Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-20101-5

Enhancing eyeglasses removal in facial images: a novel approach using translation models for eyeglasses mask completion

Zahra Esmaily[1] · Hossein Ebrahimpour-Komleh[1]

Received: 14 April 2024 / Revised: 16 July 2024 / Accepted: 19 August 2024

Abstract

Accurately removing eyeglasses from facial images is crucial for improving the performance of various face-related tasks such as verification, identification, and reconstruction. This paper presents a novel approach to enhancing eyeglasses removal by integrating a mask completion technique into the existing framework. Our method focuses on improving the accuracy of eyeglasses masks, which is essential for subsequent eyeglasses and shadow removal steps. We introduce a unique dataset specifically designed for eyeglasses mask image completion. This dataset is generated by applying Top-Hat morphological operations to existing eyeglasses mask datasets, creating a collection of images containing eyeglasses masks in two states: damaged (incomplete) and complete (ground truth). A Pix2Pix image-to-image translation model is trained on this newly created dataset for the purpose of restoring incomplete eyeglass mask predictions. This restoration step significantly improves the accuracy of eyeglass frame extraction and leads to more realistic results in subsequent eyeglasses and shadow removal. Our method incorporates a post-processing step to refine the completed mask, preventing the formation of artifacts in the background or outside of the eyeglasses frame box, further enhancing the overall quality of the processed image. Experimental results on CelebA, FFHQ, and MeGlass datasets showcase the effectiveness of our method, outperforming state-of-the-art approaches in quantitative metrics (FID, KID, MOS) and qualitative evaluations.

Keywords Eyeglasses removal · Mask completion · Image-to-image translation · Face verification · Generative Adversarial Networks (GANs)

1 Introduction

The eyes, which store crucial biological information, [1] play a central role in facial recognition. Removing occlusion in this region is essential for improving downstream tasks such as face verification [2, 3], identification [4, 5], and reconstruction [6]. Despite the

✉ Hossein Ebrahimpour-Komleh
ebrahimpour.kashanu@gmail.com

[1] Department of Electrical and Computer Engineering, University of Kashan, Kashan, Iran

Published online: 11 September 2024

capabilities of state-of-the-art facial recognition systems in real-world applications, their accuracy decreases when faced with partially occluded facial images, mainly due to eyeglasses. This problem occurs because glasses obscure important facial information, creating mismatches in facial features such as thick glass frames obscuring the eyes [1, 7].

Worn by many, eyeglasses can significantly impact facial photographs, introducing unwanted occlusions and shadows. Consequently, the accuracy of various techniques like face verification, facial expression recognition [8, 9], fatigue detection [10], and those for aesthetic purposes can be compromised. Therefore, developing an automatic technique for removing eyeglasses in portraits proves beneficial for enhancing accuracy in these applications [1, 7, 11–13].

Within the rapidly evolving field of Generative Adversarial Networks (GANs) [14] leveraging the capabilities of advanced conditional GANs [15], numerous contemporary studies focusing on face editing [16–19] have achieved significant advancements. In the absence of paired images during training, these endeavors commonly employ cycle consistency to preserve non-edited attributes or areas [20]. Parallel to the growing trend of facial attribute manipulation, various GAN-based methods, including [7, 21, 22], have enhanced the capability to recognize faces with glasses. This improvement is achieved by training a face recognition model by synthesizing a large set of face images with glasses.

Most of these methods exclusively target eyeglasses without addressing the associated lighting effects. ByeGlassesGAN [7], on the other hand, creates paired data that incorporates various lighting effects for training. This approach utilizes parallel segmentation in eyeglasses removal, indicating the significance of mask prediction in the process. Following ByeGlassesGAN, a framework was proposed that used a "detect then remove" approach, identifying and subsequently removing not only eyeglasses but also their cast shadows from the images [11]. This framework includes a multi-step network architecture that initially detects masks for eyeglasses and their cast shadows, utilizing the estimated masks as guidance in the next steps of the eyeglasses removal process.

The methods employed in [7, 11] not only achieved superior quality results but also highlighted the crucial role of accurate eyeglass frame extraction. This accuracy is essential for effectively removing the frame and its cast shadows from facial images. Incomplete or poorly generated frames can lead to persistent artifacts on the face while also diminishing the image's overall realism and quality. Consequently, we directed our efforts towards refining the eyeglasses removal framework, explicitly addressing incompletely extracted eyeglasses masks.

This paper focuses on improving the accuracy of eyeglasses masks, especially those with incomplete extractions from the initial prediction stage. We aim to generate eyeglasses masks with fewer imperfections, leading to more effective eyeglasses and shadow removal in subsequent steps. We propose a novel approach that utilizes a training dataset designed for eyeglasses mask image completion. This dataset comprises an extensive collection of images containing eyeglasses masks in two states: damaged (incomplete) and complete (ground truth).

In contrast to traditional methods in image completion that utilize random shapes like rectangles, circles, or irregular patches, we employed the Top-Hat morphological operation when generating the damaged masks within the dataset. This approach allows us to simulate real incomplete masks more effectively, as it directly modifies the eyeglasses frame region rather than random areas of the entire mask image.

Two separate deep learning networks, U-Net and Pix2Pix, were trained on the created dataset to restore the damaged and incomplete eyeglass frames within the images.

Following a thorough evaluation, we selected the Pix2Pix network for the proposed approach due to its superior performance.

A straightforward post-processing step was implemented to eliminate the background outside the eyeglasses frame box. This step ensured that facial reconstruction was limited to areas directly associated with the eyeglasses frame and its shadows rather than encompassing background regions in facial portraits. Experiments demonstrated that the proposed method achieved better results in terms of quality and visual aspects compared to previous similar methods. This paper proposes a novel approach for eyeglasses removal that addresses the limitations of existing methods. Our key contributions are:

- **Targeted Eyeglass Mask Dataset:** We introduce a unique dataset designed for eyeglass mask image completion. This dataset is generated by applying Top-Hat morphological operations to existing eyeglasses mask datasets, creating images containing eyeglasses masks in two states: damaged (incomplete) and complete (ground truth).
- **Eyeglass Mask Completion:** We leverage a Pix2Pix image-to-image translation model trained on the newly created dataset. This model restores incomplete eyeglasses mask predictions, significantly improving the accuracy of eyeglasses frame extraction and leading to more realistic results.
- **Effective Post-processing:** Our method incorporates a post-processing step that refines the completed mask. This step ensures that only the eyeglasses frame box is targeted for removal, thereby helping prevent the formation of artifacts in the background or other parts of the image, further enhancing the overall quality of the processed image.

2 Related works

Eyeglass removal is a technique employed to mitigate the negative impact of eyeglasses on face recognition accuracy. Early studies employ statistical learning [23, 24] for eyeglasses removal, typically under the assumptions of frontal facial images and controlled environments, restricting their applicability. In general, statistical learning methods and principal component analysis (PCA) were the primary approaches before deep learning, which require less computing power but have limitations in handling eyeglasses and adapting to in-the-wild images, conditions of the environment, and head poses [25, 26].

Most subsequent studies employed deep neural networks for eyeglasses removal, achieving significant improvements in face recognition accuracy. Liang et al. [27] introduced a method using a Deep Convolutional Neural Network (DCNN) for removing eyeglasses from frontal face images. Their approach involved reconstructing the eyeglasses region, with the network trained to learn the mapping between facial images with and without eyeglasses from a significant dataset in video surveillance. Zhao et al. [28] proposed a method for eyeglasses removal that relies on attribute detection and image processing steps. This method incorporates an improved Total Variation restoration model. Their approach involves many steps, including determining eyeglasses position, identifying eyeglasses frames, extracting color information, detecting reflective areas, extracting eyeglasses templates, and removing eyeglasses.

