eyeglass reflections, but only evaluated the algorithm on limited images. Dong et al. [28] used multi-scale Laplace kernel parameters to enhance the reflection boundary information. Liu et al. [29] employed object semantic cue as the guidance. Wu et al. [10] proposed a residual network based on channel attention, so that the network can learn global information on different channels. However, these methods can only eliminate weak reflection and have poor ability to reason about and complete the contents in strong reflection regions. Moreover, although there are several datasets for natural image reflection removal, *e.g.*, Zhang et al. [8], Nature [6], SIR^2 [9] and SIR^{2+} [32], there is no publicly available dataset tailored for eyeglass reflection removal.

B. Multiple-Image-based Reflection Removal

Since the location and intensity of the reflection varies with the photographic perspective, it is possible to reason about the contents in the reflection regions using multiple images under varying illumination [15], [16], [33], [34], [35], [36], [37], [38], [39], [40], [41]. Some existing methods used the motion cues to separate and remove the reflections. Agrawal et al. [33] used two flash images with and without a printed checkerboard for reflection removal, while Fu et al. [34] took a single hyperspectral image for component separation. Schechner et al. [16] employed different focus lengths to capture multiple images and removed the interpenetration of the reflective layer. Kong et al. [15] used multiple polarized images captured by the polarization camera to solve the problem, since the polarization of reflection and transmission are usually different, which makes it easy to distinguish the reflection. Wan et al. [40] proposed the first facial images reflection removal method, which constructs a guided removal framework using two similar facial images to restore important

facial features. Although multiple image methods make a great achievement in reflection removal, the difficulty of capturing multiple images hinders the application of these kind of methods.

C. Eye Inpainting

Considering that it is difficult to reproduce the eye details perfectly with universal completion methods, many methods try to introduce the eye's prior knowledge or reference image for eye inpainting. Agarwala et al. [42] used example photos to generate final results with a mixture of patch matching and blending. However, it is not robust to lighting conditions. GAN-based methods [43] are also used in eye inpainting, and the results tend to be highly realistic but low in fidelity. Yan et al. [44] further introduced the eye aesthetic assessment and face semantic parsing for restoration of eyes, and Dolhansky and Ferrer [7] utilized exemplar information to produce high-fidelity results that are largely dependent on the quality of the reference image. In contrast, we complete the eye information by leveraging the symmetric information of the other eye and the prototypical features recorded in the memory block.

D. Memory Augmentation-Based Methods

Memory augmentation-based methods [45], [46], [47], [48], [49] have shown promising results in various deep learning tasks, including image inpainting [50], [51], anomaly detection [52], [53], and semantic segmentation [54]. Sukhbaatar et al. [45] proposed an end-to-end training memory network for natural language processing tasks, which solves the limitations of traditional neural networks in processing long-sequence data. Feng et al. [50] restored the content of corrupted regions based on known regions and the learned semantic distribution using a Style-GAN based generative memory, which is difficult to train. Xu et al. [51] designed a texture memory that records patch samples extracted from unmasked regions as a guide to generate the inpainted image. Inspired by these methods, we exploit the memory module to record the prototypical features of eyes and inpaint the contents in strong reflection regions.

E. Multi-Task Network

Multi-task learning is a learning framework to enforce shared representation and mutual influence between multiple tasks for improving generalization ability, which has been successfully used in various layer separation tasks, such as image draining [11], specular highlight removal [55]. However, these methods use the hard parameter sharing scheme which can not avoid the negative transfer between unrelated or even conflicting tasks. Addressing this issue, in this paper, we design different modules, *i.e.,* the memory module for content inpainting and the residual block for reflection elimination, for different task branches to learn the task-specific information.

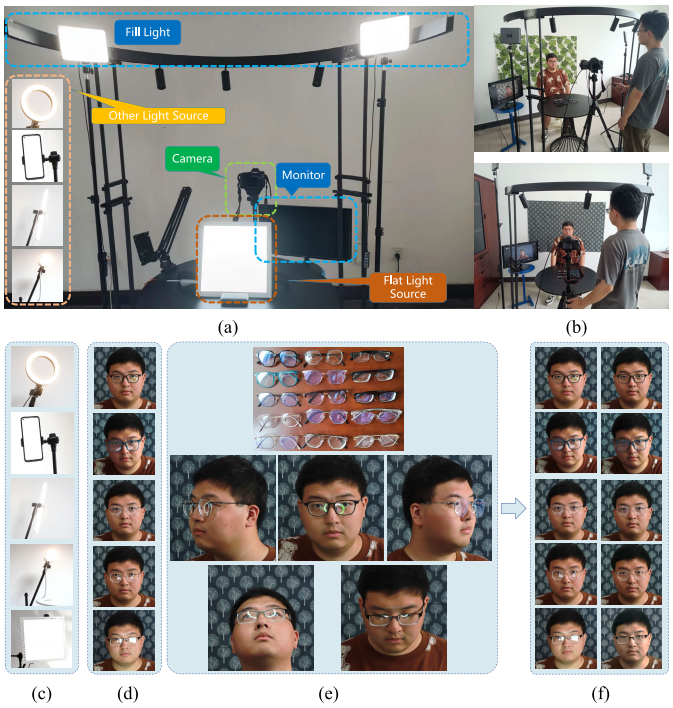


Fig. 2. **Construction overview and samples of ReyeR.** (a) The studio we constructed and each component for collecting the samples in ReyeR. (b) Data collecting process with different background. (c) Light sources used: round light, mobile phone panel light, strip light, normal desk lamp and flat light. (d) Five different light angles. (e) 24 kinds of eye glasses with different materials and five different participant angles. (f) Sample pairs of eyeglass images in our dataset with reflection (the first row) and without reflection (the second row).

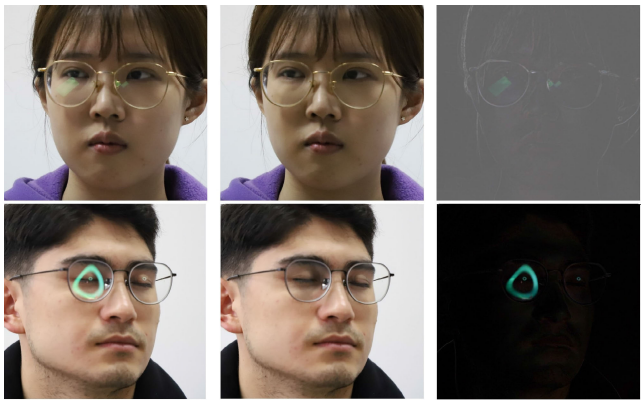
III. ReyeR Dataset

There are a few datasets for natural image reflection removal [6], [8], [9], however, they are not suitable for eyeglass reflection removal. To our knowledge, there is no publicly available dataset for eyeglass reflection removal. To train the learning-based models, we construct the first real-world eyeglass reflection image restoration dataset (ReyeR) with reflection- and reflection-free image pairs.[1]

To simulate the realistic eyeglass reflections and guarantee the strict alignment between the reflection image and the corresponding reflection-free image, we built a studio with controlled lighting conditions and different backgrounds for photography (as shown in Fig. 2 (a) and (b)). As shown in Fig. 2 (c), we use five different light sources to produce different shapes of reflections, each of which produces reflections in five different directions, as shown in Fig. 2 (d). Then we adjust the position of the participant to produce different distributions of reflections at different locations on the eyeglass. We use 24 kinds of eyeglass with different materials to obtain images with different reflection intensities, as shown in Fig. 2 (e). After that, we turn off the reflection light source and take the reflection-free image as the corresponding ground-truth image, the sample pairs can be seen in Fig. 2 (f).

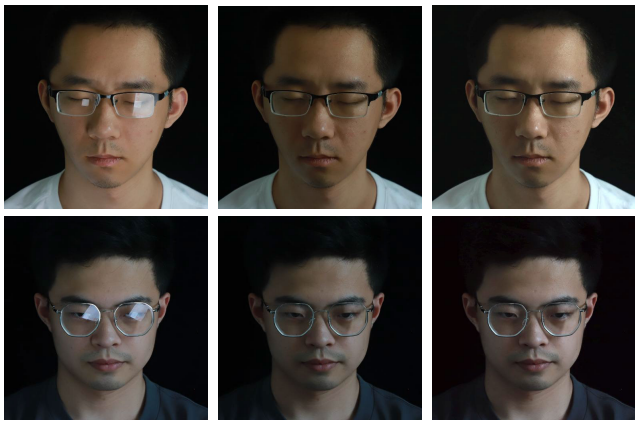
Due to slight portrait jitters, uncontrolled blinks and changes of lighting conditions may result in pixel offsets between the reflection image and the corresponding ground-truth image.

[1] All the privacy data have been permitted for research purpose only.



(a) Input (b) Captured GT (c) Difference map

Fig. 3. Two image pairs and their corresponding difference maps. We filter out the image pairs with large pixel offset (the upper sample) and only keep the image pairs with small pixel offset (the lower sample).



(a) Input (b) Captured GT (c) Color transferred GT

Fig. 4. We follow [56] to transfer the color characteristics of the input image to the captured GT image to obtain the color transferred GT.

