Appendix A: Conditional De-Identification of 3D Magnetic Resonance Images

Lennart Alexander Van der Goten[1,2]
lavdg@kth.se
Tobias Hepp[3,4]
tobias.hepp@tuebingen.mpg.de
Zeynep Akata[3,4]
zeynep.akata@uni-tuebingen.de
Kevin Smith[1,2]
ksmith@kth.se

[1] KTH Royal Institute of Technology
Stockholm, SWEDEN
[2] Science for Life Laboratory
Solna, SWEDEN
[3] Max Planck Institute for Intelligent Systems
Tübingen, GERMANY
[4] University of Tübingen
Tübingen, GERMANY

for the Alzheimer's Disease Neuroimaging Initiative[1]

1 De-Identification Methods

We compare our result with three publicly available and widely-established methods for de-identification of MRI head scans, depicted in Figure 3 in the main text. All methods have in common that they *(1)* are not deep-learning-driven, *(2)* require no additional training and *(3)*, are used on a day-to-day basis by clinical- and neuroscientists. All procedures were applied with default settings on images of resolution $128\times128\times128$ with model-agnostic preprocessing as described in Section A.3.

QUICKSHEAR [17] computes a plane to divide a given MRI into two parts: one containing facial structures, and the other containing the remainder of the scan. Voxels in the first part are set to zero.

FACE MASK [13] uses a filtering method to blur the facial features. Based on registration to an atlas, the face region is identified, normalized and filtered. The result is transformed back to the original image space.

DEFACE [2] estimates the probabilities of voxels belonging to the face based on an atlas of healthy control subjects. Intensities of voxels whose probabilities are small enough are set to zero.

[1] 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

2 Benchmark Datasets

In this work, we consider two standard publicly available and large-scale medical imaging datasets which feature T1-weighted volumetric MR images of the skull for each subject. Scanner types and acquisition protocols differ between and within the datasets.

OASIS-3 [12] contains MRI and PET scans of 1,098 patients gathered from multiple longitudinal studies. The total number of negative subjects is 605, *i.e.* patients without any signs of mental diseases, and the positive subjects 493 exhibit Alzheimer's disease (AD). In total, each of the 2,168 different MRI sessions comes with a varying modalities recorded using four different Siemens devices (BioGraph mMR PET-MR 3T, TIM Trio 3T, Sonata 1.5T, Vision 1.5T).

ADNI [25] is a large-scale dataset that comprises MRI scans of healthy, mildly cognitively impaired and AD patients recorded by six different scanners from GE (25X, Widebore 25X), Philips (R3, R5) and Siemens (20VB17, Prisma D13, Skyra E11, VE11C). For the sake of reproducibility, we choose the standardized variant [27] with 2,172 3T MRI scans over a time span of three years.

3 Model-agnostic Preprocessing

To ensure quality, comparable signal intensity distributions, and a consistent orientation of the acquired MR images, we apply various preprocessing steps. We apply standard preprocessing steps to ensure quality, comparable signal distributions, and consistent orientation (*e.g.* [9, 16]) to positively affect algorithms. These preprocessing steps are applied prior to all de-identification methods in our study.

Orientation Correction Images within MR datasets are often not consistently aligned which ultimately hampers the learning process. Thus, we leverage FSL (FMRIB Software Library) [20] to re-orient all images to the *radiological orientation convention*, the so-called *LAS* orientation.

Bias Field Correction The *bias field* is a low-frequency degradation of an MR scan due to magnetic field inhomogeneities which is typically imperceptible to humans. If unaccounted for, this degradation can induce different grayscale values on the same tissue type which, in turn, might impair downstream algorithms [10]. We use the nonparametric N4BiasFieldCorrection algorithm from the ANTs library [22].

Registration MR scans exhibit a high degree of variability primarily because of anatomical reasons, but also because patients are typically not consistently positioned within the MRI tube. We therefore apply a non-rigid, double-affine *TRSAA* registration offered by the ANTs library to mitigate these effects. The necessary registration template is chosen uniformly at random *once* for *each* dataset.

Region-of-Interest Segmentation The toolkit Robex (Robust Brain Extraction [6]) was used to separate the brain from surrounding tissues. We have chosen this method as its segmentations turned out to more robust than competing methods (*e.g.* [18]). On the flip side, this increased accuracy comes at the expense of longer execution times.

White Stripe Normalization[1] Previous work (*e.g.* [19]) has emphasized the importance of normalization for statistical evaluations. As no significant performance differences between intensity normalization schemes have been reported [15], we opt for the compara-

[1] Only done for our model as other algorithms expect no change in scaling

tively simple *white strip normalization* that estimates both the mean μ_b and the (biased) standard deviation σ_b over the image voxels $x_{u,v,w}$ belonging to the brain tissue (as indicated by $b_{u,v,w}(x) = b_{u,v,w}$):

$$\mu_b = \frac{\sum_{u,v,w\in\{1,\dots,S\}} b_{u,v,w} \cdot x_{u,v,w}}{\sum_{u,v,w\in\{1,\dots,S\}} b_{u,v,w}}$$
$$\sigma_b = \frac{\sum_{u,v,w\in\{1,\dots,S\}} b_{u,v,w} \cdot (x_{u,v,w} - \mu_b)^2}{\sum_{u,v,w\in\{1,\dots,S\}} b_{u,v,w}}$$

Finally *all* voxel values are shifted and rescaled using the *z-score* transformation $\hat{x} = \frac{x-\mu_b}{\sigma_b}$. Observe that this transformation is invertible given that ones memorizes μ_b and σ_b.

