Original Paper

Artificial Intelligence–Enabled Facial Privacy Protection for Ocular Diagnosis: Development and Validation Study

Haizhu Tan[1], PhD; Hongyu Chen[2,3*], BD; Zhenmao Wang[4*], MD; Mingguang He[5*], PhD; Chiyu Wei[1], MD; Lei Sun[6], PhD; Xueqin Wang[7], PhD; Danli Shi[5], PhD; Chengcheng Huang[1], MD; Anping Guo[8], MD

[1]Department of Preventive Medicine, Shantou University Medical College, Shantou, China

[2]Department of Optoelectronic Information Science and Engineering, Physical and Materials Science College, Guangzhou University, Guangzhou, China

[3]Han's Laser Technology Industry Group Co., Ltd, Shenzhen, China

[4]Joint Shantou International Eye Center of Shantou University and The Chinese University of Hong Kong, Shantou, China (Hong Kong)

[5]The Hong Kong Polytechnic University, Hong Kong, China (Hong Kong)

[6]Department of Ophthalmology, the Fourth Affiliated Hospital of Harbin Medical University, Haerbin, China

[7]University of Science and Technology of China, Hefei, China

[8]Department of Pharmacy, First Affiliated Hospital of University of Science and Technology of China, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China

*these authors contributed equally

Corresponding Author:

Haizhu Tan, PhD
Department of Preventive Medicine
Shantou University Medical College
22 Xinling Rd
Shantou, 515031
China
Phone: 86 13318055534
Email: linnanqia@126.com

Abstract

Background: Facial biometric data, while valuable for clinical applications, poses substantial privacy and security risks.

Objective: This paper aims to address the privacy and security concerns related to facial biometric data and support auxiliary diagnoses, in pursuit of which we developed Digital FaceDefender, an artificial intelligence–driven privacy safeguard solution.

Methods: We constructed a diverse set of digitally synthesized Asian face avatars representing both sexes, spanning ages 5 to 85 years in 10-year increments, using 70,000 facial images and 13,061 Asian face images. Landmark data were separately extracted from both patient and avatar images. Affine transformations ensured spatial alignment, followed by color correction and Gaussian blur to enhance fusion quality. For auxiliary diagnosis, we established 95% CIs for pixel distances within the eye region on a cohort of 1163 individuals, serving as diagnostic benchmarks. Reidentification risk was assessed using the ArcFace model, applied to 2500 images reconstructed via Detailed Expression Capture and Animation (DECA). Finally, Cohen Kappa analyses (n=114) was applied to assess agreement between diagnostic benchmarks and ophthalmologists' evaluations.

Results: Compared to the DM method, Digital FaceDefender significantly reduced facial similarity scores (FDface vs raw images: 0.31; FLAME_FDface vs raw images: 0.09) and achieved zero Rank-1 accuracy in Pose #2-#3 and Pose #2-#5, with near-zero accuracy in Pose #4 (0.02) and Pose #5 (0.04), confirming lower reidentification risk. Cohen Kappa analysis demonstrated moderate agreement between our benchmarks and ophthalmologists' assessments for the left eye ($\varkappa$=0.37) and right eye ($\varkappa$=0.45; both P<.001), validating diagnostic reliability of the benchmarks. Furthermore, the user-friendly Digital FaceDefender platform has been established and is readily accessible for use.

Conclusions: In summary, Digital FaceDefender effectively balances privacy protection and diagnostic use.

J Med Internet Res 2025;27:e66873; doi: 10.2196/66873

Keywords: facial biometric data; Digital FaceDefender; privacy protection; auxiliary diagnosis; artificial intelligence; ocular disease

Introduction

With the rapid advancement of artificial intelligence (AI) applications in medical imaging, a vast array of newly generated medical images now encompass a wide range of personal information, including nonbiometric, physiological, behavioral biometric, and soft biometric identifiers [1]. These images undergo digitization, storage, transmission, and retrieval by healthcare organizations for various purposes [2,3]. However, this digitalization process has brought about significant concerns regarding security and multifaceted privacy, spanning these identifiers and more. In response to these concerns, the primary objective of our study is to develop a model that effectively extracts critical features for auxiliary diagnosis in eye hospitals while simultaneously ensuring patient privacy protection.

In their seminal work published in Nature Medicine, Yang et al [4] introduced the concept of a "Digital Mask" (DM) designed to protect patient privacy while preserving disease-relevant features critical for diagnosis. While this innovation represents a significant step forward in privacy preservation, Meeus et al [5] subsequently raised concerns about the reidentification risks associated with the DM. Specifically, their study extracted frames from facial videos provided by Yang et al [4], with one frame serving as a reference image and another used to compute the mask. Facial regions, excluding the eyes, were masked while maintaining facial contours, thus creating the "DMface." The Faces Learned with an Articulated Model and Expressions (FLAME) model was then applied to generate a novel facial mask, termed "FLAME_DMface," by integrating the Skinned Multi-Person Linear Model body model [5-7]. Meeus et al [5] evaluated the reidentification risks by comparing FLAME_DMface with the corresponding reference images using the Additive Angular Margin Loss-based ArcFace model. Their findings revealed potential reidentification vulnerabilities [5,8].

Yang et al [9] responded by emphasizing the secure maintenance of the original clinical examination videos, arguing that reidentification attacks using the FLAME model would be rendered irrelevant if the original videos were kept inaccessible. While we acknowledge the efforts of Yang et al [9] to address patient privacy concerns and mitigate identification risks, we also concur with the feedings by Meeus et al [5] that the reidentification risks associated with DM persist following FLAME processing. The retention of facial contours, including key regions such as the face, nose, eyes, and mouth, underscores the necessity for refining deidentification techniques to further reduce the risk of unauthorized reidentification.

In clinical environments, the balance between maintaining patient privacy and enabling auxiliary diagnosis is of paramount importance, particularly in specialized fields like ocular disease. In such cases, partial exposure of the eyes and periocular region may be necessary, but DM's focus on extracting disease-relevant features from a limited eye region—comprising only the upper and lower eyelids and the iris—may be insufficient for comprehensive disease characterization, as shown in Multimedia Appendix 1.

Furthermore, previous research, such as that by Neumann et al [10], has highlighted the negative impact of extensive facial coverings in medical settings. Such coverings may hinder clinician-patient communication, diminish empathy [11], and reduce diagnostic accuracy [12] by obscuring essential facial cues [13] related to age, sex, and expression. These challenges became particularly evident during the COVID-19 pandemic and among Muslim women wearing veils [14-16]. As illustrated in Multimedia Appendix 1, DM may exacerbate these issues due to its extensive coverage, which is visually comparable to a white plaster cast.