Thus, we filter out the image pairs with high jitter by defining a threshold on the average value of the difference map (see the left part of Fig. 3) and retain high-quality image pairs without large offset (see the right part of Fig. 3). In addition, to further reduce the inevitable pixel tone deviation caused by the change of the lighting condition, we perform color correction followed by Reinhard et al. [56], which transfers the color characteristic from the input reflection image to the captured ground-truth image. Fig. 4 shows the input reflection image, the captured ground-truth image and the color transferred ground-truth image. It can be seen that there involves a certain degree of illumination error even after applying the illumination correction algorithm. To measure the difference, we calculate the MSE value between the input image and the color transferred ground-truth image by masking the reflection regions. For the images in the dataset, the MSE value is 0.54 with a standard deviation of 0.24, which indicates that the illumination error can be of little impact on the reflection removal algorithm and can be neglected.

Finally, we collect 13,610 high-quality image pairs, which are divided into the training set, testing set and validation set according to the ratio of 8:1:1. Table I shows some details about our ReyeR dataset. We have called for 356 individuals to capture portrait images covering different professionals,

TABLE I

DETAILS OF OUR REYER DATASET. P.R., L.R., A.R., C.R., F.R. ARE POINT REFLECTION, LINE REFLECTION, AREA REFLECTION, CIRCULAR REFLECTION, AND FLAT REFLECTION, RESPECTIVELY

Image res.	#Total dataset	#Training	#Testing	#Validation
1024×1024	13,610	11,046	1,282	1,282
#P.R.	#L.R.	#A.R.	#C.R.	#F.R
2,590	2,514	2,910	2,796	2,900
#light resource	#light orientation	#eyeglass	#participants	
5	5	24	356	

gender and age groups, and we make sure that the same person only appears in one of the training/testing/validation sets. Our ReyeR dataset covers different reflection intensity, and we do not clearly distinguish the strong reflection image from the weak reflection image, since it is common for varying reflection existing in the same image.

IV. PROPOSED METHOD

A. Overall Framework of ER^{2}Net

Fig. 5 shows the overview of our ER^{2}Net, which contains a multi-task network (MTNet) and a result fusion module (RFM). MTNet is a three-branches network, including a Detection Branch, an Inpainting Branch and an Elimination Branch, to predict the reflection detection result $\mathbf{D}$, the content inpainting result $\mathbf{I}_i$ and the reflection elimination result $\mathbf{I}_e$, respectively. Since the elimination branch is only effective for the weak reflection regions and the inpainting branch can restore the contents in the strong reflection regions, we design the RFM to fuse $\mathbf{I}_e$ and $\mathbf{I}_i$ according to the reflection intensity of each pixel to produce the final fine-grained result $\mathbf{I}_{out}$.

MTNet consists of a shared encoder and three decoders for each branch. Specifically, the detection branch is a plain encoder-decoder structure. The elimination branch is an encoder-decoder network followed by a residual block, which consists of gated convolution [57], dilated convolution, and channel attention layer [10]. To improve the content reasoning ability of the inpainting branch, we insert a memory module between the encoder and decoder in the inpainting branch to record the prototypical contextual features cross images. Our inspiration is that similar features in the reflection-free regions in other images may provide contextual information for restoration, which comes from the existing multiple-image-based reflection removal methods. Furthermore, to avoid visual artifacts, we propose the eye symmetry loss ℓ_{eye} using the prior information of the human eye to optimize the final result. The whole network can be trained end-to-end with the reconstruction loss ℓ_{rec}, the perception loss ℓ_p, the detection loss ℓ_{Focal}, the weight loss ℓ_{weight} and our proposed eye symmetry loss ℓ_{eye}.

B. Memory Module

Due to the lack of effective information in strong reflection regions, it is difficult to reason about the true contents in these regions. Existing multiple-image-based reflection removal methods try to exploit the similar texture information cross multiple images to restore the content information in reflection regions. Inspired by this, we design a memory

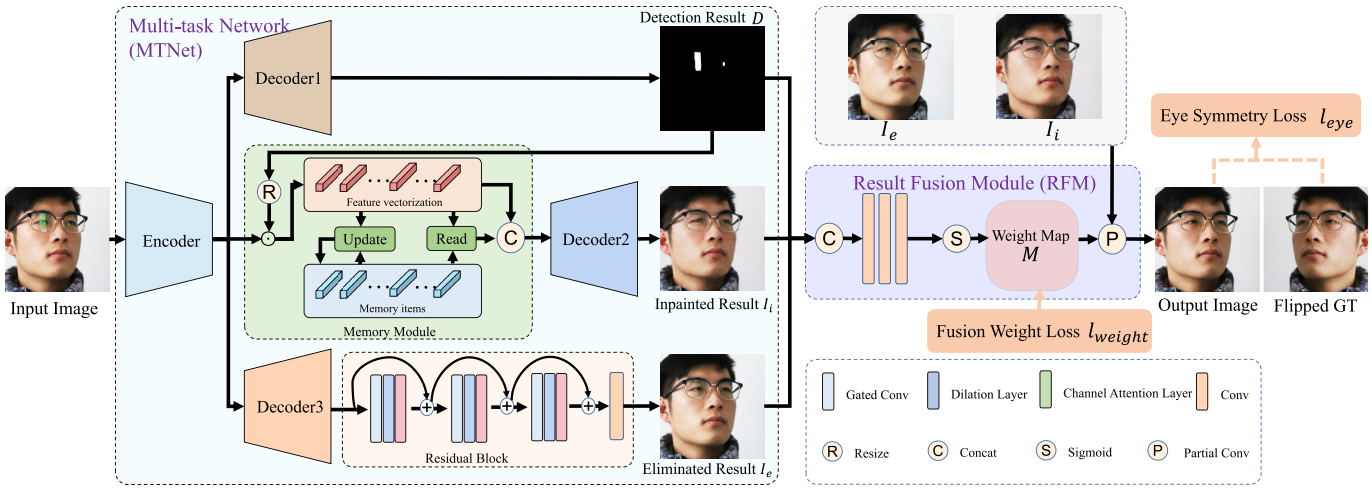


Fig. 5. **Overview of our ER^{2}Net.** ER^{2}Net consists of a multi-task network (MTNet) and a result fusion module (RFM). MTNet is a three-branches network, including a Detection Branch, an Inpainting Branch and an Elimination Branch, to predict the detection result $\mathbf{D}$, the content inpainting result $\mathbf{I}_i$ and the reflection elimination result $\mathbf{I}_e$, respectively. In RFM, $\mathbf{I}_i$ and $\mathbf{I}_e$ are adaptively fused according to $\mathbf{D}$ to produce the final fine-grained result $\mathbf{I}_{out}$.

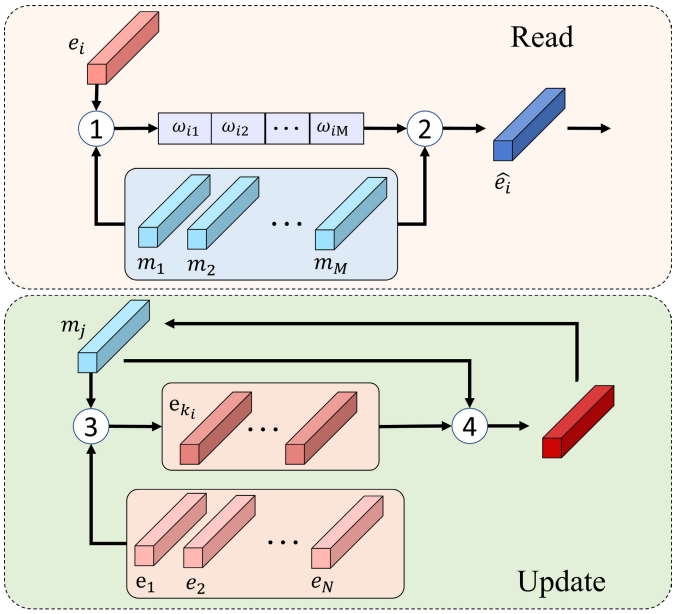


Fig. 6. Illustration of reading and updating operations in the memory module. For reading operation, we compute the correlation value in Eq. (1) between the query $\mathbf{e}_i$ and each item ($\mathbf{m}_j$), and leverage all the items to reason about the contextual feature $\hat{\mathbf{e}}_i$ in Eq. (2). For updating operation, we retrieve the most relevant queries $\mathbf{e}_{k_i}$ between memory item $\mathbf{m}_j$ and each query ($\mathbf{e}_i$) by Eq. (3), and update the memory items $\mathbf{m}_j$ based on all the queries $\mathbf{e}_{k_i}$ in Eq. (4). ①, ②, ③, ④ are the equations of Eqs. (1)-(4), respectively.

module to record the prototypical contents cross images as the auxiliary information for content restoration.

Given the input image $\mathbf{I}_{in}$, we multiply the feature map generated by the encoder with the detection result $\mathbf{D}$ to obtain the new feature map $\mathbf{E} \in \mathcal{R}^{H\times W\times C}$ in the reflection regions. Fig. 6 shows the read and update operations in the memory module. Specifically, we initialize the memory block $\mathcal{M}$ as a matrix with $C \times M$ elements, where M is the number of the items in the memory. The memory block performs reading to reason about the contextual content of the reflection regions. It is also updated following the network training to learn to record the important contextual features across images.

