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

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

Abstract

We propose a client-server system which allows for the analysis of multi-centric medical images while preserving patient identity. In our approach, the client protects the patient identity by applying a pseudo-random non-linear deformation to the input image. This results into a proxy image which is sent to the server for processing. The server then returns the deformed processed image which the client reverts to a canonical form. Our system has three components: 1) a flow-field generator which produces a pseudo-random deformation function, 2) a Siamese discriminator that learns the patient identity from the processed image, 3) a medical image processing network that analyzes the content of the proxy images. The system is trained end-to-end in an adversarial manner. By fooling the discriminator, the flow-field generator learns to produce a bi-directional non-linear deformation which allows to remove and recover the identity of the subject from both the input image and output result. After end-to-end training, the flow-field generator is deployed on the client side and the segmentation network is deployed on the server side. The proposed method is validated on the task of MRI brain segmentation using images from two different datasets. Results show that the segmentation accuracy of our method is similar to a system trained on non-encoded images, while considerably reducing the ability to recover subject identity.

1 Introduction

Convolutional neural networks (CNNs) are the *de facto* solutions to a large number of medical image analysis tasks, from disease recognition, to anomaly detection, segmentation, tumor resurgence prediction, and many more [6, 15, 16, 33]. While solutions to these decade long problems are flourishing, a consistent obstacle to their deployment has been privacy protection. Despite being essential to preserve human rights, privacy protection rules are nonetheless a break on the development of machine learning methods, and in particular to cloud-based medical image analysis solutions. However, cloud-based solutions have great benefits, such as preventing clinics from having to purchase and maintain specialized hardware. As such, if these systems are to prosper in the medical world, they will have to integrate privacy protection policies to their processes.

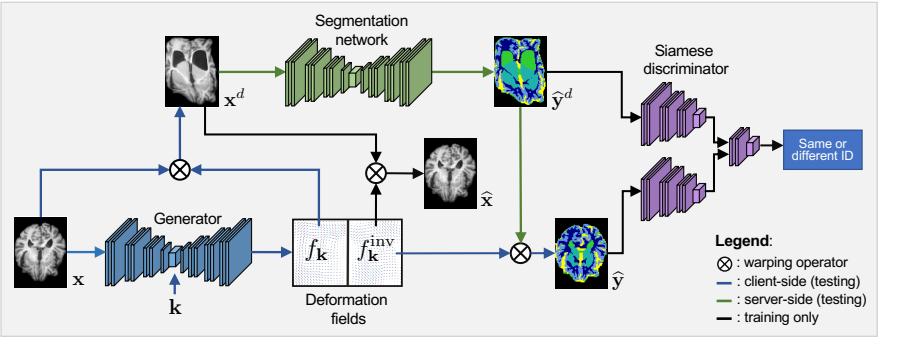


Figure 1: Diagram of the proposed system. Once deployed (testing), the client performs operations identified by blue lines, and the server operations corresponding to green lines.

The simplest privacy protection protocol is anonymization. For medical images, this means removing patient tags from DICOM images or converting it into identity-agnostic formats such as TIFF. Unfortunately, patient identity can be recovered just by inspecting raw images [12, 14]. Results reported in Section 4.1 show that the identity-recognition F1-scores can go up to 98%. Needless to say, data exchanged between the client and the server can be encrypted. While this ensures protection against outside cybercriminals, it does not protect against malicious people from within the organization. Alternatively, one can use homomorphic encryption which allows to perform forward and backward passes of encrypted data without having to decrypt it [22, 45]. Although these methods perform well in some applications, homomorphic cryptosystems typically incur high computational costs [10, 22, 24] and are mostly restricted to simple neural networks.

In this paper, we propose a novel client-server system which can process medical images while preserving patient identity. As shown in Fig. 1, instead of sending an image $\mathbf{x}$ to the server, the client deforms the image with a non-linear spatial deformation field $f_{\mathbf{k}}$ conditioned on a client-specific private key $\mathbf{k}$. The warped image $\mathbf{x}^d$ is then sent to the server where it is processed and sent back to the client. At the end, the deformed result $\mathbf{y}^d$ is unwarped with the inverse transformation function f^{inv}. Results obtained on the task of 3D MRI brain image segmentation reveal that the patient identity is preserved both on the MRI image and the segmentation map while keeping a high segmentation accuracy.

2 Related works

Homomorphic encryptions One way of preserving privacy is via homomorphic-encryption (HE) [7, 10, 22], which allows neural networks to process encrypted data without having to decrypt it. However, HE is not void of limitations. First, it has non-negligible communication overhead [29]. Furthermore, being limited to multiplications and additions, non-linear activation functions have to be approximated by polynomial functions, which makes CNNs prohibitively slow (Nandakumar *et al.* [22] report processing rates of 30 min per image). Thus, homomorphic networks have been relatively simplistic [9] and it is not clear how state-of-the-art deep neural nets [28] can accommodate this approach.

Federated learning Another solution for multi-centric deep learning data analysis is federated learning. [13, 19, 20, 32, 39, 41]. The idea of this approach is to train a centralized model by keeping the data of different clients decentralized and exchanging model parameters or back-propagated gradients during training. While it improves privacy by not sharing data, it requires significant network bandwidth, memory and computational power, and is

susceptible to data leakage from specialized attacks like model inversion [36, 44].

