Received: 26 July 2023 | Revised: 9 April 2024 | Accepted: 6 May 2024

DOI: 10.1002/hbm.26721

TECHNICAL REPORT

WILEY

Fast refacing of MR images with a generative neural network lowers re-identification risk and preserves volumetric consistency

Nataliia Molchanova[1,2,3,4,5] | Bénédicte Maréchal[1,4,5] | Jean-Philippe Thiran[1,5] | Tobias Kober[1,4,5] | Till Huelnhagen[1,4,5] | Jonas Richiardi[1,3] | the Alzheimer's Disease Neuroimaging Initiative

[1]Department of Radiology, Lausanne University Hospital (CHUV), Lausanne, Switzerland

[2]Institute of Informatics, University of Applied Sciences and Arts of Western Switzerland (HES-SO), Sierre, Switzerland

[3]Faculty of Biology and Medicine, University of Lausanne (UNIL), Lausanne, Switzerland

[4]Advanced Clinical Imaging Technology, Siemens Healthineers International AG, Lausanne, Switzerland

[5]Laboratory of Signal Processing 5, Ecole Polytechnique Fédérale de Lausanne, (EPFL), Lausanne, Switzerland

Correspondence
Nataliia Molchanova, Department of Radiology, Lausanne University Hospital (CHUV), Lausanne, Switzerland.
Email: nataliia.molchanova@unil.ch

Abstract

With the rise of open data, identifiability of individuals based on 3D renderings obtained from routine structural magnetic resonance imaging (MRI) scans of the head has become a growing privacy concern. To protect subject privacy, several algorithms have been developed to de-identify imaging data using blurring, defacing or refacing. Completely removing facial structures provides the best re-identification protection but can significantly impact post-processing steps, like brain morphometry. As an alternative, refacing methods that replace individual facial structures with generic templates have a lower effect on the geometry and intensity distribution of original scans, and are able to provide more consistent post-processing results by the price of higher re-identification risk and computational complexity. In the current study, we propose a novel method for anonymized face generation for defaced 3D T1-weighted scans based on a 3D conditional generative adversarial network. To evaluate the performance of the proposed de-identification tool, a comparative study was conducted between several existing defacing and refacing tools, with two different segmentation algorithms (FAST and Morphobox). The aim was to evaluate (i) impact on brain morphometry reproducibility, (ii) re-identification risk, (iii) balance between (i) and (ii), and (iv) the processing time. The proposed method takes 9 s for

Till Huelnhagen and Jonas Richiardi have an equal contribution.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

face generation and is suitable for recovering consistent post-processing results after defacing.

KEYWORDS
brain morphometry, conditional generative adversarial networks, defacing, de-identification, magnetic resonance imaging, privacy, re-identification risk

1 | INTRODUCTION

Due to its superior soft tissue contrast, magnetic resonance imaging (MRI) has become the modality of choice for clinical brain imaging. MRI is widely used for the diagnosis and monitoring of various diseases, such as dementia, Alzheimer's disease, multiple sclerosis, brain cancer, and others. It plays an equally important role in research on these diseases and more broadly in neuroscience.

The field of view of MRI head scans typically includes the subject face, which raised privacy concerns about sharing this data already a decade ago and have been gaining increasing attention in recent years (Mazura et al., 2012; Prior et al., 2009; Schwarz et al., 2019). This is in part due in part to the availability of automated face recognition software that achieves very high accuracy (Deepface python library, 2021; National Institute of Standards and Technology, 2021). Indeed, identification of a person from a routine head MR scan was shown to be a feasible task (Abramian & Eklund, 2019; Christopher, 2022; Schwarz et al., 2019).

While the research environment becomes more demanding in terms of providing open access to the data, for example, following FAIR principles (Wilkinson et al., 2016), the possibility of subject re-identification might collide with privacy regulations. Currently, data privacy protection regulations often require removal of any identifiable features from the data to avoid the possibility of mapping particular people with certain diseases; in some cases, such rulings also imply facial de-identification (U.S. Department of Health and Human Services Office for Civil Rights, 2013).

Various techniques were proposed for face de-identification, which are split between three approaches: (i) defacing, that is, completely or partially removing facial features (Cox, 1996; Cox & Hyde, 1997; Gulban et al., 2022); (ii) refacing, that is, changing the facial features (Mikulan et al., 2021) or defacing and inserting a new face (Cox, 1996; Cox & Hyde, 1997; Huelnhagen et al., 2020; Schwarz et al., 2021); (iii) blurring using spatial filters (Milchenko & Marcus, 2012).

Despite the availability of de-identification tools and respective studies, the question of their impact on the outcomes of downstream image processing remains unclear. Potentially inconsistent post-processing results can be caused for example, by a failure of a skull-stripping procedure, which is usually a part of in brain segmentation software to solely process brain tissue in subsequent steps, and can be sensitive to either intensity histogram changes or head shape deformations (Kalavathi & Surya Prasath, 2015). Also, image registration steps, which are part of many image analysis workflows, are sensitive to image alterations induced by image anonymization techniques. Defacing is claimed to provide the best privacy protection among (i–iii). However, different studies that comparing various image de-identification tools report that defacing can alter the results of brain tissue segmentation, subcortical segmentation, cortical thickness estimation, or atrophy estimation, yet deriving different conclusions about significance of this impact (Bhalerao et al., 2022; de Sitter et al., 2020; Gao et al., 2022; Gao et al., 2023; Huelnhagen et al., 2020; Rubbert et al., 2022; Schwarz et al., 2021; Theyers et al., 2021).

In comparison to defacing, the refacing mitigates the impact on image post-processing results, yet sufficiently protects from re-identification of individuals (Huelnhagen et al., 2020; Mikulan et al., 2021; Schwarz et al., 2021). First, the refacing has a smaller effect on the intensity distribution of an image, as compared to defacing. Second, refaced images show realistic facial features that are closer to the shape-related assumptions built-in within downstream processing steps such as brain extraction or image registration (see Figure 1). While blurring causes the smallest alteration to image intensity distribution, it may also be insufficient in terms of de-identification, as the possibility to reconstruct facial features from blurred scans was previously shown (Abramian & Eklund, 2019). Therefore, refacing must achieve a trade-off between destructing or altering the original image data and the residual re-identification potential.

Existing refacing solutions comprise multi-step processing pipelines using common processing tools, like FreeSurfer, ANTS, AFNI, and other. Using population-average templates of faces for conducting refacing is a common solution (Cox, 1996; Cox & Hyde,1997; Schwarz et al., 2021). Face replacement with a template, requires defacing, correct registration and additional contrast adjustments of the template face. The *Anonymi* (Mikulan et al., 2021) tool uses a different approach from population-average templates that includes reconstruction of skin and skull surfaces with *watershed* algorithm, determining potentially identifiable areas and filling the space between skull and skin within identifiable areas with random values. Based on this implementation, the outputs provided by *Anonymi* are closer to facial blurring.

Despite the promise of existing refacing techniques, we hypothesize that conditional Generative adversarial networks (cGAN) (Goodfellow et al., 2014) constitute a more suitable basis for a de-identification tool. cGANs are used for conditional generative modeling in deep learning and have been widely explored for image-to-image translation tasks in medical imaging (Yi et al., 2019). In

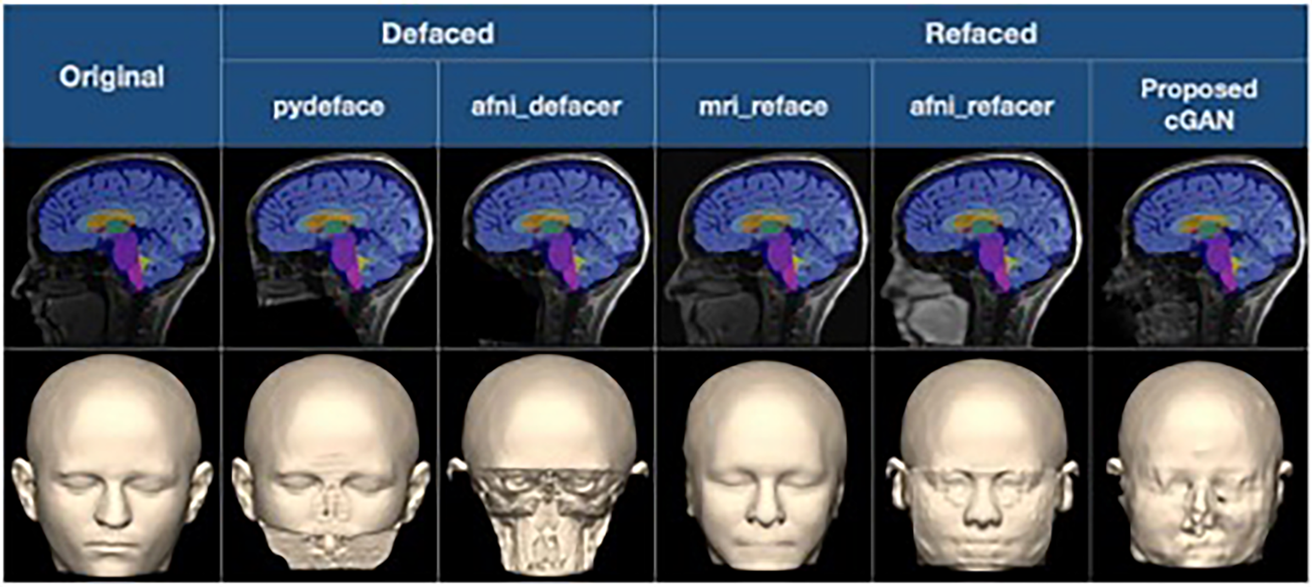

FIGURE 1 Examples of the de/refacing techniques on one subject obtained on in-house data. More examples of the refacing with cGAN on the in-domain ADNI data are available in Appendix F.

application to the refacing task, they are able to avoid multi-step processing pipelines that require sufficient resources for parallel processing, might be computationally greedy and exhibit potential points of failure in each processing step. Instead, a cGAN can simultaneously take care of the de-identification and factors contributing to consistent post-processing within an inference. Internal noise injection via dropout or noise layers contribute to de-identification of faces, while the adversarial loss may provide similarity of faces necessary for consistent post-processing. A previous study (Huelnhagen et al., 2020) proposed a 2D *pix2pix* cGAN for the generation of a random face on defaced images. This approach was shown to increase the reproducibility of volumetric brain measurements compared to original, non-deidentified images. Nevertheless, this study did not assess the re-identification risk after refacing, and performed the assessment of the morphometry results consistency in comparison to only one defacing tool.