Specifically, let $\mathbf{m}_j \in \mathcal{R}^{1\times 1\times C}$ be the j^{th} item of $\mathcal{M}$, and the i^{th} feature vector $\mathbf{e}_i \in \mathcal{R}^{1\times 1\times C}$ in $\mathbf{E}$ can be seen as a query. We first calculate the correlation value between each query $\mathbf{e}_i$ and each item $\mathbf{m}_j$ as:

$$w_{i,j} = \frac{\exp(\mathbf{m}_j^T \mathbf{e}_i)}{\sum_{j'=1}^{M} \exp(\mathbf{m}_{j'}^T \mathbf{e}_i)}. \tag{1}$$

Then, for each query $\mathbf{e}_i$, the memory block performs reading to leverage all the items in $\mathcal{M}$ to reason about the contextual feature $\hat{\mathbf{e}}_i$ as follows:

$$\hat{\mathbf{e}}_i = \sum_{j=1}^{M} w_{i,j} \mathbf{m}_j. \tag{2}$$

We apply the reading operator to all the queries and obtain a transformed feature map $\hat{\mathbf{E}}$. We then concatenate it with the original feature map $\mathbf{E}$ along the channel dimension and feed them to the decoder.

To make the memory learning to record the most similar contextual contents cross images, we only choose the queries having the maximum correlation values with the items for updating. To do that, we retrieve the most relevant queries $\mathbf{e}_{k_i}$ between memory item $\mathbf{m}_j$ and each query $\mathbf{e}_i$ by

$$k_i = \arg\max_i w_{i,j}. \tag{3}$$

Then, we update the memory items $\mathbf{m}_j$ based on all the queries that are most relevant with $\mathbf{m}_j$:

$$\mathbf{m}_j \leftarrow f(\mathbf{m}_j + \sum_{i=1}^{N} \mathbf{1}(k_i = j) w_{i,j} \mathbf{e}_i), \tag{4}$$

where $f(\cdot)$ is the ℓ_2 norm, $\mathbf{1}(\cdot)$ is the indicator function, and $N = H \times W$ is the number of feature vectors in $\mathbf{E}$.

Unlike other memory modules [50], [51], we update the memory items to record prototypical features in both the training and the testing phase to improve generalization ability, as the contextual patterns in the training and test sets may be different.

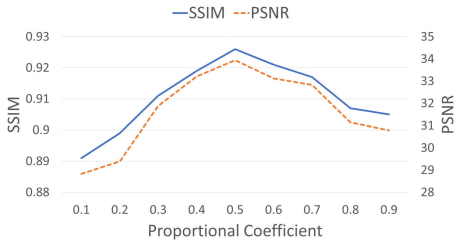


Fig. 7. Quantitative results of reflective intensity threshold under different proportional coefficient.

C. Result Fusion Module (RFM)

Given the elimination result $\mathbf{I}_e$, the inpainting result $\mathbf{I}_i$, and the reflection mask $\mathbf{D}$, the RFM module is designed as shown in Fig. 5. Taking $\mathbf{I}_e$, $\mathbf{I}_i$ and $\mathbf{D}$ as input, the RFM generates a weight map $\mathbf{M}$ via three 1×1 convolution layers followed by a Sigmoid function. With such weight map, the model is aware of distinguishing strong and weak reflection regions. To this end, we design a weight loss according to a heuristic rule: in the strong reflection areas, we tend to trust the result of $\mathbf{I}_i$, and the corresponding weights should be approximately 0 in such areas. For areas of weak reflection and background, $\mathbf{I}_e$ provides a more accurate result, and the weight values should approximate to 1. Then, the weight loss can be expressed as

$$\ell_{weight} = \|\mathbf{M}[\mathbf{R} > t]\|_2^2 + \|\mathbf{M}[\mathbf{R} < t] - 1\|_2^2, \tag{5}$$

where $\mathbf{R} = \mathbf{I}_{in} - \mathbf{I}_{gt}$ is the difference of gray values between the input image and ground-truth image, and t is a threshold to distinguish the pixels with strong reflection from that with weak reflection, which is set adaptively proportional to the difference between the input and its ground truth. The sensitivity study on the proportional coefficient is presented in Fig.7, from which we can see that $t = 0.5 \max(\mathbf{I}_{in} - \mathbf{I}_{gt})$ is the optimal threshold.

As stated before, we design a fusion weight loss for learning to generate a weight map according to the reflection intensity. For areas with weak reflection and background, the weight values approximate to 1, and we tend to trust the elimination result, so the reliable result can be expressed as $\mathbf{I}_e \odot \mathbf{M}$, where "$\odot$" is the element-wise product operator. While for the areas with strong reflection, the weight values approximate to 0, and we tend to trust the inpainting result. So the reliable result can be expressed as $\mathbf{I}_i \odot (\mathbf{D} - \mathbf{M})$. Then, we concatenate these two reliable results $\mathbf{I}_{cat} = [\mathbf{I}_e \odot \mathbf{M}, \mathbf{I}_i \odot (\mathbf{D} - \mathbf{M})]$. Since the weight map $\mathbf{M}$ is also updated with the training of the network, to better use the weight map, we exploit the partial convolution [58], which is firstly used in the image inpainting problem. By using the result fusion module, our network can learn to update the weight map and fuse the two results simultaneously. The result fusion can be formulated as:

$$\mathbf{I}_{out} = \begin{cases} (W * \mathbf{I}_{cat}) \times \dfrac{1}{\tilde{\mathbf{M}}_{3\times3}} + b, & \text{if } \tilde{\mathbf{M}}_{3\times3} > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where "$*$" is the convolution operator, W and b are the learnable parameters of the partial convolution, and $\mathbf{I}_{out}$ is the output of RFM. $\tilde{\mathbf{M}}_{3\times3}$ is the average value of $\mathbf{M}$ in a neighborhood region of 3×3, which can be efficiently calculated by a 3×3 average pooling operation.

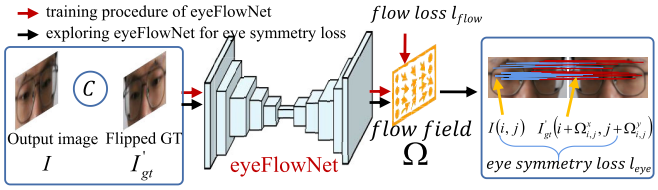


Fig. 8. Overview of the eye symmetry loss ℓ_{eye}.

D. Eye Symmetry Loss

We exploit the symmetry property of human eyes for the restoration of eye details to avoid artifacts. However, it is difficult to capture the matching relationship between the key points in two eyes in the portrait images. Inspired by the optical flow for estimating the matching relationship of a pixel in two neighbouring frames, we propose to estimate the flow field between an eye image and its flipped version to model the matching relationship between the key points of two eyes. Then, with the predicted flow field we propose the eye symmetry loss as the symmetry prior for the restoration of eye details, and the overview of which is shown in Fig. 8.

Specifically, we first design an eyeFlowNet with a simple encoder-decoder structure to estimate the flow vector. The eyeFlowNet takes the eye image $\mathbf{I}_{eye}$ and its flipped image $\mathbf{I}'_{eye}$ as inputs, and predicts a flow vector field $\Omega = (\Omega^x, \Omega^y)$. For a pixel (i, j) in $\mathbf{I}_{eye}$, $(i + \Omega^x_{i,j}, j + \Omega^y_{i,j})$ indicates the position of its matching pixel in $\mathbf{I}'_{eye}$. Since $\mathbf{I}'_{eye}$ is the flipped image of $\mathbf{I}_{eye}$, $\mathbf{I}_{eye}(i, j)$ and $\mathbf{I}'_{eye}(i + \Omega^x_{i,j}, j + \Omega^y_{i,j})$ are a pair of matching points from the left and right eyes. Then, our eye symmetry loss is defined to enforce a pixel of the predicted eye image, *i.e.*, $\mathbf{I}(i, j)$, to be the same as that of the matching point in the flipped ground-truth image, *i.e.*, $\mathbf{I}'_{gt}(i + \Omega^x_{i,j}, j + \Omega^y_{i,j})$:

$$\ell_{eye} = \sum_{\mathbf{I}} \sum_{i,j} \|\mathbf{I}(i, j) - \mathbf{I}'_{gt}(i + \Omega^x_{i,j}, j + \Omega^y_{i,j})\|_2^2 \odot \mathbf{M}_{eye}(i, j), \tag{7}$$

where $\mathbf{M}_{eye}$ is the mask of eye areas, and this loss can be imposed on the three predicted results, so $\mathbf{I} \in \{\mathbf{I}_e, \mathbf{I}_i, \mathbf{I}_{out}\}$. In Fig. 8, we give the detail procedure of exploring the eyeFlowNet for calculating the eye symmetry loss. ℓ_{eye} ensures that our network can learn to fit the ground-truth image and the information about the symmetrical structure of the eyes simultaneously, to avoid visual defects with obvious inconsistency in the structure of the left and right eyes.