Privacy preserving with adversarial learning A popular solution consists in training a generator to create perturbed images, from either a noise distribution [40] or real images [26]. Then, the generated images are employed to train a discriminator to differentiate between original and synthetic images. Nevertheless, the encoding in these frameworks is not optimized under the supervision of specific utility objectives, potentially achieving sub-optimal results and sacrificing the performance on the utility task. To overcome this limitation, recent works have integrated specific utility losses, which are jointly optimized with the privacy objectives [25, 27, 30, 37, 38, 42]. These approaches, which typically tackle simple problems (i.e., QR code classification or face recognition), resort to standard classification models for both the adversarial and task-specific objectives, where the number of classes is fixed. An alternative to alleviate the issue when the number of classes is non-fixed is to employ a Siamese architecture as the discriminator, which predicts whether two encoded images come from the same subject [12, 23].

Differences with existing methods In contrast with prior works, the proposed framework can easily scale-up to non-fixed classes scenarios. Furthermore, compared to [23], our approach presents significant differences both in the objectives and methodology. First, privacy preserving is investigated in the context of biometrical data in [23] (e.g., fingerprint), whereas we focus on volumetric medical images, which are dissimilar in nature. Second, they aim at finding the smallest possible transformation of an image to remove identity information while can be still used by non-specific applications. In contrast, our goal is to obfuscate images with the strongest possible transformation so that subject identity cannot be recovered while at the same time the encoded image can be used to train a model in the segmentation task. This results in important methodological differences, such as an additional network and different objective functions. More related is the work in [12] whose transformations in deteriorated images come in the form of intensity changes. But contrary to our method, the structural information in the segmentation results is preserved, which can be used to retrieve the patient identity.

3 Methods

3.1 Proposed architecture

As shown in Fig. 1, during training, our system consists of three components: a *transformation generator*, a *segmentation network* and a *discriminator*. We describe the role of each of these components below.

3.1.1 Transformation generator

The first component is a generator G that takes as input a 3D image $\mathbf{x} \in \mathbb{R}^{H \times W \times D}$ and a random vector $\mathbf{k} \in \mathbb{R}^{M}$ and outputs a transformation $f_{\mathbf{k}}$ that distorts $\mathbf{x}$ so that the corresponding subject's identity cannot be recovered, yet segmentation can still be performed. Vector $\mathbf{k}$ is a private key, known only by the client, that parameterizes the transformation function and ensures that this function cannot be inferred from distorted images.

In Privacy-Net [12] a generator is also used for this purpose, however, it directly outputs the distorted image. In this work, we follow a different approach where the generator outputs the transformation function $f_{\mathbf{k}}$, which is used afterwards to distort the image. Computing this function explicitly enables to perform the segmentation in the *transformed space*, where identity is obfuscated, and then reverse the transformation back to the original space. To ensure that the transformation is reversible, we could limit $f_{\mathbf{k}}$ to a specific family of functions (e.g., free-form deformation [35]). However, to add flexibility and learn a function most

suitable for the downstream segmentation, we instead enforce the generator to output both $f_{\mathbf{k}}$ and its inverse $f_{\mathbf{k}}^{\text{inv}}$, and use a reconstruction loss (see Section 3.3.3) to impose that $f_{\mathbf{k}}^{\text{inv}} \circ f_{\mathbf{k}} = I$. Given a training example $(\mathbf{x}, \mathbf{y})$, where $\mathbf{y} \in \mathbb{R}^{H \times W \times D \times C}$ is the ground-truth segmentation mask over C classes, $f_{\mathbf{k}}$ is used to compute the distorted image $\mathbf{x}^d = f_{\mathbf{k}}(\mathbf{x})$ and distorted segmentation $\mathbf{y}^d = f_{\mathbf{k}}(\mathbf{y})$. The former is sent to the segmentation network for processing, while $\mathbf{y}^d$ is used to evaluate the segmentation output. On the other hand, the inverse function $f_{\mathbf{k}}^{\text{inv}}$ is used to obtain the reconstructed image $\widehat{\mathbf{x}} = f_{\mathbf{k}}^{\text{inv}}(\mathbf{x}^d)$ and reconstructed segmentation $\widehat{\mathbf{y}} = f_{\mathbf{k}}^{\text{inv}}(\widehat{\mathbf{y}}^d)$ in the original space.

As for the generator (c.f. figure 1 in the supplementary materials) it comprises an encoder path with 4 convolution blocks that takes an input image and computes feature maps of increasingly-reduced dimensions via pooling operations, and a decoder path also with 4 convolution blocks which produces an output map of same size as the input. In this work, we use the generator to predict a flow-field f which assigns a displacement vector $f_{u,v,w} \in \mathbb{R}^3$ to each voxel (u,v,w) of the 3D image $\mathbf{x}$. More information on the transformation is given in Section 3.2. Transpose convolutions are used in the decoder path to upscale feature maps. We also preserve high-resolution information by adding skip connections between convolution blocks at the same level of the encoder and decoder paths. Moreover, to ensure that the private key $\mathbf{k}$ is used at different scales, we include another path in the model that gradually upscales $\mathbf{k}$ with transpose convolutions and concatenates the resulting map with feature maps of corresponding resolution in the decoder path.

3.1.2 Segmentation network

The segmentation network S takes as input the distorted image $\mathbf{x}^d$ and outputs a distorted segmentation map $\widehat{\mathbf{y}}^d = S(\mathbf{x}^d)$. Although any suitable network can be employed, we used a 3D U-Net [4] which implements a convolutional encoder-decoder architecture with skip-connections between corresponding levels of the encoder and decoder.