Fueled by the growing popularity of Generative Adversarial Networks (GANs), researchers have proposed various remarkable GAN-based techniques for facial attribute editing [16, 20, 29, 30]. For instance, StarGAN [16] presented a scalable approach that enables image-to-image translations across multi-domains using only one model. It was evaluated on tasks such as face attribute transfer and face expression synthesis, demonstrating its effectiveness. Another noteworthy example is the approach in [30], which utilizes a mask network and an attribute transformation network to edit facial attributes. This method maintained the identity of the original images by employing the predicted mask to delimit the editing region.

Specially designed for face images with and without eyeglasses, ERGAN [1] learns to interchange the eye region between the two faces. It introduces an unsupervised architecture that achieved notable success in eyeglasses removal through the exchange of features extracted from both a facial appearance encoder and an eye area encoder. However, this technique specifically addresses the eyeglasses not the related lighting effects.

In ByeGlassesGAN [7], a set of synthetic image pairs (with and without glasses) is crafted to train the model. This approach, while noteworthy, has two key weaknesses: it employs a 2D method for data synthesis, and it does not consider cast shadows.

Considering that the shadows, which require removal, are caused by eyeglasses, the paper [11] introduced a mask-guided multi-step architecture to enhance the understanding of the relationship between eyeglasses and cast shadows. To achieve this, the approach leverages a synthetic dataset that incorporates 3D shadows. The methodology follows a "detect then remove" principle. The approach begins by detecting a mask for the eyeglasses, followed by another mask detection for their cast shadows. These detected masks then guide a multi-step process for removing the eyeglasses [11, 31].

One way to view eyeglasses removal is as a type of face image completion. Recent advancements in deep learning have appropriately addressed image completion, as demonstrated in various works [32–35]. However, the eyeglasses removal problem differs from image completion because the glasses region could be transparent or semi-transparent.

The primary distinction between recently developed eyeglass removal methods, such as [7, 11], and conventional image completion techniques lies in their independence from a pre-defined mask for completion. These methods can leverage the original image within the eyeglasses area, thereby better preserving the facial identity in the images after eyeglasses removal [7].

Recognizing the critical importance of accurate eyeglasses mask extraction, this paper employs image-to-image translation models to complete eyeglasses masks, deviating from traditional approaches that use image completion models for direct eyeglasses removal from facial images.

Image-to-image translation involves altering a specific aspect of an image while transforming it into another one. This field has witnessed remarkable progress since the introduction of Generative Adversarial Networks (GANs). In these models, image translation from one domain to another is achieved by leveraging training data from two distinct domains [16].

On the other hand, one of the primary challenges in deep learning is the reliance on data. Additionally, data collection is expensive and complex, especially in specific specialized fields [36, 37]. By reusing existing pretrained models, we achieve significant time and storage space efficiency during processing. Building upon the work of [11], our approach addresses eyeglasses removal by incorporating an innovative eyeglasses mask completion module. This significantly enhances eyeglasses removal performance.

3 Proposed method

Both "ByeGlassesGAN" [7], employing parallel segmentation, and "Eyeglasses and Shadow Removal" (E&S-R) [11], which initially predicts the eyeglasses mask and utilizes it in a multi-step process for eyeglasses removal, have enhanced this task and also emphasized the importance of mask prediction. We propose a method that integrates an eyeglasses mask completion block and subsequently includes a post-process for mask completion during the mask prediction step in the E&S-R architecture.

Image completion, also called inpainting, is a prominent subject in computer vision. Its objective is to fill in missing areas with possible and meaningful content by leveraging the original pixel information within the image as a reference. This process addresses missing or corrupted portions, ensuring that the restored image appears seamless to the observer, devoid of any signs of damage [38]. However, as discussed in Section 2, eyeglasses removal presents a distinct problem compared to image completion due to the potential transparency of the eyeglasses region. By using the original image within the eyeglasses area, the preservation of facial identity in the images can be enhanced [7, 11].

On the other hand, one contributing factor to the decline in face recognition accuracy when individuals wear eyeglasses is the disproportionate scarcity of face images featuring subjects with glasses compared to those without eyeglasses. Training the recognition model to learn the distinctive features of various types of eyeglasses is a significant challenge [7].

This paper applies an image completion method to eyeglasses masks instead of original images. Our approach preserves the original image within the eyeglasses area. This allows for the effective restoration of extracted eyeglass frames, ultimately leading to an enhanced removal process. In order to achieve effective restoration, our approach leverages a dataset we prepared specifically for this study. This dataset comprises pairs of damaged masks and their corresponding intact masks. Subsequently, we evaluated two distinct image-to-image translation models, namely, UNet and Pix2Pix, on this dataset. We assessed which model demonstrated a better understanding of eyeglass frame shapes and more effectively learned to complete them.

We further refined the masks predicted by the high-performing model through a post-processing step, ensuring only the essential rectangular area containing the eyeglasses was retained. Then, the final mask was utilized in the subsequent stages for eyeglasses and shadow removal. Figure 1 illustrates the proposed network architecture.

3.1 Data preparation

Training the image completion model necessitated a substantial dataset of corrupted eyeglasses mask images paired with their corresponding complete versions. Lyu et al. [11] provided a synthetic facial portrait dataset. For each sample, there are portraits with and without eyeglasses, as well as an eyeglasses mask and shadow mask specifically for those with eyeglasses. They provided a variety of portraits featuring individuals wearing eyeglasses in diverse shapes, textures, and positions. These portraits also came with corresponding eyeglasses masks for each individual. We randomly selected 4500 images of these masks and corrupted them through morphological operations.

Morphological transformations involve straightforward operations that depend on the shape of an image. Morphological operations are typically applied to binary images. The

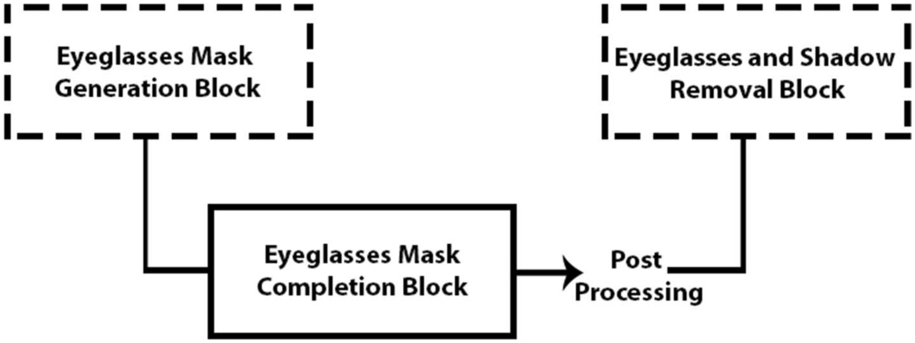
Eyeglasses Mask Generation Block
Eyeglasses and Shadow Removal Block
Eyeglasses Mask Completion Block
Post Processing

Fig. 1 Proposed Network Architecture. Dashed blocks in the upper section represent pre-trained networks from E&S-R [11]. The proposed mask completion block and post-processing stage are in the lower section

process requires two inputs: the original image and a structuring element or kernel. The latter defines the operation's characteristics, and the resulting output matches the size of the input image [39].