To train the eyeFlowNet, we detect the 68 facial keypoints [59] on the ground truth image $\mathbf{I}_{gt}$ and the flipped image $\mathbf{I}'_{gt}$, respectively. We then select the 22 keypoints related to the eyes from the 68 keypoints as the matching eye landmarks $\{(x_i^g, y_i^g)|_{i=1}^{22}\}$ and $\{(x_i^{g'}, y_i^{g'})|_{i=1}^{22}\}$ in $\mathbf{I}_{eye}$ and $\mathbf{I}'_{eye}$, respectively. Following [60], we use the landmark matching loss to constrain the eyeFlowNet:

$$\ell_{lm} = \sum_{i=1}^{22} \|\Omega^x_{x_i^{g'}, y_i^{g'}} - x_i^g\|_2^2 + \|\Omega^y_{x_i^{g'}, y_i^{g'}} - y_i^g\|_2^2. \tag{8}$$

Besides, due to the keypoints are sparse, we follow [60] to use the *TV* loss to ensure the smoothness of flow field:

$$\ell_{TV} = \|\Delta_x \Omega^x\|_2^2 + \|\Delta_y \Omega^x\|_2^2 + \|\Delta_x \Omega^y\|_2^2 + \|\Delta_y \Omega^y\|_2^2, \quad (9)$$

where Δ_x and Δ_y are the gradient operators along x and y directions, respectively.

By combining the above two terms, the loss function for constraining the eyeFlowNet becomes:

$$\ell_{flow} = \lambda_{lm}\ell_{lm} + \lambda_{TV}\ell_{TV}, \quad (10)$$

where $\lambda_{lm} = 10$ and $\lambda_{TV} = 1$ according to [60].

E. Training Loss

In addition to the weight loss mentioned above in Eq. (5) and the eye symmetry loss in Eq. (7), we also use three additional loss functions to train our ER^{2}Net.

1) Reconstruction Loss: We use ℓ_2 loss to minimize the difference between the result image and the corresponding ground-truth image $\mathbf{I}_{gt}$ on the reflection elimination result $\mathbf{I}_e$, the content inpainting result $\mathbf{I}_i$, and the final output result $\mathbf{I}_{out}$:

$$\ell_{rec}(\mathbf{I}, \mathbf{I}_{gt}) = \sum_{\mathbf{I}\in\{\mathbf{I}_e,\mathbf{I}_i,\mathbf{I}_{out}\}} \|\mathbf{I} - \mathbf{I}_{gt}\|_2^2. \quad (11)$$

2) Perception Loss: We use a pre-trained VGG-19 [61] model ϕ to improve the similarity between $\mathbf{I}_e$, $\mathbf{I}_i$, $\mathbf{I}_{out}$ and $\mathbf{I}_{gt}$:

$$\ell_p(\mathbf{I}, \mathbf{I}_{gt}) = \sum_{\mathbf{I}\in\{\mathbf{I}_e,\mathbf{I}_i,\mathbf{I}_{out}\}} \sum_l \lambda_l \cdot \|\phi_l(\mathbf{I}) - \phi_l(\mathbf{I}_{gt})\|_1, \quad (12)$$

where ϕ_l is the output of l-th layer in VGG-19 model, and λ_l is the weight factor.

3) Focal Loss: We use the focal loss [62] to constrain the detection result $\mathbf{D}$ in the reflection area, which performs well on tasks with foreground-background class imbalance problems:

$$\ell_{Focal}(\mathbf{D}_i, \mathbf{T}_i) = \begin{cases} -\alpha(1-\mathbf{D}_i)^\gamma \log \mathbf{D}_i, & \mathbf{T}_i = 1 \\ -(1-\alpha)\mathbf{D}_i^\gamma \log(1-\mathbf{D}_i), & \mathbf{T}_i = 0 \end{cases} \quad (13)$$

where $\mathbf{T} = (\mathbf{I}_{in} - \mathbf{I}_{gt}) > 0.7\max(\mathbf{I}_{in} - \mathbf{I}_{gt})$ is the reflection segmentation image, $\mathbf{I}_{in}$ is the input reflective image and the threshold 0.7 is empirically set according to [10], i is the element index in $\mathbf{D}$ and $\mathbf{T}$, and we empirically set $\alpha = 0.25$ and $\gamma = 2$ according to [62].

4) Total Loss: The learning objective for training our network can be formulated as:

$$\ell_{total} = \lambda_1\ell_{rec} + \lambda_2\ell_p + \lambda_3\ell_{Focal} + \lambda_4\ell_{weight} + \lambda_5\ell_{flow} + \lambda_6\ell_{eye}, \quad (14)$$

where we experimentally set $\lambda_1 = 1.0$, $\lambda_2 = 0.01$, $\lambda_3 = 1.0$, $\lambda_4 = 10.0$, $\lambda_5 = 1.0$ and $\lambda_6 = 1.0$.

TABLE II
QUANTITATIVE COMPARISON RESULTS ON OUR REYER DATASET. THE BEST RESULTS ARE IN **BOLD**

Methods	SSIM ↑	PSNR ↑	LPIPS ↓
Watanabe *et al.* [24]	0.963	34.40	0.026
Zhang *et al.* [8]	0.965	36.71	0.015
IBCLN [6]	0.966	35.32	0.015
SpecularityNet [10]	0.970	36.10	0.015
SGR^{2}N [29]	0.878	21.69	0.127
ExGANs [7]	0.899	28.43	0.058
Diffusion-Net [64]	0.938	27.35	0.053
ER^{2}Net	**0.974**	**37.07**	**0.012**

V. EXPERIMENTS

A. Implementation Details

Our network is implemented in PyTorch on a NVIDIA GeForce 2080Ti card. We use the Adam [63] optimizer to optimize our model with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and set the initial learning rate to 10^{-4} to decrease with an attenuation coefficient of 0.4 every 5 epochs until 10^{-5}. We train our model on the ReyeR dataset for 80 epochs. In the first 40 epochs, we train the eyeFlowNet to learn the eye symmetry prior, which is used to constrain the final result in the symmetry loss in the last 40 epochs. Besides, we collected 700 portrait images with glasses from the Internet, and added masks to the eyes for pre-training our inpainting branch.

B. Comparison Results on ReyeR

Since there are rare learning-based methods that focus on eyeglass reflection removal, we compare our method with only one eyeglass reflection removal method, *i.e.*, Watanabe and Hasegawa [24], four state-of-the-art image reflection removal, *i.e.*, Zhang et al. [8], IBCLN [6], SpecularityNet [10], and SGR^{2}N [29], an eye completion method, *i.e.*, ExGANs [7], as well as a diffusion model-based method [64]. Since the method [64] is originally proposed for shadow removal task, we modify the method by removing the mask generation condition. We train these networks on our ReyeR dataset with the optimal settings in the original published paper.

1) Metrics: For the labeled ReyeR, we calculate PSNR, SSIM and LPIPS [65] on the RGB space for evaluation, and for unlabeled images in the wild, qualitative comparisons are provided through visual observation.

The quantitative comparison results are reported in Table II. As we can see, our method performs the best on all the three metrics. Fig. 9 shows the qualitative comparison results on ReyeR. As can be seen, the results of Watanabe and Hasegawa [24], Zhang et al. [8], IBCLN [6], SpecularityNet [10] and SGR^{2}N [29] still contain some residual reflection on the glasses, and these methods can only remove weak reflection. ExGANs [7] slightly change the original eye, due to the neglect of the texture information in the reflection regions. While our method can combine the removal and inpainting results to produce realistic and fine-grained reflection-free results, and the lighting conditions, e.g., shadows, in the result images are more consistent with the ground-truth images