3.1.3 Siamese discriminator

An adversarial approach is employed to obfuscate the identity of subjects in distorted images and segmentation maps. In a standard approach, a classifier network is used as discriminator D to predict the class (i.e., subject ID) of the encoded image produced by the generator. In our context, where the number of subjects can be in the thousands and grows over time, this approach is not suitable. Alternatively, we follow a strategy similar to Privacy-Net [17] where we instead use a Siamese discriminator that takes as input two segmentation maps, $\mathbf{y}_i$ and $\mathbf{y}_j$, and predicts whether they belong to the same subject or not. Note that this differs from Privacy-Net, which applies the Siamese discriminator on the encoded images, not on the segmentation maps. For training, we generate pairwise labels s_{ij} such that $s_{ij} = 1$ if $\mathbf{y}_i$ and $\mathbf{y}_j$ are from the same subject, otherwise $s_{ij} = 0$. Since we now solve a binary prediction task, which is independent of the number of subjects IDs, this strategy can scale to a large and increasing number of subjects in the system.

3.1.4 Test-time system

At testing, the system can be used for privacy-preserving segmentation as illustrated in Fig. 1. A client-side generator is first used with the client's private key $\mathbf{k}$ to distort the 3D image to segment, $\mathbf{x}$, into an identity-obfuscated image $\mathbf{x}^d = f_{\mathbf{k}}(\mathbf{x})$, which is then sent to the server for segmentation. The server-side segmentation network takes $\mathbf{x}^d$ as input and outputs the distorted segmentation map $\widehat{\mathbf{y}}^d$. Finally, $\widehat{\mathbf{y}}^d$ is sent back to the client where the inverse transform is used to recover the segmentation map $\widehat{\mathbf{y}} = f_{\mathbf{k}}^{\text{inv}}(\widehat{\mathbf{y}}^d)$.

3.2 Transformation function

As in [2, 3], the transformation function in our model takes an image (or segmentation map) and a flow-field as input, and outputs the deformed version of the image. Similarly to the spatial transformer network [11], this geometric deformation is based on grid sampling. Let $\mathbf{b}$ be the base-grid of size $(H,W,D,3)$ containing the coordinates of image voxels, and $\mathbf{f}$ be the deformation flow-field of same size. The coordinates of the deformed grid are then given by $\mathbf{d} = \mathbf{b} + \mathbf{f}$. We obtain the deformed image $\mathbf{x}^d$ by sampling the 8 neighbor voxels around each point of $\mathbf{d}$ using tri-linear interpolation:

$$x^d_{u,v,w} = \sum_{(u',v',w')\in\Omega} x_{u',v',w'} \cdot \max\left(0, 1-|d_u - b_{u'}|\right) \cdot \max\left(0, 1-|d_v - b_{v'}|\right) \cdot \max\left(0, 1-|d_w - b_{w'}|\right) \quad (1)$$

Since Eq. (1) is differentiable, we can back-propagate gradients during optimization.

3.3 Training the proposed model

We train the transformation generator G, the segmentation network S and the discriminator jointly with the following five-term loss function :

$$\mathcal{L}_{\text{total}}(S,G,D) = \mathcal{L}_{\text{seg}}(S) + \lambda_1 \mathcal{L}_{\text{adv}}(G,D) + \lambda_2 \mathcal{L}_{\text{inv}}(G) + \lambda_3 \mathcal{L}_{\text{smt}}(G) + \lambda_4 \mathcal{L}_{\text{div}}(G) \quad (2)$$

Where λ_1, λ_2, λ_3 and λ_4 are hyper-parameters balancing the contribution of each term. In the following subsections, we define and explain the role of each term in this loss function.

3.3.1 Segmentation loss

The segmentation loss enforces that the segmentation network S learns a correct mapping from a distorted image $\mathbf{x}^d = f_{\mathbf{k}}(\mathbf{x})$ to its distorted segmentation $\widehat{\mathbf{y}}^d$. The predicted segmentation after reconstruction is $\widehat{\mathbf{y}} = f^{\text{inv}}_{\mathbf{k}}(\widehat{\mathbf{y}}^d)$. Here, we use a Dice loss [31] to measure the difference between the reconstructed predicted segmentation and its corresponding ground-truth:

$$\mathcal{L}_{\text{seg}}(S) = \min_S \mathbb{E}_{(\mathbf{x},\mathbf{y}),\mathbf{k}}\left[\ell_{\text{Dice}}(\mathbf{y},\widehat{\mathbf{y}})\right] = \min_S \mathbb{E}_{(\mathbf{x},\mathbf{y}),\mathbf{k}}\left[\ell_{\text{Dice}}\left(\mathbf{y}, (f^{\text{inv}}_{\mathbf{k}} \circ S \circ f_{\mathbf{k}})(\mathbf{x})\right)\right]. \quad (3)$$

Since this loss samples over both images $\mathbf{x}$ and random key vectors $\mathbf{k}$, the network S learns a segmentation that accounts for the variability of structures in images and their possible deformation resulting from $f_{\mathbf{k}}$.