In this work, we propose a novel solution based on a 3D cGAN for fast and effective refacing of defaced T1-weighted (T1w) MR images. We developed a methodology to assess the performance of the proposed technique that includes (i) evaluation of the impact on post-processing results, using the example of volumetric brain measurements obtained with the FSL's FAST and FIRST (Jenkinson et al., 2012) as well as MorphoBox (Schmitter et al., 2014) brain segmentation tools; (ii) approximation of the re-identification risk using modern face recognition software; (iii) assessment of the trade-off between (i) and (ii); (iv) estimation of the required processing time. Moreover, we perform a similar assessment for several common defacing and refacing software to understand where the proposed tool stands in comparison to existing solutions for de-identification. The compared de-identification tools include: *pydeface* (Gulban et al., 2022) and *afni _refacer* (both in defacing and refacing modes) (Cox, 1996; Cox & Hyde, 1997) and *mri_reface* (Schwarz et al., 2021). The workflow diagram summarizing the conducted experiments is presented in Figure 2. The results of the comparative study show that the proposed technique achieves a comparable performance in terms of (i) and (ii) among other refacing tools. However, the use of the proposed cGAN becomes feasibly beneficial for face inpainting on defaced images, as it can recover consistent post-processing results while being orders of magnitude faster than existing tools.

2 | MATERIALS AND METHODS

2.1 | Data

Seven hundred thirty-eight 3D T1-weighted scans were taken from the TADPOLE dataset (TADPOLE challenge constructed by the EuroPOND consortium, 2012-2024), a subset of Alzheimer's Disease Neuroimaging Initiative (ADNI) data (using ADNI-3-imaging protocol reported in Gunter et al. (2017), including two sessions for 185 patients with mild cognitive impairment (MCI) or Alzheimer's disease (AD) and 184 healthy controls. Patients were scanned on 1.5 T (90% of the data) or 3.0T scanners from Siemens Healthcare (Erlangen, Germany) or GE Healthcare (Chicago, Illinois, United States). The age of the control subjects ranges from 60 to 90 years with mean and standard deviation of 77.5 ± 5.4 years; patient ages span from 56 to 93 years, with mean and standard deviation of 76.7 ± 7.2 years. One hundred twenty patients have confirmed diagnoses of mild cognitive impairment, 65 patients—Alzheimer disease. Additional information is listed in Table 1.

All scans were divided into a training, a validation and a test set in the proportion of 200:8:530 and stratified the prevalence of patients

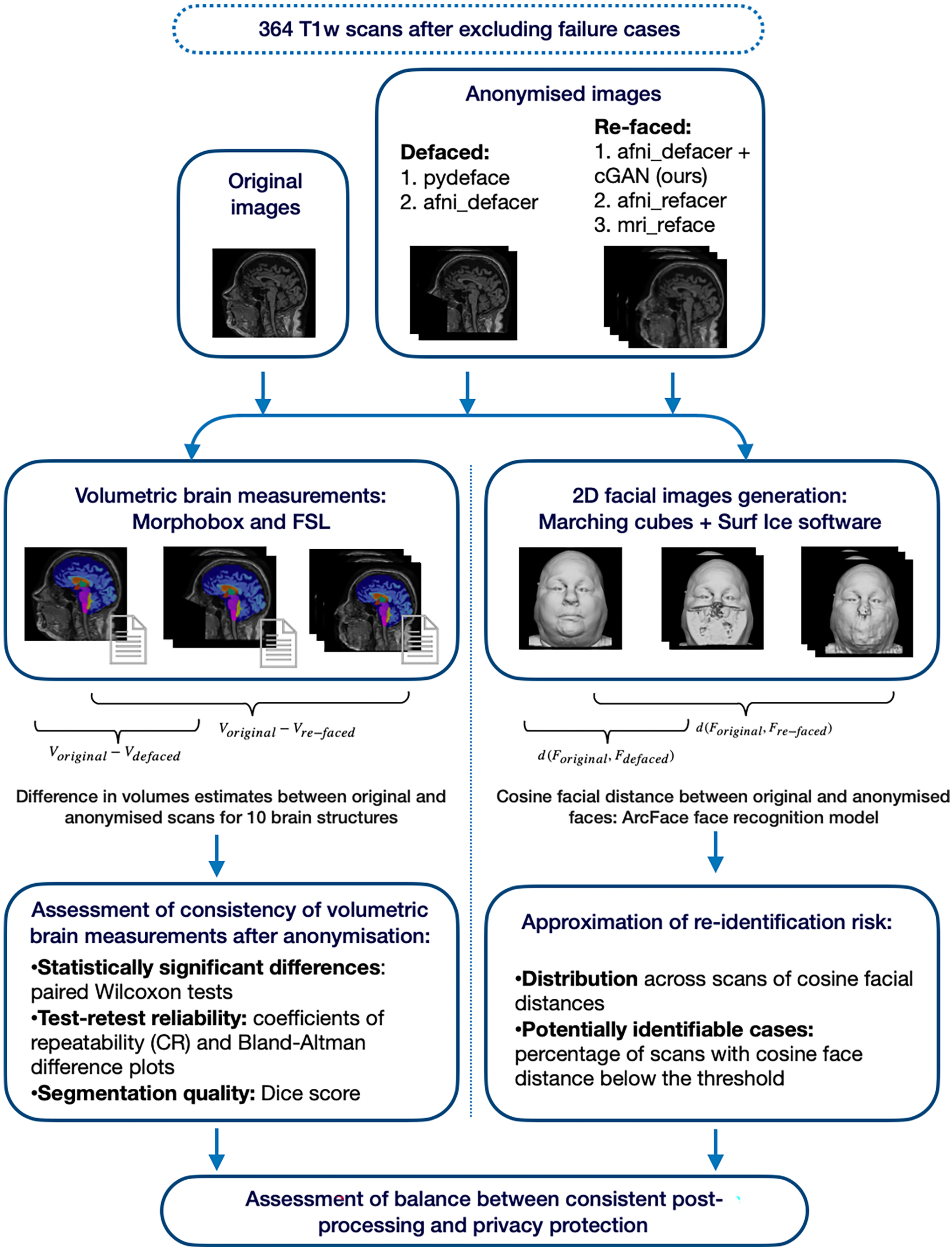
364 T1w scans after excluding failure cases
Original images
Anonymised images
Defaced:
1. pydeface
2. afni_defacer
Re-faced:
1. afni_defacer + cGAN (ours)
2. afni_refacer
3. mri_reface
Volumetric brain measurements: Morphobox and FSL
2D facial images generation: Marching cubes + Surf Ice software
$V_{original} - V_{defaced}$
$V_{original} - V_{re-faced}$
$d(F_{original}, F_{defaced})$
$d(F_{original}, F_{re-faced})$
Difference in volumes estimates between original and anonymised scans for 10 brain structures
Cosine facial distance between original and anonymised faces: ArcFace face recognition model
Assessment of consistency of volumetric brain measurements after anonymisation:
•Statistically significant differences: paired Wilcoxon tests
•Test-retest reliability: coefficients of repeatability (CR) and Bland-Altman difference plots
•Segmentation quality: Dice score
Approximation of re-identification risk:
•Distribution across scans of cosine facial distances
•Potentially identifiable cases: percentage of scans with cosine face distance below the threshold
Assessment of balance between consistent post-processing and privacy protection

FIGURE 2 Summary of experiments conducted within the current study. In addition to the presented experiments the processing time of each de-identification tool was evaluated.

TABLE 1 Dataset composition.

Scanner manufacturer	Field strength	M:F ratio	Control	AD	MCI	Total
				Number of scans		
Siemens Healthcare	1.5T	0.93	186	58	110	354
	3.0T	0.71	29	10	14	53
GE Healthcare	1.5T	1.29	136	60	116	312
	3.0T	0.9	17	2	0	19
Summary		1.05	368	130	240	738

vs. controls in each set. Both sessions belonging to one subject were put into the same set. The training and validation sets were used for training the proposed 3D cGAN architecture and the test set was used for assessing the performance of the de-identification techniques. Datasets were compiled by stratification so that significant biases in age, sex or scanner manufacturer are avoided. See more details about the data distribution in the datasets in Appendix A.1. Subject identifiers and session numbers for each of the sets are provided in the Supplementary Materials.

Pre-processing. A re-orientation to the Anterior Superior Left (ASL) orientation was performed on all images if necessary to ensure consistent processing conditions.