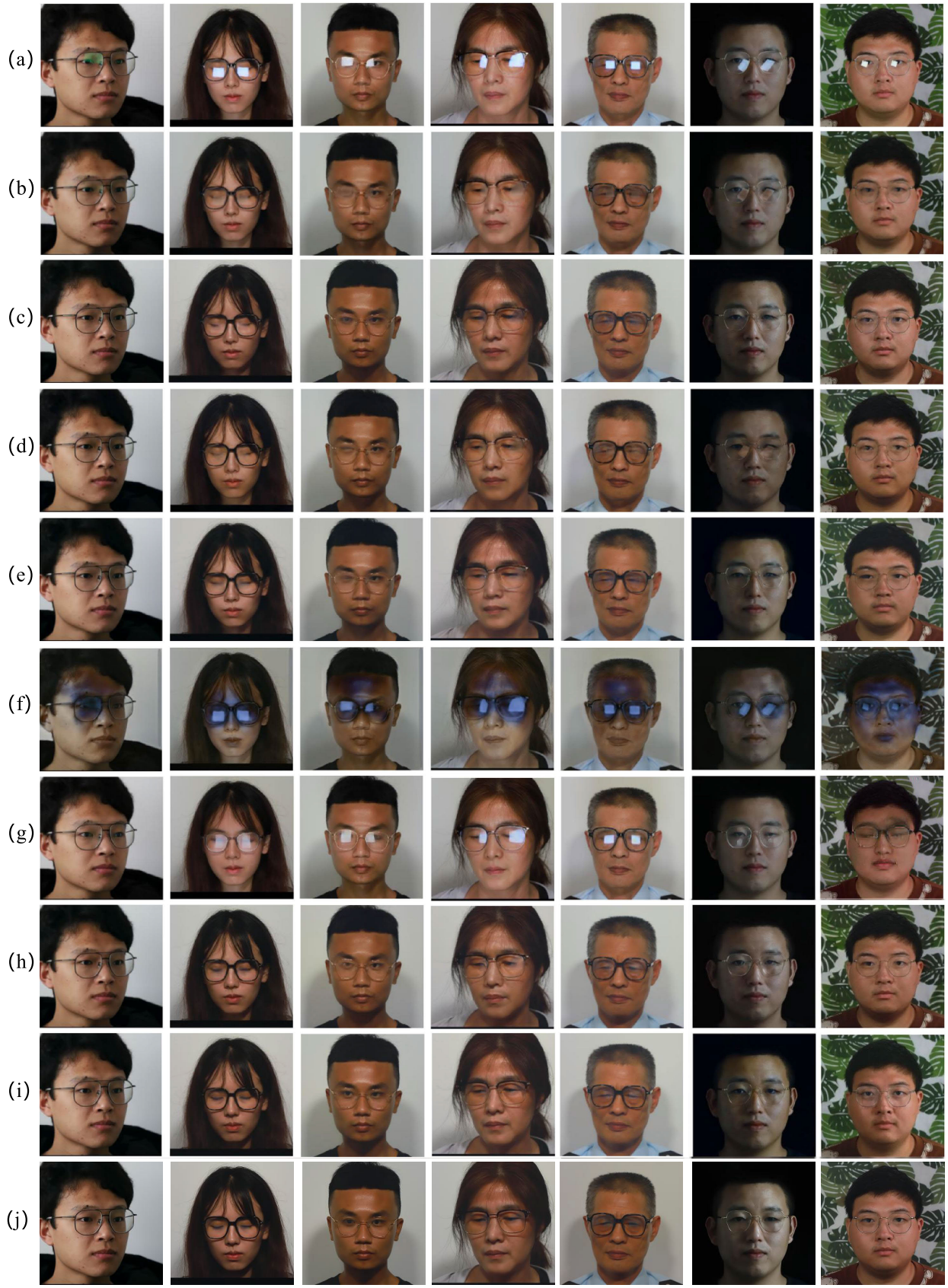


Fig. 9. Qualitative comparison with other methods on ReyeR. (a) Input image; (b) Watanabe and Hasegawa [24]; (c) Zhang et al. [8]; (d) IBCLN [6]; (e) SpecularityNet [10]; (f) SGR^{2}N [29]; (g) ExGANs [7]; (h) Diffusion-based method [64]; (i) Ours; (j) GT. Watanabe and Hasegawa [24], Zhang et al. [8], IBCLN [6], SpecularityNet [10] and SGR^{2}N [29] still have residual reflection on the glasses, ExGANs [7] slightly changes the original eye, our method can produce realistic and fine-grained reflection-free results.

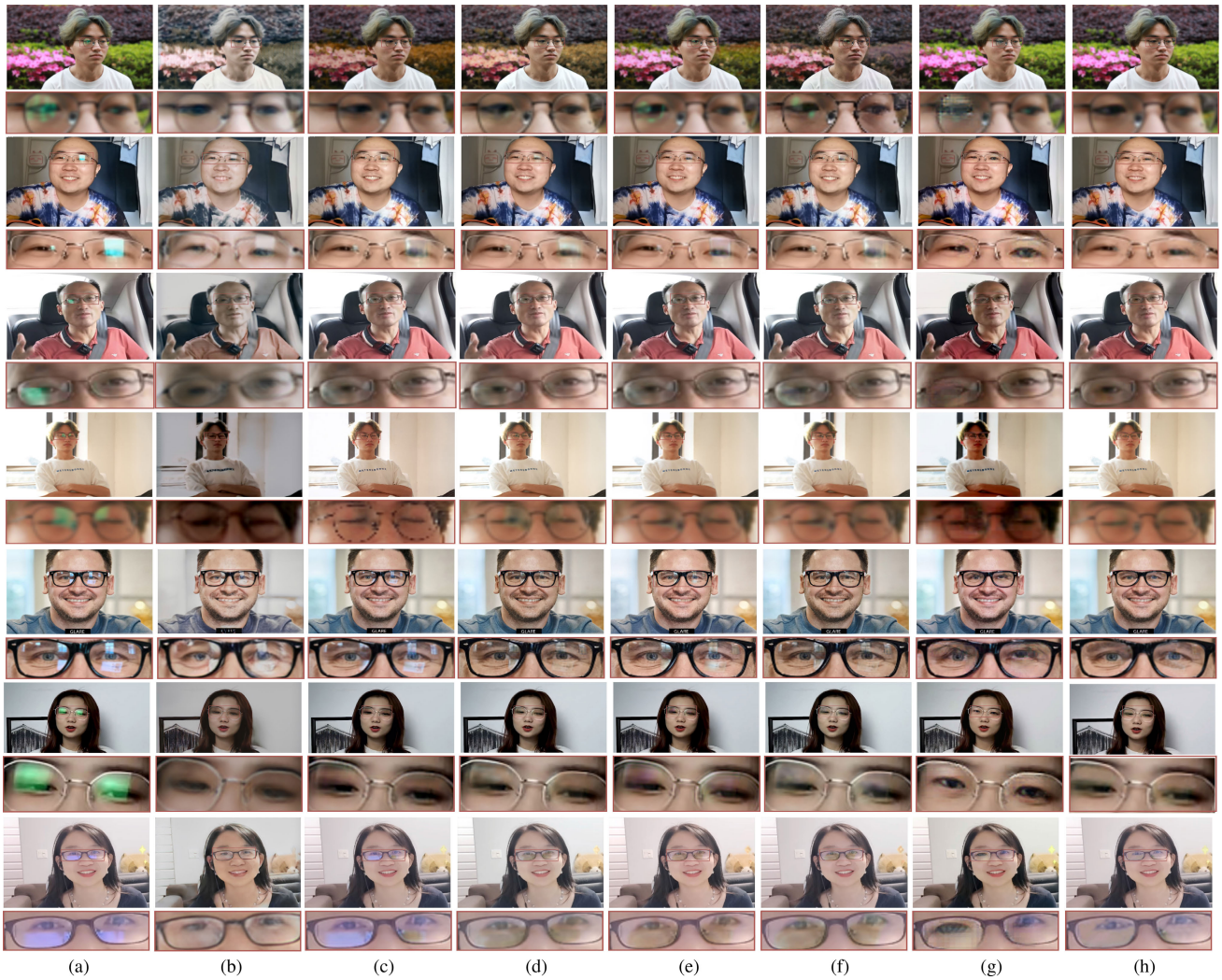


Fig. 10. Visualization results of qualitative comparisons on images in the wild with complex background. (a) Input image; (b) Watanabe and Hasegawa [24]; (c) Zhang et al. [8]; (d) IBCLN [6]; (e) SpecularityNet [10]; (f) SGR^{2}N [29]; (g) ExGANs [7]; (h) Ours. Results of Watanabe and Hasegawa [24], Zhang et al. [8], IBCLN [6], SpecularityNet [10] and SGR^{2}N [29] contain residual reflection and color distortion, ExGANs [7] has obvious visual artifacts. Our method can effectively remove the reflection and restore accurate details.

compared with the diffusion-based method. All the qualitative and quantitative results can demonstrate the effectiveness of our method.

C. Comparison on Images in the Wild With Complex Backgrounds

Fig. 10 shows the evaluation results on the images in the outdoor (the 1^{st} sample) and indoor (the $2^{nd} - 7^{th}$ samples) situations with complex backgrounds to demonstrate the generalization ability of our model. All the test images are captured with a camera and have different reflection intensities. It can be seen that the competitive methods do not clearly remove the reflection or do not recover the original eye information accurately, while our method can generalize to the images in the wild. The main reason is that our method employs the memory module for improving the generalization ability, and can restore the fine-grained eye images by leveraging the symmetry prior information. Furthermore, although the background for our dataset collection is relatively monotonous, our method can deal with the reflections in the indoor and outdoor situations very well. The main reason is that our method can detect the reflection regions more accurately, and thus can remove the reflection but not affected by the complex background.

D. Comparison on Images With Strong and Large Area Reflection

We also conduct evaluations on some images with strong reflection and large area reflection to verify the robustness of the proposed method. Fig. 11 presents several examples with strong reflection, and the 2^{nd} and the 3^{nd} rows show the examples with large area reflection. It can be seen that the competitive methods fail to accurately recover the content due to the complete information missing underneath the reflection and the large area reflection. In contrast, our method not only removes reflections, but also accurately restores the texture information as well as ensuring image quality, making the restored image more realistic. We also make a statistic analysis of the impact of reflection area in Table III, in which the large area is defined as $Ta > 0.15\%$ (Ta is the proportion of the reflection pixels in the whole image). We can see that the our method is not sensitive to the reflection area. All the results

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 11. Visualization results of qualitative comparisons on images with strong reflection. (a) Input image; (b) Watanabe et al. [24]; (c) Zhang et al. [8]; (d) IBCLN [6]; (e) SpecularityNet [10]; (f) SGR^{2}N [29]; (g) ExGANs [7]; (h) Ours. Results of Watanabe et al. [24], Zhang et al. [8], IBCLN [6], SpecularityNet [10] and SGR^{2}N [29] have some residual reflection and have obvious color distortion, ExGANs [7] has obvious visual artifacts. In comparison, our method can effectively remove the strong reflection and accurately restore the missing details.

TABLE III

STATISTIC ANALYSIS OF THE IMPACT OF REFLECTION AREA

	SSIM ↑	PSNR ↑	LPIPS ↓
Large area	0.970	37.15	0.014
Small area	0.976	37.04	0.011

show that our method can generalize to strong and large area reflection.

E. Comparison on Other Natural Image Reflection Datasets

To further validate the effectiveness of our method, we retrain our method (without the ℓ_{eye}) on the natural reflection removal datasets, including the Zhang et al. [8] and SIR^2 [9], and then compare with the state-of-the-art image reflection removal methods. Table IV shows that our method can achieve comparable results to other methods on other datasets even without the ℓ_{eye} loss.