3.3.2 Identity obfuscation loss

An adversarial loss is added to ensure that the transformation obfuscates subject identity. By maximizing the discriminator's error, the generator learns to produce transformed images from which identity cannot be recovered. However, this strategy is sensitive to noise or variation in contrast which "fools" the discriminator but still preserves structural information that can identify subjects. To alleviate this problem, we instead apply the discriminator on pairs of segmentation maps. Letting $D(\mathbf{y}_i,\mathbf{y}_j)$ be the probability that $\mathbf{y}_i$ and $\mathbf{y}_j$ are from the same subject, we define this loss as

$$\begin{aligned}\mathcal{L}_{\text{adv}}(G,D) = \min_G \max_D\ & \mathbb{E}_{\mathbf{y}_i,\mathbf{y}_j}\left[s_{ij}\log D(\mathbf{y}_i,\mathbf{y}_j) + (1-s_{ij})\log\left(1-D(\mathbf{y}_i,\mathbf{y}_j)\right)\right] \\ & + \mathbb{E}_{\mathbf{y},\mathbf{k}}\left[\log\left(1-D(\widehat{\mathbf{y}}^d,\widehat{\mathbf{y}})\right)\right] \quad (4)\end{aligned}$$

Table 1: Segmentation and re-identification results on the PPMI dataset.

Method	Segmentation DSC						Re-id. F1-score		Re-id. mAP	
	Overall	GM	WM	Nuclei	int.CSF	ext.CSF	Image	Seg.	Image	Seg.
No-Proxy	0.887	0.941	0.862	0.727	0.745	0.825	0.988	0.986	0.998	0.998
Noise (SNR=1 dB, SPP=0.1)	0.871	0.939	0.857	0.712	0.729	0.813	0.984	0.986	0.997	0.998
Noise (SNR=0.1 dB, SPP=0.5)	0.445	0.463	0.431	0.388	0.372	0.452	0.388	0.575	0.283	0.447
Voxel permutation	0.185	0.190	0.182	0.177	0.245	0.187	0.023	0.011	0.007	0.015
Privacy-Net [12]	0.812	0.925	0.824	0.580	0.598	0.752	–	–	0.189	0.632
Ours (All losses)										
$\lambda_4 = 1$	0.816	0.901	0.829	0.634	0.651	0.735	0.051	0.045	0.096	0.091
$\lambda_4 = 0.5$	0.825	0.909	0.837	0.644	0.662	0.742	0.128	0.113	0.236	0.232
$\lambda_4 = 0.25$	0.847	0.929	0.849	0.671	0.685	0.774	0.287	0.294	0.301	0.297
Ours (w/o Invertibily)	0.511	0.523	0.507	0.467	0.423	0.534	0.038	0.025	0.059	0.034
Ours (w/o Smoothness)	0.701	0.801	0.706	0.455	0.431	0.605	0.059	0.043	0.110	0.088
Ours (w/o Diversity)	0.864	0.934	0.853	0.718	0.711	0.796	0.445	0.473	0.393	0.329

with $\widehat{\mathbf{y}}^d = S(f_{\mathbf{k}}(\mathbf{x}))$ and $\widehat{\mathbf{y}} = f_{\mathbf{k}}^{\text{inv}}(\widehat{\mathbf{y}}^d)$. The first term corresponds to the cross-entropy loss on ground-truth segmentation pairs, that does not depend on the generator or segmentation network. The second term measures the discriminator's ability to recognize that a deformed segmentation and its reconstructed version (by applying the reverse transform function) are from the same subject. This second term is optimized adversarially for G and D. It can be shown using a variational bound method that optimizing the problem in Eq. (4) minimizes the mutual information between a pair $(\widehat{\mathbf{y}}^d, \widehat{\mathbf{y}})$ and the same-identity variable s_{ij} [12]. Consequently, it impedes a potential attacker from retrieving subject identity for a given distorted image by matching it with a database of existing images.

3.3.3 Transformation invertibility loss

When receiving the deformed segmentation from the server, the client needs to bring it back to the original image space. For this to be possible, the transformation function needs to be invertible, i.e. $f^{\text{inv}} \circ f = I$. To enforce this property, we minimize the $\mathcal{L}_{\text{Dice}}$ between a segmentation map and its reconstructed version. However, since the segmentation map is binary, this leads to non-smooth gradients. We avoid this problem by also minimizing the reconstruction error for input images, based on the structural similarity (SSIM) measure:

$$\mathcal{L}_{\text{inv}}(G) = \min_{G} \mathbb{E}_{(\mathbf{x},\mathbf{y}),\mathbf{k}}\Big[\ell_{\text{SSIM}}\big(\mathbf{x}, (f_{\mathbf{k}}^{\text{inv}} \circ f_{\mathbf{k}})(\mathbf{x})\big) + \ell_{\text{Dice}}\big(\mathbf{y}, (f_{\mathbf{k}}^{\text{inv}} \circ f_{\mathbf{k}})(\mathbf{y})\big)\Big]. \tag{5}$$

where $\ell_{\text{SSIM}}(\mathbf{x},\mathbf{y}) \in [0,1]$ is the SSIM loss as in [43].

The global SSIM is generated at each voxel using a $11 \times 11 \times 11$ window, and then taking the average over all voxels. In practice, we use a multi-scale structural similarity (MS-SSIM) which computes the SSIM at multiple image scales via subsampling [34].