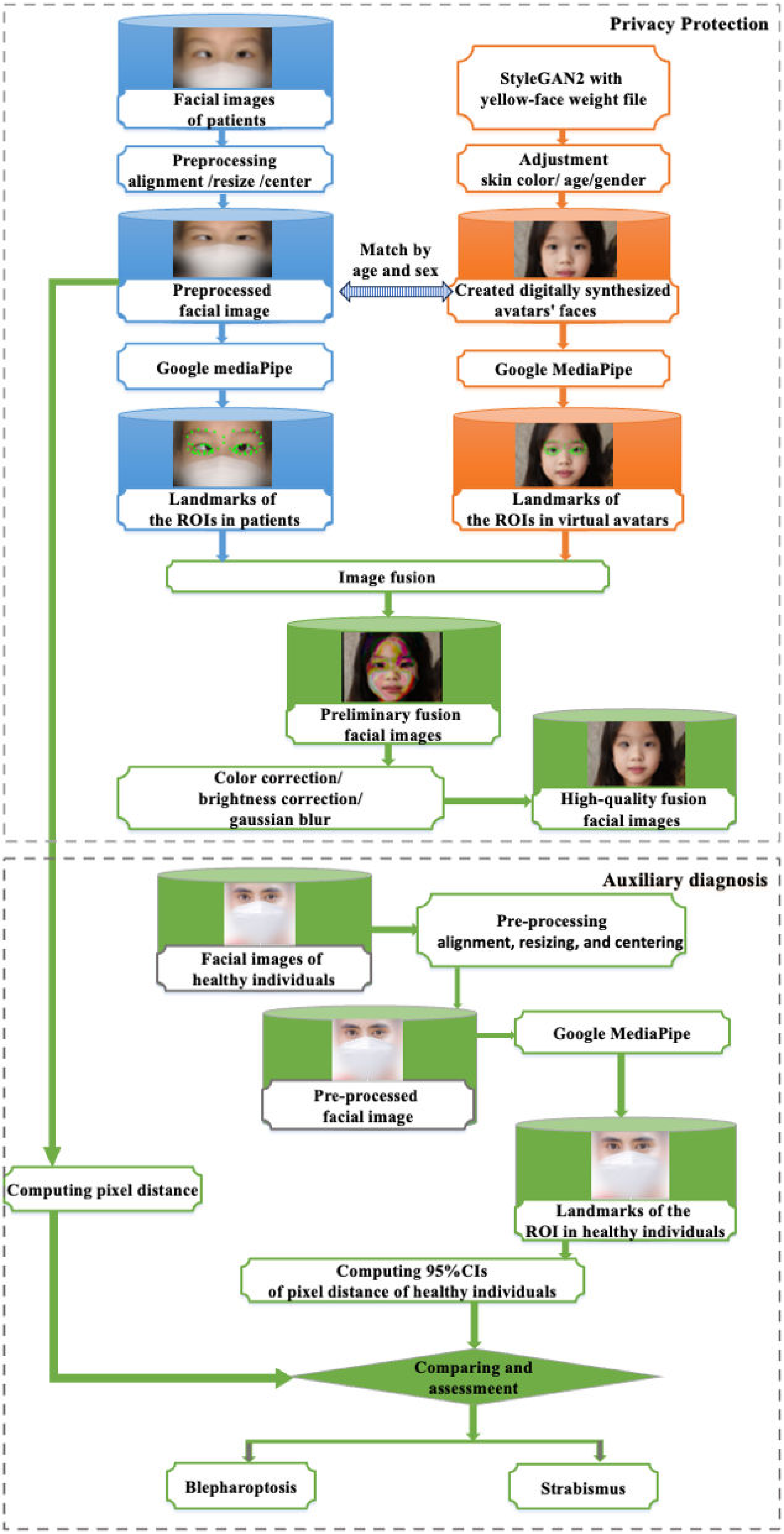
In our study, we first developed an AI-driven method—Digital FaceDefender—and assessed the reidentification risks and the mean similarity between images generated using our proposed method and the DM approach. In addition, we developed and validated an auxiliary diagnostic benchmark. We anticipate that this innovative approach will enhance the diagnostic capabilities of medical professionals, providing essential support for early-stage patient evaluation while ensuring robust patient privacy protection.

Methods

Overview

In this study, we introduced Digital FaceDefender, a method designed to enhance privacy protection while supporting auxiliary diagnosis of ocular diseases in eye hospitals. Figure 1 shows the detailed workflow, with the region of interest (ROI) clearly marked.

Figure 1. The workflow of Digital FaceDefender. ROI: region of interest.


Privacy Protection
Facial images of patients
Preprocessing alignment /resize /center
Preprocessed facial image
Google mediaPipe
Landmarks of the ROIs in patients
StyleGAN2 with yellow-face weight file
Adjustment skin color/ age/gender
Match by age and sex
Created digitally synthesized avatars' faces
Google MediaPipe
Landmarks of the ROIs in virtual avatars
Image fusion
Preliminary fusion facial images
Color correction/ brightness correction/ gaussian blur
High-quality fusion facial images
Auxiliary diagnosis
Facial images of healthy individuals
Pre-processing alignment, resizing, and centering
Pre-processed facial image
Google MediaPipe
Landmarks of the ROI in healthy individuals
Computing pixel distance
Computing 95%CIs of pixel distance of healthy individuals
Comparing and assessment
Blepharoptosis
Strabismus

Ethical Considerations

The research protocol was reviewed and approved by the Institutional Review Board/Ethics Committee of the Fourth Affiliated Hospital of Harbin Medical University (approval 2023-Ethics Review-54; see Multimedia Appendix 2). All participants provided informed consent in accordance with the principles outlined in the Declaration of Helsinki.

To protect privacy, all face images presented in the paper were generated using our image fusion method or masked rather than displaying original face images from our private database.

Dataset

In this retrospective study, we collected face images from 3 sources: 2 publicly available datasets (Flickr-Faces-HQ [FFHQ] and CASIA-FaceV5) and 1 private dataset from the Fourth Affiliated Hospital of Harbin Medical University. The images from the latter dataset were randomly selected.

Based on their extensive clinical diagnostic experience, two board-certified ophthalmologists (one Chief Physician and one Associate Chief Physician), both licensed by the Ministry of Health of the People's Republic of China, independently evaluated a random sample of facial images from 1136 healthy individuals in our private database. These ophthalmologists, affiliated with the Joint Shantou International Eye Center and the Department of Ophthalmology at the Fourth Affiliated Hospital of Harbin Medical University, assessed the images using a standardized diagnostic protocol. Only individuals exhibiting clear ocular abnormalities, including esotropia, exotropia, vertical strabismus, and ptosis, were further assessed and included in this study. To ensure diagnostic consistency, the evaluations were performed independently, and inter-rater agreement was quantified using the Cohen Kappa coefficient.

Image Preprocessing

Several preprocessing techniques were used to ensure the consistency and quality of patient images. Initially, facial alignment was performed by rotating the original images based on the line connecting the irises, ensuring the face was centered and horizontally oriented. All images were resized to 1024×1024 RGB (red, green, and blue) color space, the standard required for generating digitally synthesized avatars. If the original image size was smaller than this resolution, a white background was added; if larger, the image was cropped to maintain central facial positioning. Given that patient images are captured under variable conditions, including differences in lighting and background noise, a standard RGB color correction method [17,18] was used to adjust brightness, ensuring consistency with the digitally synthesized avatars' faces.