TABLE IV

QUANTITATIVE COMPARISON RESULTS ON OTHER NATURAL IMAGE REFLECTION DATASET. THE BEST RESULTS ARE IN **BOLD**

Datasets	Methods	SSIM ↑	PSNR ↑	LPIPS ↓
Zhang *et al.* [8]	Watanabe *et al.* [24]	0.782	21.23	0.215
	Zhang *et al.* [8]	0.792	22.42	0.195
	IBCLN [6]	0.762	21.86	0.190
	SpecularityNet [10]	0.814	22.45	0.145
	SGR^{2}N [29]	0.812	22.46	**0.127**
	ER^{2}Net (w/o ℓ_{eye})	**0.820**	**23.28**	0.129
SIR^2 [9]	Watanabe *et al.* [24]	0.791	21.34	0.207
	Zhang *et al.* [8]	0.829	21.52	0.196
	IBCLN [6]	0.886	24.71	0.187
	SpecularityNet [10]	0.881	24.57	0.156
	SGR^{2}N [29]	0.893	23.81	0.124
	ER^{2}Net (w/o ℓ_{eye})	**0.901**	**24.92**	**0.123**

F. Ablation Study

To further verify the effectiveness of each module in our method, we design seven variants, that are:

(a) (b) (c) (d) (e) (f) (g) (h) (i)

Fig. 12. Visualization of ablation study results. (a) Input image; (b) w/o Detection; (c) only Elimination; (d) only Inpainting; (e) RFM→PF; (f) w/o ResBlk; (g) w/o MM; (h) w/o ℓ_{eye}; (i) Our ER^{2}Net.

TABLE V

QUANTITATIVE COMPARISON RESULTS OF ABLATION STUDY. THE BEST RESULTS ARE MARKED IN **BOLD**

Variants	SSIM ↑	PSNR ↑	LPIPS ↓
w/o Detection	0.917	31.96	0.061
only Elimination	0.895	29.22	0.079
only Inpainting	0.877	27.24	0.083
RFM→PF	0.904	30.12	0.064
w/o ResBlk	0.901	30.09	0.060
w/o MM	0.856	25.12	0.091
w/o ℓ_{eye}	0.912	32.13	0.059
ER^{2}Net	**0.926**	**33.94**	**0.054**

(1) **w/o Detection:** MTNet without Detection Branch;
(2) **only Elimination:** MTNet only using Elimination Branch for reflection removal;
(3) **only Inpainting:** MTNet only using Inpainting Branch for content inpainting;
(4) **RFM→PF:** using a **P**lain **F**usion layer (concat+conv) instead of RFM;
(5) **w/o ResBlk:** Elimination Branch without residual block;
(6) **w/o MM:** Inpainting Branch without memory module;
(7) **w/o ℓ_{eye}:** ER^{2}Net without the eye symmetry loss;

We train and evaluate above seven variants on our ReyeR dataset. The results are summarized in Table V, from which we can observe that all modules could improve the reflection removal performance of our method, which suggests the effectiveness of our designs.

Fig. 12 provides the visual comparisons of ablated variants. From the results in Fig. 12 (b), we can observe that the detection branch can affect the final results. Moreover, without the detection branch, the memory module needs to vectorize the whole image, which will take a lot of training time and calculation.

Fig. 12 (c) illustrates that the elimination branch can effectively remove the weak reflection, but still has residual reflection or black artifacts for strong reflection. Fig. 12 (d) shows that although the inpainting branch can also remove the reflection, the restored eyes are unnatural. Fig. 12 (e) further proves that RFM is important to fusing reflection elimination and content inpainting results to generate reflection-free images. Fig. 12 (f) shows that without residual block the results exhibits more residual reflection, and Fig. 12 (g) shows that without the memory module, the results exhibits obvious eye artifacts, which demonstrate that our residual block is effective for weak reflection removal and memory module works well for eye restoration. From the results in Fig. 12 (h), we can see obvious visual artifacts, while our ER^{2}Net can remove the visual artifacts and restore the eye details effectively. It demonstrates the effectiveness of eye symmetry loss on avoiding visual artifacts and restoring the eye details.

TABLE VI

SENSITIVITY STUDY ON PARAMS OF OUR FOCAL LOSS AND WEIGHT LOSS. THE BEST RESULTS ARE MARKED IN **BOLD**

		SSIM ↑	PSNR ↑	LPIPS ↓
$\lambda_4 = 1.0, \lambda_3 =$	0.1	0.822	26.28	0.118
	0.5	0.891	27.37	0.095
	1.0	**0.895**	**29.02**	**0.079**
	2.0	0.835	28.48	0.088
	5.0	0.790	26.33	0.104
$\lambda_3 = 1.0, \lambda_4 =$	1.0	0.895	29.02	0.079
	5.0	0.896	30.15	0.071
	10	**0.912**	**32.13**	**0.059**
	15	0.901	31.28	0.064
	20	0.799	28.85	0.091

TABLE VII

SENSITIVITY STUDY ON PARAMS OF OUR FLOW LOSS AND EYE SYMMETRY LOSS. THE BEST RESULT IS MARKED IN **BOLD**

		SSIM ↑	PSNR ↑	LPIPS ↓
$\lambda_6 = 0.1, \lambda_5 =$	0.1	0.823	28.28	0.103
	0.5	0.842	28.97	0.095
	1.0	**0.911**	**30.12**	**0.060**
	2.0	0.895	29.79	0.087
	5.0	0.830	28.33	0.101
$\lambda_5 = 1.0, \lambda_6 =$	0.1	0.911	30.12	0.060
	0.5	0.910	31.15	0.058
	1.0	**0.926**	**33.94**	**0.054**
	2.0	0.917	32.37	0.058
	5.0	0.896	29.92	0.075

G. Sensitivity Study

Since there are several losses in Eq. (14), we conduct sensitivity studies to better understand the necessity of loss functions. We first consider the sensitivity study on the parameters λ_3 and λ_4 of focal loss and weight loss. Then, we perform the sensitivity study on parameters λ_5 and λ_6 of flow loss and eye symmetry loss. Table VI and Table VII report the quantitative results on different parameters of our loss functions.

(a) (b) (c) (d) (e)

Fig. 13. Visualization of three intermediate results in MTNet and final fine-grained result. (a) Input image; (b) detection result; (c) eliminating result; (d) inpainting result; (e) final result.

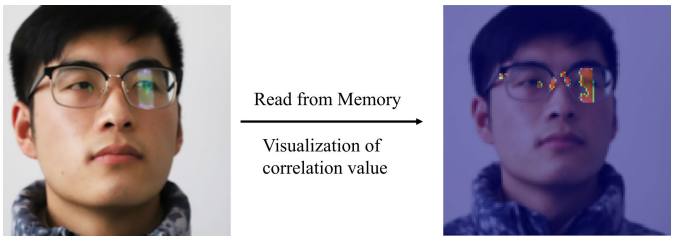


Fig. 14. Visualization of the correlation value between the input image and memory items in the reading process. The redder the color, the higher the correlation.

H. Discussions

1) Adverse Effect of Eyeglass Reflection on Downstream Tasks: We conduct the facial landmark detection following [4] on the portrait images with eyeglass reflections to empirically assess whether these reflections indeed pose a significant problem. We also implement the algorithm on the ground-truth images without reflections in the ReyeR dataset for comparison. The normalized mean error (NME) between these two results is 8.28, which demonstrates that the reflection has adverse impact on the facial landmark detection task.

2) Visualization Results of Intermediate Outputs: Our MTNet produces three intermediate outputs, the reflection detection result $\mathbf{D}$, the content inpainting result $\mathbf{I}_i$ and the reflection eliminating result $\mathbf{I}_e$, respectively. We show the three intermediate results and our final result in Fig. 13, including a weak reflection image (the 1^{st} row) and a strong reflection image (the 2^{nd} row). From the results we can see: (1) the detection branch can detect both the weak and strong reflection regions accurately; (2) the eliminating results are relatively more satisfactory in the weak reflection regions than that in the strong reflection regions, and there are still many residual reflections in strong reflection regions (see the yellow rectangle in Fig. 13 (c)); (3) the inpainting result is more effective in the regions with smooth backgrounds, but it introduces visual artifacts in the regions with texture details (see the red rectangle in Fig. 13 (d)); (4) our RFM can fully fuse the two results and bridge the disadvantages of the two methods, and generate the final fine-grained reflection-free results.

3) Discussion on the Memory Module: The mechanism of our memory module for enhancing the inpainting result is that, it records the prototypical features cross images for content reasoning by calculating the correlation between query features and memory items. In the writing process, we find the most similar item $\mathbf{m}_i$ for each query (as in Eq. (3)), and then use the queries found for updating $\mathbf{m}_i$ (as in Eq. (4)). During the reading process, the correlation values between the query features and the learned memory items are visualized in Eq. (1) in Fig. 14, which show that the missing eye contents are highly correlated with the memory items. We also analyze the effect of the size of the memory block to better understand the performance. Specifically, we try 5 different value of the memory size (s), and the performance of our method is presented in Table VIII. We choose $s = 512$ for the trade-off between the space complexity and the performance.