4 Loss Function Considerations

We initially experimented with the relativistic *average* loss *Ra-LS-GAN* variant suggested by [7]:

$$\mathcal{L}_G^{\text{Ra-LS-GAN}} = \mathbb{E}_{x_f\sim\mathcal{P}_Y}\left[(D_\Theta(x_f) - \mathbb{E}_{x_r\sim\mathcal{P}_X} D_\Theta(x_r) - 1)^2\right] + \mathbb{E}_{x_r\sim\mathcal{P}_X}\left[(D_\Theta(x_r) - \mathbb{E}_{x_f\sim\mathcal{P}_Y} D_\Theta(x_f) + 1)^2\right]$$
$$\mathcal{L}_D^{\text{Ra-LS-GAN}} = \mathbb{E}_{x_r\sim\mathcal{P}_X}\left[(D_\Theta(x_r) - \mathbb{E}_{x_f\sim\mathcal{P}_Y} D_\Theta(x_f) - 1)^2\right] + \mathbb{E}_{x_f\sim\mathcal{P}_Y}\left[(D_\Theta(x_f) - \mathbb{E}_{x_r\sim\mathcal{P}_X} D_\Theta(x_r) + 1)^2\right]$$

where $\mathcal{P}_X, \mathcal{P}_Y$ denote the original resp. the fake distribution induced by G_Φ and we drop the conditioning variable $\gamma(x)$ from the notation. Observe, however, that this loss function is incompatible with our *conditional* scenario as $\mathbb{E}_{x_r\sim\mathcal{P}_X} D_\Theta(x_r)$ and $\mathbb{E}_{x_f\sim\mathcal{P}_Y} D_\Theta(x_f)$ are computed by averaging *across* scans associated to *different* conditional information. To solve this problem, we make sure that every patient can occur at most once in a batch and get rid of the aforementioned expectations:

$$\begin{aligned}\mathcal{L}_G &= \mathbb{E}_{(x_f,x_r)\sim(\mathcal{P}_Y,\mathcal{P}_X)}\left[(D_\Theta(x_f) - D_\Theta(x_r) - 1)^2\right] + \mathbb{E}_{(x_r,x_f)\sim(\mathcal{P}_X,\mathcal{P}_Y)}\left[(D_\Theta(x_r) - D_\Theta(x_f) + 1)^2\right] \\ &= 2\mathbb{E}_{(x_r,x_f)\sim(\mathcal{P}_X,\mathcal{P}_Y)}\left[(D_\Theta(x_r) - D_\Theta(x_f) + 1)^2\right] \\ \mathcal{L}_D &= \mathbb{E}_{(x_r,x_f)\sim(\mathcal{P}_X,\mathcal{P}_Y)}\left[(D_\Theta(x_r) - D_\Theta(x_f) - 1)^2\right] + \mathbb{E}_{(x_f,x_r)\sim(\mathcal{P}_Y,\mathcal{P}_X)}\left[(D_\Theta(x_f) - D_\Theta(x_r) + 1)^2\right] \\ &= 2\mathbb{E}_{(x_r,x_f)\sim(\mathcal{P}_X,\mathcal{P}_Y)}\left[(D_\Theta(x_f) - D_\Theta(x_r) + 1)^2\right]\end{aligned}$$

Observe that this identical to the construction of *R-LS-GAN* if we follow the principles of [7].

5 Binary Downsampling

Owed to the progressive structure of CP-GAN it is necessary to downsample (*i.e.* halve their resolution) successively until some minimium resolution is attained. We suggest a simple means to downsample binary tensors that aims to preserve the sparsity degree of the input tensors. Suppose that we have some given input mask $m_0 \in \{0,1\}^{2^n\times 2^n\times 2^n}$ for some fixed $n \in \mathbb{N}$. Let $(m')_{i=1,\dots,n-p} \in [0,1]^{2^{n-i}\times 2^{n-i}\times 2^{n-i}}$ further denote the result of applying *average pooling* i times on m_0 and stopping at some minimal resolution $2^p \times 2^p \times 2^p, p \in \mathbb{N}$. A new sequence $(m')_{i=1,\dots,n-p}$ of binary representations can then be constructed by interpreting

