scientific reports

OPEN

FaceMotionPreserve: a generative approach for facial de-identification and medical information preservation

Bingquan Zhu[1], Chen Zhang[1], Yanan Sui[1]✉ & Luming Li[1,2]

Telemedicine and video-based diagnosis have raised significant concerns regarding the protection of facial privacy. Effective de-identification methods require the preservation of diagnostic information related to normal and pathological facial movements, which play a crucial role in the diagnosis of various movement, neurological, and psychiatric disorders. In this work, we have developed FaceMotionPreserve , a deep generative model-based approach that transforms patients' facial identities while preserving facial dynamics with a novel face dynamic similarity module to enhance facial landmark consistency. We collected test videos from patients with Parkinson's disease recruited via telemedicine for evaluation of model performance and clinical applicability. The performance of FaceMotionPreserve was quantitatively evaluated based on neurologist diagnostic consistency, critical facial behavior fidelity, and correlation of general facial dynamics. In addition, we further validated the robustness and advancements of our model in preserving medical information with clinical examination videos from a different cohort of patients. FaceMotionPreserve is applicable to real-time integration, safeguarding facial privacy while retaining crucial medical information associated with facial movements to address concerns in telemedicine, and facilitating safer and more collaborative medical data sharing.

Protecting the privacy of patients is a critical and ethical obligation in both healthcare practice and research[1]. Any unauthorized or malicious collection, storage, disclosure, and use of medical information would violate their human rights and dignity. Concerns about inadequate protection of personal privacy could discourage patients from seeking healthcare services and building trust with providers, especially in the case of certain clinical conditions that could lead to discrimination or stigma. Among all privacy issues, the protection of facial information is one of the biggest concerns, as the face contains the most recognizable biometric identifiers[2], strongly related to health information profiles and other personal data. Telemedicine, the emerging practice of online medical services instead of in-office visits[3], exacerbates facial privacy concerns as it often involves video conferencing or recording with patient over the internet. Placing high significance on facial privacy protection in telemedicine is important and valuable for patient benefits, promoting the willingness and access to medical services. With the increasing demand and adoption of telemedicine, adequate protection of facial privacy poses new societal and technical challenges in the era of rapid development and expanded use of face recognition and synthesis technologies.

Artificial intelligence (AI) is transforming telemedicine in health status tracking, disease screening, remote monitoring, diagnostic assistance, treatment planning and management and care services[4], helping streamline the process, improve data and service quality[5], and promoting the cost-effectiveness and accessibility[6,7]. However, AI techniques may also extract sensitive personal information beyond identity including age, gender, emotion, etc[8]. Controversies and regulations on face privacy, as well as AI tools for face de-identification, are issues of great concerns worldwide. In video-based diagnosis and telemedicine scenarios, faces also convey diagnostic information. Patients with certain neurological conditions exhibit specific facial movement disorders, such as the asymmetrical face of stroke patients. Moreover, facial movements show important social information, playing essential roles in human interactions[9]. Medical examination of facial movements is required for many disorders, including the masked faces (hypomimia) for Parkinson's disease (PD)[10]. Therefore, facial privacy

[1]National Engineering Research Center of Neuromodulation, Tsinghua University, Beijing 100084, China. [2]IDG/McGovern Institute for Brain Research, Tsinghua University, Beijing 100084, China. ✉email: ysui@tsinghua.edu.cn

protection in telemedicine practice proposes additional demands in preserving subtle facial dynamics besides de-identification of patients' faces.

Unfortunately, facial attributes and dynamics are closely entangled with identity. Conventional de-identification methods, including blurring, masking and pixelization[11], remove informative contents from entire or part of the face, making clinical facial investigation almost impossible. However, their performance of de-identification might still be challenged with current deep learning techniques in de-blurring, super resolution, image inpainting and self-supervised learning[12,13]. Deep learning based face manipulation methods provide tools for face de-identification in multiple ways. Some modified faces to cheat the identity recognition algorithms but were unable to affect human recognition performance[14]. Some others focused on changing facial features and attributes, paying little attention in tracking movement changes[15,16]. To concurrently change facial identity and keep facial attributes, a series of face swapping technologies leveraging deep learning have been developed to make deepfakes.

In deepfake videos, the original face of a person is replaced with the source face from someone else for identity alteration, while the original facial expressions are roughly retained. This technology has been deployed in film industry and is getting popular in public entertainment through social media, short-video platforms, and deepfake APPs[17]. Although the misuse of deepfakes has caused new threats of disinformation and fraud, raising concerns and criticisms on its negative impact, we argue that the technology itself can be used as a powerful tool to benefit patients in telemedicine. When applied directly to any patient video with face, deepfake is able to remove facial identity and related sensitive information whilst keeping facial movements. It is a more natural and perceptually acceptable approach to protect patients' facial privacy.

Deepfakes can be created in three major ways. Models with autoencoders as the key components[17–20] train a shared encoder and separate decoders or a decoder with separate latent embeddings for original and source faces, replacing the original identity when the source face decoder or latent embedding is chosen. These methods could swap faces while keeping facial details, but are always trained in face pairs and the decoders are mostly subject-specific, which limit their capability of generalization and scaled application. 3D vision based models, sometimes combined with neural scene representation and rendering techniques, manipulate the original identity by changing 3D face parameters and reconstructing faces with the source identity[21–23]. They allow face swapping with more synthetic flexibility but encounter troubles in precisely editing face areas like eyes and mouth and preserving subtle expressions. The performance is also affected by the underlying 3D models. Generative adversarial networks (GANs) for deepfake[24–27] use generators to create faces with new identity and discriminators to improve image quality and attribute fidelity. GANs integrated with identity extractors enable the models to generalize to arbitrary identities which help eliminate the restrictions of paired faces and reduce computational loads. However, it remains an open question whether they could preserve facial movements for diagnoses, and how the preservation should be evaluated quantitatively.

In this study, we proposed FaceMotionPreserve , a generative deep learning model-based approach for subject-agnostic de-identification and facial movement preservation of medical videos. We deployed our model to an integrated system to track and de-identify faces in real-time and demonstrated its feasibility of being a convenient tool for facial privacy protection in telemedicine. With FaceMotionPreserve , videos recorded by patients get de-identified while the facial dynamics are retained, and are then safely sent out for clinical diagnoses (Fig. 1).

FaceMotionPreserve de-identifies face images following a GAN-based pipeline. A novel facial landmark similarity module was integrated to enhance diagnostic information preservation. The model was trained only on publicly available datasets without any real patient data, and we specially constructed the training set with enough blink images to match natural blink distribution, reducing related training bias. The performance evaluation and model comparison were conducted with videos from patients with Parkinson's disease, a movement disorder for which telemedicine has already showed the efficacy in remote diagnosis and programming of neuromodulation treatment[28,29], where FaceMotionPreserve was directly applied. We have tested the de-identification performance with both human participants and a deep learning method. To comprehensively examine the preservation of

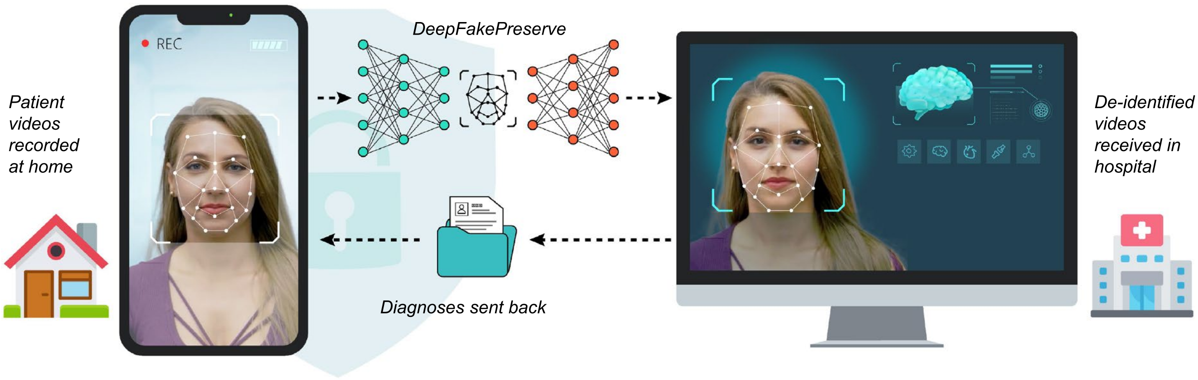

Patient videos recorded at home
REC
DeepFakePreserve
Diagnoses sent back
De-identified videos received in hospital

Figure 1. Workflow of FaceMotionPreserve for telemedicine privacy protection The system converts a patient face to a new identity in real time, providing de-identified videos to healthcare providers while keeping original facial movements with high fidelity. Medical decisions are made based on de-identified videos. The faces illustrated here are from the DeepFakeDetection Dataset.

diagnostic information, we developed a framework of different evaluation aspects quantitatively, including specialist diagnostic consistency, critical facial behavior fidelity and general facial dynamic correlation. Evaluation results from neurologists, healthy human participants and algorithms demonstrated the capability of FaceMotionPreserve for both de-identification and facial dynamics preservation, which was improved by facial landmark loss and augmented training set with annotated blink images proportional to normal human behavior. We also compared FaceMotionPreserve with other deepfake models on a different cohort of patients with PD and showed its robustness and advantages in diagnostic facial motion information preservation.