2.2 | cGAN refacing model

2.2.1 | Architecture

The refacing task can be formulated as defacing followed by generation of a new face. The latter was previously implemented using population-average templates of faces followed by procedures for anonymized face registration, contrast adjustment and/or additional removal of identifiable features (Cox, 1996; Cox & Hyde, 1997; Schwarz et al., 2021). We propose a novel approach for face generation based on conditional Generative adversarial networks (cGANs) that does not require additional processing for face positioning or contrast adjustment and allows fast generation of a new anonymized face with possible speedup using a GPU.

cGANs are used in deep learning for generative modeling and allow learning a mapping between one domain of images to another domain. Here, we learn a mapping from the domain of defaced images to the domain of refaced images by training on input–output pairs consisting of images defaced by *afni_refacer* in defacing mode as input, and original non-anonymized images as output (target). Despite learning a mapping to the space of original faces, de-identification is provided by three factors: (i) inability to recover the original face from a properly defaced image (Abramian & Eklund, 2019), (ii) early stopping of the training process to limit similarity, (iii) random noise added during inference time via dropout layers (Isola et al., 2017). The proposed method will be further referred as *cGAN afni _defacer*.

For training, we only used scans defaced by *afni_refacer* in the "defacing mode." This tool together with *pydeface* (Gulban et al., 2022) has previously shown superior performance in terms of correct face removal in comparison to other techniques (Theyers et al., 2021). While *afni_refacer* and *pydeface* showed a comparable performance, *afni_refacer*'s defacer removes a wider variety of identifiable facial features. *pydeface*, for example, does not remove ears, and more importantly, it usually leaves parts of the eyes and the nose (see Figure 1). Since our goal here was to learn a mapping to the space of the original faces, the defaced images should not contain any identifiable facial feature or their parts as there is a chance that they will be recovered by the generator.

The proposed method is based on a *vox2vox* cGAN (Cirillo et al., 2021) developed within the context of medical imaging (Yi et al., 2019). Vox2vox architecture is a 3D analog of pix2pix. (Isola et al., 2017), where in addition to replacing the convolutional layers by the 3D analogs and decreasing the U-net depth, residual blocks with dropout layers are added to the bottleneck of the architecture. The network architecture is presented in Figure 3. Compared to pix2pix, vox2vox uses an L2 loss in the objective function instead of L1 aiming to promote smoothness instead of sharpness. Several works explored the usage of the L1.5 loss (Harms et al., 2019; Liu et al., 2020), which can be seen as a way to achieve a tradeoff between smoothness and sharpness of the generated images. We adopt a similar approach, observing a minor marginal improvement in the image similarity metrics during training, such as structural similarity index and peak signal-to-noise ratio.

2.2.2 | Training

Before sending the images to the network, both defaced and original images underwent similar pre-processing steps, namely (i) intensity thresholding with the value that approximately corresponds to the 80% percentile of maximum intensity values across scans in the training set, that is, Winsorising; (ii) linear intensity scaling to $[-1,1]$ using linear transform coefficients computed on the original images and applied to both original and defaced images; (iii) division into four sub-volumes of size $128 \times 128 \times 128$. Step (i) is required to remove outliers from the intensity distributions before linear scaling. The subdivision of images on sub-volumes in step (iii) aims at reducing the memory demand of the network. This step poses a limitation on the image size that cannot exceed $256 \times 256 \times 256$, however, typical head MRI matrix sizes rarely exceed this size and the subdivision step could also easily be adjusted to allow even larger images if needed.

The modified adversarial loss is defined as follows:

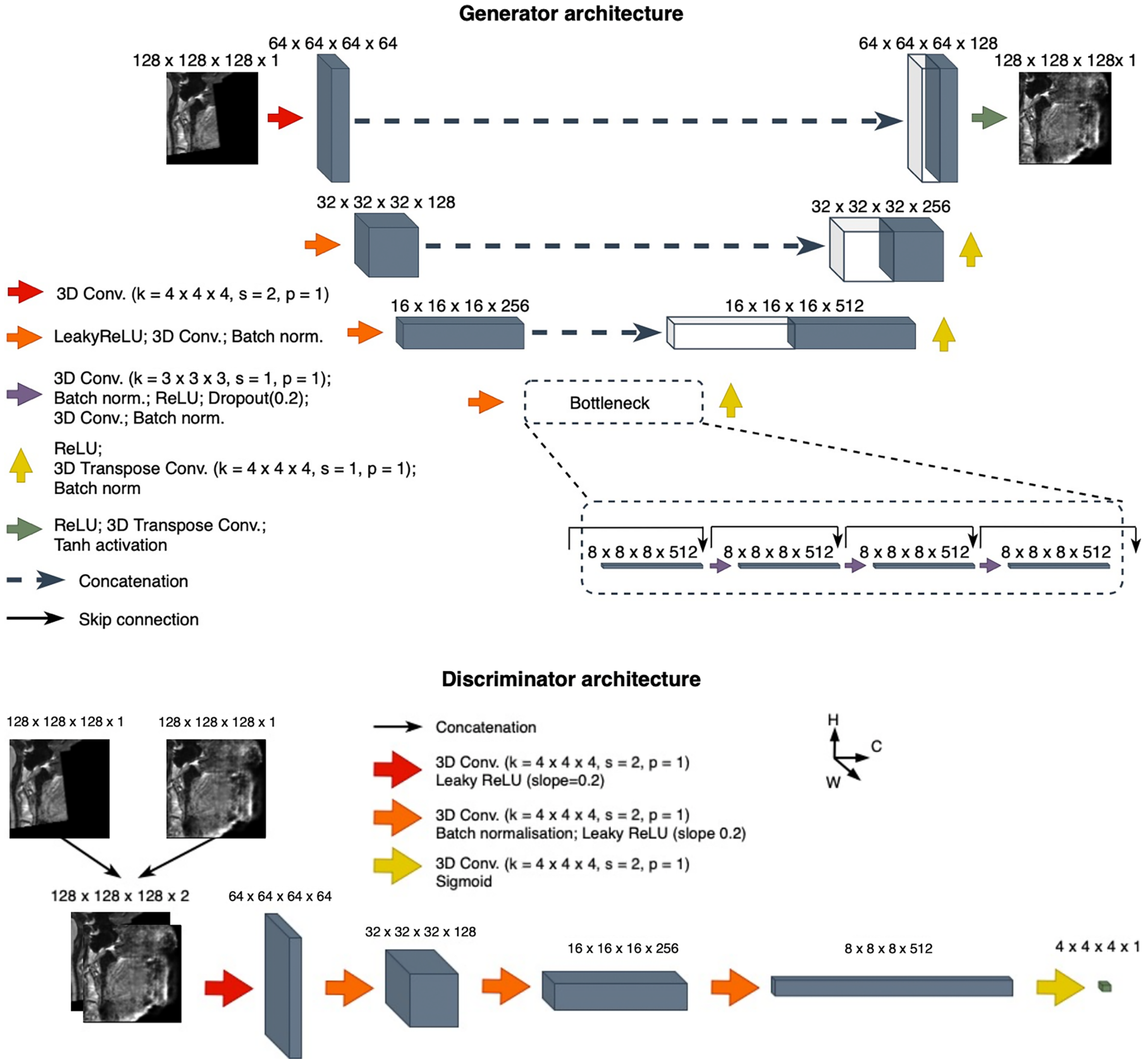

FIGURE 3 Overview of the proposed 3D cGAN including generator and discriminator architectures. 4D objects are visualized via 3D projections by collapsing the depth axis, hence only Height (H), width (W) and channels (C) axes are displayed.

$$\begin{aligned} V(D,G) &= L_{cGAN} + \lambda L_{L1.5} \\ &= E_{x,y}[\log D(\mathbf{x},\mathbf{y})] + E_{z,y}[\log(1 - D(\mathbf{y},G(\mathbf{z}|\mathbf{y})))] \\ &\quad + \lambda E_{x,y,z}\left[\|\mathbf{x} - G(\mathbf{z}|\mathbf{y})\|_{1.5}\right] \sim \min_G \max_D, \end{aligned} \tag{1}$$

where G: $\mathbf{z}$,$\mathbf{y}$ 7 $\rightarrow$ $\mathbf{x}$ is a 3D U-net generator that learns a mapping from the domain of defaced images to the domain of refaced images. The noise $\mathbf{z}$ is included in the form of dropout; $D(\mathbf{x},\mathbf{y})$ is a PatchGAN discriminator that takes as an input images from both domains and outputs the probability of $N \times N \times N$ patches of the generated image coming from the same distribution. λ is a parameter controlling the impact of the $L_{L1.5}$ term that was empirically chosen to be $\lambda = 0.015$, after experimenting on the validation dataset.

Training was performed for a total of 50 epochs with validation and weights saving on each seventh epoch and a cosine learning rate decay every 1000 iterations on an NVIDIA Tesla V100 GPU with 32GB of memory, starting with a learning rate of 0.0002. All other hyperparameters are provided in the accompanying code repository: https://gitlab.com/acit-lausanne/refacing-cgan.

2.2.3 | Inference details

During the pre-processing before the inference, defaced images underwent a preprocessing pipeline similar to the one used during training. It included (i) winsorising, (ii) linear intensity scaling to [−1,1] and (iii) division into four sub-volumes of size 128 × 128 × 128. During the inference, the dropout layers were switched on in order to provide random noise to the generator, enforcing better anonymization. After the inference, the sub-volumes were combined while averaging overlapping areas, intensities were re-scaled back using the transform coefficients saved during the pre-processing. Finally, the values within a face and air mask were copied to the defaced image to avoid unwanted changes in the brain, and additionally in the final image all non-zero values in the air mask were removed. Face and air mask were defined as the zero-values mask in the defaced image that underwent a 3D morphological closing. The air mask in the final image was formed by taking all values below 3 and applying a 3D morphological closing.