3.3.4 Transformation smoothness loss

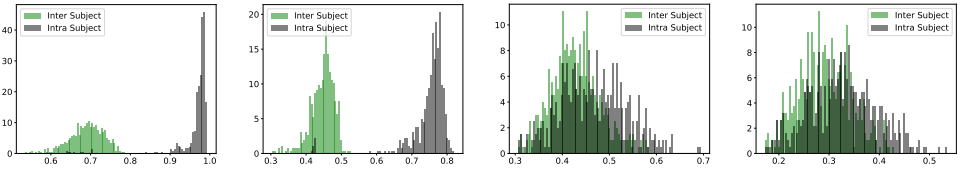
The transformation invertibility loss in Eq. (5) may sometimes lead to discontinuity in the deformation field which prevents the segmentation from being reconstructed. To regularize the deformation field produced by the generator, we include another loss that enforces spatial smoothness:

$$\mathcal{L}_{\text{smt}}(G) = \mathbb{E}_{\mathbf{x},\mathbf{k}}\left[\frac{1}{|\Omega|}\sum_{(u,v,w)\in\Omega} \|\nabla f_{u,v,w}\|_2\right] \tag{6}$$

where the spatial gradient $\nabla f_{u,v,w}$ at each voxel (u,v,w) is estimated using finite difference.

Table 2: Influence of the different terms of the loss function $\mathcal{L}_{\text{total}}$ in the reconstruction.

Method	MS-SSIM	Segmentation DSC					
		Overall	GM	WM	Nuclei	int.CSF	ext.CSF
Ours (All losses)							
$\lambda_4 = 1$	0.993	0.983	0.987	0.982	0.970	0.972	0.985
$\lambda_4 = 0.5$	0.993	0.984	0.988	0.983	0.969	0.975	0.984
$\lambda_4 = 0.25$	0.994	0.987	0.992	0.986	0.975	0.976	0.988
Ours (w/o Invertibility)	0.692	0.574	0.581	0.579	0.569	0.565	0.584
Ours (w/o Smoothness)	0.905	0.829	0.856	0.861	0.822	0.791	0.842
Ours (w/o Diversity Loss)	0.995	0.990	0.994	0.989	0.981	0.980	0.990



(a) (b) (c) (d)

Figure 2: MS-SSIM score and DSC histograms between inter- and intra-subject **(a)** undistorted MR images and **(b)** undistorted segmentation maps **(c)** deformed images and **(d)** deformed segmentation maps.

3.3.5 Transformation diversity loss

A final loss in our model is added to prevent mode-collapse in the generator where the same transformation would be generated regardless of the input private key $\mathbf{k}$. As mentioned before, having a transformation that depends on $\mathbf{k}$ is necessary to avoid an attacker learn to "reverse" the transformation by observing several deformed images or segmentation maps. To achieve this, we maximize the distortion between two deformed versions of the same image or segmentation, generated from different random private keys $\mathbf{k}$ and $\mathbf{k}'$:

$$\mathcal{L}_{\text{div}}(G) = \max_{G} \mathbb{E}_{(\mathbf{x},\mathbf{y}),\mathbf{k},\mathbf{k}'} \left[\ell_{\text{SSIM}}(f_{\mathbf{k}}(\mathbf{x}), f_{\mathbf{k}'}(\mathbf{x})) + \ell_{\text{Dice}}\big(f_{\mathbf{k}}(\mathbf{y}), f_{\mathbf{k}'}(\mathbf{y})\big) \right]. \tag{7}$$

4 Results

We start by evaluating the segmentation and re-identification performance of three different baselines. The first baseline, which we call *no-proxy baseline*, uses non-distorted images of PPMI. In the second one, named *noise baseline*, we add strong noise to the PPMI images to distort them. The third baseline, called *voxel permutation*, distorts an input by randomly shuffling the order of voxels while keeping their intensity the same. This last baseline is used to evaluate a scenario where all geometric information of the image is lost. We then evaluate our privacy-preserving segmentation method on the same data, and conduct an ablation study to measure the contribution of each loss term. Last, we assess our method's ability to generalize on MRBrainS data.

Dataset. 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 [17] and MRBrainS13 Challenge [21] 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. More details on these datasets can be found in the supplementary materials.

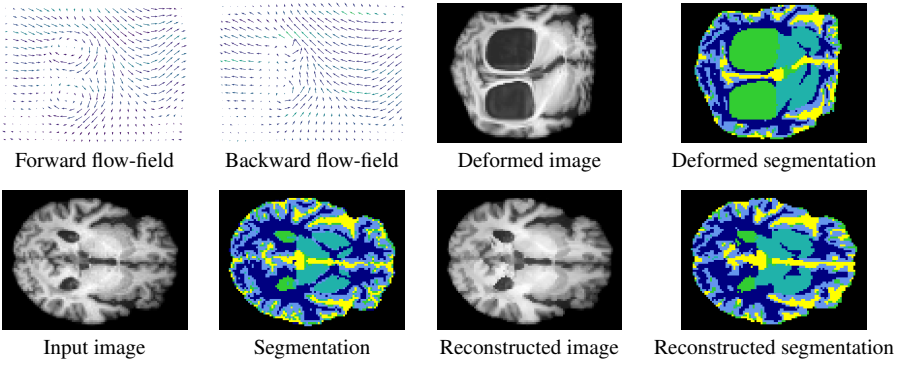


Figure 3: Visualization of forward f and backward f^{inv} deformation fields, input images with its associated ground truth map, deformed image and segmentation map and the reconstructed images and segmentation maps.