FaceMotionPreserve provides a new way to overcome difficulties in removing facial identity while keeping diagnostic information. It could alleviate worries and concerns about facial privacy in telemedicine and further promote medical data sharing and secondary use for clinical and scientific research.

Methods

Participants and evaluation datasets

We recruited two groups of patients with PD, one through telemedicine and the other in clinical and lab settings, to obtain their face related test videos, according to the Movement Disorder Society Unified Parkinson's Disease Rating Scale (MDS-UPDRS)[10], which is the most widely adopted rating scale in assessing PD symptoms and severity. Three motor examination tests, which are facial expression, speech, and rest tremor for lip/jaw were chosen because they require face videos. Facial expression is scored based on the severity of hypomimia, including the degree of movement and expression reduction in the face, such as decreased blinks, fewer movements around the mouth and parted lips during rest. Rest tremor for lip/jaw is rated by examining the maximal tremor amplitude during the entire exam. Speech is rated by evaluating the understandability, considering various factors including volume, diction, prosody and clarity. Each test is scored from 0 to 4, indicating five levels of severity, namely normal (no symptom), slight, mild, moderate, and severe.

The first cohort included 76 patients (33 females and 43 males), with an average age of 67.6 years (s.d. 9.4 years). They followed the test and recording instruction inside the telemedicine APP installed on smartphones[30], recorded required videos by themselves or their caregivers, and uploaded the videos outside clinics or healthcare service places. The APP is equipped with deep learning models for pose and light condition detection, and provides voice prompts to guide the recording. If the face position or the light condition was not satisfied, new recordings were required. We also provided every patient with a mobile phone tripod to ensure the stability. We constructed PD dataset I of 60,818 frames from 76 videos of the facial expression test and 37,859 frames from 75 videos of the speech test. Since rest tremor for lip/jaw could be scored from these recordings, we didn't ask the patients to shoot extra videos. We worked on this dataset for the evaluation of FaceMotionPreserve .

The second cohort included 35 patients (15 females and 20 males), aged at 57.0 years on average (s.d. 8.7 years). We composed PD dataset II of 10,295 frames from 35 videos of the facial expression test recorded by experienced physicians. This dataset was used for model comparison. While the patient face images were utilized as original images for model comparison, we acquired source faces from two healthy volunteers, one female recognized as a21 and one male recognized as a22, who granted their permissions to us for the usage of their face images. Their videos lasted for 20 seconds (1000 frames), including facial expressions for happy, surprise, sad and disgust.

For human evaluation of patient face de-identification performance with FaceMotionPreserve , we recruited seven adult participants as human judges to distinguish faces in the identification experiment.

This study was reviewed and approved by the Institutional Review Boards of Beijing Tsinghua Changgung Hospital and Tsinghua University Yuquan Hospital, and the Ethical Committees of Beijing Tiantan Hospital of Capital Medical University, Peking Union Medical College Hospital, and Qilu Hospital of Shandong University, Clinical Trials Identifiers NCT02937727 and NCT03053726. All experiments were performed in accordance with Declaration of Helsinki and Good Clinical Practice of China. All participants and volunteers signed their corresponding informed consents before participating. Any image or video frame involved in this study only contained one face.

FaceMotionPreserve Architecture

The architecture of FaceMotionPreserve is illustrated in Fig. 2. Following the general framework of generative adversarial network, FaceMotionPreserve modifies faces with a generator module, which was trained together with a discriminator module similar to the SimSwap model[26]. The generator module includes an encoder to encode the original image into a feature map, an ID injection step with the Adaptive Instance Normalization[31] to implicitly inject source identity (a 512-d vector extracted by an ID extractor module called Arcface[32]) information into the feature map, and a decoder to generate the de-identified image using the modified feature map. The discriminator module[33] was involved during training to improve image quality and facilitate the maintenance of facial attributes with different losses. We adopted some modules from SimSwap. During the adversarial training process, the generator get updated with feedback from the discriminator module (GAN loss, weak feature matching loss, etc.), ID extractor module (ID loss) and original images (image reconstruction loss).

To achieve facial movement and diagnostic information preservation, we extended the SimSwap, implementing a novel face dynamic similarity module, which uses a pretrained landmark extractor to quantify facial landmark discrepancy between the de-identified and original faces during training. The pretrained landmark detection module[34] detects facial landmarks on original and de-identified images. Since temporal correlation information was not available during model training, mean Euclidean distance between original and de-identified facial landmarks that directly or strongly relate to facial expressions (eyebrows, eyes and inner mouth) was calculated as the landmark similarity loss, offering an optimization goal for facial movement invariability. The loss was calculated as the following:

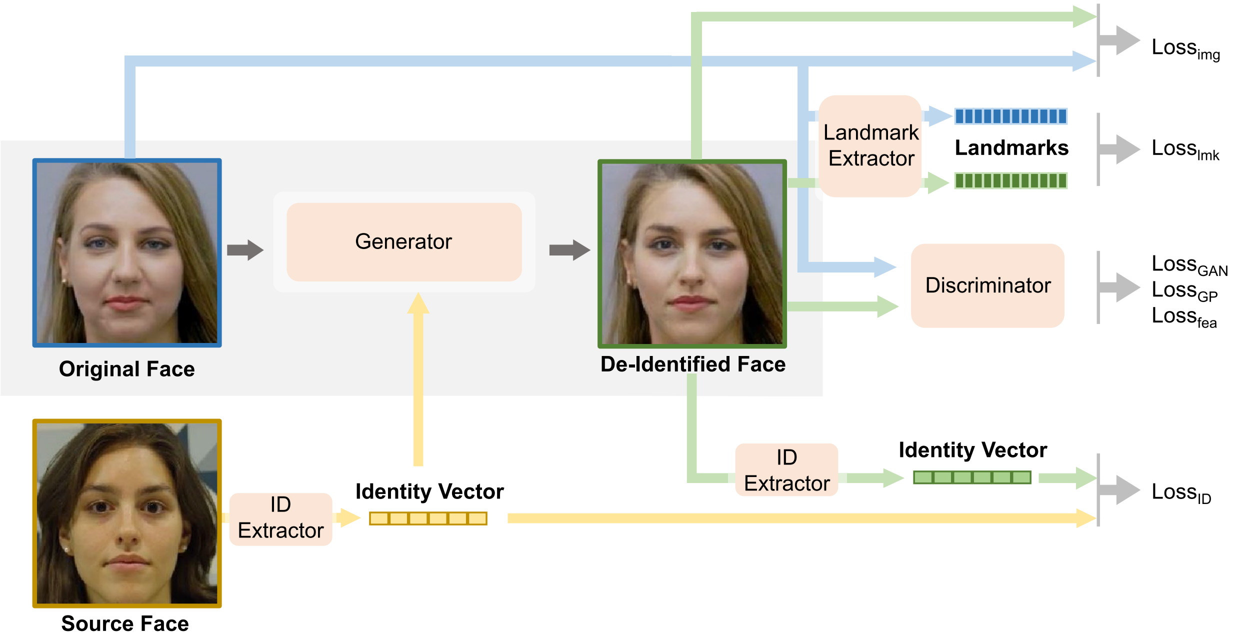


Figure 2. FaceMotionPreserve has a generator module that takes an original face and source identity as input, and outputs a de-identified face. During training, FaceMotionPreserve employs a discriminator module to improve image quality and preserve high semantic attributes, and a landmark extractor module to enhance facial dynamics preservation. Different losses are involved: identity loss ($Loss_{ID}$) enforces identity replacement, image loss ($Loss_{img}$) and discriminator's GAN loss ($Loss_{GAN}$ and $Loss_{GP}$) and feature loss ($Loss_{fea}$) improve image reconstruction, and landmark loss ($Loss_{lmk}$) enhances facial movement preservation.

$$L_{lmk} = \sqrt{\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left|l_{i,j}^{original} - l_{i,j}^{de-identified}\right|}$$

where i, j are landmark index, n is landmark total number(30), $l_{i,j}$ is Euclidean distance of landmark (i,j) pair of original and de-identified image.