2.2.4 | Hyperparameter tuning

Tuning of the inference-time dropout probability values and the epoch number was done on the validation set by optimizing the trade-off between consistency of volumetric results and degree of privacy protection. For the trade-off assessment we use the method described in Section 2.3.4.

2.3 | Performance assessment

Analogously to previous studies on face de-identification for medical images (Mikulan et al., 2021; Schwarz et al., 2019; Schwarz et al., 2021), we consider two aspects to be of the greatest importance when assessing the performance of a face de-identification method, which are: (i) consistency of image post-processing results and (ii) degree of privacy protection after the method is applied. In order to narrow the task we concentrated our effort on quantification of the effect on image post-processing results using sub-cortical and cortical brain segmentation as an example of a commonly performed target application. Segmentation typically involves skull-stripping and other steps sensitive to the changes in the intensity distribution or head deformations (Kalavathi & Surya Prasath, 2015). While cortical thickness might be an interesting target to investigate, it has been previously shown that it is less affected by defacing compared to the volume estimates when using FreeSurfer segmentation (Buimer et al., 2021). Besides, as brain extraction is performed in the beginning of the cortical thickness estimation, we consider a total intracranial volume (TIV) as a proxy. Another aspect that helps judging about the applicability of techniques to real-world tasks is processing speed, which was also estimated.

We conducted a comparative study between the proposed refacing cGAN approach and several existing face de-identification tools, that include two defacing tools (*pydeface* (Gulban et al., 2022) and *afni_refacer* (Cox, 1996; Cox & Hyde, 1997) in defacing mode) and two refacing tools (*afni_refacer* in refacing mode and *mri_reface* (Schwarz et al., 2021)). The chosen defacing tools have previously been shown to have the highest accuracy in terms of correct removal of facial features and absence of alterations in brain voxels in comparison to other defacing tools (Theyers et al., 2021). The *afni_refacer de-identification* tool from AFNI (Cox, 1996; Cox & Hyde, 1997) provides the possibility of both defacing and refacing, and both modes were explored in the current study. They will be further referred to as *afni _defacer* and *afni_refacer*. *Mri_reface* was previously compared to existing defacing tools and has shown modestly lower effect on volumetric brain measurements while providing comparable re-identification risk (Schwarz et al., 2021). Both *mri _reface* and *afni_refacer* include a first defacing stage and the subsequent insertion of a population-average face. All de-identification tools differ in terms of defaced areas, anonymized face generation, and how they place the new face (see examples in Figure 1).

2.3.1 | Reproducibility of brain volumetry

We used two different tools to obtain volumetric brain measurements and segmentation maps: *FSL* (Jenkinson et al., 2012) *fsl_anat* pipeline[1] and the in-house developed research application *MorphoBox* (Schmitter et al., 2014). Both techniques perform cortical and subcortical brain segmentation, providing segmentation maps as outputs. Absolute volumes estimates are provided by *MorphoBox* while for *fsl_anat* they can be derived from the image's geometry and segmentation map.

The two techniques segment different sets of subcortical brain structures. Thus, for the analysis we used only bilateral volume estimates evaluated by both methods. The set of structures for which the volumetric estimates were analyzed include both large and small, cortical and subcortical ones. These are grey matter (GM), white matter (WM), total intracranial volume (TIV), thalamus, caudate, putamen, pallidum, hippocampus and amygdala.

Results obtained on the original images were considered as the ground truth and were further compared with the results obtained on scans anonymized by the different techniques. Comparison aimed at (i) detecting differences in absolute volumes, (ii) assessing test–retest repeatability and (iii) assessing segmentation quality. For (i), paired Wilcoxon tests were performed for the absolute volumes of each considered brain region, controlling false detection rate using the Benjamini-Hochberg multiple comparison procedure across regions, within each tool. For (ii), Bland–Altman plots and coefficients of repeatability (CR) were computed for the difference in absolute volumes of a brain region before and after face de-identification. For (iii) Dice scores, averaged across brain regions, were computed using segmentation maps before and after de-identification.

2.3.2 | Data exclusion criteria

In order to isolate the impact of failure of the different de-identification or brain segmentation tools several criteria were

[1] https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/fsl_anat

developed to detect those scans in order to exclude them from the evaluation. These criteria include:

C1) Changes of voxels values in TIV mask generated by HD-BET (Isensee et al., 2019) after de-identification: verify that values within the brain mask are not affected by face de-identification;
C2) Visual assessment of defaced scans with outlying changes in the frontal half of the image (containing the face): verification that defacing was performed in the correct part of the head or was performed at all;
C3) Visual assessment of the original images with outlying Dice values computed between the brain tissue segmentation masks (both from *fsl_anat* and *MorphoBox*) on the original and anonymized images: detection of original scans where brain segmentation algorithms failed on the original images, and thus can no longer be used as the gold standard.

The number of scans anonymized by different tools that do not pass C1) and C2) is an indicator of the (lack of) robustness of de-identification tools on the considered dataset.

2.3.3 | Re-identification risk

Most modern face recognition software works with 2D facial images and computes a distance between two faces. If the distance between two faces is lower than some decision threshold, faces are considered to belong to the same subject. Thus, the distances between faces before and after anonymization for each of the scans can be used as a proxy for re-identification risk. More specifically, we used the average and standard deviation of these “before/after distances,” as well as the percentage of potentially identifiable cases (percentage of same-subject before-after pairs with distances below the decision threshold), to quantify the re-identification risk.

As the dataset used does not have real-world photos of individuals, we use 2D facial images generated from the original non-anonymized MRI images as a proxy for real-world photos. To generate 2D facial images from 3D T1w scans we used the marching cubes algorithm (Lorensen & Cline, 1987) for mesh generation and the *Surf Ice* (2021) software for mesh visualization (see more details in Appendix C).

Since the generated face render images do not resemble real-world photos of faces, not all common face recognition models will work correctly on these images without additional tuning. We aimed at selecting a face recognition model that can recognize faces on the generated images without additional tuning. For this purpose, the performance of seven state-of-the-art pre-trained DL-based models (Deepface python library, 2021) were compared using 2D facial images generated on the training set. The compared models include VGG-Face, Google FaceNet, OpenFace, Facebook DeepFace, DeepID, ArcFace and Dlib. The best model was the one with the best separation between the classes of correct matches (two faces belonging to one subject, but different time points) and incorrect matches (two faces belonging to different subjects) in terms of cosine facial distance. After evaluation, ArcFace (Deng et al., 2021) showed superior ability in this aspect in comparison to other models. From the distribution of distance within classes of correct and incorrect matches, we were able to determine an approximate threshold separating similar and different faces. For a more detailed description of the model and threshold selection procedure, we refer to Appendix D.

The re-identification risk is inversely proportional to the distance between the original and anonymized faces. Hence, the mean of the inverse distances between the faces generated from the original and de/refaced images across all images was used as a single measure of the re-identification risk of an de-identification technique. Lower values of the average inverse distance correspond to lower re-identification risk.

2.3.4 | Trade-off between volumetric reproducibility and re-identification risk

An ideal face anonymization method would yield both high reproducibility of volumetry and low re-identification risk. To evaluate the trade-off between these aspects, we propose a balance plot which considers a single measure summarizing the performance of both aspects. In the plot, the x-axis represents the effect on volumetric brain measurements and the y-axis re-identification risk.

As a measure of the effect on volumetric estimates of a de-identification tool a repeatability coefficient over the normalized absolute volumes was used. For this, absolute volumes for each considered brain region for different subjects were linearly scaled to [0,1] range using *min* and *max* values from the original images. The difference in normalized volumes for all brain regions were jointly used for computation of CR. We will further refer to this measure as the normalized coefficient of repeatability (nCR). Lower values of nCR correspond to better reproducibility. As a measure of spread, we used standard deviation across CR values computed across images within different normalized brain volumes.

The re-identification risk is inversely proportional to the distance between the original and anonymized faces. Hence, the mean of the inverse distances between the faces generated from the original and de/refaced images across all images was used as a single measure of the re-identification risk of a de-identification technique.

Lower values of the average inverse distance correspond to lower re-identification risk.

2.3.5 | Processing time evaluation

The processing time of a de-identification tool can give an idea about its applicability to real-world tasks that may imply processing large datasets. For evaluation of per-scan processing time, time measurements were collected during processing of eight scans from the validation set by each tool.

The *3D cGAN* is the only technique that can be executed on a GPU, thus, measurements were collected both for GPU and CPU execution. An NVIDIA Tesla V100 was used for GPU execution, while for CPU execution two 12 core 24 thread Intel Xeon Gold 6126 CPU @ 2.60GHz were used. Another difference from the rest of the techniques is that the *cGAN afni_defacer* can be applied to the whole dataset with parallel pre-processing of scans, while the rest of the techniques assume single scans as an input. In this experiment we utilize this feature and do pre-processing before the inference in parallel on separate cores.

The rest of the techniques implement inner parallelism where the number of cores is not controlled by the user. Thus, for their inner parallelism all 48 threads from the two Intel Xeon Gold 6126 CPU @ 2.60GHz were available.

