Privacy Preservation in Medical Image Analysis

by

Ngoc Bach, KIM

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE
TECHNOLOGIE SUPÉRIEURE
IN PARTIAL FULFILLMENT FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
Ph.D.

MONTREAL, DEC 21st, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

BOARD OF EXAMINERS

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Christian Desrosiers, Thesis Supervisor
Department of Software and IT Engineering, École de Technologie Supérieure, Canada

Mr. Jose Dolz, Thesis Co-supervisor
Dept. of Software and IT Engineering, École de Technologie Supérieure, Canada

Mr. Pierre-Marc Jodoin, Thesis Co-supervisor
The Computer Science Department, Université de Sherbrooke, Canada

Mr. Matthew Toews, President of the Board of Examiners
Dept. of Systems Engineering, École de Technologie Supérieure, Canada

Mr. Hervé Lombaert, Member of the Jury
Dept. of Software and IT Engineering, École de Technologie Supérieure, Canada

Mr. Guillaume-Alexandre Bilodeau, External Examiner
Dept. of Computer Engineering and Software Engineering, Polytechnique Montreal, Canada

THIS THESIS WAS PRESENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND THE PUBLIC

ON DECEMBER 6^{th}, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

Protection de la vie privée dans l’analyse d’images médicales

Ngoc Bach, KIM

RÉSUMÉ

Ces dernières années, le développement d’algorithmes basés sur l’intelligence artificielle (IA) a fait des pas de géant, notamment grâce à l’essor de l’apprentissage automatisé basé sur les réseaux de neurones profonds. Les travaux de recherche dans ce domaine se sont traduits en un large éventail d’applications, principalement en vision par ordinateur et en analyse du langage naturel. En revanche, les applications médicales de l’IA tardent toujours à s’imposer, en grande partie dû aux contraintes de confidentialité et d’accès aux données médicales.

L’objectif principal de cette thèse est de développer des modèles d’apprentissage profond destinés à l’analyse d’images médicales dans un contexte client-serveur, et ce en protégeant la vie privée des patients. Pour ce faire, trois méthodes sont proposées, chacune associée à un chapitre distinct de la thèse. Les trois méthodes suivent une stratégie similaire à haut niveau, où un client (par exemple, un hôpital) encode une image de manière à masquer l’information confidentielle, avant d’envoyer celle-ci à un serveur. Le serveur traite l’image encodée à l’aide d’un réseau de neurones spécifiquement conçu pour ces données, et renvoie ensuite le résultat au client où il est décodé. Une contrainte importante à respecter dans ce contexte est de ne jamais transmettre de données contenant des informations pouvant permettre d’identifier un patient.

La première méthode comprend trois réseaux entraînés de bout en bout: un réseau encodeur retirant d’une image les caractéristiques spécifiques à l’identité du patient correspondant, un réseau discriminateur qui tente d’identifier le patient à partir de l’image encodée, et un réseau de segmentation pouvant extraire des régions d’intérêt dans cette même image. En entraînant l’encodeur de manière à simultanément tromper le discriminateur et maximiser la performance du réseau de segmentation, celui-ci apprend à supprimer les caractéristiques confidentielles tout en conservant celles qui sont essentielles à la tâche de segmentation. La capacité de cette approche à offrir une segmentation de qualité, tout en masquant l’identité des patients, est démontrée sur la segmentation d’IRM de cerveaux provenant de la base de données *Parkinson Progression Marker Initiative* (PPMI).

Une limitation de cette première approche est de ne pas encoder les cartes de segmentation produites par le réseau pouvant également être employées pour identifier le patient. La deuxième méthode présentée dans la thèse protège l’identité du patient en appliquant une déformation spatiale non linéaire pseudo-aléatoire à l’image d’entrée. Il en résulte une image distorsionnée qui est envoyée au serveur pour traitement. Le réseau de segmentation produit alors une carte déformée de segmentation où l’identité du patient est protégée. Cette carte est renvoyée au client qui remet celle-ci sous une forme canonique. Ce système comporte trois différentes composantes: un générateur de champ de flux qui produit une fonction de déformation pseudo-aléatoire, un discriminateur siamois qui tente de prédire l’identité du patient à partir de l’image traitée et un réseau de segmentation qui analyse le contenu des images encodées. Comme dans l’approche précédente, le système est entraîné de bout en bout avec une approche d’optimisation antagoniste.

En trompant le discriminateur, le générateur de champ de flux apprend à produire une déformation non linéaire réversible pouvant supprimer l'identité du patient à la fois dans l'image et la carte de segmentation correspondante. La méthode proposée est à nouveau validée sur la tâche de segmentation d'IRM de cerveaux provenant de deux ensembles de données différents. Les résultats montrent que cette méthode offre une précision de segmentation similaire à celle obtenue sur des images non encodées et, de plus, réduit considérablement la capacité à récupérer l'identité du sujet.

Les deux premières approches reposent sur l'apprentissage antagoniste à base de discriminateurs siamois pour encoder les images à protéger. Or, l'entraînement de tels discriminateurs nécessite plusieurs images pour un même patient, ce qui est rarement possible en pratique. De plus, l'entraînement antagoniste du modèle est souvent instable, et le modèle entraîné peut être sensible aux changements dans la distribution des images. La dernière méthode explore une stratégie différente pour préserver la confidentialité. Dans cette méthode, le client protège l'image du patient à segmenter en la mélangeant à une image de référence, rendant celle-ci inexploitable par une personne non autorisée. Cette image mélangée est envoyée à un serveur pour traitement. Le serveur renvoie ensuite au client le mélange de cartes de segmentation qui la ramène enfin à la segmentation cible. Le système complet comporte deux composantes: un réseau de segmentation du côté serveur qui traite le mélange d'images, et un réseau de "démixage" qui récupère la bonne carte de segmentation à partir du mélange de segmentations. Une fois de plus, le système est entraîné de bout en bout. Les résultats d'expériences sur différents jeux de données montrent que cette méthode obtient une segmentation de qualité supérieure ou comparable aux approches précédentes, tout en étant plus simple à implémenter et nécessitant moins de calculs que celles-ci.

Mots-clés: vie privée, segmentation d'images médicales, apprentissage antagoniste, apprentissage de représentations

Privacy Preservation in Medical Image Analysis

Ngoc Bach, KIM

ABSTRACT

In recent years, the development of artificial intelligence (AI) algorithms has been the subject of tremendous progress brought namely by the rise of deep neural networks. Research works in AI have been translated into a broad range of applications, particularly, in the field of computer vision and natural language processing. In contrast, medical applications of AI have been slower to appear until now, largely due to privacy constraints on medical data.

The primary objective of this thesis is to develop novel deep learning methods for client-server medical image analysis, which also protect patient privacy. Toward this goal, three methods are proposed, each one associated with a distinct thesis chapter. These methods follow the same high-level strategy where a client (for instance, a hospital) encodes an image so that the sensitive information is obfuscated, before sending it to a server for analysis. The server processes the encoded image with a neural network designed to handle this data, and then sends the result back to the client where it is finally decoded. An important constraint to satisfy in this setting is never sending information that can be used to identify patients.

The first method, based on adversarial learning, is composed of three networks trained end-to-end: an encoder network which removes identity-specific features from the input image, a discriminator network that attempts to identify the corresponding subject from the encoded image, and a segmentation network which tries to extract regions of interest in the same image. By training the encoder to simultaneously fool the discriminator and maximize segmentation performance, it can learn to remove private features while keeping those essential for the segmentation task. The method's ability to provide a high-quality segmentation, while also obfuscating patient identity, is demonstrated on the segmentation of brain MRI from the large-scale Parkinson Progression Marker Initiative (PPMI) dataset.

A limitation of the first approach is that it does not encode the segmentation maps produced by the network, which may also be used to identify the patient. The second method presented in the thesis, which aims to alleviate this problem, protects 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 segmentation network then produces a deformed segmentation map in which the patient's identity is protected. This map is sent back to the client where it is reverted back to a canonical form. The overall system has three components: a flow-field generator which produces a pseudo-random deformation function, a Siamese discriminator that tries to recover the patient identity from the processed image, and a segmentation network that analyzes the content of the proxy images. As in the first approach, the system is trained end-to-end in an adversarial manner. By fooling the discriminator, the flow-field generator learns to produce a reversible non-linear deformation which allows to remove information related to patient identity from both the input image and resulting segmentation map. The proposed method is once again validated on the task of MRI brain segmentation using images from two different datasets. Results show this method to offer a segmentation accuracy similar to a system trained on non-encoded images, while also reducing considerably the ability to recover subject identity.

The first two approaches use an adversarial learning strategy based on Siamese discriminators to encode the images to protect. However, training such discriminators requires to have several images for the same patient, which is not always possible in practice. Moreover, the adversarial training of the model is often unstable and the trained model can be sensitive to changes in the distribution of images. The last method explores a different strategy to preserve privacy. In this method, the client protects the to-be-segmented patient image by mixing it to a reference image, making it unworkable and unrecognizable for an unauthorized person. This proxy image is sent to a server for processing. The server then returns the mixture of segmentation maps to the client, which can revert it to a correct target segmentation. The system has two components: a segmentation network on the server side which processes the image mixture, and a segmentation "unmixing" network which recovers the correct segmentation map from the segmentation mixture. Once more, the whole system is trained end-to-end. The results of experiments on different datasets show that this method achieves a high or comparable segmentation accuracy with respect to previous approaches, while also being simpler to implement and requiring less computations than these approaches.

Keywords: privacy-preserving, medical image segmentation, adversarial learning, representation learning

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

DL	Deep Learning
ML	Machine Learning
ANN	Artificial Neural Network
DNN	Deep Neural Network
AI	Artificial Intelligence
PACS	Picture Archiving and Communication System
DICOM	Digital Imaging and Communications in Medicine
CT	Computed Tomography
MRI	Magnetic Resonance Imaging
GPU	Graphics Processing Unit
PPMI	Parkinson Progression Marker Initiative
CNN	Convolutional Neural Network
FCN	Fully Convolutional Neural Network
GAN	Generative Adversarial Network
VAE	Variational Auto-Encoder
AL	Adversarial Learning
FL	Federated Learning
HFL	Horizontal Federated Learning
VFL	Vertical Federated Learning

PHE	Partial Homomorphic Encryption
SWHE	Somewhat Homomorphic Encryption
FHE	Fully Homomorphic Encryption
WM	White Matter
GM	Gray Matter
CSF	Cerebrospinal Fluid
DSC	Dice Score
MS-SSIM	Multiscale Structure Similarity
ICA	Independent Component Analysis
BSS	Blind Source Separation
NCSN	Noise Conditional Score Network
TTA	Test-time Augmentation
mAP	Mean Average Precision
ICC	Intra-class Correlation Coefficient
NCR	Necrose
ED	Edema
ET	Enhance Tumor

INTRODUCTION

0.1 Motivation

Deep learning (DL) is a branch of machine learning (ML) based on artificial neural networks (ANNs), which is particularly suited for resolving complex and high-dimensional problems. Although DL is not a novel concept, improvements in computing power, increased data availability, and higher algorithm performances have contributed to the recent rise of artificial intelligence (AI) applications. Developments in DL have enabled significant progress in a variety of fundamental domains, including computer vision, speech recognition, natural language processing and gaming. Algorithms based on deep learning have also been shown to be highly effective in strategic areas, such as healthcare, autonomous driving or surveillance, among others. In particular, in medical imaging analysis, these models have yield to improved accuracy and speed compared to more traditional methods. Their capability of detecting subtle patterns and anomalies in complex medical images make them highly useful for numerous tasks, such as diagnosis, treatment and follow-up of various diseases (Litjens *et al.*, 2017), with the potential to improve personalized medicine. While research in deep learning for medical imaging has exploded in recent years, so far, few advances have translated to clinical practice due to three main challenges:

1. Infrastructure complexity: First, the infrastructure required for running deep learning applications in clinical environments is complex. Nowadays, most healthcare organizations use PACS/DICOM servers as a standard approach to store and access data. A Picture Archiving and Communication System (PACS) server is a centralized computing system that acts as a repository for medical images created through different modalities, such as X-ray, Magnetic Resonance Imaging (MRI), and Computed Tomography (CT), among others. By using a Digital Imaging and Communications in Medicine (DICOM) viewer software, authorized PACS clients can easily access and view these images. On the other hand, deep

learning models often require specialized hardware with powerful graphical processing units (GPUs) to process large amounts of data using neural networks having millions of parameters. However, the integration of this specialized hardware into the existing PACS/DICOM systems is troublesome. Therefore, cloud-based image analysis services is an appealing solution for these issues as they allow easy trouble-shooting, immediate software updates and, most importantly, do not require a specialized hardware deployment and management on the clients' site.

