Privacy Preserving for Medical Image Analysis via Non-Linear Deformation Proxy Supplementary Materials

Bach Ngoc Kim[1]
bachknk49@gmail.com
Jose Dolz[1]
jose.dolz@etsmtl.ca
Christian Desrosiers[1]
christian.desrosiers@etsmtl.ca
Pierre-Marc Jodoin[2]
pierre-marc.jodoin@usherbrooke.ca

[1] École de Technologie Supérieure
Montreal, Quebec, Canada

[2] Université de Sherbrooke
Sherbrooke, Quebec, Canada

1 Experimental setup

1.1 Data set

We evaluate our method on the task of privacy-preserving brain MRI segmentation. Two datasets are used in our experiments: the Parkinson's Progression Marker Initiative (PPMI) dataset [6] and MRBrainS13 Challenge [7] dataset. The first dataset, which contains longitudinal data, was considered for training the Siamese discriminator to recognize same-subject brain segmentations. The second one is used to evaluate the ability of our generator trained on PPMI to generalize to another dataset.

PPMI This dataset contains T1 3D MRI scans from 350 subjects, acquired on 3T Siemens scanners from 32 different clinics. Each subject has 1–2 baseline acquisitions and 1–2 follow-up acquisitions a year later, resulting in a total of 773 images. MRI scans were registered onto a common MNI space and then reshaped to a size of $72 \times 96 \times 80$ voxels with a $2 \times 2 \times 2\,\text{mm}^3$ resolution, in order to fit the memory capacity of our hardware.

We split the data based on subjects, using 75% of subjects (269 subjects, 592 images) for the training set and remaining 25% (81 subjects, 181 images) for the testing set. As in [4], for the segmentation task, we used the labels generated automatically by Freesurfer for five classes: internal cerebrospinal fluid (CSF int), external cerebrospinal fluid (CSF ext), white matter (WM), gray matter (GM), and nuclei. To set the hyperparameters of our system, we followed a 5 fold cross-validation strategy on the training images. Once these hyperparameters were selected, we retrained the system on the entire training set and report results from the test set.

MRBrainS This second dataset comprises T1 3D MRI scans of 5 subjects obtained with a 3T Philips Achieva scanner, along with ground-truth segmentation masks for three classes: WM, GM and CSF. Note that the MRBrainS 2013 Challenge included data for 15 additional

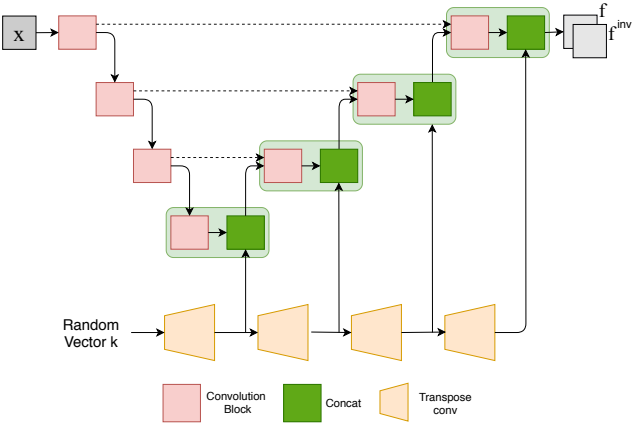

Figure 1: Network architecture of the generator.

subjects, however ground-truth segmentation was not provided for these subjects. As for PPMI, the MRBrainS images were registered to the MNI space using ANTs [2]. Since it has only 5 images, we directly report the cross-validation performance for this dataset.

1.2 Evaluation Metrics

We resort to the Dice similarity coefficient (DSC) to measure the segmentation accuracy of the different methods. Furthermore, to evaluate the methods' ability to obfuscate subject identity, we follow the retrieval-based analysis of [4] where an attacker tries to recover the identity of a subject by matching an encoded image or segmentation map against an existing database. In this analysis, mean average precision (mAP) is employed to measure the re-identification accuracy [5]. Let $\mathbf{x}_i, \mathbf{y}_i$ be an image and its corresponding segmentation map of a given subject with identity $\mathrm{id}(i)$, and denote as $\mathcal{T}_{\mathrm{id}(i)}$ the set of images and segmentation maps of this subject. Also, let $\mathcal{S}_i^k$ be the set of k most similar images to $\mathbf{x}_i$ or segmentations to $\mathbf{y}_i$ according to a given re-identification approach. The precision at cut-off k is defined as

$$\big(\text{precision@}k\big)_i = \frac{\mathcal{T}_{\mathrm{id}(i)} \cap \mathcal{S}_i^k}{k} \tag{1}$$

Considering each proxy test image / segmentation as a separate retrieval task where one must find other proxy images from the same person, the average precision (AP) for image / segmentation i is given by

$$\mathrm{AP}_i = \frac{1}{\sum_{j \neq i} s_{ij}} \sum_{k=1}^{N} \big(\text{precision@}k\big)_i \cdot \widetilde{s}_{ik}, \tag{2}$$

where $\widetilde{s}_{ik}$ is a label indicating if the k-th *most similar* image or segmentation is from the same subject as i or not. mAP is then the mean of AP values computed over all test examples. We also resort to the F1-score to evaluate the performance of the Siamese discriminator.