We used several morphological operations to alter complete forms of eyeglass frames in mask images, creating damaged ones. We then assessed which of these operations better simulated the shape of the damaged masks. For example, Erosion morphology, which thins eyeglass frames, is unsuitable for our task. It can mislead restoration models into excessively widening the frames. Ultimately, Top-Hat morphology proved to be more suitable for this purpose because the Top-Hat operation demonstrated superior performance in achieving a balance between generating realistic mask corruptions and introducing sufficient diversity for effective training of the mask completion model.

Top-Hat refers to the difference between the input image and the opening of that image. Determining the kernel size is crucial. We created a kernel in a squared shape, specifically a k×k matrix of ones. Experimenting with various values for k, we concluded that k = 5 and k = 7 would serve our purpose — creating more damage and less damage in eyeglass frames, respectively (see Fig. 2).

To train the mask completion model, we first collected 5,000 complete mask images by randomly selecting them from the E&S-R [11] eyeglasses dataset. We then employed the Top-Hat operation with varying kernel sizes: k = 5 on 2,000 images and k = 7 on 2,500 images. To ensure the model learns to identify these intact masks and avoid modifying them during completion, we intentionally left 500 images untouched. This approach

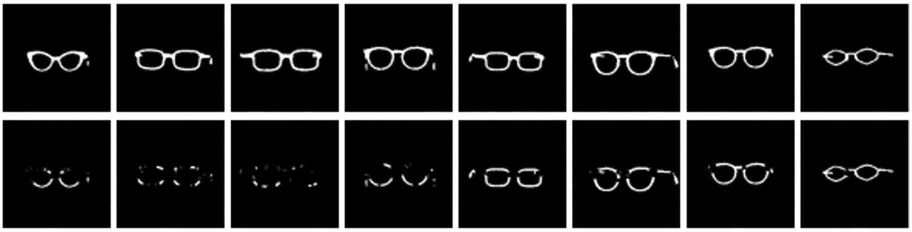

Fig. 2 Examples of eyeglasses mask images. Complete forms (top row) and damaged forms by Top-Hat operation (bottom row). Top-Hat applied with kernel size k = 5 (columns 1–4) and k = 7 (columns 5–8)

provides the model with a comprehensive dataset encompassing both intact and damaged masks with varying degrees of damage, achieved by incorporating all 5,000 images into the training process.

3.2 Image-to-image translation

The objective of image-to-image translation models is to learn a mapping between a source image and a target image. In this regard, models such as U-Net and Pix2Pix have shown exemplary performance in recording and reproducing meaningful transformations [40, 41].

U-Net is a neural network designed for image segmentation. It has two paths: one for classification (contracting path) and one for creating a segmented image (expansion path). The contracting path, or encoder, resembles a typical convolutional network. In contrast, the decoder, or expansion path, utilizes up-convolutions and concatenations with features from the contracting path. This enables the network to grasp localized classification information. The network is almost symmetrical, forming a ’U’ shape [42, 43].

Pix2Pix belongs to the class of cGANs known for their ability to generate images based on specific conditions. In this model, the output image generation is conditional upon an input, typically a source image. The discriminator compares the source image with the target image to see if the target could be a result of transforming the source. Training the generator involves adversarial loss, prompting it to produce credible images within the target domain [44]. Also, for image-to-image translation with conditional GANs like Pix2Pix, a loss function guides the training of the model to map input images to their corresponding target images [15].

Our goal was to train models that could learn to recover damaged frames of eyeglass masks. After evaluating the quality of results from both the Pix2Pix and U-Net models, we found that the Pix2Pix method outperforms U-Net, as illustrated in Fig. 3. Consequently, our work leverages the Pix2Pix model as a building block for eyeglasses mask completion in all subsequent experiments.

3.3 Our Pix2Pix network

Generator The generator transforms corrupted input mask images into corresponding completed output mask images through an encoder-decoder structure. The encoder (blocks: e1 to e8) progressively reduces the input image dimensions, extracting features at each step. Conversely, the decoder (blocks: d1 to d8) restores the encoded features to the original image size. Each block in the encoder (e) uses a convolutional layer followed by batch normalization, while each block in the decoder (d) utilizes a transposed convolutional layer with batch normalization.

Discriminator The discriminator distinguishes between real and fake mask images by evaluating two input mask images: the generated mask image and the target (real) mask image and comprising, five convolutional layers followed by leaky ReLU activations.

Implementation details This Pix2Pix model, implemented with PyTorch, utilizes the Adam optimizer with a learning rate 0.0002. We leverage the commonly accepted values for β1 (typically between 0 and 1) and β2 (often set to 0.999) to ensure stability in the optimizer. A batch size of 4 was chosen for training along with 17 epochs, which proved

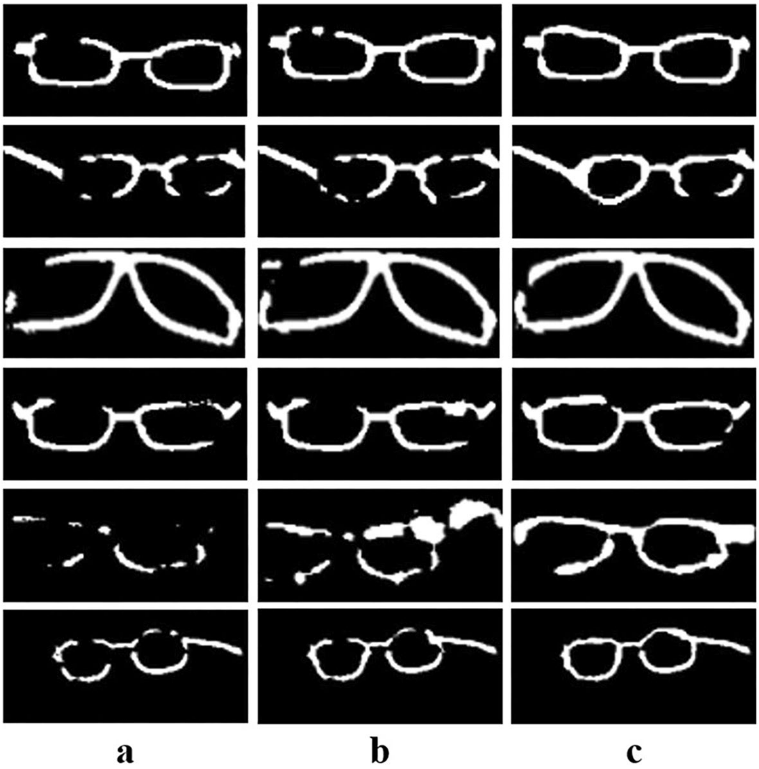

Fig. 3 Comparing models for eyeglass mask completion. (**a**) Initial eyeglasses masks. Mask completion results by (**b**) U-Net and (**c**) Pix2Pix

sufficient for achieving good results. The network operates on 256×256 input and output images. All experiments used the Google Colab environment and its free GPU resources.

3.4 Post-processing of eyeglasses masks

The predicted masks of eyeglasses may encompass not only the eyeglasses frames but also extraneous background textures or unwanted artifacts. The use of a completion block can worsen such issues. To address this, we propose a direct post-processing approach by applying functions from the OpenCV library.

Our post-processing step refines the predicted eyeglasses mask by accurately identifying the entire eyeglass frame box and removing any misclassified pixels outside it. To achieve this, we first perform a slight dilation of the white regions in the mask, effectively filling minor gaps. Next, we identify the largest contour within the dilated mask image. Based on this contour's height and coordinates, we define a rectangular region as the Region of Interest (ROI). However, to account for spectacle frames that have a significant gap between the lenses in the initial mask prediction, instead of limiting the width of this ROI to the contour itself, we consider the width of the entire image for it.