each voxel value of m'_i as a *Bernoulli* parameter determined by a *maximum likelihood* estimation over a (flattened) patch of $2^{i+1} \cdot 2^{i+1} \cdot 2^{i+1} = 2^{3(i+1)}$ (binary) realizations stemming from m_0. Accordingly, this interpretation permits us to view m'_i as a volume of *Bernoulli parameters* from which we can derive m''_i by sampling in a voxel-wise fashion. Most importantly, this construction *preserves* the non-zero $\zeta(m_0)$ proportion of m_0 *in expectation* (i.e. $\left(\frac{1}{2^n}\right)^3 \sum_{u,v,w} m_0^{(u,v,w)} = \mathbb{E}\left[\left(\frac{1}{2^{n-i}}\right)^3 \sum_{u,v,w} m_i'^{(u,v,w)}\right]$ for $i = 1, \ldots, n-p$). Although this downsampling scheme is stochastic in essence, we observe that the mapping becomes deterministic if a specific *Bernoulli* parameter is found by averaging over a region (patch) of constant realizations. Consequently, voxels of m'_i corresponding to the cerebral cortex ("brain boundary") expose higher entropy than voxels from the interior of the brain. We conjecture that this makes the proposed generator more robust as it cannot rely on not having to model certain voxels.

Method	Elapsed time per sample [sec.]
FACE MASK	91 ± 5
DEFACE	71 ± 3
CP-GAN	71 ± 3
QUICKSHEAR	28 ± 4

Table 1: *Execution Time Measurement*: Mean and standard deviations are aggregated over 200 runs on different scans.

6 Execution Time Measurement

As de-identification tools are meant to be applied within a clinical environment, it is important for them not to be overly time-consuming. Therefore, we measure the elapsed time[2] that each method takes. To put all methods on an equal footing, we decide to start measuring the time just before the *NifTi-1*[3] image is loaded and stop measuring once the de-identified *NifTi-1* scan was generated. Moreover, we decide to omit the time that was spent in *Model-Agnostic Preprocessing* since it is the same for all methods. All methods are single-threaded and we only execute one process at a time. The benchmark system is given by a Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz, 252 GiB of RAM and 1 Quadro RTX 6000 GPU (24 GiB main memory). All methods are executed exactly 200 times.

In *Table* 1 we observe that CP-GAN's execution time is in line with the other de-identification methods. It ranks second, however QUICKSHEAR is more than two times faster than CP-GAN and DEFACE, with FACE MASK taking the most time per sample. It is worthwhile to mention that both QUICKSHEAR and CP-GAN leverage the Robex [6] algorithm to compute a brain mask, which, when run in isolation, already takes 27 seconds per sample on average. This insight provides motivation to speed up either method by supplanting Robex with a faster algorithm if time constraints are a priority.

[2] Wall clock time

[3] NifTi-1 is a widely-established format for MR imagery, used to ship OASIS-3 & ADNI

7 Hyperparameters

The table below provides a list of the hyperparameter values used in the experiments appearing in the main text.

Hyperparameter	Value
Batch size	2
Number of steps (OASIS-3, ADNI)	35,000
Min. generated resolution s	4
Max. generated resolution S	128
Number of blocks N_B	6
Kernel size k	5
Leaky ReLU steepness	0.2

8 Potentially vulnerable properties

We noticed a gap between CP-GAN's performance and the theoretical optimal in the de-identification experiments above. It is our conjecture that this gap may be explained by certain properties common to all the de-identification methods considered in this work. In general, the *size* and *shape* of a patient's head, if known to an attacker, may be exploited to eliminate a large portion of candidates and narrow the search. By design, the shape of a head synthesized by CP-GAN is only determined up to the convex hull, ruling out exact (privacy-compromising) reconstructions. Yet, we hypothesize that CP-GAN preserves the *size* of the head, as the discriminator would have deemed inconsistently-sized heads as being unrealistic.

To test this, we apply a simple resizing scheme to the privacy transformed representation to observe the effect on the synthesized volume (if it indeed scales with the conditional volumes and generates realistic features). The scaling limits $(\alpha_{\min}, \alpha_{\max})$ were chosen to reflect the distribution of brain sizes present in the data. We are interested in showing that the GAN is able to realistically synthesize volumes in approximately the same conditions that it has seen during training. We identify the 5% and 95% quantiles of the brain mask volumes, η_5 resp. η_{95}. From this, we set limits $\alpha_{\min} = \eta_5/\eta_{50} \approx 0.88$ and $\alpha_{\max} = \eta_{95}/\eta_{50} \approx 1.1$. Thus, $(\alpha_{\min}, \alpha_{\max})$ reflect the extremes of head size appearing in the training data. Recall that the privacy transform $\gamma(x)$ contains the brain mask, brain data, and convex hull $[b(x), x \circ b(x), c(x)]$ and serves as the conditioning variable for CP-GAN. We scale these volumes according to $\alpha \in \mathbb{R}$ by resizing from resolution S^3 to $\lfloor \alpha S \rfloor^3$, then use a linear spline interpolation to infer intensities at integral positions (where $\lfloor \cdot \rfloor$ denotes the floor function). The resulting volume is either center cropped or evenly padded, yielding a convex hull, brain mask, and brain scaled by α within a volume of resolution S^3 as shown in the top three rows of Figure 1. As in all other experiments, we pick a resolution of $S = 128$.