Creation of Digitally Synthesized Avatars

The FFHQ dataset, consisting of 70,000 high-quality PNG images with a resolution of 1024×1024 pixels, was used to the style-based generative adversarial network architecture (StyleGAN2) [19] model, which is known for generating realistic digitally synthesized avatars. However, this dataset is biased towards Caucasian faces (69% Caucasian, 4% Black, and 27% other) raising concerns regarding generalizability to more diverse populations [20,21], raising concerns about generalizability to diverse populations.

To address this issue, we used the generator_yellow-stylegan2-config-f.pkl file, which was trained on 13,016 Asian facial images from the SeePrettyFace website [22]. We generated male and female digitally synthesized avatars with varying age progressions, ranging from 5 to 85 years old in increments of 10 years, to simulate realistic age-related facial changes.

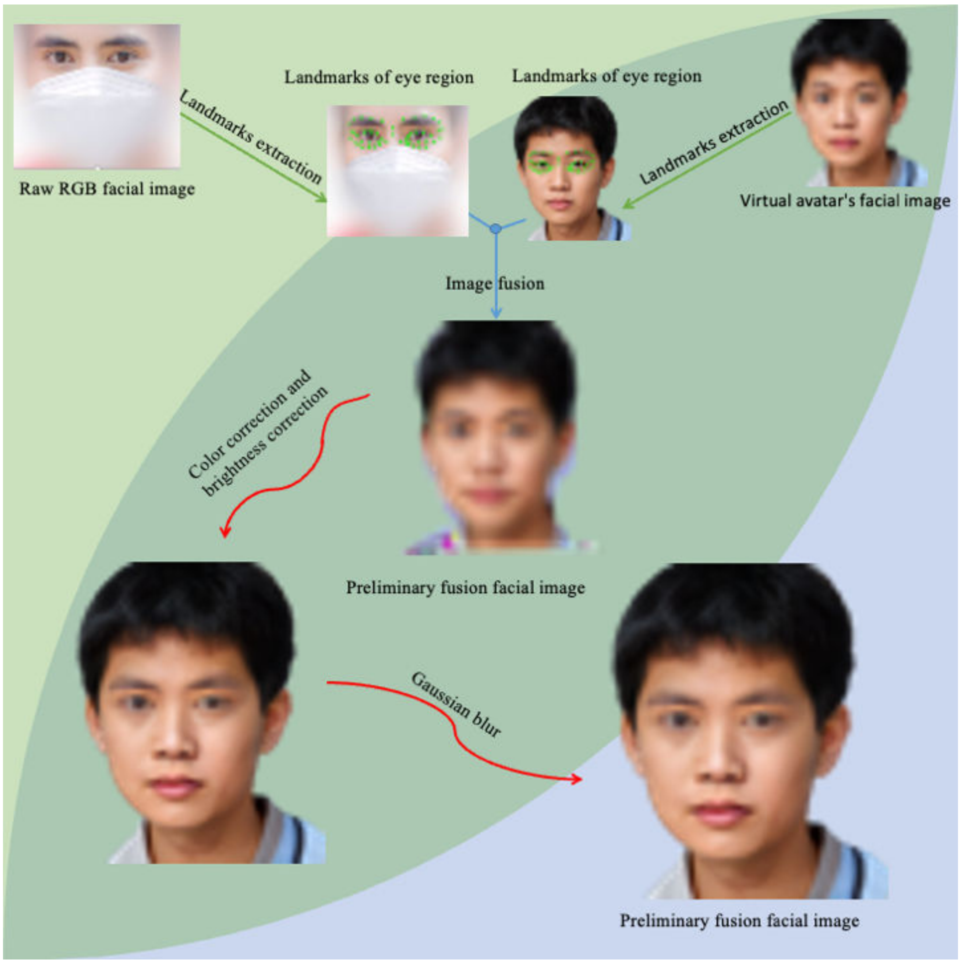
Generation and Comparison of FDface and DMface Fusion Images

Accurate generation of lesion positions and morphologies is critical for medical image generation [23]. While generative adversarial networks (GANs) excel at capturing global features, they often struggle with detailed pathological structures, as suggested by Kazeminia et al [24], and Han et al [25]. To mitigate this, we used Google's MediaPipe library [26] for facial landmark detection. The Face Mesh detector identifies 3D coordinates for 468 facial landmarks (see Multimedia Appendix 3).

Affine transformations preserve collinearity and parallelism [26], making them suitable for adjusting discrepancies in the eye regions between patient and avatar images. We applied the Open Computer Vision Library (OpenCV) warpAffine() function [27] to adjust the eye regions while preserving the original proportions and positions. These transformed images, referred to as preliminary fusion images (see Figure 1), were then used for the fusion process.

During image fusion, disparities in skin tone, brightness, and unnatural fusion boundaries can arise, potentially affecting diagnostic accuracy. To rectify this, we applied RGB color correction [17,18] to harmonize patient images with avatar facial features. In addition, a Gaussian blur [28-30] was applied to the fused image edges to smooth boundary transitions [31] and create a more natural appearance. Figure 2 shows the detailed process by which the final high-quality fusion images (referred to as FDface) are generated through the fusion of raw RGB facial images and digitally synthesized avatar facial images. The workflow begins with the extraction of landmarks from both the raw images and the digitally synthesized avatars, followed by the image fusion step. This is followed by color correction and background adjustment to ensure consistency and improve visual appeal. Finally, Gaussian filtering is applied to smooth the image, further enhancing its quality and resulting in the final high-quality fused facial images. Furthermore, we generated both DMface and FDface images by applying the DM technique and Digital FaceDefender to the same original images for comparative analysis.

Figure 2. Workflow for the image fusion process. RGB: red, green, and blue.


Raw RGB facial image
Landmarks extraction
Landmarks of eye region
Landmarks of eye region
Landmarks extraction
Virtual avatar's facial image
Image fusion
Color correction and brightness correction
Preliminary fusion facial image
Gaussian blur
Preliminary fusion facial image

Reidentification Risk Assessment

To assess the effectiveness of Digital FaceDefender in reducing reidentification risks while preserving essential facial features necessary for accurate diagnoses, we conducted a comparative analysis against DM using the Detailed Expression Capture and Animation (DECA) [8] framework. Specifically, we used the FLAME model to reconstruct facial images, generating DMface and FDface representations from the CASIA-FaceV5 dataset, which comprises 2500 facial images from 500 individuals, each captured in five different poses. These reconstructed images were labeled as FLAME_DMface and FLAME_FDface, respectively.

To evaluate the reduction in reidentification risk associated with Digital FaceDefender, we used ArcFace (IR-SE50 model) to compute facial similarity scores and Rank-1 accuracy. Specifically, we separately quantified the mean similarity scores between FDface and FLAME_FDface images and their corresponding original facial images. These similarity scores indicate the extent to which the generated images resemble the original faces. In addition, we used Rank-1 accuracy to assess whether FDface and FLAME_FDface images could be reidentified as their original counterparts using ArcFace.