The overall loss function of FaceMotionPreserve comprises losses derived from all modules, including the landmark similarity loss L_{lmk} between landmarks of original and de-identified face images from the landmark extractor, and losses from the SimSwap model (identity loss L_{ID} between source and de-identified ID vectors, adversarial loss L_{GAN}, image reconstruction loss L_{img}, weak feature matching loss[26] L_{fea} and gradient penalty loss[35] L_{GP} between original and de-identified images), as the following:

$$L = \lambda_{ID}L_{ID} + \lambda_{lmk}L_{lmk}/k + \lambda_{GAN}L_{GAN} + \lambda_{img}\delta L_{img} + \lambda_{fea}L_{fea} + \lambda_{GP}L_{GP}$$

The λs are weights of corresponding losses. δ is 1 if the original image has same identity with the source, and 0 if the identity is different. k denotes $\lambda_{ID}L_{ID}$. λ_{lmk} was set to 0.2, and $\lambda_{ID} = 10$, $\lambda_{GAN} = 1$, $\lambda_{img} = 10$, $\lambda_{fea} = 10$, $\lambda_{GP} = 1 \times 10^{-5}$ as in SimSwap. We applied adaptive weights that is inversely proportional to the identity loss and the landmark loss in order to better enforce facial landmark consistency in late training periods when all the other losses became small.

Training process

We built the training dataset with VGGFace2[36] and mEBAL[37,38], and trained FaceMotionPreserve on these publicly available datasets only. The VGGFace2 is a widely used benchmark dataset for face recognition and related tasks containing large-scale face images of different identities. However, we observed that it lacked images of natural blink phases, causing models trained on it to be biased to eye-open cases. The approximate blink frequency of human adults is 17 times per minute[39] on average and each blink lasts about 180 ms to 270 ms[40], making up 5% to 8% of the human awake time. The mEBAL presents a dataset containing annotated blink video frames. To alleviate the no-blink bias, we constructed an augmented training set by complementing the VGGFace2 with blink frames from mEBAL to make the dataset comprising around 8% blink images in total. Images larger than 256x256 pixels in size were selected. We then aligned these face images and cropped them into 224×224 pixels as the input for the model.

During training, original images and source vectors of the same or different identities were passed to our model alternately. The model was trained 500 epochs on 8 NVIDIA 2080Ti GPUs with a batch size of 32 and the ADAM optimizer with learning rate of 1×10^{-4} and $\beta_1 = 0$, $\beta_2 = 0.990$.

Model evaluation

After training, we evaluated FaceMotionPreserve on PD dataset I. Images from this dataset were referred to as original images and face images of actor 14 from the DeepFakeDetection Dataset[41,42] were used as the source. We aligned face images and cropped them into 224×224 pixels to feed into our model with tools from InsightFace[43]. For each original images in PD dataset I, FaceMotionPreserve output a corresponding de-identified face image. We evaluated the proposed approach in four aspects: the performance of face de-identification, the clinical rating consistency of face-related motor tests after de-identification, the preservation of facial dynamics, and the invariance of body motion, as shown in Fig. 3.

Evaluation for face de-identification

The performance of face de-identification was evaluated by both human judges and a widely adopted face recognition pretrained model, Arcface (model ms1mv3_r50)[32]. We recruited seven healthy adults to participate in the face re-identification task. They were showed six faces at a time, including a de-identified face, the corresponding original face from PD dataset I, the source face, and three other patients' faces as controls. The images were shuffled, standardized (color transferred and color jittered to diminish non-identity clues, then center cropped) and simultaneously displayed. The participants were not disclosed of any information about the displayed faces. They were asked to determine whether there were two faces from the same identity and if so, which two.

Arcface projects face images into 512-dimension identity vectors. The cosine value of any two identity vectors was calculated to represent the similarity of the corresponding faces. We set face verification threshold to be 0.34 (the corresponding angle between the vectors is 70°) as suggested in[44] (accuracy was about 0.9999 and false negative rate $< 5 \times 10^{-5}$ in the benchmark IJBC face dataset[45]). All frames from PD dataset I were tested with Arcface.

Neurologist rating

To evaluate the performance of our model in clinical practice, the original and de-identified videos from PD dataset I were randomly shuffled and re-ordered. Three experienced MDS-licensed neurologists specialized in movement disorders independently watched these videos and rated the subjects' facial expression, rest tremor for lip/jaw and speech. The neurologists would mark "not sure" on a video if he/she was not confident in scoring the symptom or severity. We got either a valid score or a "not sure" mark from each neurologist for every video (Fig. 5A). Some videos were recorded in noisy environment which affected the evaluation of a patient's speech and were excluded from speech rating. We analyzed the rating consistency among all three neurologists only based on the valid scores (218 for hypomimia, 206 for tremor, 120 for speech).

Analysis of facial dynamics preservation

To analyze facial dynamics, we grouped the de-identified images to make de-identified videos in accordance with the original face videos in PD dataset I. Since it's pointless to fix absolute landmark locations because they inherently vary with identity, we chose temporal sequences of facial landmark pair distance to quantify the general facial activities. We applied the SBR model[34] for facial landmark detection for its reliable performance and reduced jittering on videos. For each face video, 51 landmarks in areas of eyes, brows, nose and mouth which are closely related to facial expressions and movements were extracted (see Fig. 6A). We paired all the landmarks and got 1275 (51*50/2) facial feature sequences: each feature sequence was the time series of distance vectors between a landmark pair, denoted as $\{dx_{i,j}, dy_{i,j}\}_t$, where $dx_{i,j} = x_i - x_j$ and $dy_{i,j} = y_i - y_j$ were the differences in horizontal and vertical directions between the ith and the jth landmark, and t was the frame index. For each video, we calculated the Pearson correlation coefficients (r) for each paired original/de-identified video feature

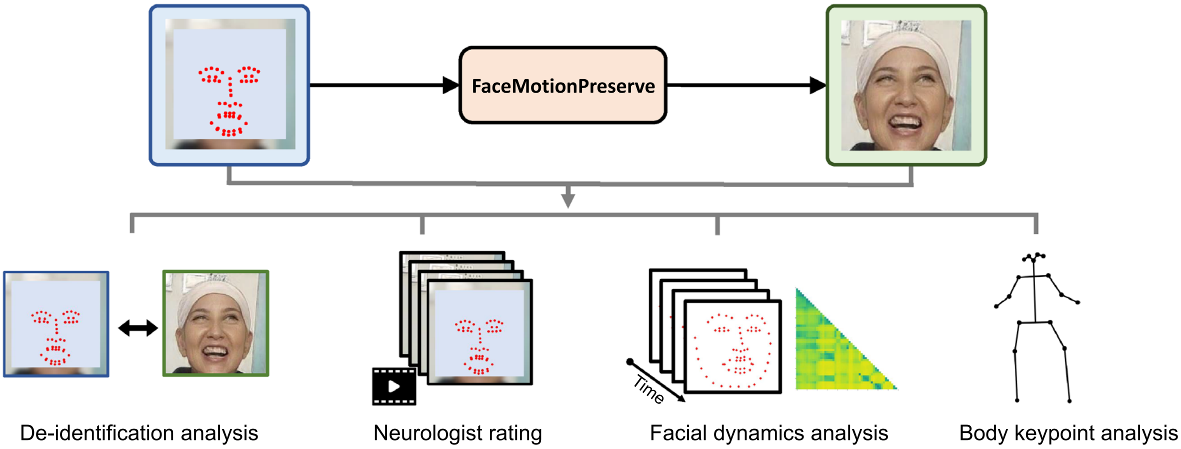


Figure 3. Evaluation of FaceMotionPreserve . We applied FaceMotionPreserve to get de-identified images for PD dataset I. We then analyzed the performance with de-identification analysis, face dynamic analysis, neurologist rating and body keypoint analysis..

sequence; and then averaged r on videos to measure invariability of a facial feature. We evaluated a model's performance for holistic preservation using the mean value of the 1275 averaged r. We also applied canonical correlation analysis (CCA) to find the most related direction z and calculated the correlation coefficient of $\{dz_{i,j}\}_t$. Facial expression videos with validated tremor ratings and hypomimia ratings from all the neurologists in PD dataset I (56 cases) were used for analysis.