The bottom row of Figure 1 contains the output of CP-GAN, which manages to account for variations in size of the conditioning information without any visible degradation of quality. In principal, this means CP-GAN shares the same potential size (but not shape) vulnerabilities as the other de-identification methods. We believe that these properties are not specific enough to represent a fundamental compromise to patient privacy. Nevertheless, future work

	USER-BASED		MODEL-BASED	
	OASIS-3	ADNI	OASIS-3	ADNI
ORIGINAL	****	****	****	****
BLURRED	****	****	****	****
FACE MASK	****	****	****	****
DEFACE	****	****	****	****
QUICKSHEAR	****	****	****	****
BLACK	**	ns	****	****
MRI WATERSHED	ns	ns	****	****

Table 2: *Analysis of statistical significance (de-identification quality)*: We perform the *Wilcoxon test* between CP-GAN and all other methods. We observe that that the superior performance of CP-GAN is generally *statistically significant* with respect to the competing methods FACE MASK, DEFACE and QUICKSHEAR. Interestingly, we find that the *user-based* de-identification performance of CP-GAN is *statistically insignificant* as compared to the (hard) control tasks BLACK and MRI WATERSHED, indicating that users find CP-GAN to be similarly difficult. The same does not apply to the *model-based* scenario in which the *Siamese* network proves to be more consistent across the aforementioned methods.

	Sørensen-Dice coefficient ↑								*Intersection-over-Union* (IoU) ↑							
	OASIS-3				ADNI				OASIS-3				ADNI			
	BRAIN	VCSF	WHITE	GREY	BRAIN	VCSF	WHITE	GREY	BRAIN	VCSF	WHITE	GREY	BRAIN	VCSF	WHITE	GREY
FACE MASK	****	****	****	****	ns	ns	**	****	****	****	****	****	ns	ns	**	****
DEFACE	****	****	****	****	****	****	****	****	****	****	****	****	****	****	****	****
QUICKSHEAR	****	****	****	****	***	***	****	****	****	****	****	****	***	***	****	****
MRI WATERSHED	****	****	****	****	****	****	****	****	****	****	****	****	****	****	****	****

Table 3: *Analysis of statistical significance (brain segmentation):* We perform the (paired) *Wilcoxon test* between CP-GAN and all other methods. We observe that the superior performance of CP-GAN is generally found to be *statistically significant* with the exception of FACE MASK on the BRAIN and VCSF regions (ADNI).

may consider avenues to address this, for example, by randomly resizing the de-identified scan.

9 Statistical significance of experimental results

In addition to the identification rate of each model, we also report in Table 2 whether the results of competing methods are *statistically significant* when compared to CP-GAN. We perform a *Wilcoxon* [26] test assuming that the *null hypothesis* is equality of all model-specific identification rates. A *Wilcoxon* test is similar to a *paired Student's t-test* but does not assume that the difference of identification rates between any two models is necessarily normally distributed. We use *Bonferroni's correction* [3] to compensate for effects that could lead to a overestimation of statistical significance in a multi-comparison scenario [1].

In the same manner, we also perform a *Wilcoxon* test for our brain segmentation experiment (Table 3). We find that CP-GAN's *Sørensen-Dice coefficient* and its *Intersection-over-Union w.r.t.* original segmentations are *significantly* different from those of other methods – with the exception of FACE MASK on the modalities VCSF and BRAIN.

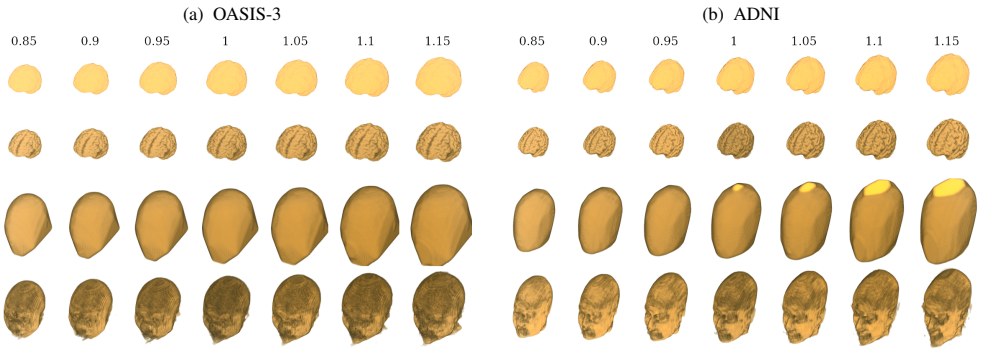

Figure 1: *Potentially vulnerable properties:* We demonstrate why CP-GAN does not achieve perfect de-identification fidelity. It is because it synthesizes heads compatible to the size of the conditional input information $\gamma(x)$, allowing users to perform a process of elimination in which dis-proportioned heads can be eliminated. In the first three rows, we manipulate $\gamma(x)$, the privacy transform CP-GAN is conditioned on to reflect natural variations appearing in the data (scale factor $\alpha \in [0.85, 1.15]$). Given these inputs, CP-GAN produces realistic and appropriately sized outputs shown in the bottom row.