2. Low availability of labeled data: Secondly, deep learning models are notoriously data hungry. In order to achieve satisfactory performances, these models often need a massive amount of labeled training data. Crowd-sourcing has emerged as a viable and effective technique for the collection and labeling of data where user expertise is not required (Su, Deng & Fei-Fei, 2012; Deng *et al.*, 2009). However, labeling medical images requires the expertise of one or more highly-trained radiologists, limiting the ability to "crowd-source" the needed information. Hence, most public labelled datasets of medical images are often small compared to other datasets in computer vision, resulting in insufficient training data. It is thus essential to deploy learning mechanisms that can pool and exploit data from various sources.
3. Data privacy: The confidential aspect of patient data is another issue impeding the creation and use of large datasets in medical image analysis. In many countries, it is illegal to access, use or share protected health information without patients' consent. For example, in 2021, the government of Quebec enacted a new law called "Law 25" – previously known as "Bill 64" (Minister Responsible for Access to Information and the Protection of Personal Information, 2021) – requiring all organizations operating in Quebec to comply with regulations on access, use, and disclosure of personal data. The acquisition of personal information for research purposes cannot be limited to legal queries alone. Access to such information is restricted and subject to a number of strict constraints. As an example, these

> constraints may include assuring the use of personal information in a manner that follows stringent confidentiality measures and communicating only what is necessary. In the context of the release of medical images, it is standard to anonymize the images by removing metadata and pixel data that could potentially reveal the identity of patients (Freymann, Kirby, Perry & Clunie, 2012). Nevertheless, this anonymization and curation procedure requires considerable time and resources, and is susceptible to human fatigue and errors (Rutherford *et al.*, 2021). In addition, previous studies (Packhäuser *et al.*, 2022; Wachinger, Golland, Kremen, Fischl & Reuter, 2015; Kumar, Desrosiers, Siddiqi, Colliot & Toews, 2017; Shamir, 2013; Esmeral & Uhl, 2022) suggest that patient information can be extracted from medical image examinations. Therefore, in order to develop a cloud-based medical image analysis application, it is critical to address the privacy restriction problem in medical data.

Although various methods have been proposed to deal with privacy issues in machine learning, none of them are fully suitable for a cloud-based medical image analysis system. Federated Learning addresses privacy restrictions by enabling the decentralized training of machine learning clients without the need to share raw data across clients (McMahan, Moore, Ramage & y Arcas, 2016). However, Hatamizadeh *et al.* (2023) demonstrated that this learning paradigm may be insecure due to its vulnerability to model inversion attacks. Another issue with Federated Learning is that the resulting trained models must also be deployed in a decentralized manner as the input data and output results are not encoded, making it unsuitable for a client-server application model. Another approach for privacy-preserving AI leverages Homomorphic Encryption (Dowlin *et al.*, 2016; Hesamifard, Takabi & Ghasemi, 2017; Nandakumar, Ratha, Pankanti & Halevi, 2019). This approach enables to perform computations on encrypted data without having to decrypt this data, thereby providing full privacy protection. However, due to its very high computational complexity, it is unfeasible for large deep learning models.

0.2 Objectives

Based on the aforementioned motivations, the main goal of this research is to *develop novel deep learning methods for client-server privacy-preserving medical image analysis*, which can perform a given image analysis task both accurately and efficiently, while also protecting the private information of users. In this thesis, we consider image segmentation as the main image analysis task to solve. This task, which consists in assigning a class label to each pixel/voxel of an input 2D/3D image, is essential in various medical applications such as surgical planning, radiotherapy treatment planning or implant design, where it enables a more precise analysis by isolating specific structures.

To achieve this main goal, three specific objectives are proposed:

1. As first objective, we explore adversarial learning as a way to encode medical images so that patient identity is obfuscated, yet encoded images can be still used to obtain accurate segmentation results. While adversarial learning has already been used for privacy-preserving image classification, this approach has never been investigated for segmentation, or in scenarios where the private information cannot be encoded as a fixed set of classes (e.g., patient IDs in our case).
2. Although it encodes images so that patent identity cannot be recovered, the above-mentioned adversarial approach does not encode the segmentation output computed on the server. This poses a potential problem since the segmentation contours in the output also contain information that can be used to identify patients. As second objective, we seek to enhance the previous adversarial approach to encode both the input image and segmentation output. Toward this objective, we propose a novel method based on a reversible geometric transformation, learned from training data, which preserves privacy while also providing an accurate segmentation.

3. The approaches proposed for the first and second objectives use an adversarial learning strategy to encode images so that patient identity is discarded or hidden, yet the information necessary for the downstream segmentation task is preserved. While useful, such approaches suffer from three problems. First, they require having several images for the same patient, which may not be feasible in some applications. Second, the adversarial training of the model is often unstable, as it involves solving an optimization problem with opposite objectives, and can be sensitive to changes in the image distribution. As a result, it may give poor results for images with different characteristics than the ones seen in training. As third and last objective of the thesis, we aim to design a simpler, yet robust approach for the privacy-preserving segmentation of medical images, which does not rely on adversarial learning. For this objective, we explore a technique based on "mix-up" where a data sample's private information is obfuscated by mixing the sample with other ones known only to the client.

0.3 Contributions

As contributions to the field, we developed three novel methods that enable client-server medical image segmentation while protecting patients' identity. These contributions are detailed below:

1. Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images: We introduce the first client-server system based on adversarial learning for medical image segmentation. Unlike previous adversarial approaches for privacy-preserving image analysis, where a standard classification network (the discriminator) tries to recover the private information (e.g., gender), the proposed method uses a Siamese discriminator for this task. As a result, it can be employed to obfuscate information which cannot be reduce to a fixed set of known classes, for instance, the ID patients in the system. As additional contribution, we analyze our method from the perspective of information theory and show

that its learning objective is related to minimizing the mutual information between the encoded image and patient ID.

2. Privacy Preserving for Medical Image Analysis via Non-Linear Deformation Proxy: As second contribution, we devise a novel approach for the privacy-preserving segmentation of medical images which can encode both the input image and segmentation output. Although it also exploits adversarial learning for encoding images to segment, this approach proposes a very different strategy: applying a pseudo-random non-linear transformation that distorts both the input image and its corresponding segmentation output, and which can be reserved to recover the true segmentation. In contrast to standard adversarial encoders that map an image to a representation, the encoder in our model generates a pseudo-random reversible flow-field, conditioned on a private key only known to the client, which is used to distort the image. This method provides a segmentation accuracy comparable to the standard adversarial approach, while offering the additional benefit of obfuscating patient identity in the segmentation output.
3. Mixup-Privacy: A simple yet effective approach for privacy-preserving segmentation: For the third contribution, we propose a straightforward, yet efficient method inspired by mix-up (Chang *et al.*, 2020) that encodes 3D patches of a target image by blending them with reference patches with known ground-truth segmentation. Unlike federated learning approaches, which require extensive training or homomorphic encryption that are computationally infeasible, our method operates within a normal training setup and incurs low computational cost. Moreover, compared to approaches based on adversarial learning, the proposed method does not require to have multiple training images for each patient (e.g., from difficult-to-obtain longitudinal datasets), provides a more stable training, and is less sensitive to distribution shifts.

0.4 List of Publications

The work presented in this thesis has resulted in the publication of three peer-reviewed papers, each one in top, yet very selective venues:

1. B. N. Kim, J. Dolz, P. -M. Jodoin and C. Desrosiers, “Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images”, in IEEE Transactions on Medical Imaging, vol. 40, no. 7, pp. 1737-1749, July 2021, doi: 10.1109/TMI.2021.3065727.
2. B. N. Kim, J. Dolz, C. Desrosiers and P. -M. Jodoin, “Privacy Preserving for Medical Image Analysis via Non-Linear Deformation Proxy”, The British Machine Vision Conference (BMVC) 2021.
3. B. N. Kim, J. Dolz, P. -M. Jodoin and C. Desrosiers, “Mixup-Privacy: A simple yet effective approach for privacy-preserving segmentation”, Information Processing in Medical Imaging (IPMI) 2023. *This paper was selected for oral* presentation.

All of these three venues are considered among the most prestigious journals and conferences in medical image computing and computer vision. The Institute of Electrical and Electronics Engineers (IEEE) publishes a monthly peer-reviewed scientific journal titled IEEE Transactions on Medical Imaging. This journal is devoted to advancing the technical aspects of medical imaging. According to Google scholar, the journal has an impact factor of 11.037, making it one of the two most influential medical imaging journals. The British Machine Vision Conference (BMVC) is an annual conference on machine vision, image processing, and pattern recognition hosted by the British Machine Vision Association (BMVA). It is one of the largest international conferences on computer vision and related topics. BMVC is a highly respected conference. The average acceptance rate of BMVC is 33.6%, and it is top-ranked in Google Scholar Metrics (11th) among all the journals and conferences in the discipline of *Computer Vision and Patter Recognition*. Information Processing in Medical Imaging (IPMI) is a conference held every two years that focuses on applied mathematics, computer science, image processing, and medical imaging analysis. IPMI is considered one of the flagship conferences in medical image processing, with a general acceptance rate of nearly 30%, and an acceptance rate for oral presentations of around 12%.

0.5 Thesis outline

The rest of the thesis is divided into five chapters, as follows:

1. Chapter 1: This chapter introduces the basic concepts to understand the various techniques presented in the thesis. The background chapter includes the basics of CNNs, supervised learning with CNNs, representation learning and adversarial learning.
2. Chapter 2: This chapter gives an up-to-date summary of research in the field of privacy-preserving learning for computer vision. Specifically, we present recent approaches based on federated learning, homomorphic encryption and adversarial learning for preserving privacy in visual tasks.
3. Chapter 3: This chapter presents our first client-server system based on adversarial learning for the privacy-preserving segmentation of medical images.
4. Chapter 4: This chapter details our second solution for the privacy-preserving analysis of multi-centric medical images, based on the generation of pseudo-random non-linear transformations.
5. Chapter 5: This chapter presents the proposed approach for privacy-preserving segmentation, which employs a mix-up strategy to obfuscate patient identity.
6. Chapter 6: In the final chapter of the thesis, we conclude by summarizing our contributions, discussing the main findings and limitations, and suggesting potential improvements in future work.