Statistical Analysis for Diagnostic Benchmark Establishment

To determine the sample size required for constructing the diagnostic benchmark, we first conducted a pilot study with $n_{pilot} = 50$ to estimate the SD (σ_{pilot}) of the difference in pixel distances along the x-axis between the left and right eyes. The margin of error was then calculated by using the formula $\left(E = Z_{\frac{\alpha}{2}} \times \frac{\sigma_{pilot}}{\sqrt{n_{pilot}}} = 1.96 \times \frac{3.31}{\sqrt{50}} = 0.92\right.$, where $Z_{\frac{\alpha}{2}} = 1.96$ correspond to a 95% confidence level). Based on these estimates, the final required sample size (n) was determined using the standard sample size formula:

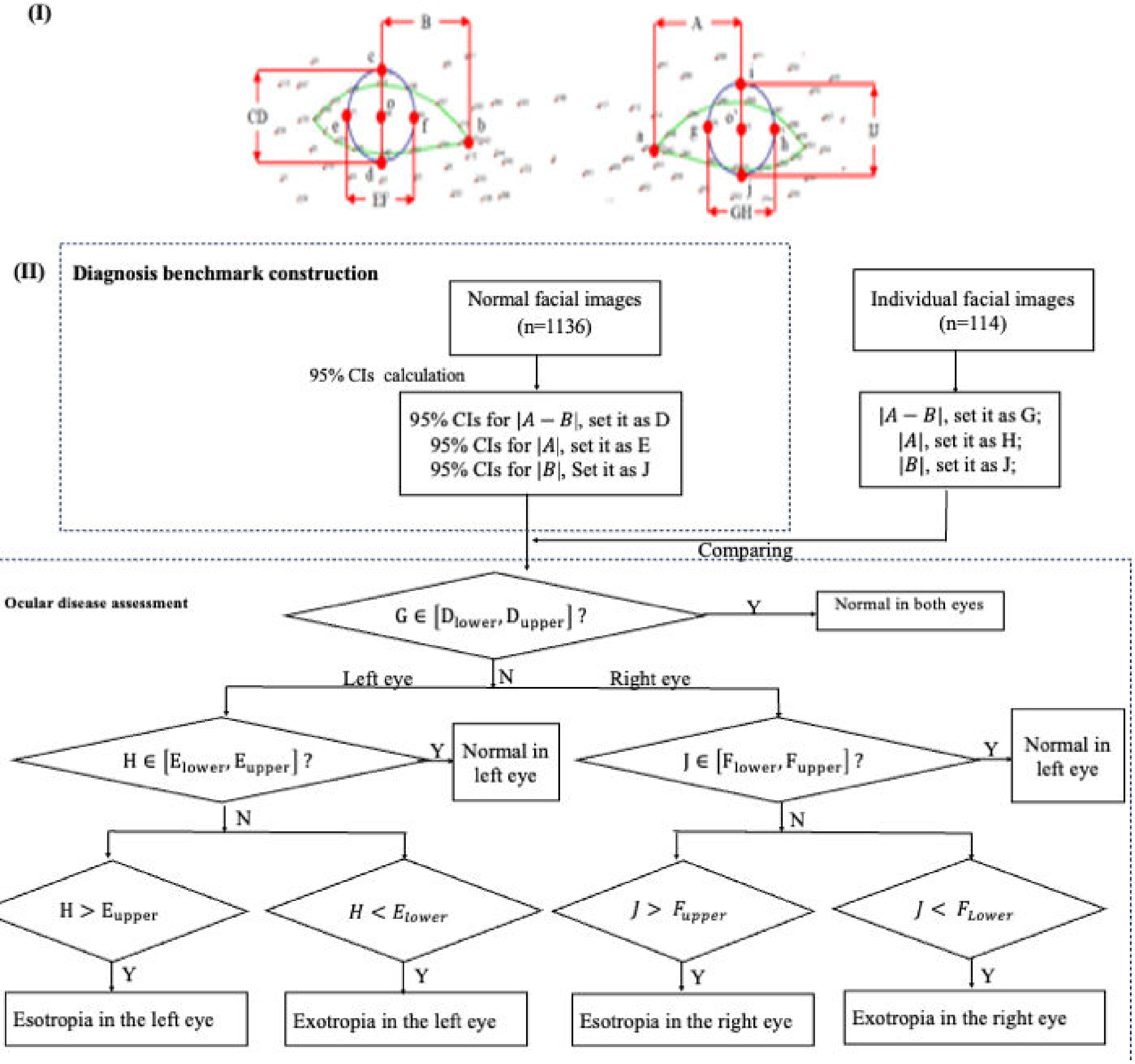
$$n = \left(\frac{Z_{\frac{\alpha}{2}} \cdot \sigma_{\text{pilot}}}{E}\right)^2 = \left(\frac{1.96 \cdot 3.63}{0.92}\right)^2 = 59.81$$

Finally, a total of 1136 healthy individuals were included in our study, exceeding the estimated sample size (n=60) and ensuring sufficient statistical power for establishing the diagnostic benchmark.

Pixel distances (|A|, |B|, |CD|, |EF|, |GH|, and |IJ|) along the x- and y-axes for both eyes were calculated between the iris center (o and o') and reference points (a, b, c, d, e, f, g, h, i, j) in 1136 healthy individuals (see Figures 3II). The difference in pixel distances along the x-axis between the left and right eyes was denoted as |A-B|. Outlier detection was performed using the IQR method, excluding data points outside 1.5 times the IQR from the lower and upper quartiles. After outlier removal, 95% CIs for the pixel distances within the eye region were calculated. For variables that followed an approximately normal distribution (assessed by visual inspection and normality tests), we used the normal distribution method to calculate CIs. For variables that deviated from normality, we applied the bootstrap method to obtain more robust CI estimates. These diagnostic benchmarks were validated on an independent dataset of 114 individuals not included in the original sample. Inter-rater consistency between the diagnostic benchmarks and ophthalmologists' assessments was evaluated using Cohen Kappa. The significance level was set to .05. Statistical analysis was performed using RStudio (version 1.1.463; Posit), which is developed and maintained by Posit.

The established diagnostic benchmarks were then used to identify ocular abnormalities, such as esotropia and exotropia (see Multimedia Appendix 4).

Figure 3. Measurement and comparison of pixel distances along the x-axis between the left and right eyes.