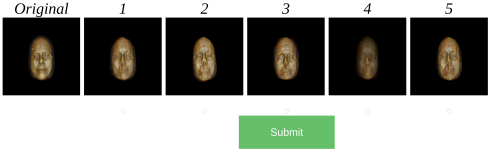

Figure 2: *User study:* We show the original rendering and 5 renderings of different patients to AMT users. Their task is to defeat the de-identification method by selecting the rendering matching the query. Here, OASIS-3 patients are de-identified using CP-GAN. "5" is the remodeled original.

10 De-identification quality user study

An exemplary question asked on *Amazon Mechanical Turk* can be found in Figure 2. A worker is said to have correctly answered a question if the original rendering was correctly associated to the *de-identified* rendering belonging to the same scan resp. person. The identification rate is defined as the percentage ratio of correctly identified pairings and total number of questions. Guessing therefore results in a success rate of $\frac{1}{5} = 20\%$.

11 De-identification quality model-based study

We use a batch size equal to 8, the Adam optimizer [11] with a learning rate of 10^{-3} ($\beta = [0, 1 - 10^{-2}]$) and train the *Siamese* network for 20 epochs. Individual layers are made up of *Convolution – Dropout* [21] – *Swish* [14] – *Instance Norm* [24] sub-layers. We have noticed that the number of epochs does not seem to play a huge role, as we were unable to make out any differences opting for 10, 15 or 25 epochs. The data is shuffled after each epoch.

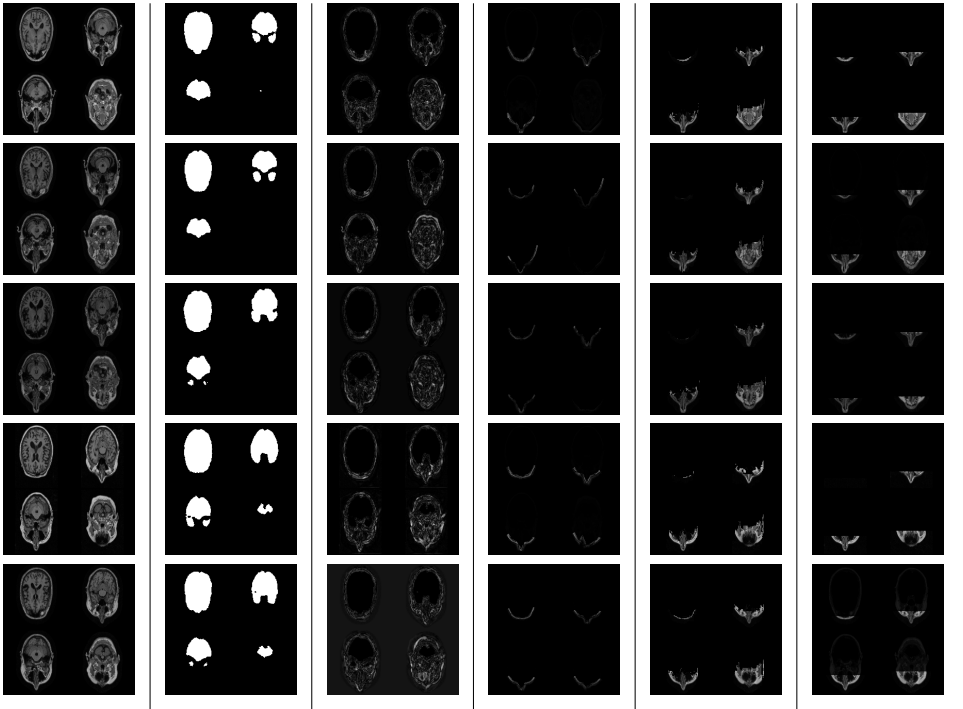

Figure 3: *Visualizing the regions altered by de-identification*: We show the regions that have been altered through de-identification by visualizing the *absolute difference* between de-identified slices and original slices. We include BRAIN in column 2 to indicate the medically relevant brain region that should *not* be altered via de-identification. CP-GAN and all other methods perfectly preserve the brain – we observe that the pixels corresponding to the brain are zero (indicated by black color). Furthermore, we observe that CP-GAN is provides a more comprehensive modification of the scan, as can be seen by the larger differences surrounding the brain including regions around the skull as well as ocular, nasal, and oral cavities. By design, regions where BRAIN is black can be modified by CP-GAN.

12 Visualizing de-identification changes

To visualize the regions that have been changed the various de-identification methods, we provide *absolute difference* maps in Figure 3. We observe that CP-GAN de-identifies in a much more global sense than the removal-based methods, modifying regions around the whole skull as well as ocular, nasal, and oral cavities. In all cases, the clinically-relevant brain region (shown in the 2^{nd} column) remains unaltered.

13 Deep learning-based age prediction

Machine learning algorithms can be trained to estimate brain age from MRI scans, and the difference between predicted and chronological age is shown to have links to aging and brain disease [8]. We investigate whether de-identification adversely affects brain age estimation.