TABLE VIII
SENSITIVITY STUDY ON SIZE OF THE MEMORY BLOCK

	$s = 256$	$s = 400$	$s = 512$	$s = 800$	$s = 1,024$
SSIM ↑	0.970	0.970	0.971	0.971	0.972
PSNR ↑	35.15	35.47	35.62	35.50	36.37
LPIPS ↓	0.017	0.016	0.016	0.016	0.015

TABLE IX
ANALYSIS ON DETECTION ACCURACY AND ITS IMPACT ON THE FINAL REMOVAL RESULT

Methods	BER ↓	RBER ↓	NBER ↓
U-Net	1.17	0.98	1.27
Ours	0.98	0.87	1.12
Methods	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**
ER^{2}Net w. U-Net	0.936	35.85	0.049
ER^{2}Net	0.974	37.07	0.012

4) Effects of the Detection Accuracy: We calculate the BER (Balance Error Rate) value on the reflection regions (denoted as RBER), non-reflection regions (denoted as NBER), and the whole image (denoted as BER), and compare them with that of a plain U-Net to evaluate the performance of our reflection detection branch. The results are presented in the upper part of Table IX. We can see that although our detection branch is a simple encoder-decoder network, it performs much better than that of U-Net. The main reason is that the learning procedure of the multi-task network promotes the training of the detection branch. We further evaluate the effects of the reflection detection accuracy on the final removal results. To this end, we first remove the detection branch in the MTNet, and replace the detection result with the result of the U-Net, which is denoted as **ER^{2}Net w. U-Net**. The results are presented in the lower part of Table IX. It can be seen that the detection result is of great importance to the final result, and our ER^{2}Net achieves much better performance on reflection removal due to the better performance on reflection detection.

5) Discussion on the Replacement Approach of the Eye Symmetry Loss: We use the adversarial loss following [66] instead of ℓ_{eye}, which is denoted as $\ell_{eye} \rightarrow \ell_{adv}$, to understand the effectiveness of ℓ_{eye} for avoiding visual effects, and the results are presented in Table X. It can be seen that our eye symmetry loss can achieve better performance compared with the adversarial loss.

6) Complexity Analysis: Our ER^{2}Net consists of the MTNet and the RFM. The RFM produces the final result requiring the result of the three results produced by MTNet. So the computational complexity of our method is defined as the

TABLE X
COMPARATIVE ANALYSIS OF THE EYE SYMMETRY LOSS AND ITS REPLACEMENT APPROACH

Methods	SSIM ↑	PSNR ↑	LPIPS ↓
ER^{2}Net w. $\ell_{eye} \rightarrow \ell_{adv}$	0.906	31.40	0.026
ER^{2}Net	0.974	37.07	0.012

TABLE XI
COMPARISON RESULTS ON COMPUTING COMPLEXITY

Methods	#Params(M)	FLOPs(G)	Infer time(s)
Watanabe *et al.* [24]	9.15	147.12	0.39
Zhang *et al.* [8]	3.16	132.26	0.12
IBCLN [6]	21.61	312.15	0.90
SpecularityNet [10]	22.84	322.37	0.84
SGR^{2}N [29]	21.58	308.79	0.85
ExGANs [7]	49.26	343.22	1.02
ER^{2}Net	14.82	146.25	0.39

Fig. 15. Performance on colorful reflections, 1^{st} row: the input images, 2^{nd} row: the result images.

summation of all the three branches in MTNet and RFM. We make comparisons with SOTA methods on params count, FLOPs and inference time in Table XI. Although Watanabe and Hasegawa [24] and Zhang et al. [8] adopt a simple U-net architecture to remove reflection with lower parameters and FLOPs, it performs much poorer compared with the other methods. On the contrary, our method can not only effectively remove reflection and produce a fine-grain result, but also has a relatively lower computing complexity.

7) Discussion on the Dataset: (a) About the colorful light source. The images in ReyeR are mainly collected under white light sources, with a small amount of images captured using yellow light sources. Due to the different materials of the lenses and the different materials of the coating film on the lenses, different glasses display different colors for reflection, including purple, blue, green, and etc. Our algorithm can effectively remove reflections with different colors (shown in Fig. 15). (b) Our method is not effective enough to the reflection caused by the sunlight, and this limitation is indeed caused the dataset where the images are collected in the artificial lighting environments. Since the sunlight is difficult to control, we cannot simulate this complex lighting condition in the lab.

I. Advantage Scenarios and Limitations

1) Advantage Scenarios: Our method can deal with both indoor and partial outdoor eyeglass image reflection removal (see Fig. 9, Table II, and Fig. 10), including reflection of various colors (see Fig. 15), various areas (see Fig. 11 and Table III), various intensity (see Fig. 11). It also performs well on natural image reflection removal (see Table IV) compared with other SOTA methods.

(a) Input (b) Ours (c) Watanabe *et al.* [24] (d) Zhang *et al.* [8]
(e) IBCLN [6] (f) SpecularityNet [10] (g) SGR^{2}N [29] (h) ExGANs [7]

Fig. 16. A failure case of our method.

2) Limitations: Our method does not perform well on the outdoor complex reflections caused by the sunshine and other complex refracted light sources. As shown in Fig. 16, the problem of residual reflection is serious. The main reason is that our dataset is mainly collected in the lab with artificial light sources under controllable lighting conditions. However, even in this case, our method still achieves better results than other methods. Addressing complex reflections is left as our future work.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have constructed a real-world eyeglass reflection dataset (ReyeR) and presented the ER^{2}Net for eyeglass reflection removal. Our key idea for dealing with both weak and strong reflection is to learn the reflection elimination and content inpainting jointly, and then fuse the results of both branches adaptively. The eye symmetry loss is introduced for avoiding visual artifacts. Extensive experimental results in our ReyeR data set demonstrate that our ER^{2}Net is effective to deal with previous challenging eyeglass reflections, and generalizes well on in-the-wild images.

FUTURE WORK

In future work, we will further capture the real eyeglass reflection dataset in outdoor scenes to simulate a more realistic reflection distribution and enhance the generalization ability of our method.

VII. DECLARATIONS

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

REFERENCES

[1] L. Du and H. Hu, "Cross-age identity difference analysis model based on image pairs for age invariant face verification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2675–2685, Jul. 2021.
[2] R. Sharma, V. K. Sharma, and A. Singh, "A review paper on facial recognition techniques," in *Proc. 5th Int. Conf. I-SMAC (IoT Social, Mobile, Analytics Cloud) (I-SMAC)*, Nov. 2021, pp. 617–621.
[3] S. Ge, C. Li, S. Zhao, and D. Zeng, "Occluded face recognition in the wild by identity-diversity inpainting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3387–3397, Oct. 2020.
[4] H. Jin, S. Liao, and L. Shao, "Pixel-in-pixel net: Towards efficient facial landmark detection in the wild," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3174–3194, Dec. 2021.

[5] L. Liu et al., "A face alignment accelerator based on optimized coarse-to-fine shape searching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2467–2481, Aug. 2019.

[6] C. Li, Y. Yang, K. He, S. Lin, and J. E. Hopcroft, "Single image reflection removal through cascaded refinement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3562–3571.

[7] B. Dolhansky and C. C. Ferrer, "Eye in-painting with exemplar generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7902–7911.

[8] X. Zhang, R. Ng, and Q. Chen, "Single image reflection separation with perceptual losses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4786–4794.

[9] R. Wan, B. Shi, L. Duan, A. Tan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3942–3950.

[10] Z. Wu et al., "Single-image specular highlight removal via real-world dataset construction," *IEEE Trans. Multimedia*, vol. 24, pp. 3782–3793, 2022.

[11] X. Zhang, K. Xing, Q. Liu, D. Chen, and Y. Yin, "Single image reflection removal based on dark channel sparsity prior," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 11, pp. 6431–6442, Jun. 2023.

[12] B. Song, J. Zhou, and H. Wu, "Multistage curvature-guided network for progressive single image reflection removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6515–6529, Oct. 2022.

[13] Y. Li, Q. Yan, K. Zhang, and H. Xu, "Image reflection removal via contextual feature fusion pyramid and task-driven regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 553–565, Feb. 2022.

[14] R. Wan, B. Shi, H. Li, L.-Y. Duan, and A. C. Kot, "Face image reflection removal," 2019, *arXiv:1903.00865*.

[15] N. Kong, Y.-W. Tai, and J. S. Shin, "A physically-based approach to reflection separation: From physical modeling to constrained optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 2, pp. 209–221, Feb. 2014.

[16] Y. Y. Schechner, N. Kiryati, and R. Basri, "Separation of transparent layers using focus," in *Proc. 6th Int. Conf. Comput. Vis.*, Jun. 1998, pp. 1061–1066.

[17] N. Arvanitopoulos, R. Achanta, and S. Süsstrunk, "Single image reflection suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1752–1760.

[18] A. Levin and Y. Weiss, "User assisted separation of reflections from a single image using a sparsity prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1647–1654, Sep. 2007.

[19] T. Sandhan and J. Y. Choi, "Anti-glare: Tightly constrained optimization for eyeglass reflection removal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1675–1684.

[20] R. Wan, B. Shi, L.-Y. Duan, A.-H. Tan, W. Gao, and A. C. Kot, "Region-aware reflection removal with unified content and gradient priors," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2927–2941, Jun. 2018.

[21] R. Wan, B. Shi, T. A. Hwee, and A. C. Kot, "Depth of field guided reflection removal," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 21–25.