We also evaluated facial diagnostic information after de-identification, including eye blinks and the rest tremor. We set a state threshold $state_{thre}$ of eye aspect ratio (EAR), defined as the Euclidean distance between upper and lower eyelids normalized by eye width, to determine the open/close state of eyes[46]. Large EAR values indicate opened eyes and small values indicate closed eyes. The consecutive eye-close states that lasted for 60 ms to 700 ms was considered as a blink. The original and de-identified EAR determined blinks were considered a true positive if two blink phases overlapped. We calculated the recall-precision curve of $state_{thre}$ ranging from 0.05 to 0.30. Videos in which the patient squinted most of the time were excluded.

We evaluated rest tremor for lip/jaw using the temporal sequence of inner-lower-lip midpoint vertical position re-aligned to full image frames. The temporal sequences were high-pass filtered by a second order Butterworth filter with the critical frequency at 3 Hz. Spectrograms were obtained using a Tukey window with shape parameter of 0.25 and segment length of 8 second.

Body motion invariance

We also verified the invariance of body motion after facial de-identification as FaceMotionPreserve would be integrated in AI-based telemedicine systems where body motion are also analyzed with deep learning models. We detected 18 body keypoints (ankles, knees, hips, neck, shoulders, elbows, wrists, nose, eyes, ears) from video frames with OpenPose[47]. Keypoints in the original image were treated as ground truth and deviations of corresponding keypoints in the de-identified image were calculated. We used object keypoint similarity (OKS)[48] as the measuring metric, which evaluated keypoint deviation normalized by body size:

$$\mathrm{OKS} = \frac{\Sigma_i\left[\exp\left(-\frac{d_i^2}{2s^2\kappa_i^2}\delta(v_i > 0)\right)\right]}{\Sigma_i[\delta(v_i > 0)]}$$

d_i is the Euclidean distance between the i-th keypoint in original and de-identified images, v_i is the original visibility of the i-th keypoint ($v_i > 0$ if visible and $= 0$ otherwise), s is body size, and κ_i denotes keypoint-related coefficient. OKS ranges from 0 to 1 and quantifies keypoint similarity normalized by human annotator variance. When OKS=0.95, deviation is around 2% of body size on facial keypoints, 5% on hands/feet and 7% on hip joints. When OKS=0.5, deviation value expands 4 times. When OKS=1, keypoint positions are identical. $AP^{0.5}$ denotes average precision with OKS above threshold 0.5. $AP^{0.5:0.95}$ is the primary metric where superscript 0.5:0.95 means averaging AP on OKS thresholds ranging from 0.5 to 0.95 with steps of 0.05. We also created masked images with a black mask covering face area, blurred images with average blur of kernel size 1/4 of face area and pixelized images to 16×16 grids for comparison. All video frames from PD dataset I were used for body keypoint detection analysis.

Model comparison

We qualitatively and quantitatively compared model performance under different settings with images in PD dataset II as the original and face images from the two volunteers as the source. We use d-l- to represent the ablated model without the landmark extractor and loss L_{lmk}, and trained without mEBAL blink data; d-l+ for model trained without mEBAL blink data but with the same model architecture as FaceMotionPreserve and training loss; and d+l+ for the ultimate version of FaceMotionPreserve trained with blink data and landmark similarity block. We also compared our model with two state-of-the-art and widely adopted deepfake models, Faceswap[18] and MegaFS[27]. Since the Faceswap model is subject-specific, for each patient in the dataset, we trained one model on the original face and the sex-matched source face with 1500 images (10% were blink images) from both faces (trained with 12,000 iterations with batch size 64, learning rate 5×10^{-5}). And we used the pretrained FTM version of MegaFS and the d+l+ version model for comparison. All 10,295 frames from the 35 patients in PD dataset II were measured with Pearson correlation coefficient and 28 videos were used in blink detection comparison (videos with large portion of frames of closed or squint eyes were excluded).

Statistics

We used paired t-test for face identity similarity analysis. Both two-sided and one-sided t-test have p value lower than 1×10^{-10} (0.0 in SciPy v1.5.4).

Results

Face de-identification

We first examined whether FaceMotionPreserve could prevent face re-identification from both human participants and learning algorithms for facial privacy protection. Human judges participated in a total of 532 trials and their re-identification results are shown in Fig. 4. Although the original and the de-identified faces were always presented, more than 59.0% identifications declared no repetition of identity. In 16.7% of the trials, human participants mapped one of the control faces with original, source, or other control identities, and 22.6% of the time they mapped the source face to the de-identified face. The re-identification rate of de-identified and original pairs was only 1.7%.

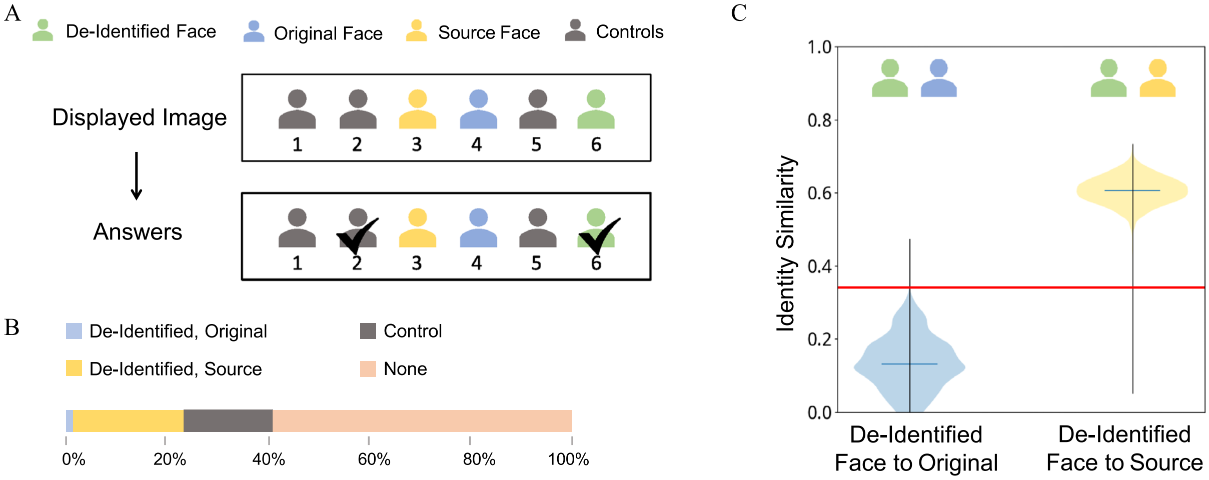


Figure 4. Face re-identification. We recruited participants to manually evaluate FaceMotionPreserve 's de-identification performance. We also calculated the identity similarity between the de-identified faces and the original/source faces using the Arcface model. (**A**) Design of face re-identification experiment. (**B**) Results from seven human judges. (**C**) Distribution of identity similarity between de-identified and original face pairs, and de-identified and source face pairs. The red horizontal line indicates face verification threshold..

We used Arcface to further evaluate identity similarity between de-identified faces and original/source faces on all the 98,677 face frames in PD dataset I. As Fig. 4C shows, the average similarity score of the de-identified faces and their corresponding original faces was 0.13 (s.d. 0.08), lower than the face verification threshold of Arcface which is 0.34. Only 0.04% faces of which the similarity scores exceeded the threshold and might be recognized as the original identity. The average identity similarity between the de-identified faces and the source face is 0.61 (s.d. 0.04), higher than the face verification threshold, with only 0.01% outliers fell below. De-identified faces were much closer to the source faces than the original faces with statistical significance ($p < 1 \times 10^{-10}$) in both two-sided and one-sided t-test, and only 0.008% de-identified faces were more closer to the original faces.

Both algorithmic and human identification significantly changed after FaceMotionPreserve processing and the chance of re-identifying the original identity was very low. These results demonstrate that FaceMotionPreserve could prevent facial re-identification and therefore protects facial privacy.