	Mean $\lvert n_D - n_O \rvert$ [yr] ↓	
	OASIS-3	ADNI
ORIGINAL	0.000	0.000
FACE MASK	**2.07**	**0.38**
DEFACE	2.39	5.81
QUICKSHEAR	2.48	12.30
CP-GAN	2.38	1.80
MRI WATERSHED	33.0	$2.9 \cdot 10^6$

Table 4: A network (trained on original imagery) predicts brain age on the original scan (n_O) and a de-identified scan (n_D) for each subject in the test set. We show the mean absolute difference $\lvert n_D - n_O \rvert$ in years, computed over 5 runs. Differences near 0 are better as they indicate less adverse effects from de-identification.

We train a three-dimensional convolutional feed-forward network with an L_1-loss function to estimate brain age in MRI scans (ground truth: chronological age). We assess how the network's predicted age n_D on the de-identified scans compares to the predicted age on the originals n_O by measuring the absolute difference $\lvert n_D - n_O \rvert$ between the two in years.

The results appear below in Appendix Table 4. We find that our model consistently outperforms DEFACE and QUICKSHEAR, with notably little bias for both ADNI and OASIS-3. Not surprisingly, FACE MASK shows the least bias, this can be explained with the fact that it only blurs the face and therefore retains almost all of the age information. It is worth noting that an uncertainty of 3-4 years is typical, as chronological age is a noisy label. The deviations reported by the de-identification methods are in the range of similar age estimation studies [4, 5, 23], suggesting that the effect on age estimation is acceptable. The performance of CP-GAN is surprisingly good considering that the model exploits age cues in regions outside of the brain, suggesting CP-GAN may implicitly model age information from the brain it conditions on.

We leveraged the Adam optimizer [11] with a learning rate of 10^{-3} whereas the batch size was chosen to be 16. We trained the model for 20 epochs. More details can be found in the supplied code base.

14 Proofs

Sparsity Preservation of Binary Downsampling. Assume that $m \in \{0,1\}^{2S \times 2S \times 2S}$ denotes some arbitrary binary image. Let $m' \in \{0,1\}^{S \times S \times S}$ the result of performing $2 \times 2 \times 2$ average pooling on m. We interpret each voxel value $m'_{i,j,k}$ as a parameter to a *Bernoulli* distribution estimated by averaging over 8 voxel values from m allowing us to draw a sample from each voxel. The sampled result, denoted by m'', is binary again and preserves the degree of sparsity $\zeta(m)$ of m in expectation, *i.e.*:

$$\zeta(m'') = \mathbb{E}\left[\frac{1}{S^3}\sum_{i,j,k} m''_{i,j,k}\right] = \frac{1}{(2S)^3}\sum_{i,j,k} m_{i,j,k} = \zeta(m)$$

Proof:

For:

$$m''_{i,j,k} \sim \mathcal{B}\left(p = 1/8 \sum_{i_0=0}^{1}\sum_{j_0=0}^{1}\sum_{k_0=0}^{1} m_{2i+i_0,2j+j_0,2k+k_0}\right)$$

we obtain:

$$\begin{aligned}
\zeta(m'') &= \mathbb{E}\left[\frac{1}{S^3}\sum_{i,j,k=0}^{S-1} m''_{i,j,k}\right] \\
&= \frac{1}{S^3}\sum_{i,j,k=0}^{S-1} \mathbb{E}\left[m''_{i,j,k}\right] \\
&= \frac{1}{S^3}\sum_{i,j,k=0}^{S-1} 1/8 \sum_{i_0=0}^{1}\sum_{j_0=0}^{1}\sum_{k_0=0}^{1} m_{2i+i_0,2j+j_0,2k+k_0} \\
&= \frac{1}{(2S)^3}\sum_{i,j,k=0}^{2S-1} m_{i,j,k} = \zeta(m)
\end{aligned}$$

where the last step follows from the observation that every element in m occurs exactly once in the summation.