3 | RESULTS

3.1 | Failed cases analysis

Using the pre-defined criteria described in Section 2.3.2, out of the initial 530 scans only 364 scans were left for further analysis, meaning that 166 were excluded under different conditions. Details on the number of excluded scans based on different criteria are listed in Table 2. Most scans (143 of 165) were excluded because of *pydeface* failure. These failures are caused either by *pydeface* cutting frontal parts of the brain or because defacing was performed in the incorrect part of the head. The characteristic examples of the different types of failures are listed in the Appendix A.2.

3.2 | Reproducibility of volumetry

Results of statistical testing are shown in Table 3 as *p*-values obtained from Wilcoxon paired tests to detect statistically significant differences in volumetric results before and after de-identification. Overall, for *fsl_anat* most regions (7–9 out of 10) were significantly different independent of which face de-identification tool was used. For *MorphoBox* there was more variation across the different de-identification methods, with *afni_refacer* and *cGAN afni_defacer* showing the lowest number of different regions (0 and 1 out of 10 respectively). *Afni_defacer* showed the greatest number of structures where statistically significant differences were detected, which is 7 out of 10 structures.

Results of the reproducibility assessment in a form of CRs computed for the differences between absolute volume estimates before and after de-identification are shown in Table 4. Briefly, *mri reface* and *cGAN afni _defacer shows* the overall best reproducibility having the lowest and comparable CRs both for *fsl_anat* and *MorphoBox* brain segmentation results. *Afni_refacer* has higher CRs in all brain structures, however, in most of the brain structures the CRs computed

TABLE 2 Number of failed cases in the test dataset chosen according to criteria described in section.

Criteria	Original	pydeface	afni_defacer	afni_refacer	mri_reface	cGAN afni_defacer
C1	-	64	0	0	9	0
C2	-	79	0	-	-	-
C3	14	-	-	-	-	-

TABLE 3 Corrected *p*-values and number of brain regions with significantly different absolute volume (paired Wilcoxon tests on results based on original scans and scans anonymized by different tools).

	pydeface		afni_defacer		afni_refacer		mri_reface		cGAN afni_defacer	
Brain regions	FSL	MB	FSL	MB	FSL	MB	FSL	MB	FSL	MB
TIV	1×10^{-9}	6×10^{-10}	4×10^{-9}	0.16	0.17	0.14	0.01	1×10^{-17}	0.47	4×10^{-8}
CSF	3×10^{-11}	0.42	8×10^{-4}	1×10^{-8}	1×10^{-21}	0.42	0.23	9×10^{-5}	2×10^{-5}	0.14
GM	5×10^{-19}	0.04	6×10^{-5}	3×10^{-6}	3×10^{-34}	0.95	0.01	0.07	2×10^{-21}	0.17
WM	0.16	2×10^{-5}	8×10^{-4}	0.06	1×10^{-51}	0.15	1×10^{-9}	2×10^{-4}	5×10^{-22}	0.13
Thalamus	1×10^{-5}	0.83	7×10^{-14}	0.07	2×10^{-55}	0.56	4×10^{-17}	0.71	2×10^{-38}	0.67
Caudate	0.01	0.93	1×10^{-3}	6×10^{-9}	0.54	0.15	0.38	0.03	0.40	0.81
Putamen	6×10^{-8}	0.13	9×10^{-9}	1×10^{-4}	3×10^{-8}	0.14	4×10^{-4}	0.03	2×10^{-5}	0.13
Pallidum	1×10^{-4}	0.83	9×10^{-6}	1×10^{-4}	5×10^{-11}	0.14	0.04	0.01	4×10^{-7}	0.24
Hippocampus	0.18	0.02	0.81	3×10^{-7}	2×10^{-10}	0.95	0.01	0.26	0.26	0.13
Amygdala	0.10	0.93	3×10^{-7}	1×10^{-15}	1×10^{-20}	0.15	7×10^{-4}	0.43	1×10^{-14}	0.13
# significant	7	4	9	7	8	0	8	6	7	1

Note: Brain volumetry was performed with *fsl_anat* (FSL) and with *MorphoBox* (MB).

TABLE 4 Coefficients of repeatability obtained for absolute volumes in milliliters estimated on original scans and scans anonymized by different tools (lower is better).

	pydeface		afni_defacer		afni_refacer		mri_reface		cGAN afni_defacer	
Brain regions	FSL	MB	FSL	MB	FSL	MB	FSL	MB	FSL	MB
TIV	75.41	72.64	114.82	599.83	66.26	24.10	45.79	*14.57*	**44.10**	22.91
CSF	37.69	28.00	47.88	163.73	34.42	18.57	**23.52**	*17.98*	**23.53**	18.18
GM	23.57	36.34	49.98	312.02	27.16	20.99	**13.11**	*19.30*	15.67	20.95
WM	24.57	24.08	45.02	136.86	35.51	11.79	**15.77**	*7.44*	20.84	10.17
Thalamus	0.53	0.71	0.76	9.48	0.81	0.67	**0.41**	0.69	0.50	*0.64*
Caudate	0.38	0.75	0.41	5.77	**0.25**	0.70	0.29	*0.69*	0.30	*0.68*
Putamen	0.48	1.15	0.64	7.47	0.65	1.03	**0.42**	1.12	0.47	*0.97*
Pallidum	0.21	0.42	0.27	2.03	0.30	0.42	**0.19**	0.45	0.21	*0.40*
Hippocampus	0.55	0.74	0.83	3.55	0.69	0.47	**0.31**	0.44	0.37	*0.42*
Amygdala	0.36	0.29	0.38	1.10	0.38	*0.23*	**0.27**	0.26	**0.27**	*0.24*

Note: Absolute brain measurements were obtained with *fsl _anat* (FSL) and with *MorphoBox* (MB). Minimal coefficients of repeatability for specific volumes are highlighted in **bold** for fsl_anat and *bold italic* for MorphoBox.

with *MorphoBox* volumetric estimates are closer to *mri reface* and *cGAN afni_defacer* than to the ones of defacing tools; on the contrary, for *fsl_anat* results, the CRs are more comparable to the defacing *pydeface* tool. The worst performance in terms of repeatability is given by *afni_defacer* that has the highest and outlying values CRs for most of the volumes estimated with *fslanat* or *MorphoBox*.

A more detailed analysis of the repeatability is given by Bland–Altman plots shown in Figure 4 for TIV and hippocampus. For the rest of the brain tissues and structures Bland–Altman plots are shown in Appendix E. For the particular brain structures, the greatest effect on the repeatability is provided by outlying volume estimates for some subjects, rather than by systematic biases in volume estimates after de-identification. For *fsl_anat* results the greatest impact is provided by outliers with the positive difference, due to underestimated volume sizes after de-identification by all tools. On the contrary, *MorphoBox* results show less number of outliers outside the limits of agreement for all de-identification tools, except for *afni_defacer* where both TIV and hippocampus have many outliers with positive differences in volume estimates.

Additional evaluation of the segmentation quality with the Dice score between brain segmentation maps before and after de-identification are provided in Appendix E. While those results give more details about the performance per brain structure, it overall complements the conclusions derived from Bland–Altman plots about the role of outliers on average performance of de-identification tools.

3.3 | Re-identification risk

The distribution of the cosine facial distances across the different de-identification techniques are shown in Table 5. All de-identification techniques showed non-zero distances between the faces before and after anonymization, suggesting a certain level of protection against re-identification. For the defacing techniques, *afni_refacer* and *cGAN afni_defacer* resulted in comparable mean values of cosine facial distances, while *mri_reface* resulted in the lowest facial distance, suggesting the highest similarity between original and anonymized faces.

Based on the performed face recognition model selection to find the best model for the given task, we were able to determine an approximate threshold for the cosine facial distance estimated by ArcFace that separates the classes of correctly and incorrectly matched faces (more detailed results are given in Appendix D) and is equal to 0.4. There is still, however, an overlap between those classes that results in 1.2% of false detection rate, that is, the percentage of faces belonging to different subjects that were misclassified as a correct match by the model. This value can be perceived as an error of the current approach for re-identification risk approximation using the percentage of potentially identifiable cases. The results presented in Table 5, show that *afni_defacer* and *afni_refacer* yield the lowest amount of potentially identifiable cases. Results for the defacing technique *pydeface* and *cGAN afni_defacer* are comparable, while *mri_reface* has the highest percentage of cases that can potentially be identified.