Finally, any areas outside the defined rectangular ROI are set to black. This effectively removes any remaining artifacts or misclassifications outside the eyeglasses frame, resulting in a more accurate mask. Notably, in portrait images, artifacts are less common in the frame width, minimizing the negative impact of expanding the width on removing unwanted areas. Some examples showcasing the effectiveness of this post-processing step are presented in Fig. 4.

4 Experiments

This section begins by outlining the test datasets and evaluation metrics used. We then compare the proposed method with several state-of-the-art eyeglasses removal works in terms of qualitative and quantitative measures.

4.1 Test datasets

CelebA The CelebA dataset is a large-scale face dataset comprising 202,599 images of 10,177 celebrities. Each image has five landmarks and 40 binary attributes. Leveraging the Eyeglasses attribute label, we partitioned the dataset into subsets with and without eyeglasses. Following this, we randomly selected images from the subsets, resulting in two distinct subsets: one with 5,550 images with eyeglasses and another with 50,000 images without eyeglasses. To ensure consistency in our experiments, we used the official aligned and cropped version of CelebA, resized all images to 256×256, and performed all experiments on this scale.

FFHQ The FFHQ dataset comprises 70,000 portrait images showcasing various accessories, including eyeglasses, sunglasses, hats, and more. In this study, the 128×128

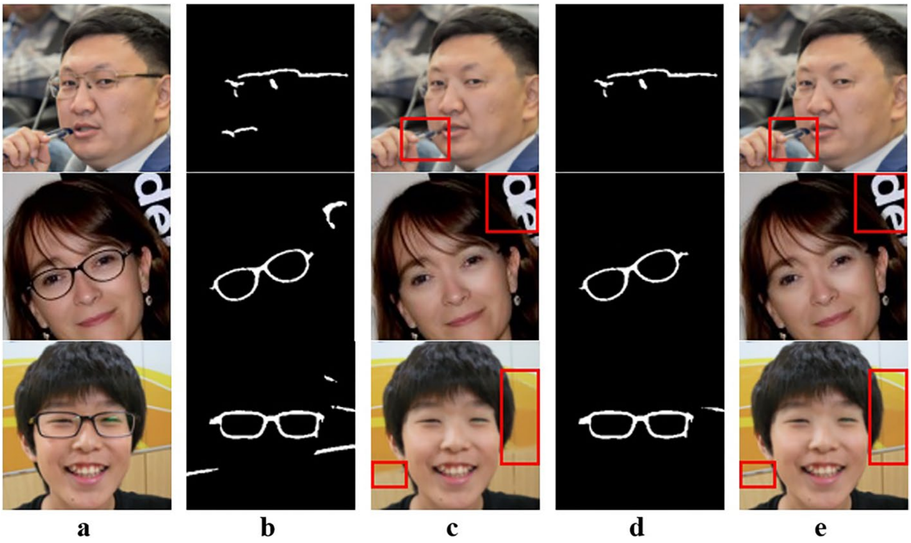

Fig. 4 Post-processing of eyeglasses mask images. (**a**) Images with eyeglasses, (**b**) Eyeglasses masks before post-processing, (**c**) Eyeglasses removed images without post- processing, (**d**) Eyeglasses masks after post-processing, and (**e**) Eyeglasses removed images after mask post- processing

thumbnail version is used. From the dataset, we manually categorized the first 15,000 images, identifying 2,396 with glasses and 12,294 without eyeglasses. All images underwent a uniform resize to 256×256.

MeGlass The MeGlass dataset initially designed to evaluate eyeglass face recognition. It contains 47,817 images from 1,710 distinct individuals. Each individual is represented by various facial images, some with and some without eyeglasses.

We employed the MeGlass testing set, which offers distinct categories for individuals with and without eyeglasses. Each category encompasses both gallery and probe images, ensuring all 1,710 identities are included. While the dataset provided a version with cropped images sized 120×120 pixels, we resized all images to a uniform dimension of 256×256 for our analysis.

4.2 Evaluation metrics

FID The *Fréchet Inception Distance* [45] computes the distance between two distributions, the distribution of real images, and the distribution of generated images concerning, the feature space of Inception embedding. FID is used to quantify the realism and similarity of generated images to real ones.

The FID considers a Gaussian distribution for the hidden activations of each distribution and subsequently calculates the Fréchet distance between those Gaussians. FID has gained popularity due to its straightforward computation and effectiveness. However, it has two limitations: it may lack robustness in the face of minor variations in evaluation methods, and it can also be subject to bias [45, 46]. Therefore, we have also used the KID metric, which is similar to FID but is more robust and lacks the bias issue.

KID The *Kernel Inception Distance* [46] is the squared Maximum Mean Discrepancy (MMD) between Inception representations. Unlike the FID, which assumes a parametric form, KID uses a polynomial kernel to distribute activations. KID estimates are unbiased, and when using the cubic kernel, this metric compares skewness in addition to the mean and variance.

TAR@FAR *True Accept Rate at False Accept Rate;* TAR is the probability of correctly accepting an identified person, while FAR is the probability of mistakenly accepting a person who does not share the same identity. TAR@FAR measures the model's accuracy under a particular level of false acceptance.

MOS The *Mean Opinion Score* is a metric used in user studies to compare different method results quantitatively. We randomly selected six images with eyeglasses from our three testing datasets. After applying various eyeglasses removal methods to these images, we asked 30 individuals to evaluate and assign a score between 1 (worst) and 5 (best) to the generated images of each method.

4.3 Comparison methods

Our comparison includes several eyeglasses removal methods: E&S-R [11], HiSD [47], SAGAN [30], and ERGAN [22]. We evaluate all methods on the same datasets. To facilitate comparisons, we directly employed their pre-trained models. HiSD is trained on the CelebA-HQ dataset [48], while the others utilize the CelebA dataset. E&S-R is additionally trained on its synthetic dataset.

4.3.1 Qualitative evaluation

A qualitative comparison of our removing eyeglasses method with three prior works is performed on different portrait images sourced from the CelebA, FFHQ, and MeGlass datasets. As shown in Fig. 5, our approach achieves higher quality when compared with earlier methods. SAGAN struggles to effectively remove eyeglasses in most of the test images. HiSD generally performs well in removing eyeglasses. However, it encounters difficulties in removing specific eyeglass shapes (second row) or eyeglasses in specific head positions (third row), sometimes leaving traces of the frame removal on the face (fourth to sixth rows). Additionally, in some images, it also manipulates areas outside the eyeglass regions,