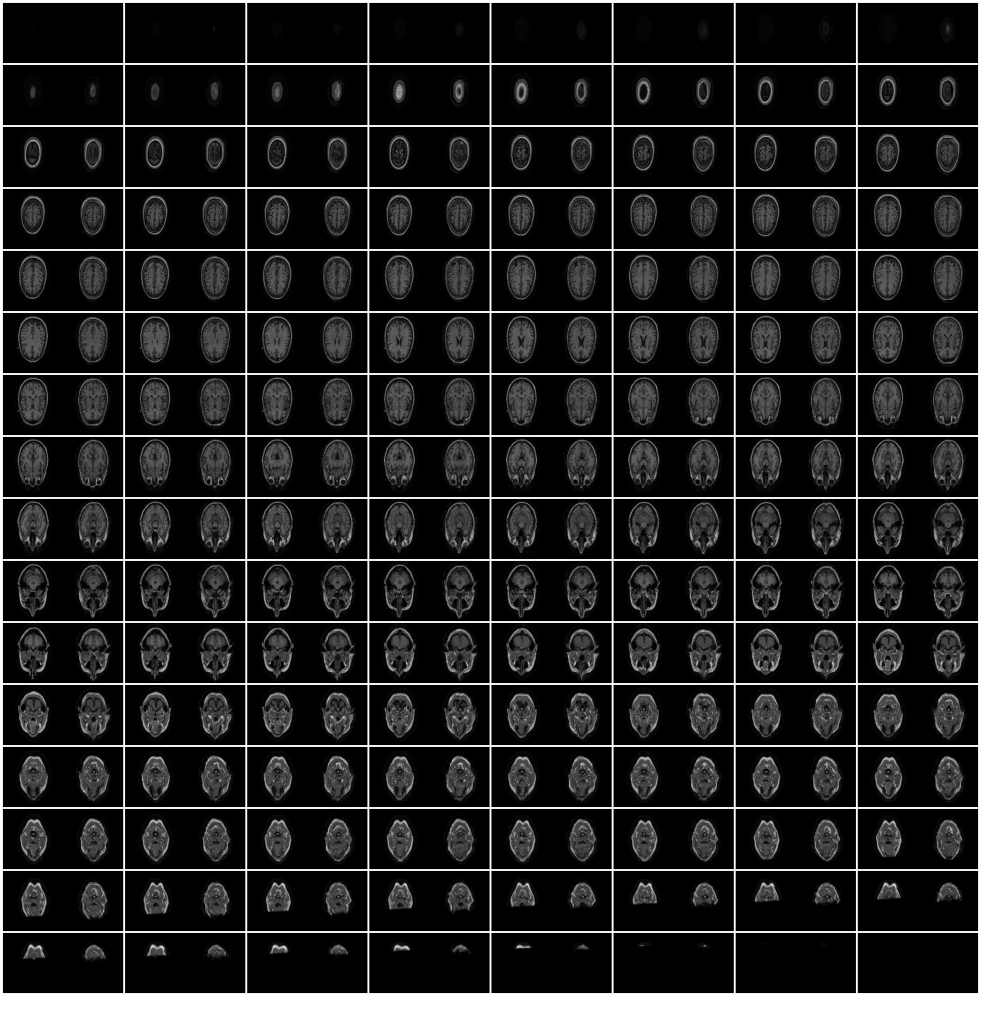

Figure 4: *Original MRI slices and those generated from* CP-GAN (OASIS-3): Slices (of a single patient) run from left to right and from top to bottom. Each box corresponds to one slice index and contains the original on the left and the synthesized counterpart on the right.

References

[1] Hervé Abdi et al. Bonferroni and šidák corrections for multiple comparisons. *Encyclopedia of measurement and statistics*, 3:103–107, 2007.

[2] Amanda Bischoff-Grethe, I. Burak Ozyurt, Evelina Busa, Brian T. Quinn, Christine Fennema-Notestine, Camellia P. Clark, Shaunna Morris, Mark W. Bondi, Terry L. Jernigan, Anders M. Dale, Gregory G. Brown, and Bruce Fischl. A technique for the deidentification of structural brain MR images. *Human Brain Mapping*, 28(9): 892–903, sep 2007. ISSN 10659471. doi: 10.1002/hbm.20312.

[3] E Carlo. Bonferroni. teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8 (3-62):4, 1936.

[4] James H Cole, Rudra P K Poudel, Dimosthenis Tsagkrasoulis, Matthan W A Caan, Claire Steves, Tim D Spector, and Giovanni Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *Neuroimage*, 163:115–124, December 2017.

[5] T Huang, H Chen, R Fujimoto, K Ito, K Wu, K Sato, Y Taki, H Fukuda, and T Aoki. Age estimation from brain MRI images using deep learning. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 849–852, April 2017.

[6] Juan Eugenio Iglesias, Cheng Yi Liu, Paul M. Thompson, and Zhuowen Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, sep 2011. ISSN 02780062. doi: 10.1109/TMI.2011.2138152.

[7] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard GAN. *7th International Conference on Learning Representations, ICLR 2019*, jul 2018. URL http://arxiv.org/abs/1807.00734.

[8] Benedikt Atli Jónsson, Gyda Bjornsdottir, TE Thorgeirsson, Lotta María Ellingsen, G Bragi Walters, DF Gudbjartsson, Hreinn Stefansson, Kari Stefansson, and MO Ulfarsson. Brain age prediction using deep learning uncovers associated sequence variants. *Nature communications*, 10(1):1–10, 2019.

[9] D. Jude Hemanth and J. Anitha. Image pre-processing and feature extraction techniques for magnetic resonance brain image analysis. In Tai-hoon Kim, Dae-sik Ko, Thanos Vasilakos, Adrian Stoica, and Jemal Abawajy, editors, *Computer Applications for Communication, Networking, and Digital Contents*, pages 349–356, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-35594-3.