FaceMotionPreserve for facial diagnostic dynamic preservation

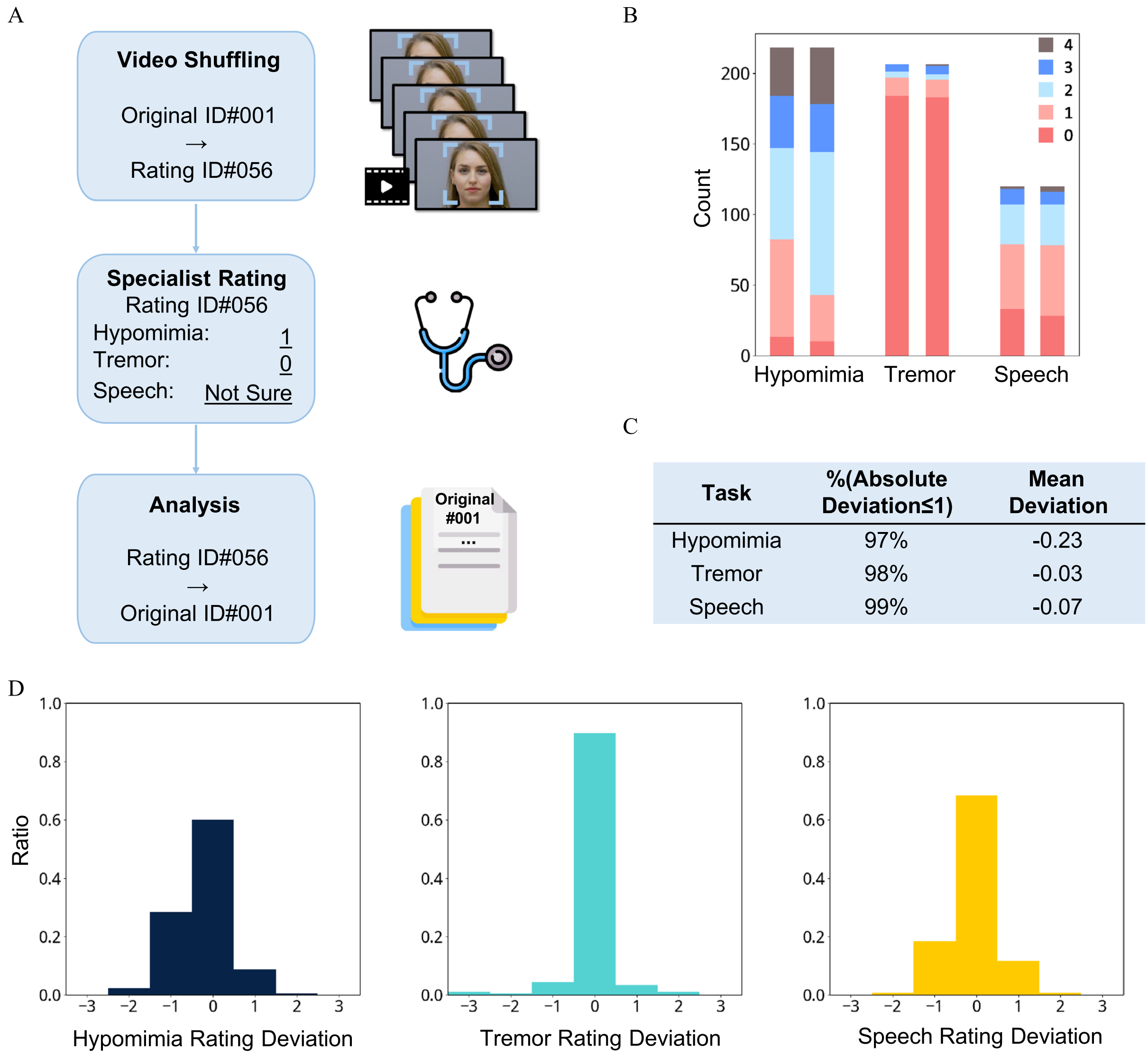
Neurologist ratings of facial symptom severity were consistent before and after FaceMotionPreserve de-identification

We invited three experienced neurologists to watch the shuffled original and de-identified videos and rated three items (facial expression or hypomimia, rest tremor for lip/jaw and speech) in MDS-UPDRS, ranging from 0 to 4 for symptom severity. Speech abnormality was included because human interpretation of speech strongly correlates with visual cues from mouth[49]. Figure 5B depicts rating score distributions. Inter-rater score deviation of original data was {0.63, 0.74, 0.10} for hypomimia, {0.27, 0.05, 0.20} for tremor, and {0.77, 0.44, 1.10} for speech. We then compared rating score deviation of paired original and de-identified videos and found high consistency between tremor and speech ratings: both showed symmetric deviation distribution and mean deviation close to 0 (−0.03 and −0.07). 98% tremor score deviation was less than or equal to 1, and 90% was identical. 99% of speech deviation was less than or equal to 1, and 68% was identical (see Fig. 5C, D). Hypomimia scores changed more than the other two items: mean deviation was −0.23, indicating that facial expression features slightly shifted after de-identification. Nevertheless, rating of hypomimia still showed high consistency with 97% deviations ≤ 1.

Critical facial behaviors were invariant

One of key symptoms of PD hypomimia is decreased blinking; and facial tremor of PD is characterized by the 4 to 6 Hz tremor at lip and jaw. Blink was defined as closing eyes (small EAR values) with a reasonable duration. We observed that modifying facial identity also altered eye outlines and led to the baseline of EAR value to change, which could consequently shift EAR threshold for blink detection (see Fig. 6C). To eliminate the shifting affects, we sampled various EAR thresholds to calculate blink invariability (overlap of detected blinks between original and de-identified images). Figure 6B shows that at threshold 0.243, the best performance yielded precision of 0.94 and recall of 0.81. The false negatives and false positives of blink detection might come from differences of EAR values, whose slight deviation near threshold could drastically change detection results. The de-identified faces had precision higher than recall, indicating that they had smaller eye blink movement than the original faces, which might explain the increased hypomimia scores.

For lip/jaw rest tremor, spectrograms of lower lip location showed typical PD tremor at 4 to 6 Hz and were highly similar in frequency and intensity between original/de-identified data (see Fig. 6D).



Task	%(Absolute Deviation≤1)	Mean Deviation
Hypomimia	97%	-0.23
Tremor	98%	-0.03
Speech	99%	-0.07

Figure 5. Neurologist evaluation experiments and results. Three experienced neurologists rated face-related MDS-UPDRS scores on 76 PD patients' original/de-identified videos. (**A**) We shuffled and renamed the original/de-identified videos for evaluation and analyzed rating scores. (**B**) Rating score distribution of original (left bars) and de-identified (right bars) videos. Scores were included if both the paired original and de-identified ratings were valid. (**C**) Summary of score deviation. (**D**) Distribution of hypomimia/tremor/speech rating score deviation. Scores were highly identical, and most scores changed $\leq$ 1.

General facial activity dynamics are invariant

The coherence of facial dynamics was evaluated by the Pearson correlation coefficient r of original/de-identified faces. We found that the y components had an average $r = 0.736$ and x components had an average $r = 0.580$ (see Fig. 6E). The difference in x and y components indicated that facial movements were directional, and we applied canonical correlation analysis (CCA) on distance vector sequences to find the best direction. The CCA vectors had an average $r = 0.844$, indicating that facial dynamics were strongly correlated. Higher CCA vector correlation (>0.9) was observed among facial parts that had relative movements in upper-lower mouth, lower mouth and eye, lower mouth and nose, upper eye and nose, upper eye and mouth.

FaceMotionPreserve for body movement preservation

Body pose estimation systems[47] are increasingly aiding movement assessments, but they usually have large receptive fields that might be affected by facial changes. Hence, we validated body keypoint invariability by comparing keypoints distances between FaceMotionPreserve or traditionally de-identified (black masked, blurred, pixelization) videos and their original videos. Identical keypoint locations have OKS = 1 and OKS decays exponentially with distance. Almost all images after FaceMotionPreserve processing had nearly invariant keypoints: over 97% video frames had OKS larger than 0.95 and over 85% had OKS larger than 0.995, while traditional de-identification methods had significant keypoint changes not only in face but also in limb and torso. Masking, blurring and pixelization on facial area removed nose and eye information and greatly influenced ear