CHAPTER 1

BACKGROUND

In this chapter, we cover several basic concepts used in this work, including convolutional neural networks and their application in image classification and image segmentation, contrastive learning, autoencoders, and generative adversarial networks.

1.1 Convolutional Neural Network

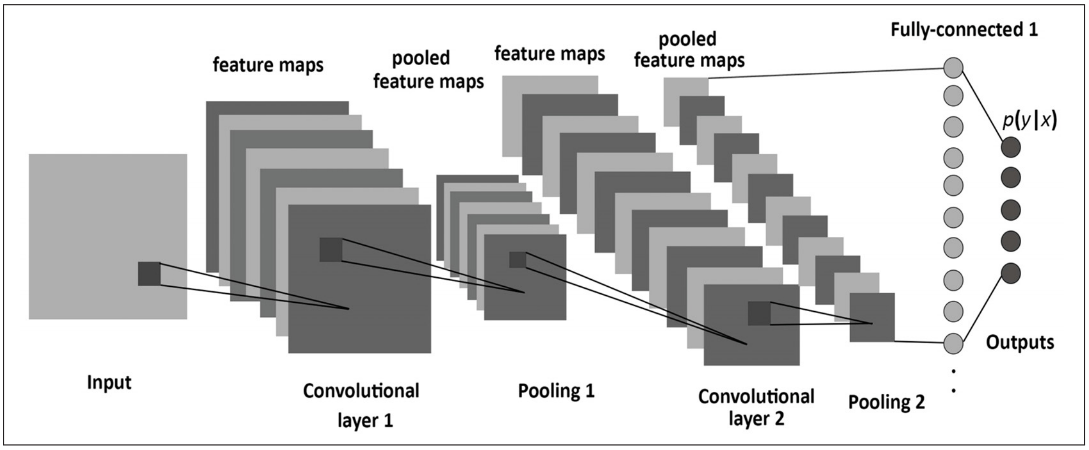


Figure 1.1 The basic structure of a CNN, consisting of convolutional, pooling, and fully-connected layers. Taken from (Albelwi & Mahmood, 2017)

A Convolutional Neural Network (CNN) (Lecun, Bottou, Bengio & Haffner, 1998) is a type of neural network that utilizes the spatial arrangement of the inputs. It has a typical design consisting of repetitive convolution and pooling layers, with a few fully-connected layers at the end. The final layer is often a softmax classifier, as depicted in Fig 1.1. CNNs are usually trained using back-propagation with Stochastic Gradient Descent (SGD). During training, the network's parameters (i.e., weights of convolution filters and fully-connected layers) are adjusted to minimize a loss function measuring the discrepancy between the network's prediction and desired output. The fundamental components of a CNN are the following.

1.1.1 Convolution Layer:

This type layer, which is the corner stone of a CNN, is comprised of a set of learnable kernels or filters that are trained to extract local features from the input. These kernels are employed to construct a feature map. Only a small portion of the input, known as the receptive field, is connected to the units on the feature map. To generate a new feature map, the filter is slid across the input, and the convolution operation is computed as follows:

$$y_{i,j,k}^{(\ell+1)} = \sum_{h=0}^{H_\ell-1} \sum_{w=0}^{W_\ell-1} \sum_{d=0}^{D_\ell-1} f_{h,w,d,k}^{(\ell)} \cdot x_{i+h,j+w,d}^{(\ell)} + b_k^{(\ell)}, \quad k = 0, \dots, D_{\ell+1}-1 \tag{1.1}$$

Here, $x^{(\ell)}$ is the input to the convolution operation at layer ℓ, comprised of D_ℓ channels, $f^{(\ell)}$ is the set of $D_{\ell+1}$ filters of size $H_\ell \times W_\ell$ and depth D_ℓ, $b^{(\ell)}$ the biases, and $y^{(\ell+1)}$ the output having $D_{\ell+1}$ channels. Moreover, d denotes the channel of the input/output, l the l^{th} layer, (i, j) the spatial position and $b^{(\ell)}$ is the biases. In this equation, (i, j) is a spatial position corresponding to pixel (or voxel in 3D), while d and k are channel indexes.

Convolution layers exploit the strategy of parameter sharing, where all units in a feature map use the same weights (filters), which reduces the number of parameters and allows the detection of the same feature regardless of its position in the input.

1.1.2 Non-linear Activation:

The output of the convolution is followed by the application of a non-linear activation function to introduce non-linearity in the model. Initially, the logistic activation, also known as the sigmoid was used to simulate biological neurons. The logistic function can be defined as follows:

$$\text{sigmoid}(x) = \frac{1}{1 + e^{-x}} \tag{1.2}$$

Because the derivative of the sigmoid becomes very small for input values far from the origin, networks based on this activation tend to suffer from the vanishing gradient problem which impedes their training. Nowadays, non-saturating activations like the Rectified Linear Unit (ReLU) or its variants (leaky ReLU, eLU, parametric ReLU, etc.) are preferred as they do not have the same problem. The ReLU also has the advantage of being simple to compute:

$$\mathrm{ReLU}(x) = \max(0, x) \tag{1.3}$$

1.1.3 Pooling Layer:

This layer, also known as down-sampling layer, decreases the resolution of the previous feature maps which reduces the number of parameters to learn. It also helps to make the network invariant to small changes or distortions of the input image. Pooling involves dividing the input into non-overlapping regions with a size of $(R \times R)$ and producing a single output from each region. Due to the destructive nature of the pooling layer, the size of the pooling region is typically chosen as 2×2. There are two main types of pooling, namely, max-pooling and average-pooling.

1.1.4 Fully-connected Layer:

The last part of a CNN is normally comprised of one or more fully-connected layers, similar to those of a feed-forward neural network, where units are connected to every neurons in the previous layer. In CNNs for classification, fully-connected layers are typically implemented by a simple linear operation,

$$y = Wx + b \tag{1.4}$$

where x is the output of previous layer, y is the output vector, W is the weights and b is the biases.

In CNNs designed for dense prediction tasks such as image segmentation, fully-connected layers are often implemented with a series of convolutions with a kernel size of 1×1. By employing this strategy, we obtain what is called a fully-convolutional neural network (FCN).

1.1.5 Softmax function:

In many prediction tasks, such as classification, the output of the network corresponds to a probability distribution over a set of K classes. In those cases, a softmax function is usually employed in the final layer to obtain probability values, as follows:

$$[\text{softmax}(x)]_i = \frac{e^{x_i}}{\sum_{j=1}^{K} e^{x_j}} \tag{1.5}$$

where $x = (x_1, ..., x_K) \in \mathbb{R}^K$.

1.2 Supervised Learning with CNN

1.2.1 Image classification

Image classification is a fundamental problem in most modern computer vision tasks (Chen *et al.*, 2021). This problem is essential for numerous applications, including object recognition, scene understanding, and content-based image retrieval. Prior to deep learning, the accuracy of image classification approaches was limited by their need to extract features manually. In contrast, deep learning methods leverage the representational capability of deep neural networks to learn complex image representations that can be used for classification. Over the years, many deep CNN architectures have been proposed for image classification, including LeNet (Lecun *et al.*, 1998), Inception-Net (Szegedy *et al.*, 2015), ResNet (He, Zhang, Ren & Sun, 2016) and VGGnet (Simonyan & Zisserman, 2015).

The selection of a suitable loss function is another crucial aspect of deep learning-based image classification. This function measures the difference between the predicted label and the actual label and is used to update the network's training parameters. Cross-entropy is a common choice for image classification, as it measures the difference between the predicted class probabilities and the true class label. In the case of binary classification, this loss can be defined as

$$\mathcal{L}_{CE}(y,\hat{y}) \;=\; -y\log(\hat{y}) \;-\; (1-y)\log(1-\hat{y}) \tag{1.6}$$

where $y \in \{0,1\}$ is the target (ground-truth) label and $\hat{y} \in [0,1]$ the predicted probability of the foreground class. This loss can be extended to the K-class prediction setting as follows:

$$\mathcal{L}_{CE}(y,\hat{y}) \;=\; -\sum_{k=1}^{K} y_k \log(\hat{y}_k) \tag{1.7}$$

In this case, $y_k \in \{0,1\}$ and $\hat{y}_k \in [0,1]$ are the ground-truth label and predicted probability for class k.

1.2.2 Image segmentation

Image segmentation is another important task of computer vision that involves dividing a given image into multiple regions with distinct meaning. This task is key to a wide range of applications related to autonomous vehicles, robotics, agriculture, gaming, and biomedical imaging. In the past, image segmentation was performed using approaches that are not based on machine learning, for instance, active-contours (Hemalatha *et al.*, 2018; Qian, Wang, Guo & Li, 2013; Kass, Witkin & Terzopoulos, 1988), level-sets (Lin, Zheng, Yang & Gu, 2004; Jiang, Zhang & Nie, 2012) and graph-cuts (Boykov & Jolly, 2000; Yi & Moon, 2012). However, these approaches often require the manual design of features by experts and tend to suffer from a slow iterative computation.

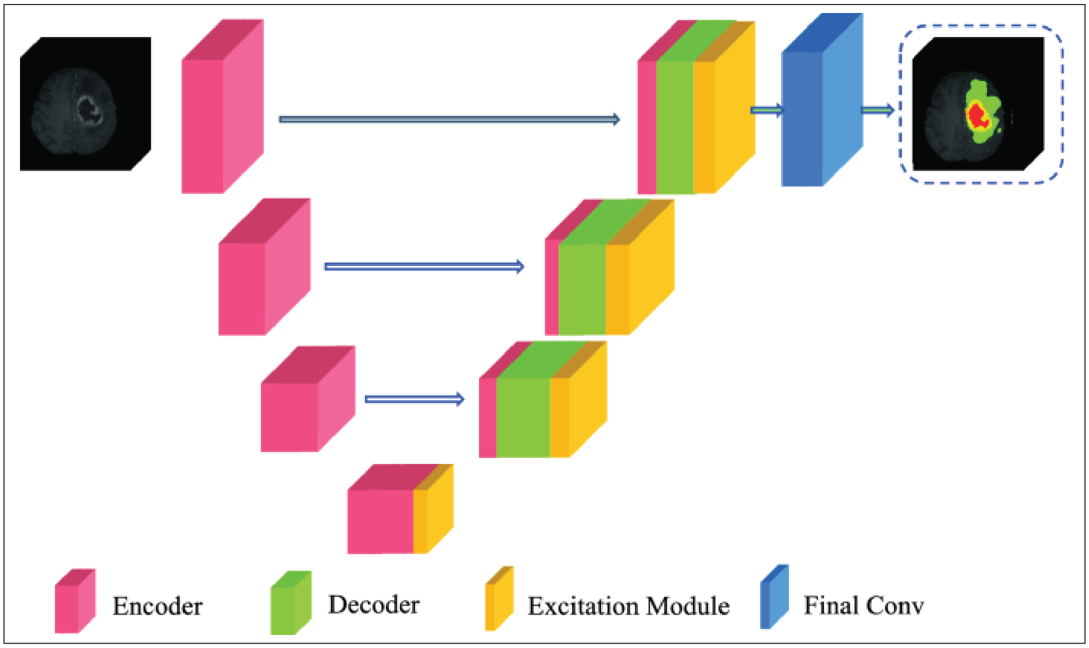


Figure 1.2 U-Net architecture for image segmentation.
Taken from (Islam *et al.*, 2020)

In recent years, deep learning has emerged as the *de facto* solution for image segmentation. As in classification, deep learning methods for segmentation utilize deep neural networks to learn image representations directly from data. A key approach for this problem is based on fully convolutional neural networks (FCNs). Specifically designed for segmentation, FCNs consist of a series of convolutional, down-sampling and up-sampling layers. Unlike CNNs, which were developed for image classification and only use pooling layers to gradually reduce the spatial resolution of the input image, FCNs also employ up-sampling layers to increase the spatial resolution of the feature maps. This allows the network to produce segmentation masks with high spatial resolution. Moreover, by replacing fully-connected layers with 1×1 convolutions, as described before, FCNs can process input images of arbitrary size.