(I)
B
A
CD
c
o
e
f
b
d
EF
i
o'
a
g
h
U
j
GH
(II)
Diagnosis benchmark construction
Normal facial images (n=1136)
95% CIs calculation
95% CIs for |A − B|, set it as D
95% CIs for |A|, set it as E
95% CIs for |B|, Set it as J
Individual facial images (n=114)
|A − B|, set it as G;
|A|, set it as H;
|B|, set it as J;
Comparing
Ocular disease assessment
G ∈ [Dlower, Dupper] ?
Y
Normal in both eyes
N
Left eye
Right eye
H ∈ [Elower, Eupper] ?
Y
Normal in left eye
J ∈ [Flower, Fupper] ?
Y
Normal in left eye
N
N
H > Eupper
H < Elower
J > Fupper
J < FLower
Y
Y
Y
Y
Esotropia in the left eye
Exotropia in the left eye
Esotropia in the right eye
Exotropia in the right eye

Digital FaceDefender Platform

We developed the Digital FaceDefender platform to support ophthalmologists and researchers in their work, offering a user-friendly interface for privacy protection and auxiliary diagnostic assistance.

Results

Image Processing

Figure S1 in Multimedia Appendix 5 illustrates certain challenges due to low resolution (640×480), nonideal shooting angles, and intricate backgrounds, such as a blue curtain. These factors could hinder accurate representation of critical eye regions. As a representative example, we enhanced Figure S1A to a high-resolution version (2436×1944) shown in Figure S1B, both in Multimedia Appendix 5. Additional corrections included aligning the pupils horizontally in Figure S1C in Multimedia Appendix 5 and refining the cropping to emphasize the central facial region more clearly in Figure S1D in Multimedia Appendix 5.

Creation of Digitally Synthesized Avatars

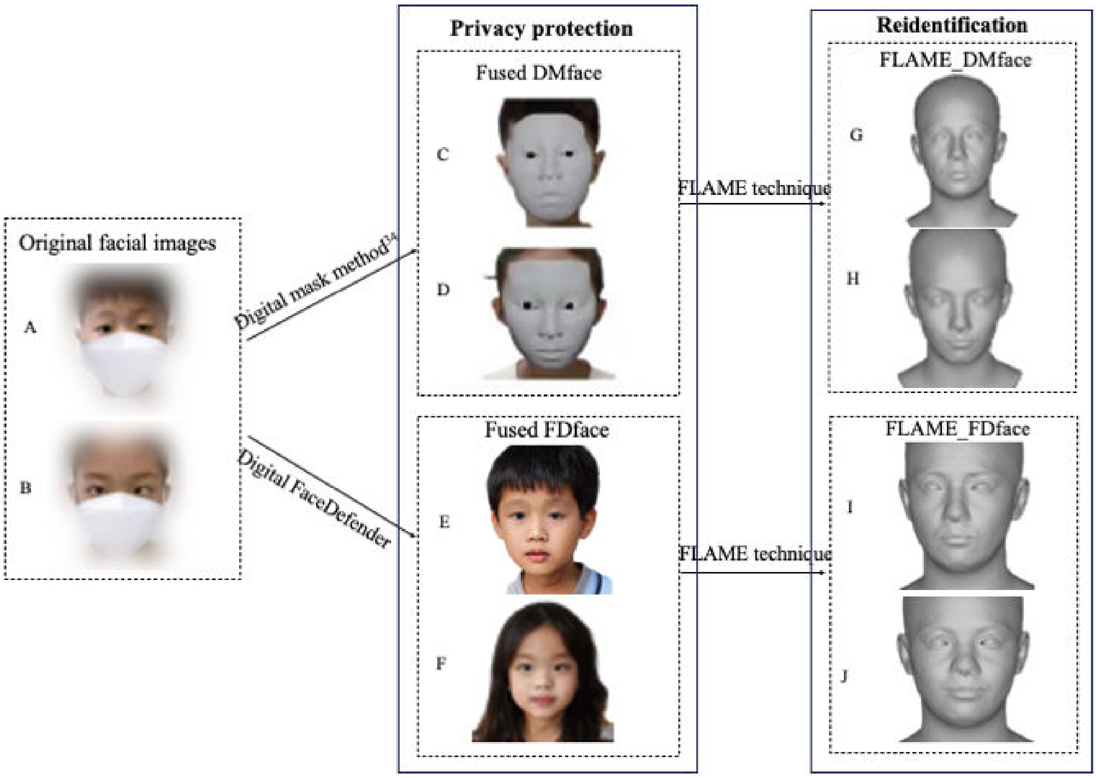
As depicted in Multimedia Appendix 6, we generated 2 sets of digitally synthesized avatars representing Asian males and females aged from 5 to 85 years at 10-year intervals. These avatars are used as templates in subsequent image fusion processes to ensure the preservation of critical facial features essential for medical diagnoses. Furthermore, a set of digitally synthesized avatars representing Caucasian individuals was also created.

Generation and Comparison of FDface and DMface Fusion Images

Using Google's MediaPipe library, we focus on 52 key landmarks within the eye and periocular regions, including the eyelids, eye sockets, irises, and brow arches, for both patients and avatars. These landmarks are used to define closed ROIs for subsequent image fusion. OpenCV is then applied to adjust the eye regions, resulting in transformed images, referred to as preliminary fusion images (see Figure 1), which are subsequently used in the fusion process.

Figure 4 presents the final fused images generated by both the DM method and our proposed Digital FaceDefender approach. The DMface images produced using the DM method are shown in Figure 4C and D, while the FDface images created by Digital FaceDefender are depicted in Figure 4E and F. The difference in fused images generated by DM and Digital FaceDefender highlights that Digital FaceDefender not only meets the privacy protection requirements but also preserves essential features related to ocular diseases. The resulting natural-looking fused images enhance clinician-patient empathy, which is crucial for accurate diagnosis and personalized treatment. Furthermore, the final fused images generated using the various ethnic digitally synthesized avatars, as presented in Multimedia Appendix 7, demonstrate the generalizability of the proposed Digital FaceDefender methodology.

Figure 4. Comparison of privacy protection and reidentification risk: DM versus Digital FaceDefender (the 2 facial images are sourced from Yang et al [4]). DM: Digital Mask; FLAME: Faces Learned with an Articulated Model and Expressions.



Reidentification Risk Assessment

Figure 4 illustrates the difference in reidentification risk between the DM method and the Digital FaceDefender approach. When applying the FLAME reconstruction technique to both DMface and FDface images, the FLAME_DMface images (Figure 4G and H) exhibit a higher reidentification risk compared to FLAME_FDface images (Figure 4I and J). This increased risk stems from the preservation of key facial contours, including the face shape, nose, eyes, and mouth, which enables effective reconstruction through FLAME in the DM method.

As shown in Table 1 and Multimedia Appendices 8 and 9, the mean similarity score between FDface and raw facial images is 0.31. While this value suggests that FDface retains some identifiable facial features, potentially beneficial for auxiliary diagnosis, it remains below the reidentification threshold established by leading facial recognition systems. For example, Microsoft's Face API reports that similarity scores above 0.5 indicate a potential match [32]. The significantly lower mean similarity score of 0.09 for FLAME_FDface versus raw facial images suggests that reconstructing the original face from FDface images is highly challenging, confirming that Digital FaceDefender significantly reduces reidentification risk while enhancing privacy protection.

Table 1. Evaluating the mean similarity between FDface and FLAME_FDface images and raw facial images.