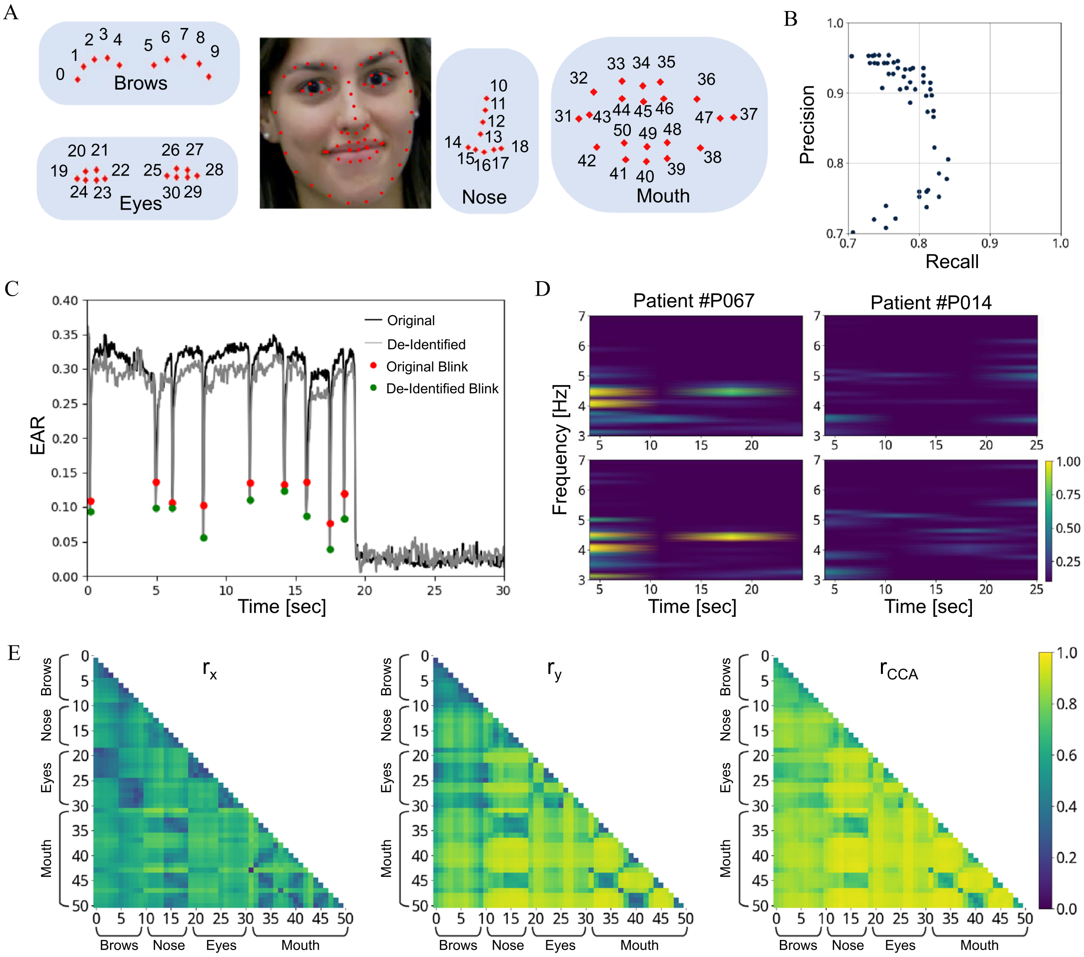
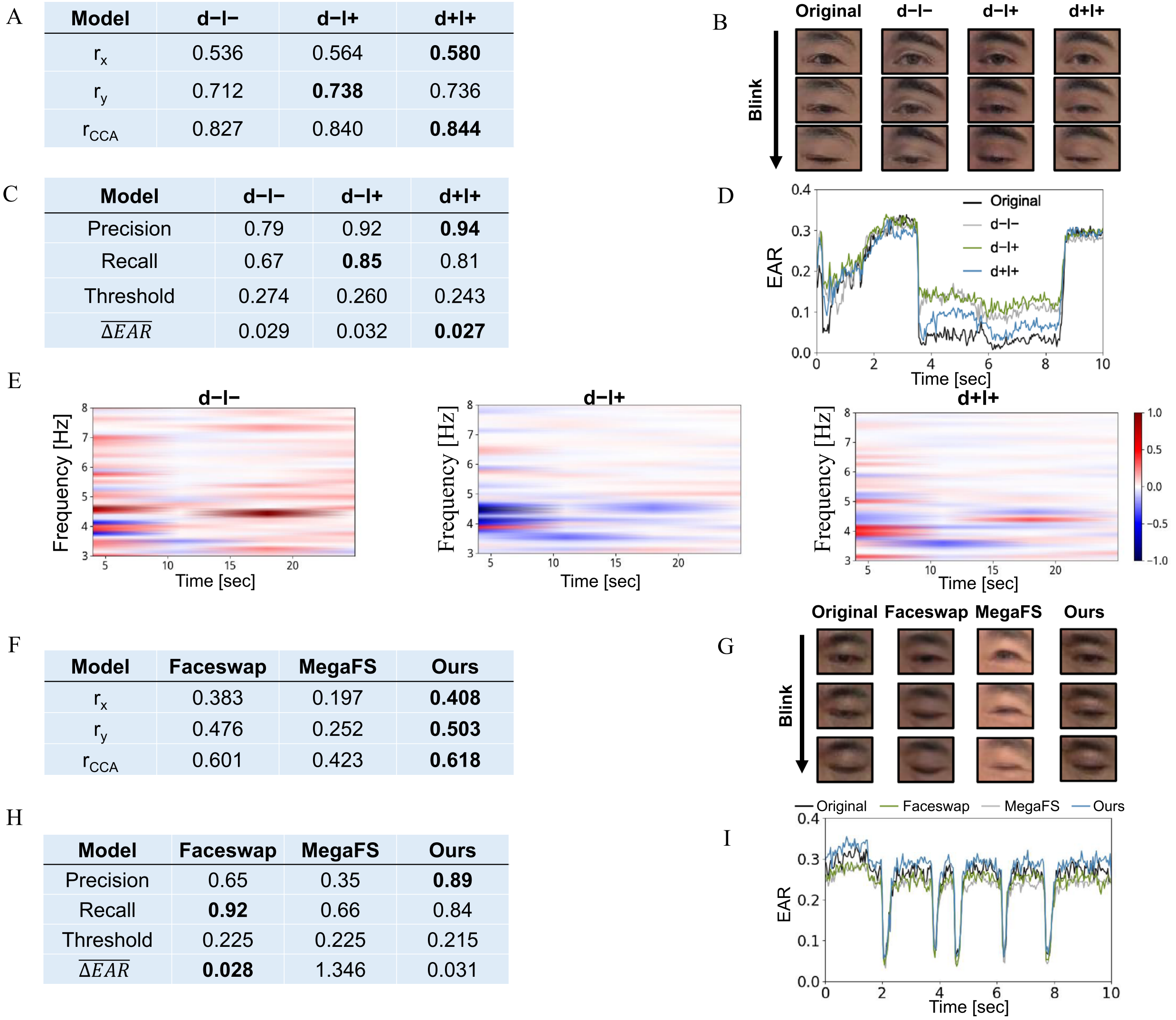


Figure 6. Facial dynamics invariability. We measured facial movements and compared the differences before/after FaceMotionPreserve processing. (**A**) Illustration of facial landmarks. (**B**) Precision–recall curve for blink detection. (**C**) EAR and detected blinks of a patient before/after FaceMotionPreserve . The de-identified EAR values were temporally similar to the original ones with simultaneous blinks. (**D**) Mouth movement spectrogram of two patients that all neurologists diagnosed with tremor. Left: a patient whose original and de-identified video both had mean tremor score 2.3. Right: a patient whose original and de-identified video both had mean tremor score 2.0. The upper graphs are original spectrograms and the bottom graphs are de-identified spectrograms. The graphs have peak power in 4–6 Hz and show high similarity between the paired data. (**E**) Mean correlation of facial landmark pairs. Row i and column j depicts landmark pair (i,j) distance vector's Pearson correlation coefficient before/after FaceMotionPreserve .

detection, leaving much fewer images (mask: 2.9%, blur: 42.0%, pixelization: 75.7%) with OKS larger than 0.95. FaceMotionPreserve induced smaller changes to keypoint detection and minimized information loss for body movement analysis, which is import in further integration of AI-powered system.

Model comparison

To locate dominant factors of model performance, we compared three model formations: one ablated model, trained without mEBAL blink data and landmark similarity block (d–l–); one trained without mEBAL blink data but with landmark similarity block (d–l+) and one trained with both mEBAL blink data and landmark similarity block (d+l+, FaceMotionPreserve model). Without extra blink data, the d–l+ model had slight inferior correlation on general facial dynamics with lower Pearson's r ($r_x = 0.564, r_y = 0.738, r_{CCA} = 0.840$) compared to the d+l+ ($r_x = 0.580, r_y = 0.736, r_{CCA} = 0.844$) (Fig. 7A). The EAR curves (Fig. 7B–D) demonstrated that d–l+ had incomplete eye closing compared to d+l+ (mean EAR deviation 0.032 and 0.027, respectively), while maintaining overall blink consistency (precision 0.92 and recall 0.85 compared to d+l+'s precision 0.94 and recall 0.81). The d–l– model further removed landmark similarity block, which reduced facial correlation ($r_x = 0.536, r_y = 0.712, r_{CCA} = 0.827$) and worsened blink tracking (Fig. 7C, precision and recall dropped from 0.92 and 0.85 for d–l+ model to 0.77 and 0.62 for d–l– model). Rest tremor for lip/jaw at frequency of 4 to 6 Hz was better preserved by d+l+, with smaller average spectrogram power deviation (0.003) from the original compared to d–l– (0.106) and d–l+ (−0.078) (Fig. 7E). In short, the landmark similarity module substantially improved facial dynamical invariability and extra blink data considerably improved blink reconstruction.