Various FCNs architectures have been proposed for image segmentation such as SegNet (Badrinarayanan, Kendall & Cipolla, 2017) and PSP-Net (Zhao, Shi, Qi, Wang & Jia, 2017). The majority of these networks use an encoder-decoder architecture. As the first component of the network, the encoder gradually reduces the spatial resolution of the feature maps and obtains a high-level contextual representation of the input image. In contrast, the decoder reconstructs the spatial information of the input and then predicts the probability of segmentation classes at each pixel as its output.

1.2.3 U-Net Architecture for Image Segmentation

The U-Net model (Ronneberger, Fischer & Brox, 2015), depicted in Fig 1.2, is one of the most popular networks for medical image segmentation. This network employs a symmetric architecture for both the encoder and the decoder. Each block consists of multiple layers of ReLU-activated convolutions and a down-sampling or up-sampling module. By processing input image at multiple scales, this model promotes segmentation results that combine both contextual and spatial information.

An important innovation of U-Net is the skip connection which connects blocks of the encoder and decoder with the same spatial resolution. This type of connection offers two advantages. First, segmenting regions frequently necessitates boundary information. A simple FCN is incapable of providing such high-resolution information because it is lost during the spatial compressing-reconstruction process. Skip connections resolve this issue by passing the information directly from the encoder to the decoder at higher spatial resolutions. Secondly, the gradient can better flow from the decoder to the encoder using skip connections, which helps update the parameters in the network's shallow layers and accelerates convergence.

1.2.4 Loss Functions for Segmentation:

Image segmentation can be regarded as a pixel-wise classification problem. Therefore, classification losses like cross-entropy loss can also be used to train the segmentation network:

$$\mathcal{L}_{CE}(y, \hat{y}) = -\sum_{k=1}^{K} \sum_{p} y_{p,k} \log(\hat{y}_{p,k}) \tag{1.8}$$

In this equation, $y_{p,k} \in \{0, 1\}$ and $\hat{y}_{p,k} \in [0, 1]$ respectively denote the ground-truth label and predicted probability for class k at pixel p.

A problem with cross-entropy for segmentation is that it gives a greater importance to classes corresponding to large regions in the image. The Dice loss is another popular loss for the same task, which avoids this issue by measuring the overlap between the predicted segmentation mask and the ground-truth mask. In the case of binary segmentation (foreground versus background), it can be defined as

$$\mathcal{L}_{Dice}(y, \hat{y}) = 1 - \frac{2 \sum_p y_p \hat{y}_p}{\sum_p y_p + \sum_p \hat{y}_p} \tag{1.9}$$

where $y_p \in \{0, 1\}$ is the segmentation ground-truth of pixel p and $\hat{y}_p \in [0, 1]$ the predicted probability for the foreground class at the same pixel. In some cases, the Dice loss can result in a slower convergence compare to the cross-entropy loss. Therefore, a common practice is to combine both Dice and cross-entropy to train the network.

1.3 Contrastive Learning

The objective of contrastive learning (Oord, Li & Vinyals, 2018) is to find a good representation of data points in a high-dimensional feature space, with the least amount of supervision. A well-known method based on this principle is to compare pairs of similar and dissimilar examples and learn representations so that similar examples are close to one another in the feature space,

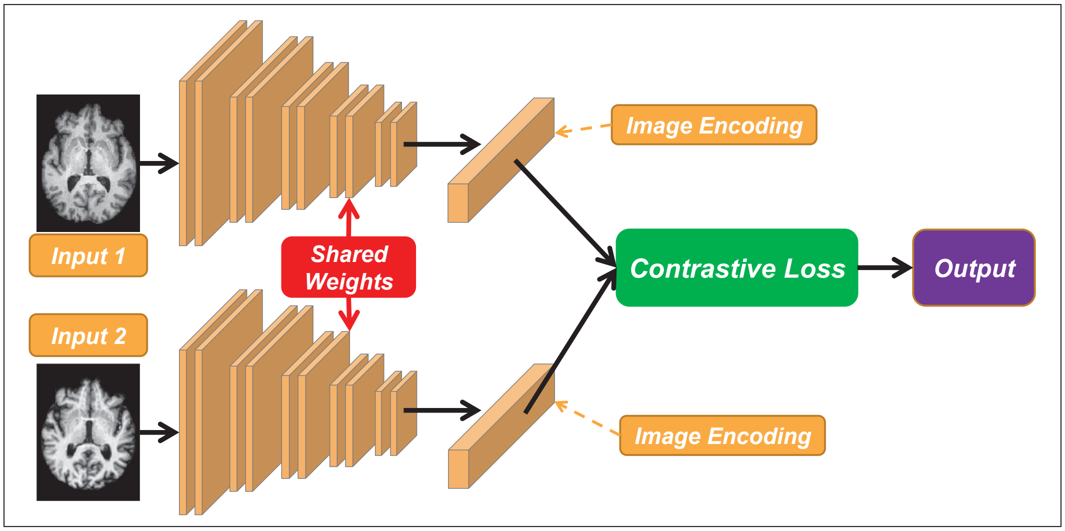


Figure 1.3 A basic siamese network architecture implementation accepts two input images, has identical CNN subnetworks for each input to compute the image embedding, and training with contrastive loss

while dissimilar examples are far apart. This method has been used with success in various computer vision tasks, including image classification, object detection and segmentation.

Contrastive learning is typically implemented using a deep neural network architecture known as a Siamese network (Bromley *et al.*, 1993) as shown in Fig 1.3. A Siamese network consists of two identical sub-networks that share the same weights and architecture. During training, pairs of examples are fed into the network, and the weights are optimized using a contrastive loss function measuring the difference between the representations of similar and dissimilar examples. An example of such loss, stemming from metric learning, is the following:

$$\mathcal{L}_{contrastive} = \sum_{i,j} y_{i,j} \cdot \|x_i - x_j\|^2 + (1 - y_{i,j}) \cdot \max\left(0,\ m - \|x_i - x_j\|^2\right) \tag{1.10}$$

In this loss, x_i, x_j are the embedding vectors of sample pairs, and $y_{i,j} \in \{0, 1\}$ is a label equal to 1 if the two samples are from the same class. Moreover, m is a hyperparameter defining the lower bound distance between dissimilar samples. Another popular loss for contrastive learning, used in recent methods such as SimCLR (Chen, Kornblith, Norouzi & Hinton, 2020a), is as follows:

$$\mathcal{L}'_{contrastive} = -\sum_{i,j} y_{i,j} \log \frac{e^{\mathrm{sim}(x_i,x_j)/\tau}}{\sum_{k \neq j} e^{\mathrm{sim}(x_i,x_k)\tau}} \tag{1.11}$$

Here, sim$(\cdot, \cdot)$ is a vector similarity function and τ is a hyper-parameter known as temperature-scaling, which sharpens or smooths the softmax probability distributions.

The losses in (1.10) and (1.11) require to have class labels for the training samples. Recent work in self-supervised learning (Chen *et al.*, 2020a; Grill *et al.*, 2020; Caron *et al.*, 2021) has focused on developing contrastive learning approaches that do not have this constraint. A typical strategy for unsupervised contrastive learning consists in generating two augmented version of training images and, for a pair of embedding vectors x_i, x_j, having $y_{ij} = 1$ if the embedding are from the same image.

One key advantage of contrastive learning is its ability to learn compact and discriminative representations of data points. The method has been shown to outperform traditional supervised learning methods, such as supervised classification, in various computer vision tasks. This is due to the fact that the method can learn representations that capture the underlying relationships between data points, rather than simply memorizing the training data. Contrastive learning also improve the performance when the number of classes is large but the number of training samples for each class is low.

1.4 Autoencoders

An autoencoder (Rumelhart & McClelland, 1987; Baldi, 2012; Hinton & Salakhutdinov, 2006; Hinton, Osindero & Teh, 2006), as shown in Fig 1.4, is a neural network architecture that is trained to reconstruct the input data from a compressed representation. This architecture, which is particularly well-suited for unsupervised learning tasks, consists of two main components: an encoder network, which maps the input data to a lower-dimensional representation, often

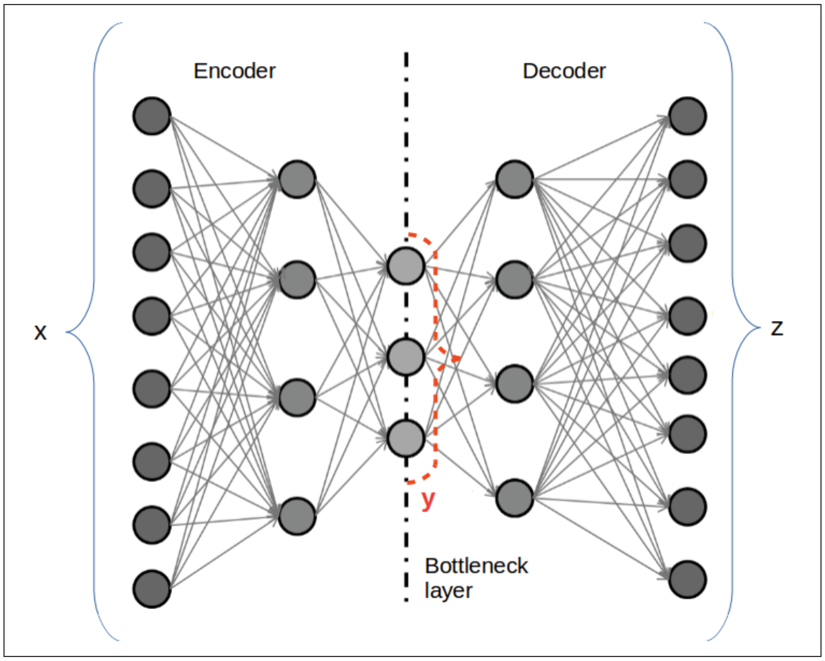


Figure 1.4 The basic structure of an autoencoder, including encoder, decoder, and bottle-neck layer. Taken from (Vachhani *et al.*, 2017)

called latent variable, and a decoder network, which maps the lower-dimensional representation back to the original input data. During the training process, the encoder and decoder are trained simultaneously, with the encoder network trying to produce a more compact representation of the input data, and the decoder network trying to reconstruct the input data from the compressed representation.