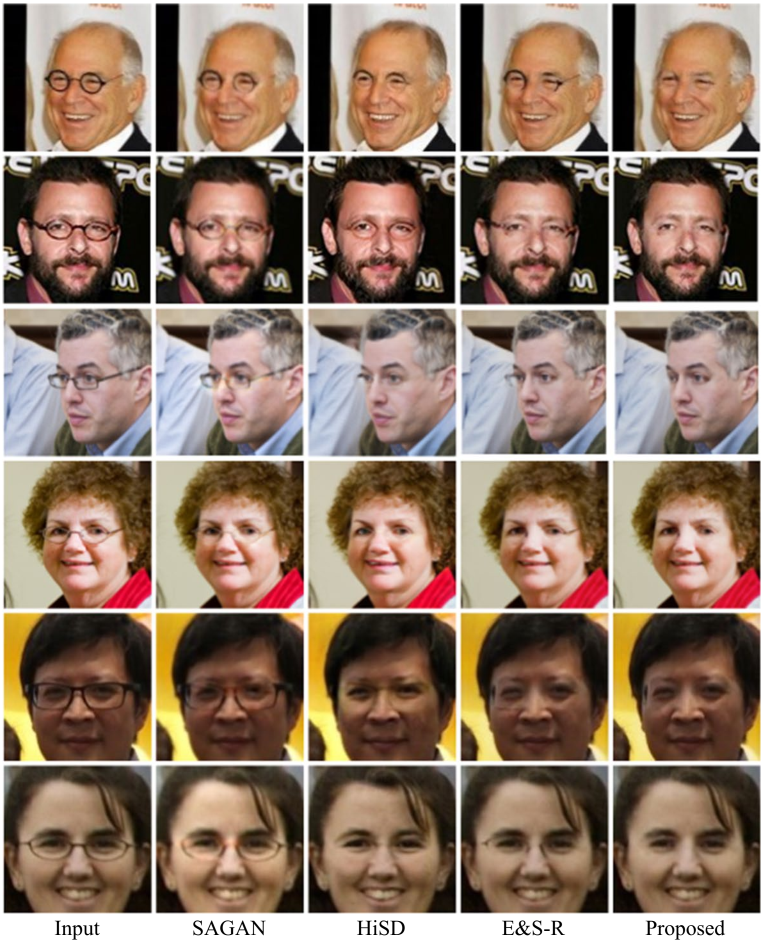

Fig. 5 Qualitative results. Eyeglasses Removal Results were obtained using various methods on different datasets. from top to bottom: CelebA dataset, FFHQ dataset, MeGlass dataset

such as hair (first row) and eyebrows (fourth and sixth rows), which can affect the realism of the generated images.

E&S-R also achieves good results in eyeglass removal. It prioritizes the eyeglass area to avoid unintended modifications in other facial regions. However, in some cases, particularly in lower-quality images, this method may leave behind residual eyeglass frame parts due to its inability to completely extract the eyeglass frame mask. Our approach has significantly addressed this issue by adding an eyeglass frame restoration step. Thus, besides maintaining the realism of facial images, our method generates higher-quality images after eyeglass removal.

4.3.2 Quantitative evaluation

To assess the realism of our eyeglasses-removed images, we employed the FID and KID metrics on the FFHQ and CelebA test datasets. It's important to consider that calculating FID and KID involves many steps, potentially introducing inconsistencies in the final metric. Various implementations utilize different low-level image quantization and resizing functions. However, some implementations exhibit errors in the way they perform resizing. To address this issue, we used a standardized library, clean-fid [49], to ensure that FID and KID scores remain comparable across different methods.

The process involved comparing images after removing eyeglasses from them with images initially without eyeglasses (free-eyeglasses images). The results (Table 1, FID and KID columns) demonstrate that our method achieves the minimum FID and KID scores on both the CelebA and FFHQ datasets. This suggests that the generated images by the proposed method have a distribution closer to that of real portrait images without eyeglasses than other methods.

However, realism could be a subjective assessment that FID and KID may not fully represent. To further compare the visual quality of eyeglasses removal methods, we employed a user study to collect mean participant opinions scores (MOS), as mentioned earlier. As evidenced by the highest MOS score in Table 1 (last column), our method surpasses existing approaches. This achievement suggests a greater fidelity in the eyeglasses-removed images produced by our method compared to others.

To assess identity preservation, we calculated TAR@FAR on the MeGlass test dataset, as shown in Table 2. We applied eyeglasses removal methods to each probe image within the "with glasses" category. We compared it with its corresponding identity in the gallery images within the "without glasses" category (rows 2 to 5). Additionally, we conducted the

Table 1 Quantitative results for realism. Comparison of FID and KID scores (lower is better) for different methods applied to the CelebA and FFHQ datasets. Also, an MOS score (higher is better) is provided for the same methods applied to the CelebA, FFHQ, and MeGlass datasets, evaluating the realism of generated eyeglasses-removed images

Methods	FID ↓		KID ↓ × 10		MOS ↑
	CelebA	FFHQ	CelebA	FFHQ	
E&S-R [11]	36.682	35.620	0.42	0.26	4.1
HiSD [7]	35.223	79.701	0.41	0.80	3.6
SAGAN [30]	80.803	43.047	0.85	0.32	2.4
Proposed	**34.462**	**35.189**	**0.39**	**0.25**	**4.7**

Table 2 Quantitative results for face identification. Comparing Tar@Far (higher is better) calculated for different methods applied to the MeGlass datasets, evaluating the effectiveness of eyeglasses removal methods in identity preservation

	Tar@Far = 0.1 ↑	Tar@Far = 0.01 ↑	Tar@Far = 0.001 ↑
With-glasses	0.7877	0.4450	0.1690
E&S-R [11]	**0.8109**	**0.4802**	**0.1979**
HiSD [7]	0.7940	0.4488	0.1877
SAGAN [30]	0.7979	0.4359	0.1720
ERGAN [22]	0.7653	0.4004	0.1374
Proposed	**0.8162**	**0.4855**	**0.2058**
No-glasses	0.8748	0.6672	0.4081

same comparison without applying eyeglasses removal to the "with glasses" images (row 1). Furthermore, we explored comparisons between each identity in the probe images and its corresponding identity in the gallery images within the same "without glasses" category (last row).

The images without glasses (row 1) achieve the highest authentication success since they represent real images providing complete identity details. However, the authentication success decreases when comparing images with and without glasses (last row), indicating the adverse impact of eyeglasses on face authentication.

ERGAN and SAGAN introduce further degradation in face recognition performance after eyeglasses removal. In contrast, HiSD leads to a slight improvement in authentication success. E&S-R and our method are highly competitive, exhibiting the most improvement both in removing eyeglasses and preserving identity.

As shown in Table 2, the results of the E&S-R method and our proposed method are very close, but our method is slightly better. To examine the extent of this difference within the images, we conducted another experiment, focusing solely on cosine similarity between the generated images of eyeglass removal and the original images (without glasses) for both the E&S-R method and the proposed method. Subsequently, we calculated the difference in cosine similarity between the methods, considering a threshold due to the proximity of the results of both methods in finding similarities (We only considered cases where the difference in cosine similarities exceeded 0.1). The analysis identified 37 images where the proposed method outperformed the E&S-R method. Figure 6 illustrates some of these comparisons.

5 Conclusion

This paper introduces a novel mask completion technique explicitly tailored to enhance eyeglasses removal. We leveraged the Top-Hat morphological transform with varying kernel sizes (e.g., 5 and 7) to generate "destructed" versions of the full-frame eyeglasses masks. These paired images, containing both intact and manipulated masks, served as training data for the Pix2Pix network. Using this dataset facilitated the network to accurately understand the shape of the full-frame eyeglasses and significantly improved its capability to complete the distorted eyeglasses frames. Incorporating this mask completion

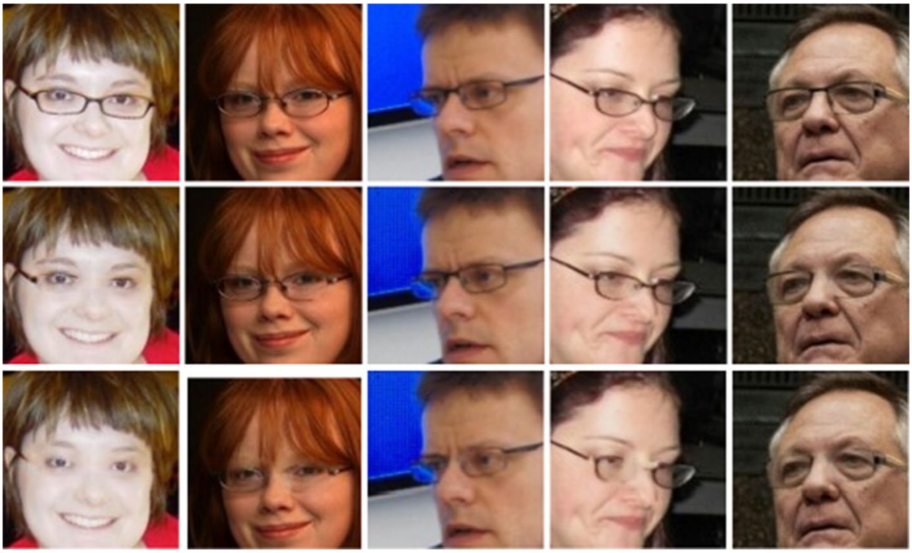

Fig. 6 Comparison of eyeglass removal methods: Original Images (1st Row), E&S-R Method (2nd Row), Proposed Method (3rd Row). Images from the MeGlass dataset

step, following the initial eyeglasses mask extraction, significantly impacts subsequent stages like eyeglasses and shadow removal.