3.4 | Trade-off between volumetric reproducibility and re-identification risk

The trade-off plots in Figure 5 relating re-identification risk and post-processing consistency, help summarizing the results for both sets of experiments. Both plots show that *afni_defacer* provides the lowest re-identification risk, at the expense of less consistent volumetric brain measurements. For the *MorphoBox* brain segmentation the *afni_defacer* has an outlying effect on the morphometry quality, not comparable to the rest of de-identification tools. The performance of other tools in terms of consistency of morphometry varies between *fsl_anat* and *MorphoBox*. While *mri reface* has the lowest impact on

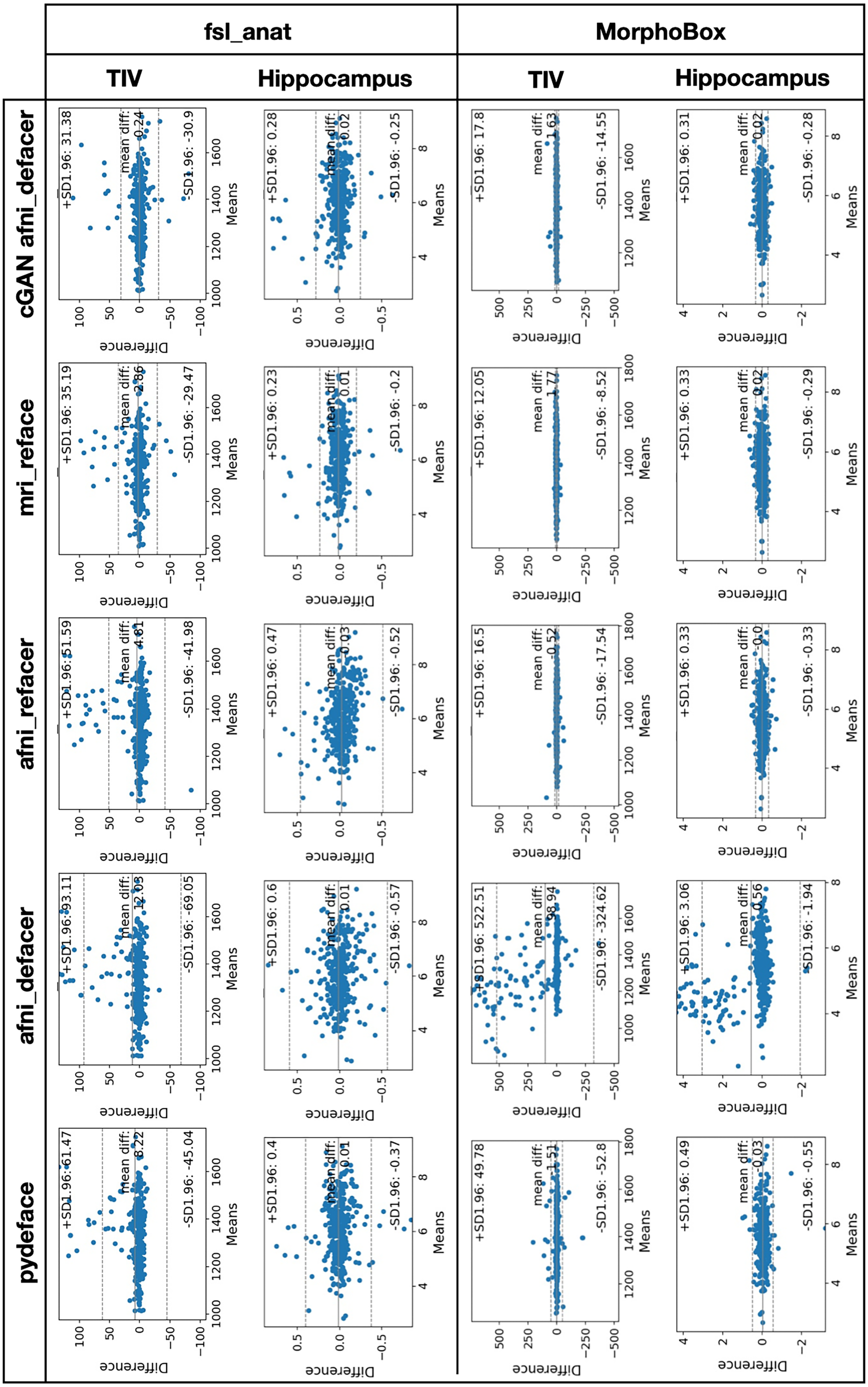

FIGURE 4 Bland–Altman difference plots for the volumetric results of the original images in comparison to the ones of the de/refaced images. Plots one big region (TIV) and the hippocampus as pars pro toto for a small region. Note that vertical axes scaling differs for *fsl _anat* and *MorphoBox*.

TABLE 5 Re-identification risk assessment for the different de-identification techniques, using the ArcFace face recognition model.

Measure	pydeface	afni_defacer	afni_refacer	mri_reface	cGAN afni_defacer
Cosine facial distance mean (st. dev.)	62.78 (16.30)	77.90 (14.89)	57.96 (11.71)	48.55 (11.11)	57.60 (14.72)
Potentially identifiable cases [%]	8.2%	3.4%	5.8%	23.5%	9.8%

Note: Mean and standard deviation of cosine facial distance and percentage of potentially identifiable cases are shown. ArcFace is used to calculate the cosine distances between the original and de/refaced faces that is inversely proportional to re-identification risk.

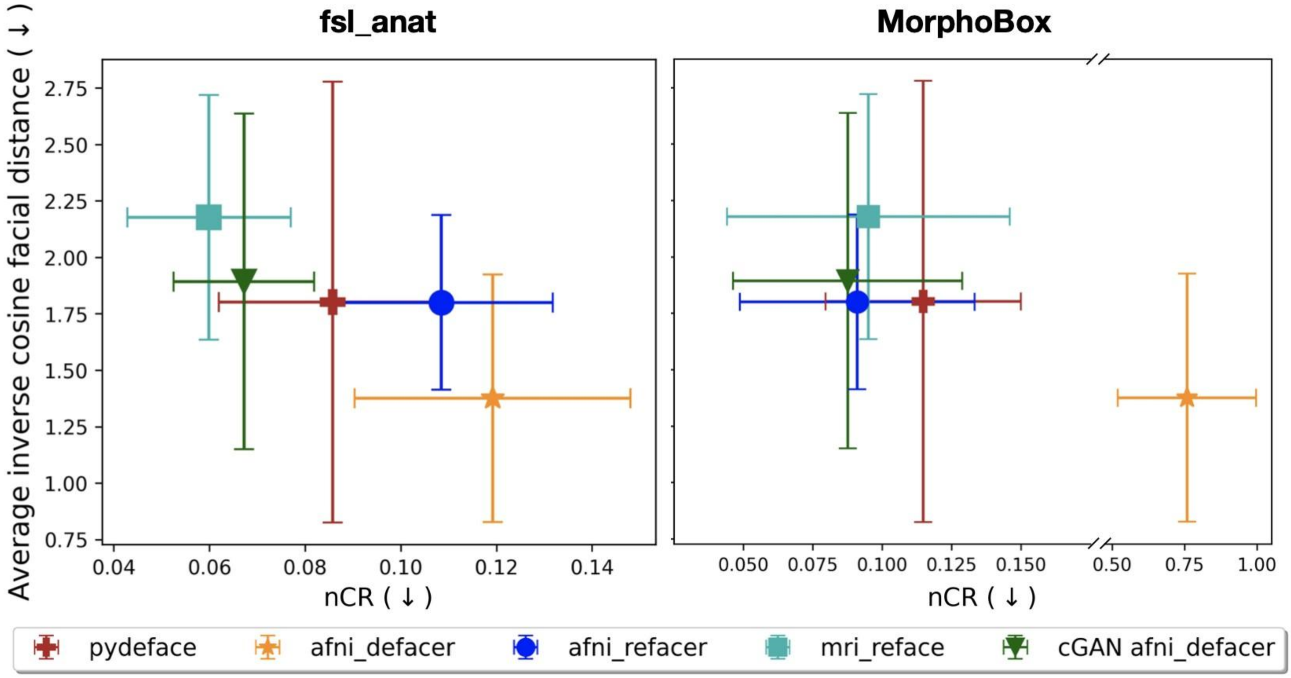

FIGURE 5 Trade-off plots for the joint evaluation of the re-identification risk and reproducibility of the morphometry results after de/refacing. The inverse face distance averaged across subjects is plotted on the y-axis as a measure of the re-identification risk. The vertical whiskers' length is the standard deviation of the inverse distances. The normalized coefficients of repeatability (nCR) were calculated using the normalized estimated volumes (averaging both across scans and across brain structures) are displayed on the x-axis as a measure of the inconsistency in the volumetric results after face de-identification. Horizontal whiskers have the length of the standard deviation of the CR values, calculated across scans for different normalized brain structure volumes.

volumetric brain measurements obtained by *fsl_anat*, for the measurements made by *MorphoBox cGAN afni defacer* anf *afni_refacer* have the lowest impact. For *MorphoBox*, however, effect on volumetry is comparable across all refacing tools, which is different from *fsl_anat*.

3.5 | Processing time

Average processing time for each re-identification tool are shown in Table 6. It is important to mention that *cGAN afni_defacer* operates on already defaced images and for application to full-head scans, its times should be added to the time taken by *afni _defacer* defacing.

Pydeface is the fastest technique for the applications to original full-head images. Timings taken from *afni _refacer* and *cGAN afni_defacer* are similar when considering applications to whole-head scans.

For applications to already defaced images for face generation, *cGAN afni_defacer* provides significant speed up in comparison to the rest of the refacing tools, both considering CPU and GPU applications.

4 | DISCUSSION

4.1 | Failed cases analysis

Most of the excluded scans were omitted because of a failure of *pydeface*. This can potentially be explained by the fact that *pydeface* was previously shown to have a higher failure rate on scans belonging to older cohorts (44–85 years) (Theyers et al., 2021) than on other cohorts, and ages in the present dataset vary from 53 to 93 years. The rest of the techniques have few or zero failures.

4.2 | Consistency of brain volume measurements

The results confirm that not only volume estimates of superficial brain structures can be affected by de/refacing, as also deep brain structures, like thalamus, putamen or pallidum showed significantly different absolute volume estimates made by *fsl_anat*.

TABLE 6 Average processing time in seconds taken by each of the techniques to process one scan.

				cGAN afni defacer	
pydeface	afni_defacer	afni_refacer	mri_reface	CPU	GPU
92	126	131	917	9	5

Note: *cGAN afni defacer* operates on already defaced images, so defacing time should be added if full images are used as input.