As with unsupervised contrastive learning methods, autoencoders can be used to learn useful representations of the input data without the need for labeled data. This makes autoencoders useful for tasks such as dimensionality reduction (Wang, Huang, Wang & Wang, 2014), feature extraction (Liu, Li, Yu & Qin, 2017) and anomaly detection (Chen, Yeo, Lee & Lau, 2018b), among others.

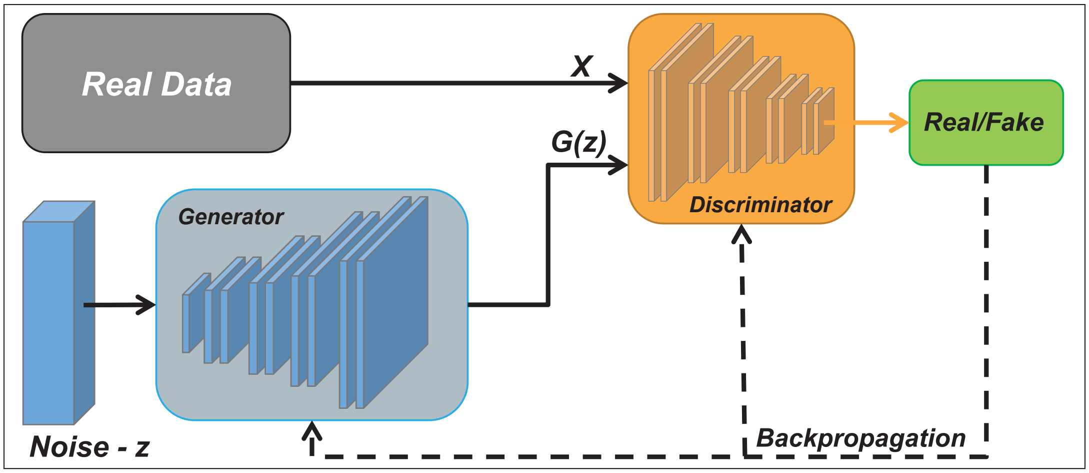


Figure 1.5 Generative Adversarial Network training consists of two networks, a generator and a discriminator

1.5 Generative Adversarial Network

As depicted in Fig 1.5, a Generative Adversarial Network (GAN) (Goodfellow *et al.*, 2014) is a deep learning model with two main components, a generator network and a discriminator network. The generator network is trained to generate new samples that are similar to a given set of training data, while the discriminator network is trained to distinguish between the generated samples and the real samples. This is achieved by training the generator network to “fool” the discriminator (i.e., maximize its error), while the discriminator is trained to correctly identify the generated samples. The process of training the generator and discriminator networks is done simultaneously, with the generator network trying to produce samples that are increasingly similar to the real samples, and the discriminator network trying to correctly identify the generated samples.

Formally, GANs are a structured probabilistic model with latent variables z and observed variables x. The discriminator is a function D that takes x as input, and the generator is defined as a function G whose input is z. Both functions are differentiable with respect to their inputs and parameters. The cost function for the discriminator is typically defined as follows:

$$J = -\mathbb{E}_{x \sim P_{data}(x)}[\log D(x)] - \mathbb{E}_{z \sim P_z(z)}[\log(1 - D(G(z)))] \tag{1.12}$$

By treating the two-player game as a zero-sum game (or mini-max game), the solution involves minimization/maximization in an outer/inner loop, yielding the objective function for discriminator D and generator G:

$$L(D, G) = \min_{G} \max_{D} \mathbb{E}_{x \sim P_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim P_z(z)}[\log(1 - D(G(z)))] \tag{1.13}$$

The main advantage of GANs is their ability to generate new samples that are similar, but not identical to the training data. This makes them useful for tasks such as image synthesis (Karras, Aila, Laine & Lehtinen, 2018), image-to-image translation (Isola, Zhu, Zhou & Efros, 2017), and speech synthesis (Kong, Kim & Bae, 2020).

However, these models also have important limitations. One of the main challenges in training GANs is that the generator and discriminator networks can get stuck in a sub-optimal equilibrium, where the generator produces samples that are not very realistic, and the discriminator is unable to tell them apart from the real samples. Another problem, known as mode collapse (Bhagyashree, Kushwaha & Nandi, 2020; Bau *et al.*, 2019), arrives when the generator learns to output the same, very plausible image for any input. To overcome these problems, several techniques have been proposed such as using a different loss function (Arjovsky, Chintala & Bottou, 2017; Gulrajani, Ahmed, Arjovsky, Dumoulin & Courville, 2017), using different architectures for the generator and discriminator networks (Zhang, Li & Yu, 2018b; Li, Fan, Wang, Ma & Cui, 2021), and using regularization techniques (Tran, Bui & Cheung, 2018; Bang & Shim, 2021).

1.6 Summary

In this chapter, we presented key concepts in deep learning, such as Convolutional Neural Networks (CNNs), image classification, image segmentation, contrastive learning, representation learning, and adversarial learning, which serve as the basis for the research conducted in this thesis. In our work, CNNs are employed predominantly for image segmentation, image transformation and predicting the identity of patients from images. Our research also leverages a strategy based on contrastive learning and Siamese networks to compare encoded features of different patients. Last, adversarial learning is at the core of two of our methods proposed for privacy-preserving segmentation.

CHAPTER 2

LITERATURE REVIEW

In this chapter, we review the three main categories of privacy-preserving methods for machine learning, which are based on federated learning, homomorphic encryption and adversarial learning.

2.1 Federated Learning

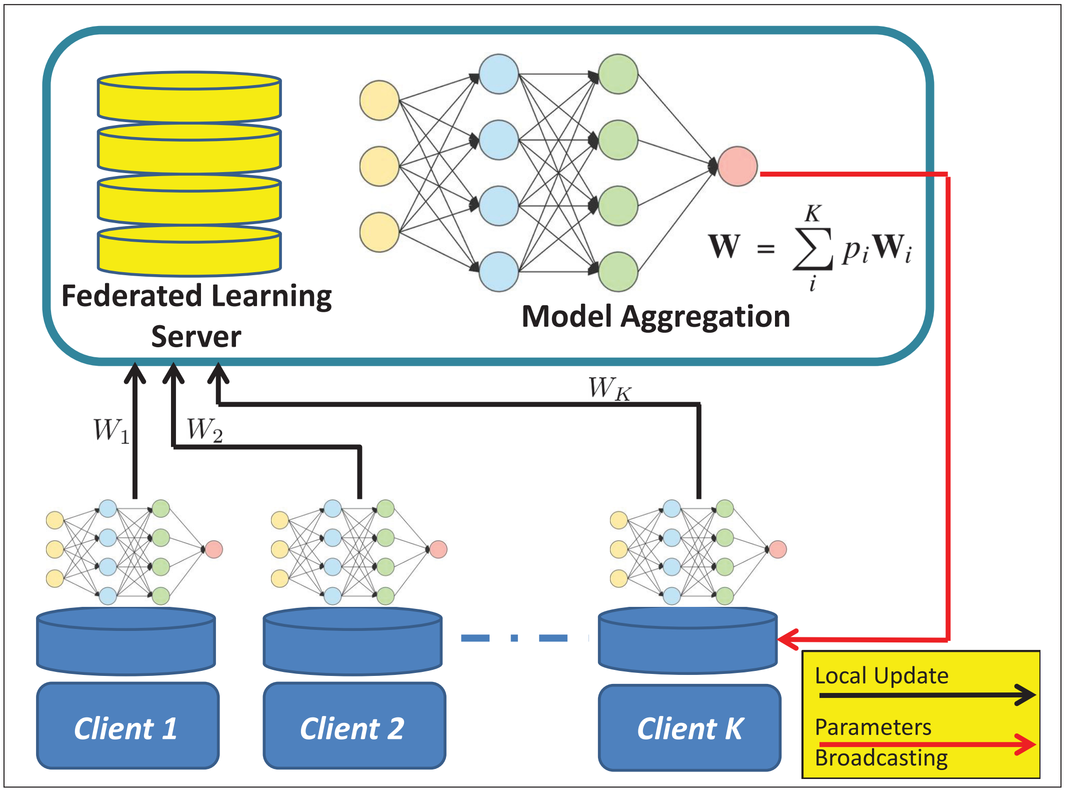


Figure 2.1 Federated Learning approach for iterative learning process where multiple clients work together to learn a model that is aggregated on a federated learning server without the requirement for client data to be sent over the network

Despite being a relatively new technology, Federated Learning (FL) has demonstrated significant promise and outstanding progress since it was first introduced by Google in 2015 (Mohri, Sivek & Suresh, 2019; Gao *et al.*, 2022; Kang *et al.*, 2020; Wang, Yurochkin, Sun, Papailiopoulos & Khazaeni, 2020; Konecný, McMahan, Ramage & Richtárik, 2016; McMahan *et al.*, 2016; McMahan, Moore, Ramage, Hampson & Arcas, 2017).

Federated Learning is a machine learning setting in which different clients (or users) collaborate to learn a model stored on a central server without the need for client data to be transmitted over the network (Kairouz *et al.*, 2021). This setting is useful in scenarios where data is distributed across multiple devices, organizations, or even countries, and where privacy concerns make it infeasible to collect and centralize the data. As depicted in Fig 2.1, the basic idea behind federated learning is to have each participating party train a local model on their own data, and then send updates (gradients or updated weights) to a central server, which aggregates the updates and improves a global model. The global model can then be used to make predictions or perform other computations on new data. Formally, the server aggregates the weights sent from the K clients (Wei *et al.*, 2020) as follows:

$$\mathbf{W} = \sum_{i}^{K} p_i \mathbf{W}_i \tag{2.1}$$

In this formulation, $\mathbf{W}_i$ contains the weights trained by client i, $\mathbf{W}$ is the aggregated weights at the server, K is the total number of clients, and $p_i = \frac{|\mathcal{D}_i|}{|\mathcal{D}|}$, with $\sum_i p_i = 1$ and $|\mathcal{D}|= \sum_i |\mathcal{D}_i|$ is the total size of data samples. Based on this definition, the optimization problem can be formulated as:

$$\mathbf{W}^* = \underset{\mathbf{W}}{\operatorname{argmin}} \sum_{i}^{K} p_i \mathcal{L}_i(\mathbf{W}) \tag{2.2}$$

where $\mathcal{L}_i(\cdot)$ is the local loss function of the i-th client.

Similarly, in case of gradients aggregation (McMahan *et al.*, 2017), the server first sends the global model weights w to all participant clients, which then compute one-step gradients $g^i = \nabla_{\mathbf{W}} \mathcal{L}_k(\mathbf{W}, D_i)$. Next, the server collects the gradients of all clients and applies the update using a weighted average:

$$\mathbf{W}^* \leftarrow \mathbf{W} - \mu \sum_{i}^{K} p_i g^i \tag{2.3}$$

where μ is the global learning rate.