A

Model	d−l−	d−l+	d+l+
r_x	0.536	0.564	**0.580**
r_y	0.712	**0.738**	0.736
r_{CCA}	0.827	0.840	**0.844**

C

Model	d−l−	d−l+	d+l+
Precision	0.79	0.92	**0.94**
Recall	0.67	**0.85**	0.81
Threshold	0.274	0.260	0.243
$\overline{\Delta EAR}$	0.029	0.032	**0.027**

F

Model	Faceswap	MegaFS	Ours
r_x	0.383	0.197	**0.408**
r_y	0.476	0.252	**0.503**
r_{CCA}	0.601	0.423	**0.618**

H

Model	Faceswap	MegaFS	Ours
Precision	0.65	0.35	**0.89**
Recall	**0.92**	0.66	0.84
Threshold	0.225	0.225	0.215
$\overline{\Delta EAR}$	**0.028**	1.346	0.031

Figure 7. Model comparison. We qualitatively and quantitatively compared model performance under different settings and with other deepfake models. Original: original data; d–l–: output from the ablated model trained without mEBAL blink data and without landmark similarity block; d–l+: output from the model trained without mEBAL blink data, with landmark similarity block; d+l+: output from FaceMotionPreserve model trained with blink data and landmark similarity block. (**A**) Pearson and CCA correlation coefficients of facial dynamics between original/de-identified pairs. Augmented training data and landmark similarity module both helped preserve facial dynamics. (**B**) Video frames of a patient's blink. The d+l+ model output best matched the original. (**C**) Comparison of blink behaviors. The d–l+ improved blink detection accuracy and the d+l+ further improved eye closing completeness by lower eye blink threshold and smaller EAR deviation. (**D**) EAR curves of a patient. (**E**) Comparison of tremor spectrogram deviation of a patient. From left to right: d–l–, d–l+, d+l+ with mean deviation to the original spectrogram 0.106, −0.078, 0.003 in 4–6 Hz, respectively. (**F–I**) Comparison to Faceswap and MegaFS..

We compared FaceMotionPreserve with other deepfake methods (Faceswap, a subject-specific auto-encoder model, and MegaFS, a subject-agnostic GAN model) on facial movement preservation performance. We conducted the evaluation on PD dataset II. Both FaceMotionPreserve and Faceswap consistently converted original faces to de-identified faces, whilst MegaFS sometimes encountered failure due to its noise-sensitive generator. MegaFS's occasional failures caused apparent difficulty in facial landmark detection and disruption in both general facial dynamics correlation (lowest Pearson's r among three models: $r_x = 0.197$, $r_y = 0.252$, $r_{CCA} = 0.423$, Fig. 7F) and blink detection (largest EAR deviation 1.346 and lowest precision 0.35 and recall 0.66, Fig. 7H). FaceMotionPreserve demonstrated superior performance in facial dynamics preservation ($r_x = 0.408$,$r_y = 0.503$,$r_{CCA} = 0.618$) compared to Faceswap ($r_x = 0.383, r_y = 0.476, r_{CCA} = 0.601$) and MegaFA (($r_x = 0.197$, $r_y = 0.252$, $r_{CCA} = 0.423$) as shown in Fig. 7F. Although FaceMotionPreserve had higher EAR deviation (0.031) than Faceswap (0.028), it followed blinks better and captured intermediate blink status (Fig. 7G &I). FaceMotionPreserve had precision of 0.89 and recall of 0.84, which is more balanced. Faceswap had low precision of 0.65 and much higher recall of 0.92 due to too many false blinks from de-identified faces, indicating unsteadiness in eye movement reconstruction. Thus, FaceMotionPreserve outperformed Faceswap and MegaFS on facial movement preservation in general.

Discussion

Telemedicine delivers more convenient and accessible service via connecting patients and healthcare providers remotely. It frees the requirements of onsite visits at clinics, saves time and financial expenses, and promotes treatment compliance, with particular benefits to patients with movement disorders or cognitive impairments[50]. The privacy concerns on video streams could discourage patients to get involved. In this work, we developed FaceMotionPreserve , a generative model-base approach, to preserve medical information in de-identified patient videos for telemedicine. FaceMotionPreserve disentangles facial behavior from facial identity and preserves facial movements.

We validated FaceMotionPreserve for de-identification and built a comprehensive analysis framework to evaluate its capability of diagnostic information preservation. Our analyses revealed that the detected features (blinks, tremor, general facial movement features and body keypoints) had high similarity before and after de-identification, and the clinicians' ratings were largely consistent. Compared to other deepfake models, FaceMotionPreserve is more robust on generating faces and preserving facial movements. Ablation studies demonstrated improved facial movement correlation with our novel landmark similarity module and proposed loss function, and improved eye movement reconstruction with augmented and blink-balanced training dataset. FaceMotionPreserve reconstructed natural face images with rich information that could retain data usage for future use. We also deployed the approach into a real-time face identity changing system, which could easily be integrated to telemedicine software as a de-identification module.

Besides data usability, we also carefully validated FaceMotionPreserve for privacy protection. We examined human participants' perception of de-identified faces, which proved to be recognized more as the source face other than the original face. Although 1.7% identifications mapped de-identified face to the original face, it's worth noting that the choices were made within very limited options (six faces) and the re-identification rate was far below chance level. In real world, human identities are enormous and the original identity hides far from the vast face collection span around the de-identified face, leaving very little chance to re-identify the original face. Worries might arise if powerful automatic face recognition tools could unearth the hidden original identity and threaten privacy. Our analysis revealed that the representative face recognition tool couldn't match the de-identified face back to the original one, thereby assuring privacy protection against machine recognition attacks.

Facial identity is the most direct clue of personal identity and is one of the most sensitive privacy information, but eliminating facial identity alone can not guarantee absolute anonymous. Besides face, people might be recognized by gait, voice and other factors. However, it is unpractical to do thorough anonymization: some identifiers (like gait feature) are the core information to be evaluated under clinical protocol. There is a trade-off between information completeness and strength of privacy protection. Since other face obfuscation methods like blurring has been widely accepted, we believe FaceMotionPreserve offers adequate protection of privacy.

Our model has limitation on static image frame inputs: each single frame is a static snapshot of face, which might impair dynamical information preservation. However, it is interesting that the dynamical movements are kept by this image-based model. It suggests that FaceMotionPreserve generates rather identical facial expressions from still faces so that the resulting sequence still preserves facial dynamics. Video-based models may work better in preserving dynamical information in video-based telemedicine and might help resolve the problem of slightly higher hypomimia score. Another problem is heavy computing budget for running FaceMotionPreserve on real-time. Although FaceMotionPreserve could run on personal laptops once trained, it requires fairly intense computing resource for training and could not run fluently on some mobile phones. As mobile devices become more prevalent, it is crucial to empower users to secure identity information at their own hands. Hence, future work might compress the model to run on-device while retaining performance. FaceMotionPreserve mainly focuses on the preservation of facial movements, while some other facial symptoms caused by conditions such as jaundice and dermatologic lesions are not included. Future work could explore the potential of the model for retaining non-movement features that are not privacy-sensitive. Lastly, FaceMotionPreserve was tested only on patients with Parkinson's disease and lacks validation on other facial movement disorders. We believe that the model is not confined to PD, because arbitrary facial movements could be decomposed into linear and non-linear combinations of facial landmark movements, which have been shown highly similar before and after de-identification. Yet adapting FaceMotionPreserve for other conditions needs a closer look on different characteristic symptoms and might require extra informative data, adaptive loss or model regularization.

FaceMotionPreserve provides powerful privacy protection for telemedicine that preserves rich medical information from faces. Changing facial identity provides strong impression of privacy protection that encourages data acquisition from participants, and paves a new way for secured telemedicine. In the future, more efficient and effective methods that adopt lighter model and absorb temporal information could further improve privacy protection in telemedicine.

Data availability

The datasets used and analyzed during the current study and the FaceMotionPreserve model are available on reasonable request.