[10] Jaber Juntu, Jan Sijbers, Dirk Dyck, and Jan Gielen. Bias Field Correction for MRI Images. pages 543–551. Springer, Berlin, Heidelberg, oct 2008. doi: 10.1007/3-540-32390-2_64.

[11] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR, dec 2015.

[12] Pamela J LaMontagne, Tammie L.S. Benzinger, John C. Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei Vlassenko, Marcus E. Raichle, Carlos Cruchaga, and Daniel Marcus. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. *medRxiv*, page 2019.12.13.19014902, dec 2019. doi: 10.1101/2019.12.13.19014902.

[13] Mikhail Milchenko and Daniel Marcus. Obscuring surface anatomy in volumetric imaging data. *Neuroinformatics*, 11(1):65–75, jan 2013. ISSN 15392791. doi: 10.1007/s12021-012-9160-3. URL http://www.ncbi.nlm.nih.gov/pubmed/22968671http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3538950.

[14] Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. *CoRR*, abs/1710.05941, 2017. URL http://arxiv.org/abs/1710.05941.

[15] Jacob C. Reinhold, Blake E. Dewey, Aaron Carass, and Jerry L. Prince. Evaluating the impact of intensity normalization on MR image synthesis. In *Proceedings of SPIE–the International Society for Optical Engineering*, volume 10949, page 126. SPIE-Intl Soc Optical Eng, mar 2019. ISBN 9781510625457. doi: 10.1117/12.2513089.

[16] Sudipta Roy, Sanjay Nag, Indra Kanta Maitra, and Samir Kumar Bandyopadhyay. A review on automated brain tumor detection and segmentation from MRI of brain. *CoRR*, abs/1312.6150, 2013. URL http://arxiv.org/abs/1312.6150.

[17] Nakeisha Schimke, Mary Kuehler, and John Hale. Preserving privacy in structural neuroimages. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 6818 LNCS, pages 301–308. Springer, Berlin, Heidelberg, 2011. ISBN 9783642223471. doi: 10.1007/978-3-642-22348-8_26.

[18] F. Ségonne, A. M. Dale, E. Busa, M. Glessner, D. Salat, H. K. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22(3):1060–1075, jul 2004. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.03.032.

[19] Russell Shinohara, Elizabeth Sweeney, Jeff Goldsmith, Navid Shiee, Farrah Mateen, Peter Calabresi, Samson Jarso, Dzung Pham, Daniel Reich, and Ciprian Crainiceanu. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage. Clinical*, 6:9–19, 12 2014. doi: 10.1016/j.nicl.2014.08.008.

[20] Stephen M. Smith, Mark Jenkinson, Mark W. Woolrich, Christian F. Beckmann, Timothy E.J. Behrens, Heidi Johansen-Berg, Peter R. Bannister, Marilena De Luca, Ivana Drobnjak, David E. Flitney, Rami K. Niazy, James Saunders, John Vickers, Yongyue Zhang, Nicola De Stefano, J. Michael Brady, and Paul M. Matthews. Advances in functional and structural MR image analysis and implementation as FSL. In *NeuroImage*, volume 23, pages S208–S219. Academic Press, jan 2004. doi: 10.1016/j.neuroimage.2004.07.051.

[21] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

[22] Nicholas J. Tustison, Brian B. Avants, Philip A. Cook, Gang Song, Sandhitsu Das, Niels van Strien, James R. Stone, and James C. Gee. The ANTs cortical thickness processing pipeline. In John B. Weaver and Robert C. Molthen, editors, *Medical Imaging 2013: Biomedical Applications in Molecular, Structural, and Functional Imaging*, volume 8672, page 86720K. SPIE, mar 2013. doi: 10.1117/12.2007128. URL http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2007128.

[23] M Ueda, K Ito, K Wu, K Sato, Y Taki, H Fukuda, and T Aoki. An age estimation method using 3D-CNN from brain MRI images. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 380–383, April 2019.

[24] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance Normalization: The Missing Ingredient for Fast Stylization. jul 2016. URL http://arxiv.org/abs/1607.08022.

[25] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, John C. Morris, Ronald C. Petersen, Jennifer Salazar, Andrew J. Saykin, Leslie M. Shaw, Arthur W. Toga, and John Q. Trojanowski. The Alzheimer's Disease Neuroimaging Initiative 3: Continued innovation for clinical trial improvement, may 2017. ISSN 15525279. URL http://www.ncbi.nlm.nih.gov/pubmed/27931796http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5536850.

[26] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. ISSN 00994987. URL http://www.jstor.org/stable/3001968.

[27] Bradley T. Wyman, Danielle J. Harvey, Karen Crawford, Matt A. Bernstein, Owen Carmichael, Patricia E. Cole, Paul K. Crane, Charles Decarli, Nick C. Fox, Jeffrey L. Gunter, Derek Hill, Ronald J. Killiany, Chahin Pachai, Adam J. Schwarz, Norbert Schuff, Matthew L. Senjem, Joyce Suhy, Paul M. Thompson, Michael Weiner, and Clifford R. Jack. Standardization of analysis sets for reporting results from ADNI MRI data, may 2013. ISSN 15525279.