[22] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2752–2759.

[23] D. Conte, P. Foggia, G. Percannella, and M. Vento, "Removing object reflections in videos by global optimization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 11, pp. 1623–1633, Nov. 2012.

[24] S. Watanabe and M. Hasegawa, "Reflection removal on eyeglasses using deep learning," in *Proc. 36th Int. Tech. Conf. Circuits/Systems, Comput. Commun. (ITC-CSCC)*, Jun. 2021, pp. 1–4.

[25] A. Amanlou, A. A. Suratgar, J. Tavoosi, A. Mohammadzadeh, and A. Mosavi, "Single-image reflection removal using deep learning: A systematic review," *IEEE Access*, vol. 10, pp. 29937–29953, 2022.

[26] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3258–3267.

[27] S. Kim, Y. Huo, and S. Yoon, "Single image reflection removal with physically-based training images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5163–5172.

[28] Z. Dong, K. Xu, Y. Yang, H. Bao, W. Xu, and R. W. H. Lau, "Location-aware single image reflection removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5017–5026.

[29] Y. Liu, Y. Li, S. You, and F. Lu, "Semantic guided single image reflection removal," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 3s, pp. 1–23, Oct. 2022.

[30] Y.-C. Chang, C.-N. Lu, C.-C. Cheng, and W.-C. Chiu, "Single image reflection removal with edge guidance, reflection classifier, and recurrent decomposition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 2032–2041.

[31] B. H. Pawan Prasad, K. S. Green Rosh, R. B. Lokesh, K. Mitra, and S. Chowdhury, "V-DESIRR: Very fast deep embedded single image reflection removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2390–2399.

[32] R. Wan, B. Shi, H. Li, Y. Hong, L.-Y. Duan, and A. C. Kot, "Benchmarking single-image reflection removal algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1424–1441, Feb. 2023.

[33] A. Agrawal, R. Raskar, S. K. Nayar, and Y. Li, "Removing photography artifacts using gradient projection and flash-exposure sampling," in *Proc. ACM SIGGRAPH Papers*, Jul. 2005, pp. 828–835.

[34] Y. Fu, A. Lam, I. Sato, T. Okabe, and Y. Sato, "Separating reflective and fluorescent components using high frequency illumination in the spectral domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 5, pp. 965–978, May 2016.

[35] X. Guo, X. Cao, and Y. Ma, "Robust separation of reflection from multiple images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2195–2202.

[36] T. Xue, M. Rubinstein, C. Liu, and W. T. Freeman, "A computational approach for obstruction-free photography," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 1–11, Jul. 2015.

[37] B.-J. Han and J.-Y. Sim, "Reflection removal using low-rank matrix completion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3872–3880.

[38] C. Sun, S. Liu, T. Yang, B. Zeng, Z. Wang, and G. Liu, "Automatic reflection removal using gradient intensity and motion cues," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 466–470.

[39] Y. Hong, Q. Zheng, L. Zhao, X. Jiang, A. C. Kot, and B. Shi, "Panoramic image reflection removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7758–7767.

[40] R. Wan, B. Shi, H. Li, L.-Y. Duan, and A. C. Kot, "Face image reflection removal," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 385–399, Feb. 2021.

[41] T. Li, Y.-H. Chan, and D. P. K. Lun, "Improved multiple-image-based reflection removal algorithm using deep neural networks," *IEEE Trans. Image Process.*, vol. 30, pp. 68–79, 2021.

[42] A. Agarwala et al., "Interactive digital photomontage," in *Proc. ACM SIGGRAPH Papers*, Aug. 2004, pp. 294–302.

[43] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, vol. 2. Cambridge, MA, USA: MIT Press, 5555, pp. 2672–2680.

[44] B. Yan, Q. Lin, W. Tan, and S. Zhou, "Assessing eye aesthetics for automatic multi-reference eye in-painting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 13506–13514.

[45] S. Sukhbaatar et al., "End-to-end memory networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 28–37.

[46] Q. P. Thi, A.-C. Pham, N.-H. Ngo, and D.-T. Le, "Memory-based method using prototype augmentation for continual relation extraction," in *Proc. RIVF Int. Conf. Comput. Commun. Technol. (RIVF)*, Dec. 2022, pp. 1–6.

[47] C. Yin, J. Tang, Z. Xu, and Y. Wang, "Memory augmented deep recurrent neural network for video question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3159–3167, Sep. 2020.

[48] A. Gupta and J. Berant, "GMAT: Global memory augmentation for transformers," 2020, *arXiv:2006.03274*.

[49] Z. Lai, E. Lu, and W. Xie, "MAST: A memory-augmented self-supervised tracker," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 6479–6488.

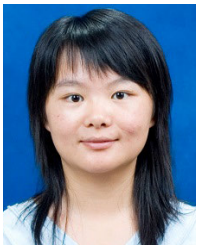
[50] X. Feng, W. Pei, F. Li, F. Chen, D. Zhang, and G. Lu, "Generative memory-guided semantic reasoning model for image inpainting," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7432–7447, Nov. 2022.

[51] R. Xu, M. Guo, J. Wang, X. Li, B. Zhou, and C. C. Loy, "Texture memory-augmented deep patch-based image inpainting," *IEEE Trans. Image Process.*, vol. 30, pp. 9112–9124, 2021.

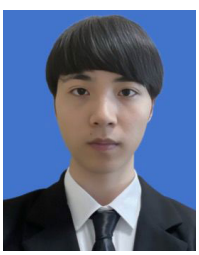
[52] D. Gong et al., "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1705–1714.

[53] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14372–14381.
[54] T. Han, W. Xie, and A. Zisserman, "Memory-augmented dense predictive coding for video representation learning," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2008, pp. 312–329.
[55] G. Fu, Q. Zhang, L. Zhu, P. Li, and C. Xiao, "A multi-task network for joint specular highlight detection and removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7748–7757.
[56] E. Reinhard, M. Adhikhmin, B. Gooch, and P. Shirley, "Color transfer between images," *IEEE Comput. Graph. Appl.*, vol. 21, no. 4, pp. 34–41, Apr. 2001.
[57] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4471–4480.
[58] G. Liu et al., "Partial convolution based padding," 2018, *arXiv:1811.11718*.
[59] Z. Zhang, P. Luo, C. Change Loy, and X. Tang, "Learning deep representation for face alignment with auxiliary attributes," 2014, *arXiv:1408.3967*.
[60] X. Li, M. Liu, Y. Ye, W. Zuo, L. Lin, and R. Yang, "Learning warped guidance for blind face restoration," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 1804, pp. 272–289.
[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[62] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[63] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[64] L. Guo et al., "ShadowDiffusion: When degradation prior meets diffusion model for shadow removal," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14049–14058.
[65] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[66] B. Ding, C. Long, L. Zhang, and C. Xiao, "ARGAN: Attentive recurrent generative adversarial network for shadow detection and removal," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10212–10221.

Wentao Zou received the master's degree from the School of Computer Science, Wuhan University, Wuhan, China. His research interests include deep learning, computer vision, and image highlight removal.



Xiao Lu received the Ph.D. degree in pattern recognition and intelligent systems from Hunan University, Changsha, China, in 2015. She is currently an Associate Professor with the College of Engineering and Design, Hunan Normal University, Changsha. Her research interests include machine learning, computer vision, image processing, and robot perception and decision.



Zhilv Yi is currently pursuing the degree with the School of Computer Science, Wuhan, China. His research interests include deep learning, computer vision, and specular reflection removal.

Ling Zhang received the Ph.D. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2017. She is currently an Associate Professor with the School of Computer Science and Technology, Wuhan University of Science and Technology. Her research interests include computer graphics, computer vision, and computational photography.

Gang Fu received the Ph.D. degree from the School of Computer Science, Wuhan University, China, in 2022. He is currently a Post-Doctoral Researcher with the Department of Computing, The Hong Kong Polytechnic University, China. He has published more than 20 research papers in prestigious international journals and conference proceedings, including *International Journal of Computer Vision*, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON MULTIMEDIA, PR, CVPR, ICCV, ACM MM, and AAAI. His research interests include illumination processing and editing in computer vision and graphics, involving multiple key sub-problems such as specular highlight detection and removal, shadow removal, intrinsic decomposition, and relighting.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Assistant Professor with the Department of Computing and an Assistant Professor with the School of Design, The Hong Kong Polytechnic University, Hong Kong. He has published over 200 top-tier scholarly research articles, pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the ACM TechNews, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts to research fields, addressing societal needs and contributed tremendously to the people concerned. His current research interests include image/video stylization, colorization, artistic rendering and synthesis, realism in non-photorealistic rendering, computational art, and creative media.

Chunxia Xiao (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from Hunan Normal University, Changsha, in 1999 and 2002, respectively, and the Ph.D. degree in applied mathematics from the State Key Laboratory of CAD&CG, Zhejiang University, Hangzhou, in 2006. He became an Assistant Professor with Wuhan University, Wuhan, in 2006 and a Professor in 2011. From October 2006 to April 2007, he was a Post-Doctoral Researcher with the Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong. From February 2012 to February 2013, he visited the University of California at Davis, Davis, for one year. He is currently a Professor with the School of Computer Science, Wuhan University. He has published more than 150 papers in journals and conferences. His main research interests include computer graphics, computer vision, virtual reality, and augmented reality.