We also see that measurements obtained by different brain segmentation software are affected differently by face de-identification procedures. Particularly, *fsl_anat* post-processing results are more affected by face de-identification than *MorphoBox* as reflected by the amount of brain regions where statistically significant differences were detected. *MorphoBox* performs soft tissue labeling based on an intensity model whereas *fsl_anat* uses a shape and appearance model. The shape-based algorithms tend to be more sensitive to noise. Thus, slight changes in the boundary regions, which may result from the de-identification, can cause large errors. This could hence explain the lower reproducibility for *FSL*.

While results of the statistical testing are more related to systematic biases in the volume estimates, CRs are related to the spread of the values and thus provide insight into the stability of the de-identification techniques with regard to providing consistent results. In the volumetric brain measurements obtained with *MorphoBox*, we see the lowest amount of statistically significant differences after applying *afni_refacer* and *cGAN afni_defacer*, and comparable CR values for *cGAN afni_defacer*, *mri_reface*, *afni_refacer* (CR in small structures are also comparable for *pydeface*). For the *fsl_anat* results, the amount of statistically significant differences is relatively high for all de/refacing tools (in 7–9 out of 10 brain structures), while the CR values are the lowest and comparable for *mri_reface* and *cGAN afni_defacer*.

With regard to the proposed *cGAN afni_defacer* technique, the results show that it is able to recover differences in volumetric brain measurements introduced by *afni_defacer* defacing. In particular, it mitigates the number of brain structures showing significantly different volume estimates, from 9 to 7 regions for *fsl_anat* and from 7 to 1 for *MorphoBox*. It also significantly reduces the CR values in comparison to *afni_defacer*, suggesting more consistent volumetric brain measurements. Image quality did not affect these differences, except for the TIV ROI (see Appendix G).

4.3 | Re-identification risk

Judging by the distributions of the cosine facial distances, defacing techniques yield the lowest similarity between original and anonymized faces. In fact, this is mostly explained by face detection failure preceding face recognition. Yet, *pydeface* has higher similarity of faces than *afni_defacer*, because it removes a smaller portion of the head leaving intact distinctive facial features, such as ears, partially eyes and nose septum. While the face is left defaced, it does not raise any issues, however if refacing is applied to such faces, the face detection algorithm will no longer fail and residual facial features will be possibly picked up by the algorithm. This implies that it is also important to ensure proper defacing of all facial features. This was one of the primary considerations for the proposed *cGAN afni_defacer* being trained particularly with *afni_defacer* images.

Mri_reface resulted in the lowest mean value of cosine facial distances and the highest percentage of potentially identifiable cases, however this might be explained by the fact that it is based on population-average face templates and produces the most realistic faces. These results reflect the limitations of the proposed approach for approximation of the re-identification risk. The relatively high number of potentially re-identifiable cases may also indicate that, apart from the facial features, also parameters like overall head shape and size may have a significant contribution to subject identification.

The Bland–Altman plots give an intuition about different sources of errors affecting repeatability. For the particular cases of TIV and hippocampus, the outlying values of the difference and the amount of such outliers are more likely to be affecting the repeatability than the presence of systematic biases introduced by de-identification techniques.

4.4 | Trade-off plots

While trade-off plots should indicate the techniques that achieve the optimal trade-off between privacy protection and consistency of volumetric brain measurements after de-identification, our results do not give a clear answer to this question. We see that different post-processing tools, that is, *MorphoBox* and *fsl_anat*, disagree on the ranking of the techniques, except for *afni_defacer*. *Afni_defacer* has the overall lowest re-identification risk, however at the cost of a higher effect on post-processing results. For the rest of the techniques, the differences with regard to the effect on volumetric brain measurements obtained with *MorphoBox* might be insignificant, however brain segmentation results obtained with *fsl_anat* may suggest some ranking of the techniques in terms of their effect on volumetric brain measurements. In general, the plots confirm our hypothesis that there seems to be a trade-off between privacy protection and consistent post-processing, showing that for the investigated de-identification tools more consistent post-processing is achieved at the expense of higher re-identification risk. It also shows that the quality of post-processing after de-identification, also depends on the post-processing tool used, not just the de-identification tool itself.

4.5 | Processing time

Considering applications to whole-head scans, we see that time taken by the defacing tools is not substantially smaller than the time taken by refacing tools. Therefore, the processing time is not a reason for choosing defacing over refacing.

The proposed cGAN refacing tool gives significant speed advantages only in the scenario when one wants to recover consistent volumetric brain measurements from already defaced images. This may, however, change with the development of faster defacing tools.

5 | CONCLUSION

In this study, we propose a new refacing technique based on a 3D cGAN that operates on the defaced T1w images. We compared the proposed technique to two defacing (*pydeface* and *afni_refacer* in defacing mode) and two refacing techniques (*afni_refacer* in refacing mode and *mri_reface*) in terms of (i) their degree of privacy protection; (ii) their impact on volumetric brain measurements obtained with *MorphoBox* and *fsl_anat* software, as an example of a common image post-processing and analysis workflow; (iii) their required processing time. We showed that the proposed technique achieves a good trade-off between (i) and (ii) independently of the brain segmentation technique used. Its processing speed brings a significant advantage for applications to already defaced scans and has a comparable processing time with other refacing techniques even if defacing is taken into account. These results, in addition to all, suggest that the proposed de-identification method is a viable technique for ensuring consistent volumetric results from defaced images by face inpainting.

Through our comparative study we were able to confirm that there exists a trade-off between the degree of privacy protection and consistent post-processing results, meaning that one can not achieve both superior face de-identification and low impact on the post-processing results at the same time. Complete defacing with accurate removal of all facial features leads to face detection and/or face recognition failure, and was also shown to corrupt the brain tissue and subcortical segmentation and volumetric brain measurements. Our results suggest that refacing is a better alternative in terms of providing consistent post-processing results in comparison to defacing. Moreover, the face generation does not necessarily need to produce highly realistic faces after the face reconstruction as long as volumetric results remain consistent, as for the proposed refacing cGAN or afni_refacer. As an exception, *pydeface* has a comparable effect on the volumetric brain measurements obtained with MorphoBox software, however, it does not provide such deep defacing as *afni defacer* often leaving parts of eyes, cheeks or nose and, thus, has a re-identification risk comparable to refacing tools. Summarizing the obtained results, for the best privacy protections we would suggest to choose defacing tools that properly remove all facial features, including eyes, nose, ears, cheeks. However, for data re-usability refacing should be the method of choice.

There are several limitations of this study that we would like to address. First of all, we investigate the impact on post-processing results only on example of volumetric brain measurements and only with two existing tools with specific tools parameters. While further investigation of this impact is preferable, our results show that a consistency of volumetry after de-identification is specific to the post-processing tool. Thus, it is crucial for any study to verify on their own how the post-processing results of interest are affected by de-identification of any kind. Second, there are generalization limits of the provided trained cGAN, as it was trained on an older cohort of subjects and solely on T1-weighted MR images, defaced with a specific technique. Based on previous work, showing that the impact of de-identification procedures on structural brain measures are comparable across different age groups (Buimer et al., 2021), there is a reason to consider our results being transferable to other age groups. Nevertheless, being a trainable approach our defacing tool can be potentially adapted to any type of data with a reduced computational cost considering a fine-tuning scenario where existing weights are used for initialization. With this said, we recommend using defacing tools that do not provide complete facial features removal as a basis, as a properly trained cGAN is able to recover some original facial features from their residuals meaning a significant increase in the re-identification risk. Furthermore, the development of generative AI in recent years gives a promise for further improvement of refacing algorithms through integration of more advanced techniques, such as transformer architectures or diffusion models.

ACKNOWLEDGMENTS

The authors wish to thank Dr Oscar Esteban for his suggestion on edits on this paper. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research

Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

CONFLICT OF INTEREST STATEMENT

BM, TK and TH are employed by and hold stock of Siemens Healthineers. All other authors have no conflict of interest with regard to the subject matter of this study.

DATA AVAILABILITY STATEMENT

The MRI data from the TADPOLE challenge of ADNI that was used for training of the proposed method and for the comparative study between different face de-identification tools are publicly available. All subjects identifiers for the data used, volumetric brain measurement obtained with FSL on the test set, code for training and testing of the proposed method, as well as weights of the trained models are available online at https://gitlab.com/acit-lausanne/refacing-cgan.

ORCID

Nataliia Molchanova https://orcid.org/0000-0002-7211-8863

REFERENCES

Abramian, D., & Eklund, A. (2019). Refacing: Reconstructing anonymized facial features using GANS. In *2019 IEEE 16th international symposium on biomedical imaging (ISBI 2019). pp. 1104–1108*. IEEE. https://doi.org/10.1109/ISBI.2019.8759515

Bhalerao, G. V., Parekh, P., Saini, J., Venkatasubramanian, G., John, J. P., & ADBS consortium. (2022). Systematic evaluation of the impact of defacing on quality and volumetric assessments on T1-weighted MR-images. *Journal of Neuroradiology*, *49*(3), 250–257. https://doi.org/10.1016/j.neurad.2021.03.001 https://www.sciencedirect.com/science/article/pii/S0150986121000559

Buimer, E. E., Schnack, H. G., Caspi, Y., van Haren, N. E., Milchenko, M., Pas, P., Pol, H. E. H., & Brouwer, R. M. (2021). De-identification procedures for magnetic resonance images and the impact on structural brain measures at different ages. *Human Brain Mapping*, *42*(11), 3643–3655. https://doi.org/10.1002/hbm.25459

Cirillo, M. D., Abramian, D., & Eklund, A. (2021). Vox2Vox: 3D-GAN for brain tumour Segmentation. In *Lecture notes in computer science* (Springer, Cham, Vol. 12658, pp. 274–284).