Furthermore, our post-processing stages refine the repaired masks and ensure that facial changes outside the eyeglasses box and alterations in background textures are minimized. This approach achieves a twofold benefit: it enhances the quality and realism of eyeglasses-removed images, while demonstrably preserving facial identity integrity as evidenced by the success rate of identity verification. By addressing the challenge of accurately extracting eyeglasses masks, our method significantly advances the field of eyeglasses removal techniques, with implications for various face-related applications, including verification, identification, and reconstruction. Future research may explore further refinements and extensions of this approach to tackle other aspects of facial attribute manipulation and enhancement. Additionally, exploring parsing facial features, particularly eyebrows, could lead to preserving better areas that intersect with the eyeglasses frame.

6 Ablation study

This section evaluates the effectiveness of individual components within our proposed method through ablation studies. Qualitative comparisons with and without the eyeglasses mask completion module are illustrated in Fig. 7. The second column of Fig. 7 displays examples of initial predicted eyeglasses masks that are incomplete. Simulating realistic eyeglasses mask degradation significantly impacts the learning of the mask completion model.

To train a model capable of accurately reconstructing the shape of eyeglass frames, an operation was required that could generate masks resembling the broken or incomplete

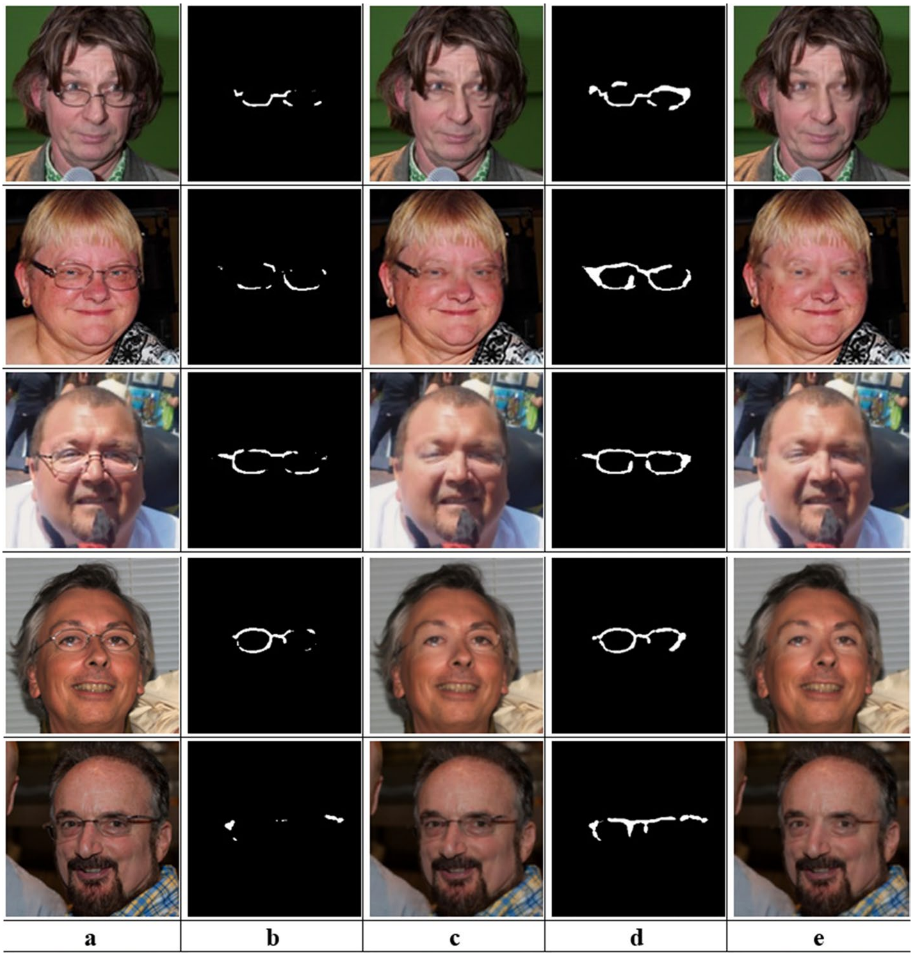

Fig. 7 Mask completion impact on eyeglass removal. (**a**) Images with eyeglasses or inputs, (**b**) Predicted mask by E&S-R Method, (**c**) Generated images of eyeglass removal by E&S-R, (**d**) Mask completion by proposed method and (**e**) Generated images of eyeglass removal by the proposed method

forms of eyeglasses from intact masks. Figure 8 showcases the effect of using three types of morphological transformations: Top-Hat, Black-Hat, and Erosion (odd kernels 1–9).

Black-Hat and Top-Hat at low kernels (k = 1, 3) severely damage masks. While high kernel Top-Hat (k = 9 +) preserves the overall structure, it may not introduce enough incompleteness. Erosion, in contrast, excessively degrades masks at high kernels but offers minimal destruction at low kernels, only slightly thinning the shape of the glasses. Notably, kernels 5 and 7 in Top-Hat achieve a good balance, mimicking real-world imperfections. Moreover, these kernels perform better than others in Erosion as well. Therefore, only kernels 5 and 7 were used to generate incomplete masks.

To demonstrate the impact of the morphological transformation used to create the dataset for training the mask completion model, three separate datasets were created with