Comparison	Mean similarity
FDface versus raw facial images	0.31
FLAME_FDface versus raw facial images	0.09

Further validation comes from Rank-1 accuracy analysis across different facial poses (see Tables 2 and 3). Notably, for Pose #2 and Pose #3 (see Table 2) and Poses #2-#5 (see Table 3), the Rank-1 accuracy is 0, indicating that no successful reidentifications occurred in these conditions. Even for Pose #4 (0.02) and Pose #5 (0.04), the Rank-1 accuracy remains near zero, confirming that Digital FaceDefender effectively mitigates reidentification risks across various facial angles.

These results underscore the robustness of our proposed Digital FaceDefender in balancing privacy protection with retention of diagnostic features.

Table 2. Evaluating the Rank-1 accuracy between FDface images and raw facial images with different poses.

Pose	Rank-1 accuracy
Pose #2	0
Pose #3	0
Pose #4	0.02
Pose #5	0.04

Table 3. Evaluating the Rank-1 accuracy between FLAME_FDface images and raw facial images with different poses.

Pose #	Rank-1 accuracy
Pose #2	0
Pose #3	0
Pose #4	0
Pose #5	0

Diagnostic Benchmark Establishment

Using the developed diagnostic benchmarks, we evaluated their agreement with ophthalmologists' assessments for ocular disease detection (see Figure 5). As shown in Table 4, Cohen Kappa analysis demonstrated a fair agreement for the left eye ($\varkappa$=0.37) and a moderate agreement for the right eye ($\varkappa$=0.445). Both agreements were statistically significant (*P*<.001), indicating the reliability of the benchmarks. However, factors such as head posture and image quality may influence diagnostic accuracy.

Figure 5. Clinical diagnoses of ocular diseases based on fused images.

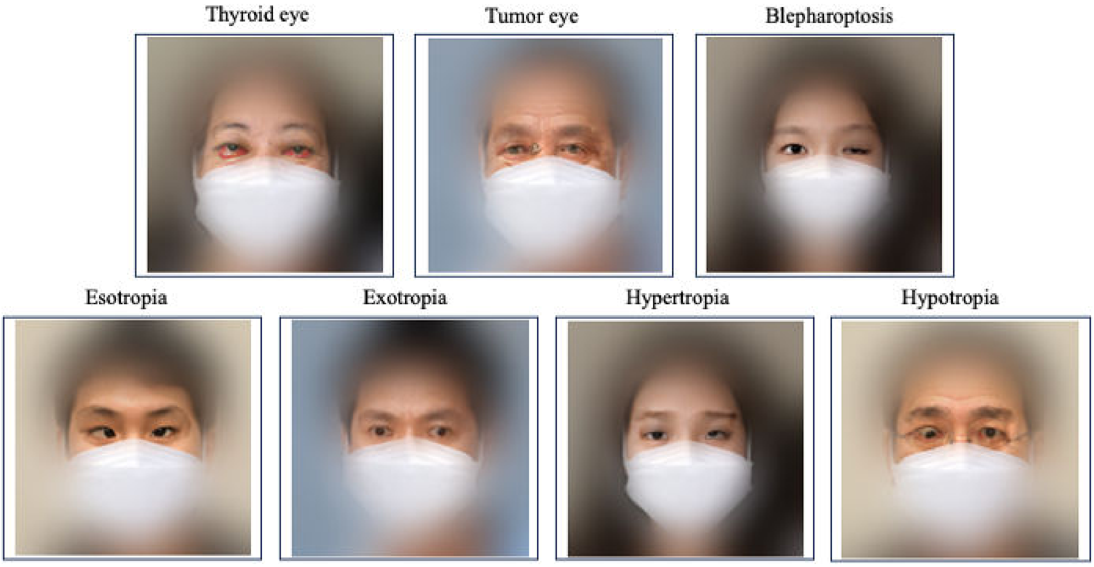


Table 4. Cohen Kappa analysis in classifying right and left eye strabismus.

	Kappa	*P* value
Left eye	0.37	<.001
Right eye	0.45	<.001

Digital FaceDefender Platform

The Digital FaceDefender platform is available as a website [33]. Multimedia Appendix 10, along with the accompanying Multimedia Appendix 11, demonstrates the effectiveness of the Digital FaceDefender Platform in protecting patients' facial privacy while simultaneously supporting the diagnosis of ocular diseases. In addition, the Digital FaceDefender platform utilizes the developed diagnostic benchmarks to identify raw images with suboptimal camera angles, providing text prompts to flag them as poor-quality images.

Discussion

This study introduces Digital FaceDefender to safeguard patients' privacy and achieve auxiliary diagnosis. Collectively, the comparison of DMface images, FDface images, FLAME_DMface images, and FLAME_FDface images, along with the evaluation metrics including Rank-1 accuracy, mean similarity, and Cohen Kappa, demonstrates that Digital FaceDefender effectively facilitates auxiliary diagnosis while concurrently safeguarding patient privacy.

Quantitative analysis, including mean similarity scores and Rank-1 accuracy, demonstrates that Digital FaceDefender substantially reduces reidentification risks while effectively preserving critical diagnostic features. In contrast, existing anonymization methods exhibit certain limitations. For example, Shawn Shan et al [34] introduces subtle perturbations to facial images to prevent unauthorized facial recognition. However, these perturbations can be removed using super-resolution models, adversarial training, or denoising techniques, thereby restoring identifiable facial features and significantly weakening privacy protection. Similarly, FaceObfuscato [35] disrupts reidentification attacks, making the method resistant to adversarial optimization techniques. However, it remains vulnerable to non–gradient-based and adaptive attacks that leverage auxiliary information or brute-force reconstruction. In addition, its transformations may distort critical eye-region details, potentially degrading performance in ocular auxiliary diagnosis. Face Deidentification [36] uses GAN-based face synthesis to generate realistic, high-resolution anonymized faces. While this method enhances image fidelity, it can be susceptible to GAN inversion attacks. In contrast, Digital FaceDefender achieves lower reidentification and similarity scores, demonstrating superior robustness in privacy protection. Unlike diffusion-based anonymization methods, such as the approach proposed by Kung et al [37], which suffer from uncontrollable anonymization levels due to the nonlinear latent space, Digital FaceDefender enables precise control over the strength of anonymization, ensuring that key ocular features remain intact for auxiliary diagnosis. Furthermore, diffusion-based anonymization can degrade expression recognition accuracy, whereas Digital FaceDefender generates visually natural and diagnostically meaningful fused images. Compared to DeepPrivacy [38],which uses a StyleGAN-like architecture to synthesize high-resolution, photorealistic faces. However, it remains vulnerable to adversarial attacks. Digital FaceDefender preserves clinically relevant facial features while minimizing reidentification risks.