Thus, the training process of a federated learning system typically includes the following four steps:

1. Local training: All active clients compute gradients or parameters locally and send them to the server.
2. Model aggregating: The server performs a secure aggregation of parameters or gradients uploaded by K clients without knowing their local data.
3. Parameters broadcasting: The server transmits aggregated parameters to the K clients for next training round.
4. Model updating: All clients update their model with the aggregated parameters and test its performance.

The ability to train models on much larger and more diverse datasets than would be possible if the data were centralized is one of the primary advantages of federated learning. This can result in more accurate models with better generalization capabilities (Nguyen *et al.*, 2022). Moreover, because the data remains decentralized, there are fewer privacy and security concerns compared to conventional centralized machine learning techniques.

2.1.1 Types of Federated Learning

Federated learning can be categorized according to the distribution of the data. Assuming that the data matrix D_i denotes the information possessed by individual data owners, where each sample and characteristic are respectively represented by rows and columns in the matrix. Hereby, we define a collection of samples is referred to as the sample space which is the rows of the data matrix D_i, and a set of features is called feature space which is the columns of the data matrix D_i. The training dataset includes a set of samples with their features. Federated learning

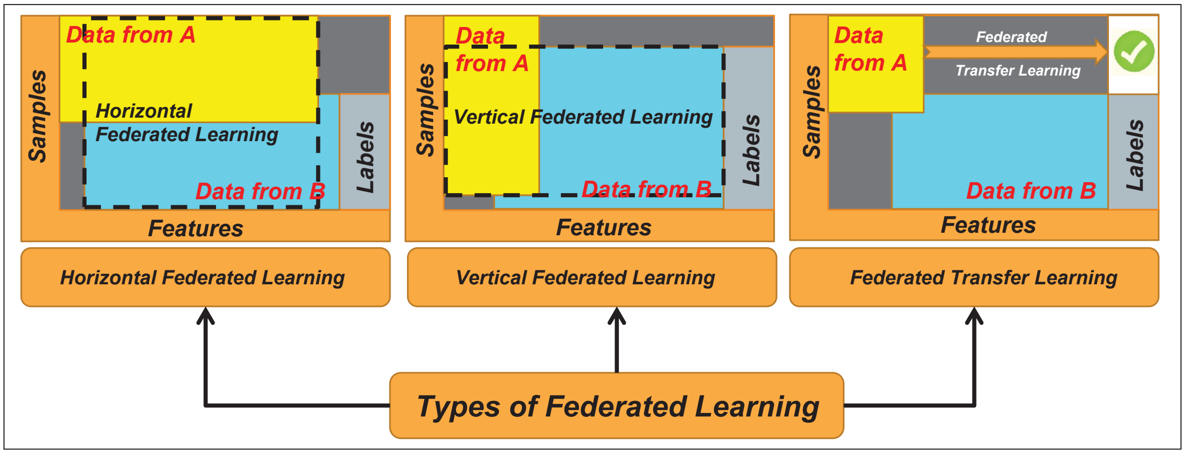


Figure 2.2 Illustration of three Federated Learning types including Horizontal Federated Learning, Vertical Federated Learning and Transfer Federated Learning

is categorized depending on how the data is dispersed among different parties in the feature space and sample space.

As illustrated in Fig 2.2, there are three main federated learning architectures: Vertical Federated Learning, Horizontal Federated Learning, and Federated Transfer Learning (Yang, Liu, Chen & Tong, 2019; Prayitno *et al.*, 2021; Liu, Zhang, Ge & Li, 2020b).

2.1.1.1 Vertical Federated Learning:

Also known as feature-based federated learning (Yang *et al.*, 2020c), this architecture is typically employed in situations where two or more client datasets share a similar sample space but distinct input feature spaces. This type of learning enables clients to aggregate the feature information they have for a specific sample by utilizing a third party to ensure that no information about the unique sample is shared during the feature sharing process. Consequently, vertical federated learning computes the cost function and gradients of a machine learning model while preserving the privacy of the unique sample data and collaboratively sharing the sample's unique features collected under various clients.

2.1.1.2 Horizontal Federated Learning:

Often referred to as sample-based federated learning (Yang *et al.*, 2020b), this architecture is implemented when client datasets have different sample spaces but share the same feature space. Horizontal federated learning is accountable for sharing information with cross-users in different clients in order to enrich the dataset for the machine learning model. It can also be utilized in a multitasking system in which multiple clients are permitted to learn distinct tasks while sharing knowledge across their various samples and protecting the confidentiality of sensitive data in the datasets. Horizontal federated learning ensures that no data leaks occur during client-to-client sharing, thereby protecting and preserving data. In (Aledhari, Razzak, Parizi & Saeed, 2020), authors state that horizontal federated learning focuses on security and its primary benefit to allow for independence in learning across clients.

2.1.1.3 Federated Transfer Learning:

This last architecture is typically applied to situations where two client datasets share a very small sample space but have different feature spaces (Yang *et al.*, 2020a). Unlike vertical federated learning which is only applicable for an entire intersecting dataset sample space, Federated Transfer Learning provides an intermediate solution through transfer learning that enables learning across the entire dataset even if only a small intersection exists across a similar sample.

2.1.2 Challenges in Federated Learning

There are a number of challenges to be overcome in order to make federated learning practical, such as dealing with data heterogeneity, managing communication and computation overheads, and maintaining the privacy and security of the data (Banabilah, Aloqaily, Alsayed, Malik & Jararweh, 2022).

The first challenge in federated learning is dealing with the varying data distributions across different clients. Since each client trains its model using local data, variability in data distributions may lead to sub-optimal performance of the global model. The assumption that training data are independent and identically distributed (IID) is frequently made in standard machine learning. However, this assumption rarely holds true in FL systems. In (Duan *et al.*, 2019), the authors proposed a framework called *Astraea* to address the issue of imbalanced distributed training data which causes accuracy degradation in FL settings. This framework eliminates global imbalance by runtime data augmentation. It also creates a mediator to reschedule the training of clients based on the Kullback–Leibler divergence across the different client data distributions in order to calculate average local imbalances.

Managing the communication and computation costs associated with FL is another challenge. Because each party must send updates to the central server, and the server must aggregate the updates before broadcasting the updated global model back to its clients, the amount of data transmitted can be substantial. Additionally, as long as the data is decentralized, the computation must be performed on each client's device, which can be a burden. In (Yao, Huang & Sun, 2018), the authors demonstrate that a two-stream model with Maximum Mean Discrepancy constraints decreases FL communication costs by 20%. In (Chen, Sun & Jin, 2020b), an asynchronous strategy is proposed for model update with a weighted aggregation technique. The outcomes demonstrated that the proposed architecture outperformed standard algorithms. *Yao et al.* proposed two solutions to address the communication and performance issues taken by the FL algorithm (Yao, Huang, Wu, Zhang & Sun, 2019). The first solution, called FedMMD, employs a two-stream model with Maximum Mean Discrepancy as opposed to a single model like Federated Averaging. The second solution is FL with FedFusion, which aggregates features from local and global models, resulting in increased precision and decreased communication costs. To solve the problem of communication overhead in FL, Wang, Wang & Li (2019b) introduced an algorithm called Communication Mitigated Federated Learning, which eliminates irrelevant client-side updates when training with client-specific, biased data.

In (Hatamizadeh *et al.*, 2023), the authors demonstrated that FL may be unsafe due to its vulnerability to model inversion attacks. Hence, maintaining the privacy and security of the data is also a major concern in FL. Since the data remains decentralized, it is important to ensure that the updates sent to the central server do not reveal any sensitive information about the data. In (Xu, Li, Liu, Yang & Lin, 2020), the authors used VerifyNet to address two issues with deep neural networks in FL settings: maintaining the confidentiality of user data during the training process, and verifying the accuracy of the model outcomes (or predictions) broadcasted by the server. The authors proposed a method called FedLDA that employs FL in Latent Dirichlet Allocation frameworks to mitigate the risks associated with data collection. Lu, Liao, Lio & Hui (2020) introduced a Privacy-Preserving Asynchronous Federated Learning Mechanism (PAFLM) for edge network computing, which achieves a more effective federated training without sharing private user data. Differentially Private Federated Learning (DP-FL) (Huang *et al.*, 2020) is another framework proposed for imbalanced data scenarios, which operates on the cloud. Truex, Liu, Chow, Gursoy & Wei (2020) presented LDP-Fed, a novel FL system with a formal privacy guarantee based on local differential privacy. The authors of (Liu, Li, Xu, Lu & He, 2020a) observe that if security is high in differential privacy settings, accuracy will be compromised. Consequently, they proposed an Adaptive Privacy-preserving Federated Learning (APFL) framework that maintains the privacy and precision of trained models in FL settings. Wang *et al.* (2019c) presented a solution for FL model privacy leakage. They use Generative Adversarial Networks with a multi-task discriminator to retrieve private client information and perform invisible updates on the server-side, unlike traditional federated learning which operates on the client-side.

In (Bhagoji, Chakraborty, Mittal & Calo, 2019), it is shown that federated model training is susceptible to data poisoning, also known as model poisoning, small malicious attempt to destroy the global model by making it misclassify specific inputs, resulting in negative effects on other participating clients. One of the primary challenges of federated learning is the malicious participation of clients who may inject the model with false input in order to corrupt the global model. Chen *et al.* (2020c) proposed a training-integrity protocol for Trusted Execution Environment in to detect and eliminate malicious attacks in early stages. Mowla, Tran, Doh & Chae (2020) developed a security architecture for flying ad-hoc networks that detects on-device jamming attacks based on frequency hopping. When updating the global model, the method can better identify client groups.

In summary, despite the numerous studies conducted on federated learning in recent years, several challenges remain open. It offers an attractive way to train a neural network with data hosted over different clients and provides an additional layer of privacy protection. However, this approach does not permit the use of a centralized cloud-based model for making test-time predictions without transmitting patient data, as the trained global model will eventually deployed on the client-side.

2.2 Homomorphic Encryption

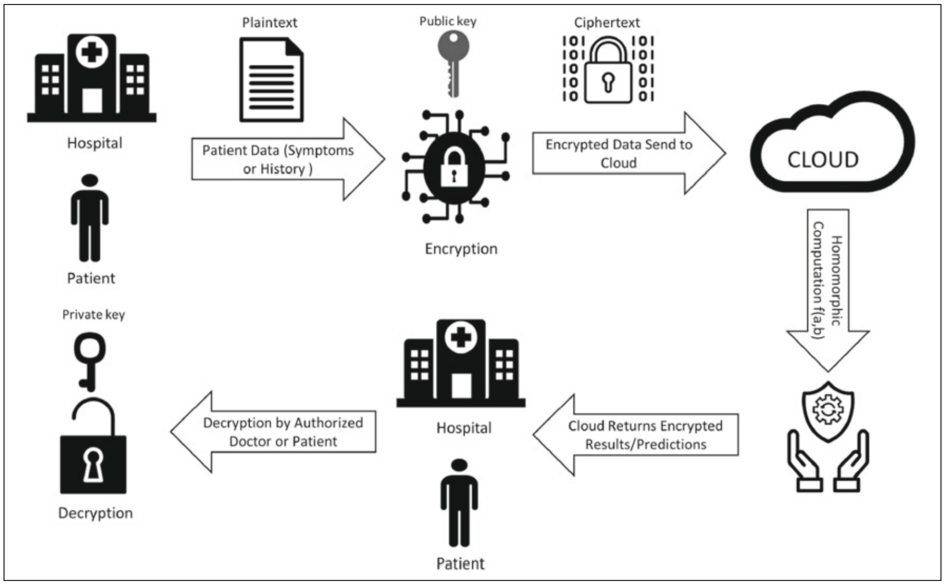

Hospital
Patient
Plaintext
Patient Data (Symptoms or History)
Public key
Encryption
Ciphertext
Encrypted Data Send to Cloud
CLOUD
Homomorphic Computation f(a,b)
Cloud Returns Encrypted Results/Predictions
Hospital
Patient
Decryption by Authorized Doctor or Patient
Private key
Decryption

Figure 2.3 Homomorphic encryption in healthcare.
Taken from Munjal & Bhatia (2022)

Homomorphic encryption (Dowlin *et al.*, 2016; Hesamifard *et al.*, 2017; Nandakumar *et al.*, 2019) is a type of encryption that allows computations to be performed on ciphertext. It produces an encrypted result that, when decrypted, matches the result of the operations as if they had been performed on plaintext.