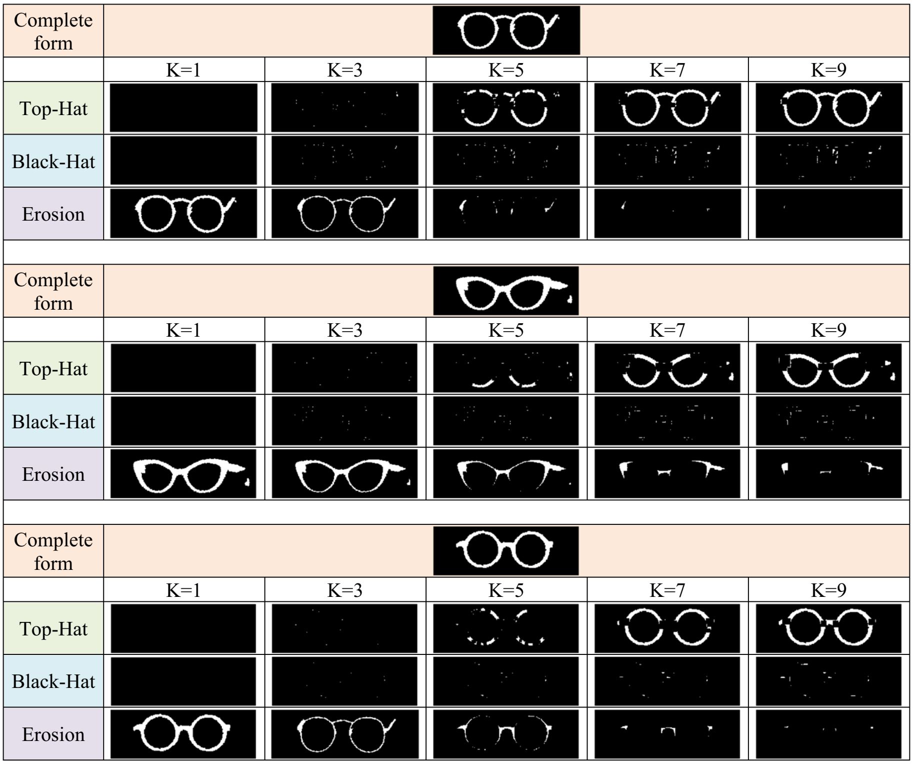

Fig. 8 Effect of different morphological transformations: Top-Hat, Black-Hat, and Erosion (odd kernels 1–9) on the complete form of eyeglasses masks

Top-Hat, Black-Hat, and Erosion morphologies (k = 5, 7). The Pix2Pix model was trained using pairs of corresponding degraded and complete mask images. The final results of the eyeglasses removal images after completing the masks individually with a trained mask completion model are shown in Fig. 9.

In general, the mask completion model trained on datasets generated with Black-Hat and Erosion morphologies improved the output results in some cases. However, the model trained on data generated with Erosion tends to thicken the frames in some examples unnecessarily, and the model trained on data generated with Black-Hat does not preserve the continuity of the added pixels in the frame structure. Conversely, the best performance was achieved by the mask completion model trained on the dataset generated by Top-Hat due because this operator provides a better simulation of real incomplete masks and helps the mask completion model to understand the shape of the eyeglass frames better.

Finally, Table 3 presents the FID and KID scores for the CelebA and FFHQ datasets with and without the mask completion step and with and without the post-processing step to elucidate the role of each step in enhancing image quality. The results demonstrate that the addition of the mask completion model alone plays a significant role in improving the quality and realism of eyeglasses removal images.

Fig. 9 Impact of different mask completion models on eyeglasses removal. (**a**) Original image with glasses, (**b**) Initial predicted mask, (**c**) Removal without mask completion, (**d**) Mask completed with Pix2Pix trained on Top-Hat corrupted dataset, (**e**) Removal after using the mask from *"d"*, (**f**) Mask completed with Pix2Pix trained on Erosion corrupted dataset, (**g**) Removal after using the mask from *"f"*, (**h**) Mask completed with Pix2Pix trained on Black-Hat corrupted dataset and (**i**) Removal after using the mask from *"h"*

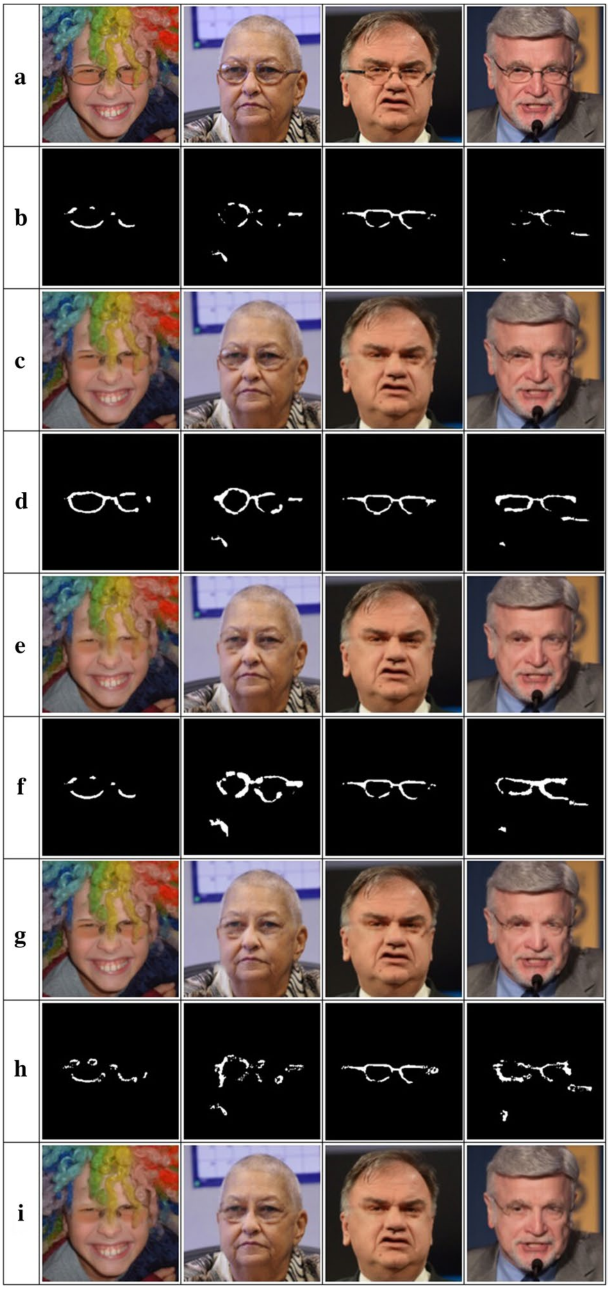

Table 3 Clarify impact of each stage in proposed method by comparison of FID and KID scores

Stages of eyeglasses removal	FID ↓		KID ↓ × 10	
	CelebA	FFHQ	CelebA	FFHQ
(no mask completion and no post-processing) [11]	36.682	35.620	0.42	0.26
just post-processing	36.836	35.595	0.43	0.25
just mask completion	34.483	35.282	0.40	0.25
mask completion + post-processing	34.462	35.189	0.39	0.25

Interestingly, applying post-processing to the FFHQ dataset without mask completion still yields positive outcomes, while it has a detrimental effect on the CelebA dataset. This discrepancy can be attributed to the increased difficulty in accurately detecting the rectangular region of the eyeglass frame in the absence of mask completion, potentially leading to the erroneous removal of areas belonging to the eyeglass frame that are not seamlessly connected to the surrounding regions. Nonetheless, incorporating post-processing into the CelebA and FFHQ datasets after mask completion consistently leads to improved results.

Funding This research is not supported by any funding schemes or governments.

Data availability The code, links for paired eyeglasses masks, and pretrained models are available at: https://github.com/z-esmaily/Eyeglasses_Removal/tree/main.

Declarations

Conflicts of interest The authors have declared that there is no conflict of interest exists.