However, there are several limitations to this study. First, while Google's MediaPipe can detect 468 facial landmarks, this number may be insufficient to accurately represent the ROI, particularly since only 16 landmarks are assigned to the eyelids and 5 to the iris for each eye. Second, MediaPipe faces challenges in accurately localizing the ROI, especially when image quality is compromised. Third, aligning pixel distances in the ROI on the digitally synthesized avatar's face with the actual ROI size on the patient's face remains problematic. A possible solution could involve using a horizontal ruler during photography as a reference for scaling. Fourth, the quality of image fusion is influenced by factors such as camera equipment, shooting angles, and environmental conditions, highlighting the need for standardized imaging protocols. Fifth, enhancing the accuracy of fusion boundaries will require larger and more diverse training datasets, which remain scarce, particularly for rare medical conditions. Finally, we explore some novel advanced techniques, like Hugging Face using Stable Diffusion (V3.5), developed by Hugging Face [35], to generate digitally synthesized avatars from various ethnic backgrounds. However, the results produced by Stable Diffusion did not surpass those generated by StyleGAN2 in terms of fusion quality (see Multimedia Appendix 7). The reason for this could be the significantly higher resolution of virtual avatars produced by Stable Diffusion compared to the raw facial images, which may have introduced challenges in achieving seamless fusion effects (see Multimedia Appendix 7).

In conclusion, unlike other privacy protection technologies, Digital FaceDefender demonstrates dual efficacy: it preserves privacy while facilitating auxiliary diagnoses. Although our preliminary findings are promising, further refinement of the image fusion process is essential to enhance the realism and accuracy of the resultant images. In addition, improving the automatic matching between the virtual avatar's face and the patient's face has the potential to reduce the workload of health care professionals. Future research should focus on these areas to advance the utility and applicability of Digital FaceDefender in clinical practice.

Acknowledgments

We appreciate Lifeng Weng for his efforts in masking some raw facial images to ensure privacy protection.

Our research was partially supported by National Natural Science Foundation of China (grants 12231017 and 72171216), and Innovative development funds of Anhui Province Federation of Social Sciences (2022CX 081).

Data Availability

The data supporting this study's findings are divided into two groups: publicly available data and restricted data. The publicly available data were obtained from open-source datasets. The restricted data, which involve identifiable information from individuals, are subject to licensing agreements that allow usage solely for analytical purposes and cannot be shared publicly. All code used in this study is available at GitHub [39].

Users interested in exploring the Digital FaceDefender platform [33] can request trial access by contacting the administrators via email at linnanqia@126.com.

Authors' Contributions

HT contributed to writing – review & editing, writing – original draft, visualization, supervision, resources, project administration, methodology, investigation, formal analysis, data curation, and conceptualization. HC contributed to writing – original draft, visualization, validation, software, methodology, investigation, and formal analysis. ZW contributed to writing – original draft, validation, software, methodology, investigation, data curation, and conceptualization. MH contributed to writing – review & editing, validation, and conceptualization. CW contributed to writing – original draft, software, and methodology. LS contributed to writing – original draft, validation, and data curation. XW contributed to writing – review & editing and funding acquisition. DS contributed to writing – review & editing and conceptualization. CH contributed to writing – original draft and methodology. AG contributed to writing – original draft and funding acquisition.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Images from "A digital mask to safeguard patient privacy."

[PNG File (Portable Network Graphics File), 151 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Ethical approval.

[PDF File (Adobe File), 484 KB-Multimedia Appendix 2]

Multimedia Appendix 3

The landmarks of face and periocular region.

[DOCX File (Microsoft Word File), 535 KB-Multimedia Appendix 3]

Multimedia Appendix 4

The detection of blepharoptosis in the right eye and left eye.

[DOCX File (Microsoft Word File), 438 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Image preprocessing for images with low-quality.
[DOCX File (Microsoft Word File), 1356 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Digitally synthesized avatars of Asian males and females representing various age groups.
[DOCX File (Microsoft Word File), 4815 KB-Multimedia Appendix 6]

Multimedia Appendix 7

The final fused images with various ethinic digital avatars.
[PNG File (Portable Network Graphics File), 551 KB-Multimedia Appendix 7]

Multimedia Appendix 8

Comparison of similarity rates between FDface and original facial images.
[DOCX File (Microsoft Word File), 98 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Comparison of similarity rates between FLAME_FDface and original facial images.
[DOCX File (Microsoft Word File), 98 KB-Multimedia Appendix 9]

Multimedia Appendix 10

Illustration of the Digital FaceDefender platform.
[DOCX File (Microsoft Word File), 565 KB-Multimedia Appendix 10]

Multimedia Appendix 11

Implementation video of the Digital FaceDfender platform.
[MP4 File (MP4 video File), 811 KB-Multimedia Appendix 11]

References

1. Mason J, Dave R, Chatterjee P, Graham-Allen I, Esterline A, Roy K. An investigation of biometric authentication in the healthcare environment. Array. Dec 2020;8:100042. [doi: 10.1016/j.array.2020.100042]
2. Alhammad N, Alajlani M, Abd-Alrazaq A, Epiphaniou G, Arvanitis T. Patients' perspectives on the data confidentiality, privacy, and security of mHealth apps: systematic review. J Med Internet Res. May 31, 2024;26:e50715. [doi: 10.2196/50715] [Medline: 38820572]
3. Nigam D, Patel SN, Raj Vincent PMD, Srinivasan K, Arunmozhi S. Biometric authentication for intelligent and privacy-preserving healthcare systems. J Healthc Eng. 2022;2022:1789996. [doi: 10.1155/2022/1789996] [Medline: 35368929]
4. Yang Y, Lyu J, Wang R, et al. A digital mask to safeguard patient privacy. Nat Med. Sep 2022;28(9):1883-1892. [doi: 10.1038/s41591-022-01966-1] [Medline: 36109638]
5. Meeus M, Jain S, de Montjoye YA. Concerns about using a digital mask to safeguard patient privacy. Nat Med. Jul 2023;29(7):1658-1659. [doi: 10.1038/s41591-023-02439-9] [Medline: 37464037]
6. Li T, Bolkart T, Black MJ, Li H, Romero J. Learning a model of facial shape and expression from 4D scans. ACM Trans Graph. Dec 31, 2017;36(6):1-17. [doi: 10.1145/3130800.3130813]
7. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ. SMPL: a skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries. Vol . 2023:2. 851-866.
8. Deng J, Guo J, Xue N, Zafeiriou S. ArcFace: additive angular margin loss for deep face recognition. Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 15-20, 2019; Long Beach, CA, USA. [doi: 10.1109/CVPR.2019.00482]
9. Yang Y, Lyu J, Wang R, Xu F, Dai Q, Lin H. Reply to: concerns about using a digital mask to safeguard patient privacy. Nat Med. Jul 2023;29(7):1660-1661. [doi: 10.1038/s41591-023-02435-z] [Medline: 37464038]
10. Neumann M, Moore ST, Baum LM, et al. Politicizing masks? Examining the volume and content of local news coverage of face coverings in the US through the COVID-19 pandemic. Polit Commun. Jan 2, 2024;41(1):66-106. [doi: 10.1080/10584609.2023.2239181]
11. McCrackin SD, Capozzi F, Mayrand F, Ristic J. Face masks impair basic emotion recognition: group effects and individual variability. Soc Psychol. 2022;54. [doi: 10.31234/osf.io/2whmp]
12. Kumari P, Seeja KR. A novel periocular biometrics solution for authentication during COVID-19 pandemic situation. J Ambient Intell Humaniz Comput. 2021;12(11):10321-10337. [doi: 10.1007/s12652-020-02814-1] [Medline: 33425055]