4.1 Baseline

4.1.1 No-proxy Baseline

Re-identification We measure the ability of the Siamese discriminator trained independently to correctly recover the identity of a patient with the original, non-distorted images and segmentation maps of PPMI. These *no-proxy* results are reported in the first row of Table 1, where we observe that the F1-scores of the discriminator are above 98% and the mAP is close to 100%. To further compare the inter / intra-subject similarity, we computed a MS-SSIM score between each pair of MRI images and each pair of segmentation maps and put the inter / intra histograms in Fig. 2 (a) and (b). As can be seen, when considering non-encoded images, the intra-subject MS-SSIM scores (grey curves) are significantly larger than that of the inter-subjects (green curves). This demonstrates that identity can be recovered easily from non-distorted images.

Segmentation Result In the *no-proxy* row of Table 1, we also report the segmentation Dice scores of our segmentation method trained on the undistorted images. The overall Dice score is the average Dice across regions weighted by the region size. These results correspond roughly to those obtained in a recent publication for a similar architecture [5]. Note that the nuclei and the internal CSF have a lower Dice due to their smaller sizes.

4.1.2 Added Noise Baseline

For this baseline, we added two types of noise to the MRI images: a Rician noise with its associated SNR, and a salt and pepper (SPP) noise, where the noise level is measured as the probability of setting a voxel to 0 or 1.

The results for this baseline are reported in the second and third rows of Table 1. In terms of re-identification performance, with a Rice noise SNR $1dB$ and salt pepper noise density of 0.1, the Siamese discriminator can easily re-identity subjects, obtaining F1-scores and mAP values above 98%. However, for a Rice noise SNR of $0.1dB$ and salt pepper noise density of 0.5, the image content is almost destroyed. In this case, the Siamese discriminator fails to re-identify the subject and shown by the very low F1-score and mAP.

Looking at the segmentation performance of the method on noisy images, we see that the segmentation network is also robust to moderate noise. Thus, for a Rice noise SNR $1dB$ and salt pepper noise density of 0.1, the Dice score of the method is similar to the *non-proxy*

baseline. The segmentation however collapses when the image is corrupted by pepper noise with a density of 0.5.

4.1.3 Voxel Permutation Baseline

As mentioned above, this baseline randomly shuffles the order of voxels in an image and its corresponding segmentation ground-truth. We train a Siamese discriminator to re-identify the shuffled images, and a U-Net to segment the shuffled images. The results of this baseline are reported in the fourth row of Table 1. As expected, this strong distortion removes the Siamese discriminator's ability to re-identify subjects, which obtains F1-scores and mAP values lower than 2%. Moreover, the segmentation network cannot segment the shuffled images correctly and obtains catastrophically low Dice scores.

4.2 Results on PPMI

Re-identification Here, we measure the ability of the generator to obfuscate the identity of a patient. Quantitative results for our system are reported on the *All losses* ($\lambda_4 = 1$) row of Table 1. We can see that the F1-scores drop to 5% and the mAP to 9% for both the distorted image and distorted segmentation maps. This indicates that most information on patient identity has been removed from these data. Fig. 2 (c) and (d) gives the inter / intra-subject MS-SSIM score histograms between of deformed images and deformed segmentation map. We observe that the grey and green curves overlap almost entirely, showing that same-subject images are as different as those from separate subjects. Figure 3 depicts an input image and segmentation ground-truth, together with their associated flow-fields, distorted and reconstructed images. Despite the large and variable deformation applied to images, the segmentation network can precisely delineate the complex-shaped brain regions.

Reconstruction The first row of Table 2, i.e., *All losses*, reports the reconstruction accuracy obtained with both input images and segmentation maps. MS-SSIM is used to evaluate the similarities on the raw inputs, whereas we employ the Dice score to measure differences on the segmentation maps. Particularly, we observe that our system is capable of reconstructing both distorted images and segmentation masks, with a MS-SSIM value near to 100% and an overall Dice above 0.98.

Segmentation The segmentation DSC achieved by our method is reported in the *All losses* row of Table 1. The obfuscation procedure being lossy by its very nature, the segmentation scores are slightly below that of the *no-proxy* approach. However, the reported Dice score is higher than 0.80 which is suitable for several clinical applications. This is supported by observations in the clinical literature, where authors report DSC values of 0.70 to be acceptable [1, 8, 46, 47] while others, more conservative, suggest minimum DSC values of 0.80 [18]. That said, if an application requires a larger Dice score, one can improve it by reducing the λ_4 Diversity coefficient (c.f. Eq.(2)). The segmentation results for different values of λ_4 are reported in the *All losses* rows of Table 1 . By doing so, one would improve the overall Dice score all the way to 0.86, i.e. almost on par with No-Proxy. Of course, doing so would result into a slightly larger re-identification F1-Score and mAP. At worst, the F1-score could reach 0.44 which is still much smaller than the 0.988 reported on the first line of Table 1.

Comparison to the state-of-art We also compared our system to the recently-proposed Privacy-Net [12]. As can be seen, while the segmentation Dice scores are globally similar to those of our approach (*all loss* row), our re-identification mAP values are significantly lower both on images and segmentation maps. Note that both the system in [12] and the proposed framework resort to UNet as backbone segmentation architecture. This demonstrates that

Table 3: Segmentation result on the MRBrainS13 test set.