Received: 12 January 2024; Accepted: 18 July 2024
Published online: 27 July 2024

References

1. HIPAA. *The HIPAA Privacy Rule*. *https://www.hhs.gov/hipaa/for-professionals/privacy/index.html*. Accessed 19 Sep 2022 (2022).
2. Sun, Y. Chen, Y. Wang, X. & Tang, X. Deep learning face representation by joint identification-verification. In *Proceedings of the International Conference on Neural Information Processing Systems, 1988-1996* (MIT Press, 2014).
3. Marin, A. Telemedicine takes center stage in the era of COVID-19. *Science* **4**, 731–733 (2020).
4. Schünke, L. C. *et al.* A rapid review of machine learning approaches for telemedicine in the scope of COVID-19. *Artif. Intell. Med.* **129**, 102312. https://doi.org/10.1016/j.artmed.2022.102312 (2022).
5. Vodrahalli, K. *et al.* Development and clinical evaluation of an artificial intelligence support tool for improving telemedicine photo quality. *JAMA Dermatol.* **159**, 496–503. https://doi.org/10.1001/jamadermatol.2023.0091 (2023).
6. Teo, Z. L. & Ting, D. S. W. AI telemedicine screening in ophthalmology: Health economic considerations. *Lancet Glob. Health* **11**, e318–e320 (2023).
7. Liu, H. *et al.* Economic evaluation of combined population-based screening for multiple blindness-causing eye diseases in China: A cost-effectiveness analysis. *Lancet Glob. Health* **11**, e456–e465 (2023).
8. Liu, Z. Luo, P. Wang, X. & Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3730–3738 (IEEE, 2015).
9. Cowen, A. S. *et al.* Sixteen facial expressions occur in similar contexts worldwide. *Nature* **589**, 251–257 (2021).
10. Goetz, C. G. *et al.* Movement disorder society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results. *Mov. Disord.* **23**, 2129–2170 https://doi.org/10.1002/mds.22340. https://movementdisorders.onlinelibrary.wiley.com/doi/pdf/10.1002/mds.22340 (2008).
11. Yang, K. Yau, J. Li, F.-F. Deng, J. & Russakovsky, O. A study of face obfuscation in ImageNet. In *Proceedings of the International Conference on Machine Learning* (PMLR, 2022).
12. Shen, Z. Lai, W.-S. Xu, T. Kautz, J. & Yang, M.-H. Deep semantic face deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2018).
13. Chen, Y. Tai, Y. Liu, X. Shen, C. & Yang, J. FSRNet: End-to-end learning face super-resolution with facial priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2492–2501 (IEEE, 2018).
14. Gafni, O. Wolf, L. & Taigman, Y. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9378–9387 (IEEE, 2019).
15. Mahajan, S. Chen, L.-J. & Tsai, T.-C. SwapItUp: A face swap application for privacy protection. In *2017 IEEE 31st International Conference on Advanced Information Networking and Applications*. 46–50 (IEEE, 2017).
16. Gross, R. Sweeney, L. Cohn, J. Torre, F. D. l. & Baker, S. Face de-identification. In *Protecting Privacy in Video Surveillance*. 129–146 (Springer, 2009).
17. Perov, I. *et al. Deepfacelab: Integrated, Flexible and Extensible Face-Swapping Framework*. arXiv preprint arXiv:2005.05535 (2020).
18. Torzdf & Andenixa. *Faceswap*. https://github.com/deepfakes/faceswap. Accessed 30 May 2022 (2019).
19. Zhu, B. Fang, H. Sui, Y. & Li, L. Deepfakes for medical video de-identification: Privacy protection and diagnostic information preservation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 414–420 (ACM, 2020).
20. Naruniec, J. Helminger, L. Schroers, C. & Weber, R. M. High-resolution neural face swapping for visual effects. Comput. Graph. Forum **39**, 173–184. https://doi.org/10.1111/cgf.14062. https://onlinelibrary.wiley.com/doi/pdf/10.1111/cgf.14062 (2020).
21. Peng, B., Fan, H., Wang, W., Dong, J. & Lyu, S. A unified framework for high fidelity face swap and expression reenactment. *IEEE Trans. Circuits Syst. Video Technol.* **32**, 3673–3684 (2021).
22. Gafni, G. Thies, J. Zollhofer, M. & Nießner, M. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8649–8658 (2021).
23. Guo, Y. *et al.* AD-NeRF: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5784–5794 (2021).
24. Nirkin, Y. Keller, Y. & Hassner, T. FSGAN: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7184–7193 (2019).
25. Li, L. Bao, J. Yang, H. Chen, D. & Wen, F. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5074–5083 (IEEE, 2020).
26. Chen, R. Chen, X., Ni B. & Ge, Y. SimSwap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2003–2011 (ACM, 2020).
27. Zhu, Y. Li, Q. Wang, J. Xu, C. & Sun, Z. One shot face swapping on megapixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4834–4844 (2021).
28. Chen, Y. Hao, H. Chen, H. Tian, Y. & Li, L. The study on a real-time remote monitoring system for Parkinson's disease patients with deep brain stimulators. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 1358–1361 (IEEE, 2014).
29. Vedam-Mai, V. *et al.* Proceedings of the eighth annual deep brain stimulation think tank: Advances in optogenetics, ethical issues affecting DBS research, neuromodulatory approaches for depression, adaptive neurostimulation, and emerging DBS technologies. *Front. Hum. Neurosci.* **169** (2021).
30. Wei, Y. Zhu, B. Hou, C. Zhang, C. & Sui, Y. Interactive video acquisition and learning system for motor assessment of Parkinson's disease. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. 5024—5027 (International Joint Conferences on Artificial Intelligence Organization, 2021).
31. Huang, X. & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the International Conference on Computer Vision*. 1501–1510 (IEEE, 2017).
32. Deng, J. Guo, J. Niannan, X. & Zafeiriou, S. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019).
33. Wang, T.-C. et al. High-resolution image synthesis and semantic manipulation with conditional GANs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8798–8807 (IEEE, 2018).
34. Dong, X. et al. Supervision-by-registration: An unsupervised approach to improve the precision of facial landmark detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 360–368. https://doi.org/10.1109/CVPR.2018.00045 (IEEE, 2018).
35. Gulrajani, I. Ahmed, F. Arjovsky, M. Dumoulin, V. & Courville, A. C. Improved training of Wasserstein GANs. *Adv. Neural Inf. Process. Syst.* **30** (2017).
36. Cao, Q. Shen, L. Xie, W. Parkhi, O. M. & Zisserman, A. VGGFace2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*. 67–74 (IEEE, 2018).
37. Daza, R. Morales, A. Fierrez, J. & Tolosana, R. mEBAL: A multimodal database for eye blink detection and attention level estimation. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 32–36 (ACM, 2020).
38. Hernandez-Ortega, J. Daza, R. Morales, A. Fierrez, J. & Ortega-Garcia, J. edBB: Biometrics and behavior for assessing remote education. In AAAI Workshop on Artificial Intelligence for Education (ACM, 2019).
39. Bentivoglio, A. R. *et al.* Analysis of blink rate patterns in normal subjects. *Mov. Disord.* **12**, 1028–1034 (1997).
40. Caffier, P. P., Erdmann, U. & Ullsperger, P. Experimental evaluation of eye-blink parameters as a drowsiness measure. *Eur. J. Appl. Physiol.* **89**, 319–325 (2003).

41. Dufour, N. *et al. DeepFakes Detection Dataset by Google & JigSaw*.
42. Rössler, A. *et al.* FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the International Conference on Computer Vision* (IEEE, 2019).
43. Guo, J. Deng, J. Lattas, A. & Zafeiriou, S. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations* (2021).
44. Deng, J. Guo, J. Liu, T. Gong, M. & Zafeiriou, S. Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the European Conference on Computer Vision*. 741–757 (Springer, 2020).
45. Maze, B. *et al.* IARPA Janus Benchmark-C: Face dataset and protocol. In *International Conference on Biometrics*. 158–165 (IEEE, 2018).
46. Soukupova, T. & Cech, J. Real-time eye blink detection using facial landmarks. In *Proceedings of the 21st Computer Vision Winter Workshop* (Slovenian Pattern Recognition Society, 2016).
47. Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E. & Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 172–186 (2019).
48. Lin, T.-Y. *et al.* Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. 740–755 (Springer, 2014).
49. McGurk, H. & MacDonald, J. Hearing lips and seeing voices. *Nature* **264**, 746–748 (1976).
50. Zhang, J. *et al.* Implementation of a novel bluetooth technology for remote deep brain stimulation programming: The pre- and post-COVID-19 Beijing experience. *Mov. Disord.* (2020).

Acknowledgements

This study was funded by the National Key R &D Program of China (2021YFE0111800), the National Natural Science Foundation of China (No. 62206152), Beijing Municipal Science and Technology Program (Z211100003521006), STI 2030 - Major Projects (2022ZD0209400), and Shuimu Tsinghua Scholar Program (2022SM143).

Author contributions

B.Z., C.Z., Y.S. and L.L conceptualized the research, B.Z., C.Z., Y.S. and L.L developed the methods, B.Z., C.Z., Y.S. conducted the investigation, B.Z. visualized the results, Y.S. and L.L. supervised the research. B.Z. wrote the original draft. All authors reviewed and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.S.

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

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