Encryption methods typically prohibit the processing of encrypted data, meaning that data are always processed in their original form. In contrast, homomorphic encryption permits computation on encrypted data and provides the user with encrypted results. Homomorphic encryption not only enables the processing of encrypted data, but also maintains privacy in the process.

Homomorphic encryption is founded on the mathematical concept of homomorphism. A homomorphism is a structure-preserving map between two algebraic structures, such as groups, in abstract algebra (Yi, Paulet & Bertino, 2014). There are two main types: group homomorphisms and ring homomorphisms. Let $(G, \star)$ and $(H, \diamond)$ be groups, the map $\varphi : G \longrightarrow H$ is a group homomorphism if

$$\varphi(x \star y) \ = \ \varphi(x) \diamond \varphi(y), \ \forall x, y \in G \tag{2.4}$$

On the other hand, a ring homomorphism is defined as follows. Let R and S be rings with addition and multiplication, the map $\varphi : R \longrightarrow S$ is a ring homomorphism if

1. φ is a group homomorphism on the additive groups $(R, +)$ and $(S, +)$
2. $\varphi(x, y) = \varphi(x)\varphi(y), \ \forall x, y \in R$

One of the primary benefits of homomorphic encryption is that it enables computations to be performed on sensitive data without first decrypting it. This is especially useful in situations where data must be shared across multiple parties for the purpose of computation, but the data itself must remain private. For example, in Fig 2.3, *Munjal et al.* describe an application of homomorphic encryption for a cloud-based healthcare system in which a user sends sensitive data to the cloud in order to predict certain outcomes. Homomorphic encryption allows the users to encrypt the data, send it to the cloud, have the service provider perform the computations on the encrypted data, and then receive the encrypted result back without the service provider ever having access to the plaintext data. According to (Munjal & Bhatia, 2022), the application protocol consists of four distinct steps, as follows:

1. Step 1: The client (patient or physician) must initially encrypt the data. The client then transmits the encrypted data to the cloud-server for processing.

2. Step 2: The cloud-service executes operations using the homomorphic encryption property with some function $f(\cdot)$.
3. Step 3: The cloud-service returns the client's encrypted results.
4. Step 4: Client performs decryption at its end using the decryption function and recovers the encrypted results.

2.2.1 Types of Homomorphic Encryption

Homomorphic encryption seeks to develop an algorithm that permits an arbitrary number of additions and multiplications on encrypted data. The final result should be the ciphertext that would be generated if the same operations were performed on the corresponding plaintexts and then encrypted. Designing such an encryption algorithm is a difficult problem. Existing homomorphic encryption approaches can be categorized based on how close they are to achieving this objective. According to (Munjal & Bhatia, 2022), there are three types of homomorphic encryption: partially homomorphic encryption (PHE), somewhat homomorphic encryption (SWHE), and fully homomorphic encryption (FHE). Partially homomorphic encryption allows for a specific type of computation, such as addition or multiplication, to be performed on ciphertext. Somewhat homomorphic encryption supports a limited amount of operations, i.e., it evaluates the circuit up to a certain depth or limit. Fully homomorphic encryption allows for any computation to be performed. Fully homomorphic encryption is considered to be most powerful, but also the most computationally expensive and less practical approach to implement in practice.

2.2.1.1 Partially homomorphic encryption:

Partially homomorphic encryption algorithms enable the infinite repetition of a specific operation. An algorithm may be additively homomorphic, meaning that adding two ciphertexts produces the same result as encrypting the sum of two plaintexts.

Designing partially homomorphic encryption algorithms is relatively simple. There are numerous available partial homomorphic encryption schemes. Rivest, Adleman & Dertouzos (1978) first proposed the term “privacy homomorphism” when they introduced an asymmetric encrypted system that supports the multiply operation over ciphertext. Their proposed system relies on the difficulty of the prime factorization problem. In (Goldwasser & Micali, 1982), the authors introduced the first scheme with semantic security proof, based on the hardness of quadratic residuosity assumption described by Kaliski (2011). Diffie & Hellman (1976) presented a key exchange algorithm that minimizes the need for secure key distribution channels and supplies the equivalent of a written signature. This key exchange algorithm was then improved by Elgamal (1985). In (Fousse, Lafourcade & Alnuaimi, 2011), the authors proposed dense probabilistic encryption, a homomorphic encryption scheme over the addition operator with an improved expansion factor. Compared to prime residuosity, Paillier (1999) offered a “trapdoor” technique based on composite residuosity classes that are advantageous to public-key cryptosystems. Originally, this cryptosystem allowed additions to be done over encrypted data, but subsequent enhancements to the method demonstrated that multiplications could also be accomplished. Due to the difficulty of lattice operations, the author of (Kawachi, Tanaka & Xagawa, 2007) presented a homomorphic cryptosystem, dubbed *pseudohomomorphic*, with addition over a huge cyclic group. In (Galbraith, 2001), authors presented a more natural adaptation of the cryptography method in Paillier (1999). This approach is applied to elliptic curves, while retaining the other homomorphic characteristics of Paillier (1999).

2.2.1.2 Somewhat homomorphic encryption:

Somewhat homomorphic encryption is the next step over partially homomorphic encryption. A relatively homomorphic encryption technique enables a finite number of operations, compared to an infinite number of a single operation as in partially homomorphic encryption. Each addition operation increases noise, and each multiplication operation multiplies noise. Similar to fully homomorphic encryption, somewhat homomorphic encryption is a unique but incomplete solution. For instance, a somewhat homomorphic encryption technique for encrypting data may support any combination of up to five additions or multiplications. A sixth operation of either type, however, would produce an invalid result.

Boneh, Goh & Nissim (2005) introduced a mechanism for adding and multiplying encrypted data that is semantically secure. This mechanism permits an infinite amount of additions with a single multiplication on ciphertext of a defined length, and supports quadratic formula evaluation over ciphertexts. The proposed method hardness depends on the subgroup decision problem as described in (Gjøsteen, 2005), in which one must determine if a given element g_1 of a finite group G is a member of subgroup G_1. The work in (Yao, 1982) is an early attempt to perform function operations on the ciphertext. As a solution to the Millionaires' Issue, authors devised a two-party communication protocol that compares the wealth of two affluent individuals without revealing the exact amounts. Moreover, in this method, the ciphertext grows linearly at most when each gate in the circuit is calculated. Ishai & Paskin (2007) presented an encryption method based on the evaluation of a branching algorithm on encrypted data. In (Sander, Young, Yung & Inc, 2001), the authors introduced the first somewhat homomorphic encryption scheme over a semi-group for NC1 is a class of circuits with poly-logarithmic depth and polynomial dimension. With a single OR/NOT gate, the proposed method enabled polynomially many ANDing of ciphertexts. After each multiplication, the ciphertext grows in size.

On the path to completely homomorphic encryption, somewhat homomorphic encryption methods are crucial stepping stones. Even for a fixed number of operations, it is more difficult to build an algorithm that supports both addition and multiplication of ciphertexts than it is to design one that permits limitless addition or multiplication of ciphertexts.

2.2.1.3 Fully homomorphic encryption:

A fully homomorphic encryption algorithm allows to perform an infinite number of ciphertext additions and multiplications while still producing valid results. This type of encryption has a great potential of making privacy and functionality compatible by keeping information both secure and accessible. In contrast to other forms of homomorphic encryption, it is capable of performing arbitrary computations on the ciphertexts. Fully homomorphic encryption methods can be divided into four distinct groups: lattice based cryptography, learning with errors (LWE/RLWE), integer based and NTRU. These sophisticated cryptographic methods serve as prerequisites for fully homomorphic encryption algorithms (Munjal & Bhatia, 2022).

Lattice-based encryption is an advanced form of cryptography. Cryptosystems based on this approach are built on difficult tasks such as the Shortest Vector Problem to discover the shortest non-zero vector in the grid, or the Closest Vector Problem to identify the lattice vector that is closest to the provided vector. A lattice in $n - dimensional$ space is any regular grid of points. Technically, n independent vectors known as the lattice's basis $\{b1, b2, b3, ..., bn\}$ are formed as a vector set follows:

$$\mathcal{L}(b_1, b_2, b_3, ..., b_n) = \sum_{j=1}^{N} x_j b_j,\ x_j \in \mathbb{Z} \tag{2.5}$$

Gentry (2009) introduced the first fully homomorphic encryption algorithm, which is based on ideal lattices with the bootstrappable property.

Learning with errors is a fundamental principle that is at the core of advanced cryptographic methods. It generalization the problem of Learning Parity with Noise. Learning with errors can be conceptualized as two closely connected sub-problems. Decision Learning with Errors and Search Learning with Errors. Learning with errors is a potent cryptography tool because its difficulty is equivalent to that of lattice-based problems, namely the shortest vector problem and the nearest vector problem. In (Regev, 2009), the hardness of worst-case lattice problems was reduced from Shortest Vector Problem to Learning with Errors, indicating that if an algorithm can solve the Learning with Errors problem in a reasonable amount of time, it can also handle the Shortest Vector Problem. The work in (Lyubashevsky, Peikert & Regev, 2010) introduces the Ring Learning with Errors problem, which is a significant improvement to the Learning with Errors problem and enables the development of new applications. It showed that Ring Learning with Errors problems might be reduced to worst-case problems on ideal lattices, which is a difficult task for polynomial-time quantum algorithms. Using squashing and bootstrapping techniques as in (Gentry, 2009), a more realistic fully homomorphic encryption scheme was developed in (Brakerski & Vaikuntanathan, 2011a).This scheme uses a straightforward form of Ring Learning with Errors, called Polynomial Learning with Errors, which can also be reduced to worst-case situations similar to Shortest Vector Problem on perfect lattices. The authors of (Brakerski, Gentry & Vaikuntanathan, 2012) developed a leveled homomorphic encryption technique to limit the noise development associated with each operation. In contrast to the multiplicative depth of the circuit, the noise's expansion is linear. The proposed strategy was based on a mechanism for modulus switching known as the BGV scheme. In (Gentry, Halevi & Smart, 2012), this mechanism was also employed with AES to achieve a better performance. However, it requires many versions of the public key, necessitating a huge amount of memory on the system. Brakerski (2012) introduced a scale-invariant approach for leveled homomorphic encryption algorithms. In contrast to modulus switching, the modulus of the ciphertext is maintained during the homomorphic evaluation. Hence, only one copy of the scale-invariant evaluation key must be kept. This scheme was transformed from Learning with Errors to Ring Learning with Errors through a comprehensive investigation of numerous subroutines, such as multiplication, bootstrapping, and re-linearization. In (Brakerski & Vaikuntanathan, 2011b), a re-linearization strategy that creates a somewhat homomorphic encryption without assuming ideal hardness was introduced. With the intent of converting somewhat homomorphic encryption to fully homomorphic encryption without using squashing and Sparse Subset Sum Problem, a dimensional-modulus reduction was also provided. In prior Learning with Errors schemes, the multiplication step was difficult and costly to perform due to re-linearization. In (Gentry, Sahai & Waters, 2013), the authors proposed a novel methodology for fully homomorphic encryption, the approximate eigenvector method, which does all addition and matrix operations through matrices, making it asymptotically faster. Without the re-linearization stage, matrices were simply added and multiplied.