Setting	Overall	GM	WM	CSF
Non-distorted images	0.881	0.879	0.887	0.883
Distorted images	0.839	0.832	0.840	0.835

i) our approach preserves the segmentation capabilities shown in [12], and also *ii)* it can drastically improve the obfuscation of identity.

4.3 Ablation study

To examine the importance of each loss term, we proceeded to the following ablation study.

Invertibility loss We trained the whole system without the invertibility loss of Eq. (5). Although the segmentation loss in Eq. (3) implicitly handles the reconstruction of segmentation maps, it is not sufficient for learning a reversible transformation. As can be seen from Tables 1 and 2, the reconstruction accuracy and the segmentation Dice score for this setting are catastrophically low. This is further illustrated in Fig. 2 of Supplementary Materials were the reconstructed image and segmentation map of a deformed brain are plagued with artifacts.

Smoothness loss We trained the system without the smoothness loss of Eq. (6) that regularizes the flow-field. As shown in Fig. 3 of Supplementary Materials, the resulting flow-field has abrupt discontinuities which degrade the reconstruction accuracy and lead to a drop in accuracy as reported in Tables 1 and 2.

Diversity loss As indicated in Tables 1 and 2, removing the transformation diversity loss of Eq. (7) leads to a higher reconstruction accuracy and Dice score. While this might seem beneficial, it comes at the expense of a higher re-identification F1-score and mAP as shown in the last row of the Table 1. As mentioned before, adjusting the λ_4 coefficient allows one to compromise between strict identity preserving and large Dice score (*All losses* rows of Table 1).

4.4 Results on MRBrainS

To demonstrate the generalizability of the learned transformation for privacy-preserving segmentation, we fixed the generator pre-trained on PPMI and then only retrained the segmentation network on the MRBrainS data. Table 3 reports the segmentation accuracy for non-distorted and distorted images of MRBrainS. Similarly to PPMI, we also observe a small drop of the Dice score between the segmentation results without and with deformation. Particularly, our method achieves an overall Dice of 83.9%, which is nearly 4% lower than the performance on non-deformed images. This suggests that the proposed approach can generalize well to other datasets.

5 Conclusion

We presented a strategy for learning image transformation functions that remove sensitive patient information from medical imaging data, while also providing competitive results on specific utility tasks. Particularly, our system integrates a flow-field generator that produces pseudo-random deformations on the input images, removing structural information that otherwise could be used to recover the patient identity from segmentation masks. This contrasts with prior works, where the image deformations come in the form of intensity changes, leading to the preservation of identifiable structures. This was empirically demonstrated in our experiments, where the proposed system drastically decreased the re-identification performance based on segmentation masks, compared to competing methods. Additional numerical experiments suggest that the proposed approach is a promising strategy to prevent leakage of sensitive information in medical imaging data.

Acknowledgment

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC) as well as the Reseau de BioImagrie du Quebec (RBIQ) and thank NVIDIA corporation for supporting this work through their GPU grant program.

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/access-dataspecimens/download-data). For up-to-date information on the study, visit ppmi-info.org.

PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including list the full names of all of the PPMI funding partners found at www.ppmi-info.org/about-ppmi/who-we-are/study-sponsors.

References

[1] Lisanne Anders, Florian Stieler, Kerstin Siebenlist, Jörg Schäfer, Frank Lohr, and Frederik Wenz. Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. *Journal of the European Society for Therapeutic Radiology and Oncology*, 102:68–73, 09 2011. doi: 10.1016/j.radonc.2011.08.043.

[2] Guha Balakrishnan, Amy Zhao, Mert R. Sabuncu, John V. Guttag, and Adrian V. Dalca. Voxelmorph: A learning framework for deformable medical image registration. *CoRR*, abs/1809.05231, 2018. URL http://arxiv.org/abs/1809.05231.

[3] Krishna Chaitanya, Neerav Karani, Christian F. Baumgartner, Olivio Donati, Anton S. Becker, and Ender Konukoglu. Semi-supervised and task-driven data augmentation. *CoRR*, abs/1902.05396, 2019. URL http://arxiv.org/abs/1902.05396.

[4] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[5] J. Dolz, K. Gopinath, J. Yuan, H. Lombaert, C. Desrosiers, and I. Ben Ayed. HyperDense-Net: A hyper-densely connected CNN for multi-modal image segmentation. *IEEE TMI*, 38(5):1116–1126, 2019.

[6] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170:456–470, 2018.

[7] N. Dowlin, R. Gilad-Bachrach, K. Laine, K. Lauter, M Naehrig, and J. Wernsing. A reproducible evaluation of ANTs similarity metric performance in brain image registration. In *proc of ICML*, 2016.

[8] M. A. Gambacorta, C. Valentini, N. Dinapoli, L. Boldrini, N. Caria, M. C. Barba, G. C. Mattiucci, D. Pasini, B. Minsky, and V. Valentini. Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using auto-segmentation computed system. *Acta Oncol*, 52(8):1676–1681, Nov 2013.

[9] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *ArXiv*, abs/1711.10677, 2017.

[10] Ehsan Hesamifard, Hassan Takabi, and Mehdi Ghasemi. CryptoDL: Deep neural networks over encrypted data. *arXiv preprint arXiv:1711.05189*, 2017.

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. *CoRR*, abs/1506.02025, 2015. URL http://arxiv.org/abs/1506.02025.

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

[13] Jakub Konecný, H. Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016.

[14] 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.