Cox, R. W. (1996). AFNI: Software for analysis and Visualization of Functional Magnetic Resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. *NMR in Biomedicine*, *10*(4–5), 171–178. https://doi.org/10.1002/(sici)1099-1492(199706/08)10:4/5

de Sitter, A., Visser, M., Brouwer, I., Cover, K. S., van Schijndel, R. A., Eijgelaar, R. S., Müller, D. M. J., Ropele, S., Kappos, L., Rovira, Á., Filippi, M., Enzinger, C., Frederiksen, J., Ciccarelli, O., Guttmann, C. R. G., Wattjes, M. P., Witte, M. G., de Witt Hamer, P. C., Barkhof, F., … MAGNIMS Study Group and Alzheimer's Disease Neuroimaging Initiative. (2020). Facing privacy in neuroimaging: removing facial features degrades performance of image analysis methods. *European Radiology*, *30*(2), 1062–1074. https://doi.org/10.1007/s00330-019-06459-3

Deepface python library. 2021. https://github.com/serengil/deepface

Deng, J., Guo, J., Yang, J., Xue, N., Cotsia, I., & Zafeiriou, S. P. (2021). ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*, 1. https://doi.org/10.1109/TPAMI.2021.3087709

Gao, C., Jin, L., Prince, J. L., & Carass, A. (2022). Effects of defacing whole head MRI on neuroanalysis. In *Proceedings Volume 12032, Medical Imaging 2022: Image Processing; 120323W*. SPIE. https://doi.org/10.1117/12.2613175

Gao, C., Landman, B. A., Prince, J. L., & Carass, A. (2023). *Reproducibility evaluation of the effects of MRI defacing on brain segmentation*. *Journal of Medical Imaging*, 10. https://doi.org/10.1117/1.jmi.10.6.064001

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 27). Curran Associates, Inc.

Gulban, O. F., Nielson, D., Lee, J., Poldrack, R., Gorgolewski, C., Vanessasaurus, & Markiewicz, C. (2022). *poldracklab/pydeface: v2.0.0*. Zenodo. https://doi.org/10.5281/zenodo.6856482

Gunter, J., Borowski, B., Thostenson, K., Arani, A., Reid, R., Cash, D., Thomas, D., Zhang, H., DeCarli, C., Fox, N., Thompson, P., Tosun, D., Weiner, M., & Jack, C. (2017). ADNI-3 MRI protocol. *Alzheimer's & Dementia*, *13*, P104–P105. https://doi.org/10.1016/j.jalz.2017.06.2411

Harms, J., Lei, Y., Wang, T., Zhang, R., Zhou, J., Tang, X., Curran, W., Liu, T., & Yang, X. (2019). Paired cycle-GAN based image correction for quantitative cone-beam CT. *Medical Physics*, *46*, 3998–4009. https://doi.org/10.1002/mp.13656

Huelnhagen, T., Fartaria, M. J., Corredor-Jerez, R., Mahdi1, M. F. A. W., Piredda, G. F., Marechal, B., Richiardi, J., & Kober, T. (2020). Don't lose your face–refacing for improved morphometry. In *International Society for Magnetic Resonance in Medicine*, *28*. https://cds.ismrm.org/protected/20MProceedings/PDFfiles/0546.html

Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.-P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K. H., & Kickingereder, P. (2019). Automated brain extraction of multisequence MRI using artificial neural networks. *Human Brain Mapping*, *40*(17), 4952–4964. https://doi.org/10.1002/hbm.24750

Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 5967–5976). IEEE. https://doi.org/10.1109/CVPR.2017.632

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*(2), 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Kalavathi, P., & Surya Prasath, V. B. (2015). Methods on skull stripping of MRI head scan images—A review. *Journal of Digital Imaging*, *29*, 365–379.

Liu, Y., Lei, Y., Wang, T., Fu, Y., Tang, X., Curran, W., Liu, T., Patel, P., & Yang, X. (2020). CBCT-based synthetic CT generation using deep-attention CycleGAN for pancreatic adaptive radiotherapy. *Medical Physics*, *47*, 2472–2483. https://doi.org/10.1002/mp.14121

Lorensen, W. E., & Cline, H. E. (1987). Marching cubes: A high resolution 3D surface construction algorithm. In *SIGGRAPH Computer Graphics* (Vol. 21, pp. 163–169). Association for Computing Machinery. https://doi.org/10.1145/37402.37422

Mazura, J. C., Juluru, K., Chen, J. J., Morgan, T. A., John, M., & Siegel, E. L. (2012). Facial recognition software success rates for the identification of 3D surface reconstructed facial images: Implications for patient privacy and security. *Journal of Digital Imaging*, *25*(3), 347–351. https://doi.org/10.1007/s10278-011-9429-3

Mikulan, E., Russo, S., Zauli, F. M., d'Orio, P., Parmigiani, S., Favaro, J., Knight, W., Squarza, S., Perri, P., Cardinale, F., Avanzini, P., & Pigorini, A. (2021). A comparative study between state-of-the-art MRI de-identification and AnonyMI, a new method combining re-identification risk reduction and geometrical preservation. *Human Brain Mapping*, *42*(17), 5523–5534. https://doi.org/10.1002/hbm.25639

Milchenko, M., & Marcus, D. (2012). Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics*, *11*, 65–75. https://doi.org/10.1007/s12021-012-9160-3

National Institute of Standards and Technology. (2021). *Face recognition vendor test (FRVT) 1:N identification*. https://pages.nist.gov/frvt/html/frvt11.html

Prior, F. W., Brunsden, B., Hildebolt, C., Nolan, T. S., Pringle, M., Vaishnavi, S. N., & Larson-Prior, L. J. (2009). Facial recognition from volume-rendered magnetic resonance imaging data. *IEEE Transactions on Information Technology in Biomedicine*, *13*(1), 5–9. https://doi.org/10.1109/TITB.2008.2003335

Rubbert, C., Wolf, L., Turowski, B., Hedderich, D. M., Gaser, C., Dahnke, R., Caspers, J., & for the Alzheimer's Disease Neuroimaging Initiative. (2022). Impact of defacing on automated brain atrophy estimation. *Insights into Imaging*, *13*(1), 54. https://doi.org/10.1186/s13244-022-01195-7

Schmitter, D., Roche, A., Mar'echal, B., Ribes Lemay, D., Abdulkadir, A., Bach Cuadra, M., Daducci, A., Granziera, C., Kloppel, S., Maeder, P., Meuli, R., & Krueger, G. (2014). An evaluation of volume-based morphometry for prediction of mild cognitive impairment and Alzheimer's disease. *NeuroImage: Clinical*, *7*, 7–17. https://doi.org/10.1016/j.nicl.2014.11.001

Schwarz, C. G., Kremers, W. K., Lowe, V. J., Savvides, M., Gunter, J. L., Senjem, M. L., Vemuri, P., Kantarci, K., Knopman, D. S., Petersen, R. C., & Jack C. R. (2022). Face recognition from research brain PET: An unexpected PET problem. *NeuroImage*, *258*, 119357. https://doi.org/10.1016/j.neuroimage.2022.119357 https://linkinghub.elsevier.com/retrieve/pii/S1053811922004761. (visited on February 08, 2022)

Schwarz, C. G., Kremers, W. K., Wiste, H. J., Gunter, J. L., Vemuri, P., Spychalla, A. J., Kantarci, K., Schultz, A. P., Sperling, R. A., Knopman, D. S., Petersen, R., & Jack, C. (2021). Changing the face of neuroimaging research: Comparing a new MRI de-facing technique with popular alternatives. *NeuroImage*, *231*, 117845. https://doi.org/10.1016/j.neuroimage.2021.117845

Schwarz, C., Kremers, W., Therneau, T., Sharp, R., Gunter, J., Vemuri, P., Arani, A., Spychalla, A., Kantarci, K., Knopman, D., Petersen, R., & Jack, C. R. (2019). Identification of anonymous MRI research participants with face-recognition software. *New England Journal of Medicine*, *381*, 1684–1686. https://doi.org/10.1056/NEJMc1908881

Surf Ice Software. (2021). University of South Carolina, McCausland Center for Brain Imaging. https://www.nitrc.org/projects/surfice/

TADPOLE challenge constructed by the EuroPOND consortium. (2012-2024) http://europond.eu funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 666992. url: https://tadpole.grand-challenge.org.

Theyers, A., Zamyadi, M., O'Reilly, M., Bartha, R., Symons, S., Macqueen, G., Hassel, S., Lerch, J., Anagnostou, E., Lam, R., Frey, B., Milev, R., Müller, D. J., Kennedy, S., Scott, C., & Strother, S. (2021). Multisite comparison of MRI defacing software across multiple cohorts. *Frontiers in Psychiatry*, *12*, 617997. https://doi.org/10.3389/fpsyt.2021.617997

U.S. Department of Health and Human Services Office for Civil Rights. (2013). *HIPAA administrative simplification*. *Regulation* Text https://www.hhs.gov/sites/default/files/hipaa-simplification-201303.pdf

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J., Da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1). https://doi.org/10.1038/sdata.2016.18

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, *58*, 101552. https://doi.org/10.1016/j.media.2019.101552

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Molchanova, N., Maréchal, B., Thiran, J.-P., Kober, T., Huelnhagen, T., Richiardi, J., & the Alzheimer's Disease Neuroimaging Initiative (2024). Fast refacing of MR images with a generative neural network lowers re-identification risk and preserves volumetric consistency. *Human Brain Mapping*, *45*(9), e26721. https://doi.org/10.1002/hbm.26721