13. Sharma R, Ross A. Periocular biometrics and its relevance to partially masked faces: a survey. Comput Vis Image Underst. Jan 2023;226:103583. [doi: 10.1016/j.cviu.2022.103583]
14. Binka M, Adu PA, Jeong D, et al. The impact of mask mandates on face mask use during the COVID-19 pandemic: longitudinal survey study. JMIR Public Health Surveill. Jan 11, 2023;9:e42616. [doi: 10.2196/42616] [Medline: 36446134]
15. Kret ME, de Gelder B. Islamic headdress influences how emotion is recognized from the eyes. Front Psychol. 2012;3:110. [doi: 10.3389/fpsyg.2012.00110] [Medline: 22557983]
16. Kret ME, Maitner AT, Fischer AH. Interpreting emotions from women with covered faces: a comparison between a Middle Eastern and Western-European sample. Front Psychol. 2021;12:620632. [doi: 10.3389/fpsyg.2021.620632] [Medline: 34025499]
17. Switching Eds: face swapping with Python, dlib, and OpenCV. GitHub. URL: https://matthewearl.github.io/2015/07/28/switching-eds-with-python/ [Accessed 2023-10-15]
18. Color balance. Wikipedia. URL: https://en.wikipedia.org/wiki/Color_balance [Accessed 2023-10-01]
19. Tero K. Flickr faces HQ (FFHQ) 70K from stylegan. GitHub. URL: https://github.com/NVlabs/ffhq-dataset [Accessed 2023-11-12]
20. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. Presented at: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Jun 15-20, 2019; Long Beach, CA, USA. [doi: 10.1109/CVPR.2019.00453]
21. Maluleke VH, Thakkar N, Brooks T, et al. Studying bias in GANS through the lens of race. European Conference on Computer Vision. 344-360. [doi: 10.1007/978-3-031-19778-9_20]
22. SeePrettyFace. URL: http://seeprettyface.com/ [Accessed 2025-06-11]
23. Wen C, Yang Y, Xiao Q, Huang M, Pan W, Alzheimer's Disease Neuroimaging Initiative. Genome-wide association studies of brain imaging data via weighted distance correlation. Bioinformatics. Dec 8, 2020;36(19):4942-4950. [doi: 10.1093/bioinformatics/btaa612] [Medline: 32619001]
24. Yi X, Walia E, Babyn P. Generative adversarial network in medical imaging: a review. Med Image Anal. Dec 2019;58:101552. [doi: 10.1016/j.media.2019.101552] [Medline: 31521965]
25. Shin HC, Tenenholtz NA, Rogers JK, et al. Medical image synthesis for data augmentation and anonymization using generative adversarial networks. In: Simulation and Synthesis in Medical Imaging: Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018. Springer; 2018:1-11. [doi: 10.1007/978-3-030-00536-8_1]
26. Affine transformation. Wolfram Mathworld. URL: https://mathworld.wolfram.com/AffineTransformation.html [Accessed 2024-01-08]
27. WarpAffine. Nvidia. URL: https://docs.nvidia.com/deeplearning/dali/user-guide/docs/examples/image_processing/warp.html [Accessed 2024-01-13]
28. Image filters: Gaussian blur. Medium. URL: https://aryamansharda.medium.com/image-filters-gaussian-blur-eb36db6781b1 [Accessed 2025-06-20]
29. Graham PA. Epidemiology of strabismus. Br J Ophthalmol. Mar 1974;58(3):224-231. [doi: 10.1136/bjo.58.3.224] [Medline: 4834596]
30. Popkin T, Cavallaro A, Hands D. Accurate and efficient method for smoothly space-variant Gaussian blurring. IEEE Trans Image Process. May 2010;19(5):1362-1370. [doi: 10.1109/TIP.2010.2041400] [Medline: 20106740]
31. Tai YW, Jia J, Tang CK. Local color transfer via probabilistic segmentation by expectation-maximization. Presented at: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05); Jun 20-25, 2005; San Diego, CA, USA. [doi: 10.1109/CVPR.2005.215]
32. Huang R, Zheng Y, Hu Z, Zhang S, Li H. Multi-organ segmentation via co-training weight-averaged models from few-organ datasets. Presented at: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference; Oct 4-8, 2020; Lima, Peru.
33. Digital FaceDefender. URL: http://218.16.242.201:5460 [Accessed 2025-06-20]
34. Shan S, Wenger E, Zhang J. Fawkes: protecting privacy against unauthorized deep learning models. In: Proceedings of the 29th USENIX Conference on Security Symposium (SEC'20). USENIX Association; 2020:1589-1604.
35. Jin S, Wang H, Wang Z, et al. FaceObfuscator: defending deep learning-based privacy attacks with gradient descent-resistant features in face recognition. Presented at: 33rd USENIX Security Symposium (USENIX Security 24); Aug 14-16, 2024; Philadelphia, PA, USA.
36. Meden B, Gonzalez-Hernandez M, Peer P, Štruc V. Face deidentification with controllable privacy protection. Image Vis Comput. Jun 2023;134:104678. [doi: 10.1016/j.imavis.2023.104678]
37. Kung HW, Varanka T, Saha S, Sim T, Sebe N. Face anonymization made simple. arXiv. Preprint posted online on Nov 1, 2024. [doi: 10.48550/arXiv.2411.00762]

38. Hukkelås H, Mester R, Lindseth F. DeepPrivacy: a generative adversarial network for face anonymization. Presented at: International symposium on visual computing; Oct 7-9, 2019; Lake Tahoe, NV, USA. [doi: 10.1007/978-3-030-33720-9_44]

39. Digital Face Defender. GitHub. URL: https://github.com/GreatFishbig/Digital-Face-Defender [Accessed 2025-06-11]

Abbreviations

AI: artificial intelligence
DECA: Detailed Expression Capture and Animation
DM: Digital Mask
FFHQ: Flickr-Faces-HQ
FLAME: Faces Learned with an Articulated Model and Expressions
GAN: generative adversarial network
RGB: red, green, and blue
ROI: region of interest

Edited by Taiane de Azevedo Cardoso; peer-reviewed by Mirsaeed Abdollahi, Wenjun Wu; submitted 25.09.2024; final revised version received 02.04.2025; accepted 10.04.2025; published 09.07.2025

Please cite as:
Tan H, Chen H, Wang Z, He M, Wei C, Sun L, Wang X, Shi D, Huang C, Guo A
Artificial Intelligence–Enabled Facial Privacy Protection for Ocular Diagnosis: Development and Validation Study
J Med Internet Res 2025;27:e66873
URL: https://www.jmir.org/2025/1/e66873
doi: 10.2196/66873