References

1. Hu B, Zheng Z, Liu P, Yang W, Ren M (2020) Unsupervised eyeglasses removal in the wild. IEEE Transact Cybern 51(9):4373–4385
2. Guo J, Zhu X, Zhao C, Cao D, Lei Z, Li SZ (2020) Learning meta face recognition in unseen domains. Proc IEEE/CVF Conf Comput Vis Patt Recognit. pp. 6163–6172
3. Cao D, Zhu X, Huang X, Guo J, Lei Z (2020) Domain balancing: Face recognition on long-tailed domains. Proc IEEE/CVF Conf Comput Vis Patt Recognit. pp. 5671–5679
4. Gaston J, Ming J, Crookes D (2018) Matching larger image areas for unconstrained face identification. IEEE Transact Cybernet 49(8):3191–3202
5. Sun Y, Xu Q, Li Y, Zhang C, Li Y, Wang S et al (2019) Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification. Proc IEEE/CVF Conf Comput Vis Patt Recognit. pp. 393–402
6. Wang Y, Tang YY, Li L, Chen H (2019) Modal regression-based atomic representation for robust face recognition and reconstruction. IEEE transactions on cybernetics 50(10):4393–4405
7. Lee YH, Lai SH (2020) Byeglassesgan: Identity preserving eyeglasses removal for face images. Comput Vis–ECCV 2020: 16th Eur Conf Glasgow. Springer International Publishing. pp. 243–258
8. Yang H, Ciftci U, Yin L (2018) Facial expression recognition by de-expression residue learning. Proc IEEE Conf Comput Vis Patt Recognit. pp. 2168–2177
9. DyapadyAnnappa RRB (2023) A comprehensive review of facial expression recognition techniques. Multimedia Syst 29(1):73–103
10. Rangesh A, Zhang B, Trivedi MM (2020) Driver gaze estimation in the real world: Overcoming the eyeglass challenge. 2020 IEEE Intell Veh Symp (IV). pp. 1054–1059

11. Lyu J, Wang Z, Xu F (2022) Portrait eyeglasses and shadow removal by leveraging 3d synthetic data. Proc IEEE/CVF Conf Comput Vis Patt Recognit. pp. 3429–3439
12. Guo J, Zhu X, Lei Z, Li SZ (2018) Face synthesis for eyeglass-robust face recognition. Biom Recognit: 13th Chinese Conf CCBR 2018, Urumqi, China. Springer International Publishing. pp. 275–284
13. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. Proc IEEE Int Conf Comput Vis. pp. 3730–3738
14. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S et al (2014) Generative adversarial nets. Adv Neur Inf Proc Syst. 27
15. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. Proc IEEE Conf Comput Vis Patt Recognit. pp. 1125–1134
16. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. Proc IEEE Conf Comput Vis Patt Recognit. pp. 8789–8797
17. Chu W, Tai Y, Wang C, Li J, Huang F, Ji R (2020) Sscgan: Facial attribute editing via style skip connections. Comput Vis–ECCV 2020: 16th Eur Conf Glasgow, UK. Springer International Publishing. pp. 414–429
18. Liu M, Ding Y, Xia M, Liu X, Ding E, Zuo W et al (2019) Stgan: A unified selective transfer network for arbitrary image attribute editing. Proc IEEE/CVF Conf Comput Vis Patt Recognit. pp. 3673–3682
19. Wu PW, Lin YJ, Chang CH, Chang EY, Liao SW (2019) Relgan: Multi-domain image-to-image translation via relative attributes. Proc IEEE/CVF Int Conf Comput Vis. pp. 5914–5922
20. Gao Y, Wei F, Bao J, Gu S, Chen D, Wen F et al (2021) High-fidelity and arbitrary face editing. Proc IEEE/CVF Conf Comput Vis Patt Recognit. pp. 16115–16124
21. Guo J, Zhu X, Lei Z, Li SZ (2018) Face synthesis for eyeglass-robust face recognition. Chinese Conf Biomet Recognit. Springer International Publishing. pp. 275–284
22. Cheng M, Cao X (2021) ERGAN: High Perform GAN for Eyeglasses Removal. 16th Int Conf Int Syst Knowl Eng (ISKE). IEEE. pp. 406–411
23. Wong WK, Zhao H (2013) Eyeglasses removal of thermal image based on visible information. Inf Fus 14(2):163–176
24. Jin JS, Xu C, Xu M, Zhang Z, Peng Y (2013) Eyeglasses removal from facial image based on mvlr. The Era of Interactive Media. Springer, New York, pp 101–109
25. Rangesh A, Zhang B, Trivedi MM (2020) Driver gaze estimation in the real world: Overcoming the eyeglass challenge." 2020 IEEE Int Veh Symp (IV). pp. 1054–1059
26. Kang S, Hahn T (2021) Eyeglass Remover Network based on a Synthetic Image Dataset. KSII Transact Int Inf Syst. 15(4)
27. Liang M, Xue Y, Xue K, Yang A (2017) Deep convolution neural networks for automatic eyeglasses removal. DEStech Transact Comput Sci Eng
28. Zhao M, Zhang Z, Zhang X, Zhang L, Li B (2021) Eyeglasses removal based on attributes detection and improved TV restoration model. Multimed Tools Appl 80:2691–2712
29. Liu Y, Li Q, Deng Q, Sun Z, Yang MH (2023) Gan-based facial attribute manipulation. IEEE Transact Patt Anal Mach Intell
30. Zhang G, Kan M, Shan S, Chen X (2018) Generative adversarial network with spatial attention for face attribute editing. Proc Eur Conf Comput Vis (ECCV). pp. 417–432
31. Laishram L, Shaheryar M, Lee JT, Jung SK (2023) High-Quality Face Caricature via Style Translation. IEEE Access
32. Jo Y, Park J (2019) Sc-fegan: Face editing generative adversarial network with user's sketch and color. Proc IEEE/CVF Int Conf Comput Vis. pp. 1745–1753
33. Liu G, Reda FA, Shih KJ, Wang TC, Tao A, Catanzaro B (2018) Image inpainting for irregular holes using partial convolutions. Proc Eur Conf Comput Vis (ECCV). pp. 85–100
34. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2019) Free-form image inpainting with gated convolution. Proc IEEE/CVF Int Conf Comput Vis. pp. 4471–4480
35. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention. Proc IEEE Conf Comput Vis Patt Recognit. pp. 5505–5514
36. Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. Artif Neur Netw Mach Learn–ICANN 2018: 27th Int Conf Artif Neur Netw Rhodes, Greece. Springer International Publishing. pp. 270–279
37. Esmaeily Z, Rezaeian M (2023) Building roof wireframe extraction from aerial images using a three-stream deep neural network. J Electron Imaging 32(1):013001–013001
38. Shao C, Li X, Li F, Zhou Y (2022) Large Mask Image Completion with Conditional GAN. Symmetry 14(10):2148

39. Sreedhar K, Panlal B (2012) Enhancement of images using morphological transformation. arXiv preprint
40. Zhu JY, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. Proc IEEE Int Conf Comput Vis. pp. 2223–2232
41. Tanjim MM (2023) Debiasing Image Generative Models. University of California, San Diego
42. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation. Med Image Comput Comput-Ass Intervent–MICCAI 2015: 18th Int Conf Munich, Germany. Springer International Publishing. pp. 234–241
43. Siddique N, Paheding S, Elkin CP, Devabhaktuni V (2021) U-net and its variants for medical image segmentation: A review of theory and applications. IEEE Access 9:82031–82057
44. Henry J, Natalie T, Madsen D (2021) Pix2Pix GAN for Image-to-Image Translation**.** Res Gate Publication. pp. 1–5
45. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv Neur Inf Proc Syst. 30
46. Bińkowski M, Sutherland DJ, Arbel M, Gretton A (2018) Demystifying mmd gans. arXiv preprint
47. Li X, Zhang S, Hu J, Cao L, Hong X, Mao X et al (2021) Image-to-image translation via hierarchical style disentanglement. Proc IEEE/CVF Conf Comput Vis Patt Recognit. pp. 8639–8648
48. Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive growing of GANs for improved quality, stability, and variation. Int Conf Learn Represent
49. Parmar G, Zhang R, Zhu JY (2022) On aliased resizing and surprising subtleties in gan evaluation. Proc IEEE/CVF Conf Comput Vis Patt Recognit. pp. 11410–11420

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