[15] B. Lee, S.H. Chun, J.H. Hong, and et al. Deepbts: Prediction of recurrence-free survival of non-small cell lung cancer using a time-binned deep neural network. *Scientific Report*, 1952:456–470, 2020.

[16] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *MedIA*, 42:60–88, 2017.

[17] 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.

[18] Gian Carlo Mattiucci, Luca Boldrini, Giuditta Chiloiro, G. D'Agostino, Silvia Chiesa, Fiorenza de Rose, Luigi Azario, Danilo Pasini, M. Gambacorta, Mario Balducci, and Vincenzo Valentini. Automatic delineation for replanning in nasopharynx radiotherapy: What is the agreement among experts to be considered as benchmark? *Acta Oncologica*, 52:1417 – 1422, 2013.

[19] H. Brendan McMahan, Eider Moore, Daniel Ramage, and Blaise Agüera y Arcas. Federated learning of deep networks using model averaging. *CoRR*, abs/1602.05629, 2016.

[20] H.B. McMahan, E. Moore, D. Ramage, S. Hampson, and B.A. Fast trust region for segmentation. In *Proc of ICAIS*, pages 1273–1282, 2017.

[21] 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.

[22] Karthik Nandakumar, Nalini Ratha, Sharath Pankanti, and Shai Halevi. Towards deep neural network training on encrypted data. In *proc of CVPR-W*, pages 0–0, 2019.

[23] Witold Oleszkiewicz, Peter Kairouz, Karol Piczak, Ram Rajagopal, and Tomasz Trzciński. Siamese generative adversarial privatizer for biometric data. In *proc of ACCV*, pages 482–497, 2018.

[24] Pascal Paillier. Public-key cryptosystems based on composite degree residuosity classes. In *proc ICTACT*, pages 223–238, 1999.

[25] Francesco Pittaluga, Sanjeev Koppal, and Ayan Chakrabarti. Learning privacy preserving encodings through adversarial training. In *proc of IEEE WACV*, pages 791–799, 2019.

[26] Nisarg Raval, Ashwin Machanavajjhala, and Landon P Cox. Protecting visual secrets using adversarial nets. In *proc of CVPR-W*, pages 1329–1332, 2017.

[27] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 620–636, 2018.

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[29] Bita Rouhani, Sadegh Riazi, and Farinaz Koushanfar. DeepSecure: Scalable provably-secure deep learning. In *in proc of Design Auto. Conf. (DAC)*, 2018.

[30] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2586–2594, 2019.

[31] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised Dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer, 2017.

[32] Praneeth Vepakomma, Tristan Swedish, Ramesh Raskar, Otkrist Gupta, and Abhimanyu Dubey. No peek: A survey of private distributed deep learning. *CoRR*, abs/1812.03288, 2018.

[33] K. Wang, Y. Zhao, Q. Xiong, M. Fan, G. Sun, L. Ma, and T. Liu. Research on healthy anomaly detection model based on deep learning from multiple time-series physiological signals. *Sci. Program.*, 2016.

[34] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[35] George Wolberg. Image morphing: A survey. *The Visual Computer*, 14, 03 1999. doi: 10.1007/s003710050148.

[36] Bingzhe Wu, Shiwan Zhao, Guangyu Sun, Xiaolu Zhang, Zhong Su, Caihong Zeng, and Zhihong Liu. P3SGD: Patient privacy preserving SGD for regularizing deep CNNs in pathological image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2099–2108, 2019.

[37] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *proc of ECCV*, pages 606–624, 2018.

[38] Taihong Xiao, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker, and Ming-Hsuan Yang. Adversarial learning of privacy-preserving and task-oriented representations. In *AAAI*, 2020.

[39] Pengtao Xie, Misha Bilenko, Tom Finley, Ran Gilad-Bachrach, Kristin E. Lauter, and Michael Naehrig. Crypto-nets: Neural networks over encrypted data. *CoRR*, abs/1412.6181, 2014.

[40] Chugui Xu, Ju Ren, Deyu Zhang, Yaoxue Zhang, Zhan Qin, and Kui Ren. GANobfuscator: Mitigating information leakage under gan via differential privacy. *IEEE TIFS*, 14(9):2358–2371, 2019.

[41] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):12, 2019.

[42] Tsung-Yen Yang, Christopher Brinton, Prateek Mittal, Mung Chiang, and Andrew Lan. Learning informative and private representations via generative adversarial networks. In *proc of ICBD*, pages 1534–1543, 2018.

[43] H. Zhao, O. Gallo, I. Frosio, and J. Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. doi: 10.1109/TCI.2016.2644865.

[44] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, pages 14774–14784, 2019.

[45] M Tarek Ibn Ziad, Amr Alanwar, Moustafa Alzantot, and Mani Srivastava. CryptoImg: Privacy preserving processing over encrypted images. In *proc of IEEE CNS*, pages 570–575, 2016.

[46] A. P. Zijdenbos, B. M. Dawant, R. A. Margolin, and A. C. Palmer. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE TMI*, 13 (4):716–724, 1994.

[47] Kelly Zou, Simon Warfield, Aditya Bharatha, Clare Tempany, Michael Kaus, Steven Haker, William Wells, Ferenc Jolesz, and Ron Kikinis. Statistical validation of image segmentation quality based on a spatial overlap index. *Academic radiology*, 11:178–89, 02 2004. doi: 10.1016/S1076-6332(03)00671-8.