1.3 Implementation details

We used a U-Net architecture with 3D convolution kernels for the segmentation network and the modified 3D U-Net model of Fig. 1 for the transformation generator. For all experiments,

we trained the system for 100 epochs with the Adam optimizer and a learning rate of 10^{-4}. The balancing weights in Eq. (2) of the main paper were set to $\lambda_1 = 0.5$, $\lambda_2 = 1$, $\lambda_3 = 10$ and $\lambda_4 = 1$. The system was implemented in Pytorch, and experiments were performed on Intel(R) Core(TM) i7-6700K 4.0GHz CPU with a 16 GB NVIDIA Tesla P100 GPU. Additional implementation details can be found in the supplementary material. The code will be made public upon acceptance of this paper.

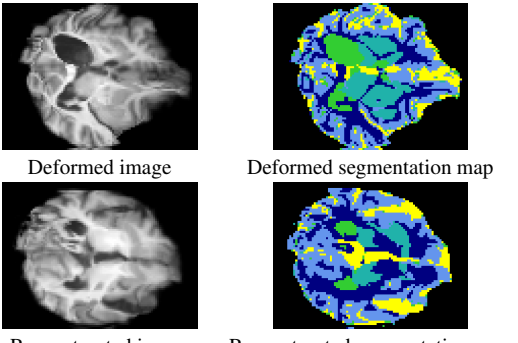

Figure 2: Reconstructed image and segmentation map of a deformed brain without the invertibility loss of Eq. (5) in the main paper.

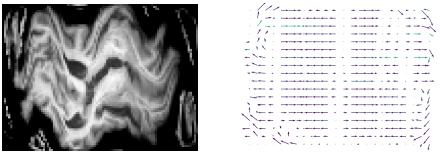

Figure 3: Artifact in the deformed image when train without smoothness.

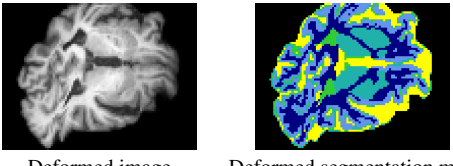

Figure 4: Weak distortion when train without diversity loss.

2 Link to mutual information minimization

Following a similar approach as in [9], we can show that optimizing the adversarial loss in Eq. (4) of the main manuscript amounts to minimizing the mutual information (MI) between encoded images and a variable representing subject identity.

The link between MI and the adversarial training of a discriminator was first shown in the InfoGAN paper [3], where a generative adversarial network conditioned on a set of

latent variables (e.g., object class, pose, orientation, etc.) is trained to maximize the MI between the generated image $\mathbf{x}$ and the latent variables C. This theoretical result relied on the principle of variational MI maximization [1], in which a lower bound on MI is derived by approximating a conditional distribution $P(C|\mathbf{x})$, intractable to compute, with an auxiliary distribution $Q(C|\mathbf{x})$.

Here, a similar approach is used to show that optimizing a discriminator D in a mini-max fashion helps remove information about subject identity. However, since our model is based on a Siamese discriminator, not a standard classifier, a slightly different strategy is needed. In our case, C is a binary variable that equals 1 if two transformed images $\mathbf{x}_1^d$, $\mathbf{x}_2^d$ are from the same subject, else it equals 0. We want to show that optimizing Eq. (4) minimizes the MI between these images and variable C, denoted as $I\big(\mathbf{x}_1^d, \mathbf{x}_2^d; C\big)$. As described in Section 3.2.2 of the main manuscript, we found that the segmentation $\widehat{\mathbf{y}}^d$ predicted from the transformed $\mathbf{x}^d$ provided as much information about the subject identity as $\mathbf{x}^d$. Thus, we have that $I\big(\mathbf{x}_1^d, \mathbf{x}_2^d; C\big) \approx I\big(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d; C\big)$ and can instead focus on minimizing the second term:

$$
\begin{aligned}
I\big(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d; C\big) &= H\big(C\big) - H\big(C \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d\big) && (3)\\
&= H\big(C\big) + \mathbb{E}_{\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d \sim P(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d)} \Big[\mathbb{E}_{C' \sim P(C|\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d)} \big[\log P\big(C' \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d\big) \big] \Big] && (4)\\
&= H\big(C\big) + \mathbb{E}_{C \sim P(C),\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d \sim P(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d|C)} \Big[\mathbb{E}_{C' \sim P(C|\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d)} \big[\log P\big(C' \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d\big) \big] \Big] && (5)
\end{aligned}
$$

where $H(X)$ being the Shannon entropy of random variable X. Since $P(C|\widehat{\mathbf{y}}, \widehat{\mathbf{y}}')$ is unknown and computationally untractable, we approximate it with an auxiliary distribution $Q(C|\widehat{\mathbf{y}}, \widehat{\mathbf{y}}')$, giving the following lower bound:

$$
\begin{aligned}
I\big(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d; C\big) &= H\big(C\big) + \mathbb{E}_{C \sim P(C),\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d \sim P(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d|C)} \Big[&& (6)\\
&\quad \mathbb{E}_{C' \sim P(s|\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d)} \big[\underbrace{D_{\mathrm{KL}}\big(P\big(C' \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d\big) \,\big\|\, Q\big(C' \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d\big)\big)}_{\geq 0} + \log Q\big(C' \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d\big) \big] \Big] && (7)\\
&\geq H\big(C\big) + \mathbb{E}_{C \sim P(C),\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d \sim P(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d|C)} \Big[\mathbb{E}_{C' \sim P(C|\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d)} \big[\log Q\big(C' \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d\big) \big] \Big] && (8)
\end{aligned}
$$

Using the fact that, for random variables X, Y and function $f(x,y)$, $\mathbb{E}_{x \sim X, y \sim Y|x}\big[f(x,y)\big] = \mathbb{E}_{x \sim X, y \sim Y|x, x' \sim X|y}\big[f(x',y)\big]$ [3], we then get

$$
\begin{aligned}
I\big(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d; C\big) &\geq H\big(C\big) + \mathbb{E}_{C \sim P(C),\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d \sim P(\widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d|C)} \big[\log Q(C \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d) \big] && (9)\\
&= H\big(C\big) + \mathbb{E}_{C \sim P(C),\, \mathbf{k}_1, \mathbf{k}_2 \sim P(\mathbf{k}_1, \mathbf{k}_2|C),\, \widehat{\mathbf{y}}_1^d \sim P(\widehat{\mathbf{y}}_1^d|\mathbf{k}_1),\, \widehat{\mathbf{y}}_2^d \sim P(\widehat{\mathbf{y}}_2^d|\mathbf{k}_2)} \big[\log Q(C \,|\, \widehat{\mathbf{y}}_1^d, \widehat{\mathbf{y}}_2^d) \big] && (10)
\end{aligned}
$$

In the last equality, $P(\widehat{\mathbf{y}}^d|\mathbf{k})$ is the distribution of segmentations for a subject identified by $\mathbf{k}$. We can sample from this distribution since the mapping from an input image $\mathbf{x}$ to the transformed segmentation $\widehat{\mathbf{y}}^d$ is deterministic.

The link between MI and the adversarial loss of our model then comes from using the output of discriminator D as auxiliary distribution Q, and observing that Eq. (10) corresponds to the minus the cross-entropy in the adversarial loss plus a constant (i.e., $H(C)$). Optimizing the adversarial thus corresponds to minimizing an maximally-tight bound to MI.

References

[1] David Barber Felix Agakov. The IM algorithm: a variational approach to information maximization. *Advances in neural information processing systems*, 16(320):201, 2004.

[2] B. Avants, N. Tustison, G. Song, PA Cook, A. Klein, and JC Gee. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.

[3] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[4] Bach Ngoc Kim, Jose Dolz, Pierre-Marc Jodoin, and Christian Desrosiers. Privacy-net: An adversarial approach for identity-obfuscated segmentation of medical images, 2020.

[5] Kuldeep Kumar, Matthew Toews, Laurent Chauvin, Olivier Colliot, and Christian Desrosiers. Multi-modal brain fingerprinting: a manifold approximation based framework. *NeuroImage*, 183:212–226, 2018.

[6] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The Parkinson Progression Marker Initiative (PPMI). *Progress in neurobiology*, 95(4):629–635, 2011.

[7] Adriënne M Mendrik, Koen L Vincken, Hugo J Kuijf, Marcel Breeuwer, Willem H Bouvy, Jeroen De Bresser, Amir Alansary, Marleen De Bruijne, Aaron Carass, Ayman El-Baz, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Comp. Intel. and Neuro.*, 2015:1, 2015.