The work in (Van Dijk, Gentry, Halevi & Vaikuntanathan, 2010) describes a fully homomorphic encryption scheme based on integers, where the hardness of the algorithm comes from the Approximate Greatest Common Divisor problem (Galbraith, Gebregiyorgis & Murphy, 2016).

While this approach was symmetric, a basic symmetric homomorphic encryption technique that can be transformed into an asymmetric homomorphic encryption strategy was presented by Van Dijk *et al.* (2010). In (Cheon *et al.*, 2013), the batch fully homomorphic encryption technique over integers was proposed as an extension of integer-based fully homomorphic encryption. Using the well-known Chinese Remainder Theorem, l plaintexts $\{m_0, m_1, ..., m_l\}$ were packed into a single ciphertext using the DGHV method. This technique is capable of encrypting not only bits but also elements from $\mathbb{Z}_Q$ rings. Further fully homomorphic encryption techniques have also been developed over integers: a scale-invariant fully homomorphic encryption over integers (Coron, Lepoint & Tibouchi, 2014), a technique with integer plaintext (Ramaiah & Kumari, 2012), a symmetric fully homomorphic encryption scheme that does not require bootstrapping (Aggarwal, Gupta & Sharma, 2014), and a somewhat homomorphic encryption method for conducting arithmetic operations on large integer values without converting to bits (Pisa, Abdalla & Duarte, 2012). All of these methods, as their names imply, improved fully homomorphic encryption over integers.

Hoffstein, Pipher & Silverman (1998) proposed the NTRU-based encryption approach, which is the first attempt at encrypting lattice problems. In (López-Alt, Tromer & Vaikuntanathan, 2012), the authors employ on-the-fly multiparty computation, wherein each user is responsible for delivering encrypted data to the cloud and decrypting it when outputs are received. Their approach employs a novel type of encryption, multikey fully homomorphic encryption, which is capable of operating on inputs encrypted with many distinct and unrelated keys. The multikey fully homomorphic encryption protocol was built on NTRU, an efficient public-key encryption protocol. Despite the fact that NTRU was not initially fully homomorphic, its transformation into a fully homomorphic structure decreases its efficiency but increases its capacity. Yet, the decisional small polynomial ratio assumption relating to the uniformity of public-key cryptography is necessary (Stehlé & Steinfeld, 2011). The authors of (Bos, Lauter, Loftus & Naehrig, 2013) avoid this additional assumption and propose a secured fully homomorphic encryption system under Ring Learning with Errors with simply circular security assumptions. In (Qin, Huang & Fan, 2021), a fully homomorphic encryption approach based on the power-of-prime cyclotomic ring was proposed. The Ring Learning with Errors assumption-based method does not require the decision small polynomial ratio assumption. The effectiveness of the approach results in enhanced noise management, which improves ciphertext storage, processing, and communication.

In summary, fully homomorphic encryption aims to enable anyone to perform useful operations on encrypted data without having access to the encryption key. This can enhance security in various cloud computing applications.

2.2.2 Problems with Homomorphic Encryption

Currently, the main issue with fully homomorphic encryption is that it is computationally intensive. By complying with the requirements of full homomorphism (i.e., permitting ciphertexts to be added or multiplied an infinite number of times without corrupting the result), these algorithms are slow and may have extremely high storage requirements.

In 2018, IBM released an updated version of its HElib C++ library for homomorphic encryption (Halevi & Shoup, 2018). This version is 25-75 times faster than its predecessor, which was 2 million times faster than the original version released three years before. However, the original version performed mathematical operations approximately 100 quadrillion times slower than the corresponding plaintexts, hence the new and improved version remains approximately one million times slower than plaintext operations on average. A million-factor slowdown is quite significant. Using the 2018 version of HElib, a calculation that would take a second using plaintexts would take an average of 11,5 days. Clearly, institutions that would otherwise be interested in homomorphic encryption cannot accept such a trade-off. However, a 100 million-fold acceleration over three years is quite impressive. Currently, homomorphic encryption may not be a viable option, but this could change in the near future.

Another problem with homomorphic encryption is the processing of activation functions in neural networks. Current homomorphic encryption method cannot handle nonlinear functions as popular activation functions, such as the sigmoid and ReLU functions (Pulido-Gaytan *et al.*, 2021; AboulAtta, Ossadnik & Ahmadi, 2019). Some methods substitute conventional functions with the nonlinear low-degree square function (Dowlin *et al.*, 2016; Brutzkus, Gilad-Bachrach & Elisha, 2019). However, the unbounded derivation of the square function induces problems during training for networks with more than two non-linear layers (Chabanne, De Wargny, Milgram, Morel & Prouff, 2017). The limitation of the polynomial approximation of the activation function is addressed by a number of methods, including Taylor series and Chebyshev polynomials (Chabanne *et al.*, 2017; Hesamifard *et al.*, 2017; Al Badawi *et al.*, 2020; Takabi, Hesamifard & Ghasemi, 2016; Shokri & Shmatikov, 2015; Bakshi & Last, 2020; Bourse, Minelli, Minihold & Paillier, 2018). Nevertheless, approximating with polynomials of the lowest possible degree remains a challenging problem.

Overall, homomorphic encryption is a powerful tool that has the potential to revolutionize the way sensitive data is shared and used for computation. While there are still challenges to overcome in terms of practical implementation and performance, it is a promising area of research that will likely continue to evolve and grow in importance in the coming years. Due to the computational cost, however, homomorphic encryption cannot be currently used in practical applications.

2.3 Adversarial Learning for Preserving Privacy

2.3.1 Introduction

The recent success of adversarial learning has led to the increased adoption of this technique for the protection of sensitive information, particularly in visual data. A large number of works (Pittaluga, Koppal & Chakrabarti, 2019; Wu *et al.*, 2018; Yang, Brinton, Mittal, Chiang & Lan, 2018; Roy & Boddeti, 2019) leveraged adversarial training to jointly optimize privacy and utility objectives. In these studies, the mapping functions for the adversarial and task-specific terms are standard classification models where the number of classes is fixed. In (Chen, Konrad & Ishwar, 2018a), a model which integrates a Variational Autoencoder (VAE) and a GAN is proposed to create an identity-invariant representation of face images. To explicitly control the features to be preserved, this model includes a discriminator which predicts the identity of the subject in a generated image. As the number of possible labels corresponds to the number of subjects to identify, this approach is not suitable for large-scale applications. To alleviate this problem, Oleszkiewicz, Kairouz, Piczak, Rajagopal & Trzciński (2018) use a Siamese architecture for the discriminator, which predicts whether two encoded images come from the same subject or not. In this previous work, an auto-encoder loss is employed as task-agnostic utility objective to avoid the encoder from generating trivial images.

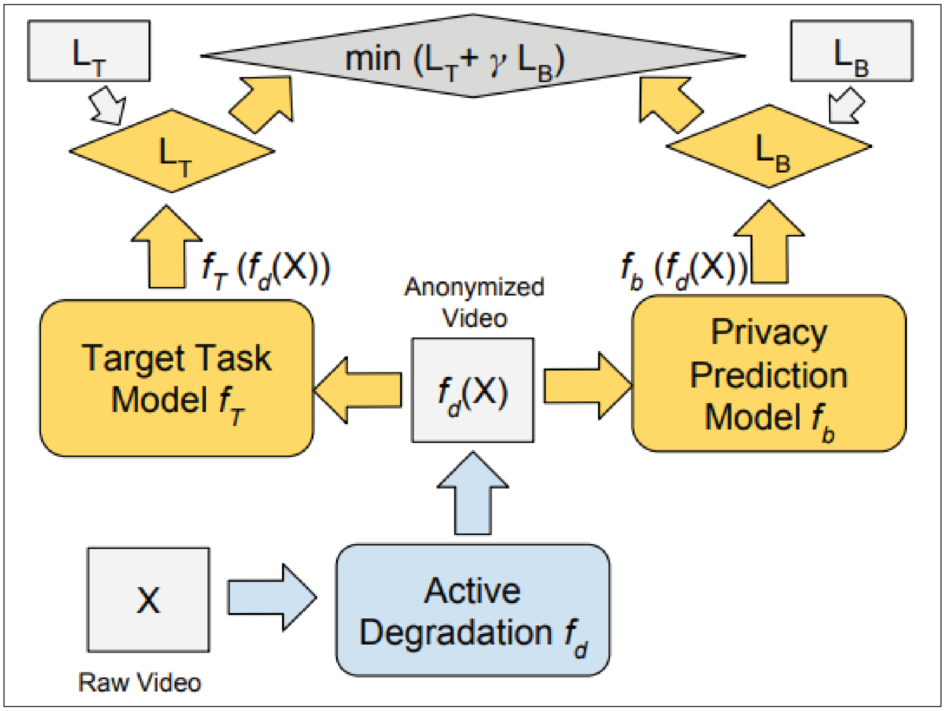


Figure 2.4 Basic adversarial training framework for privacy-preserving visual recognition.
Taken from Wu *et al.* (2018)

2.3.2 Adversarial Training Framework for Privacy-preserving Visual Recognition

In (Wu *et al.*, 2018), the authors proposed an adversarial training framework for privacy-preserving visual recognition. The proposed framework explicitly learns a degradation applied to the original inputs to optimize the trade-off between target task performance and the associated privacy budgets on the degraded video.

The proposed framework is illustrated in Fig 2.4. In this framework, X is the training data, f_T the model to perform a target task T, Y_T the data label associated with the target task, L_T is the loss function which is defined to evaluate the target task performance, L_B is a privacy budget loss function which is defined to evaluate the privacy leakage, and f_d is the active degradation function which is used to distort the input data X. The goal of privacy-preserving visual recognition is expressed as:

$$\min_{f_T, f_d} \; L_T(f_T(f_f(X)), Y_T) \; + \; \gamma L_B(f_d(X)) \tag{2.6}$$

The desired f_d visually degrades the input X as the common input for both the target task and the privacy budget in such a way that: 1) the target task performance is minimally affected, i.e.,

$$\min_{f_T, f_d} \; L_T(f_T(f_f(X)), Y_T) \; \approx \; \min_{f'_T} \; L_T(f'_T(X), Y_T) \tag{2.7}$$

and 2) the privacy budget is greatly reduced

$$L_B(f_d(X)) \; \ll \; L_B(X) \tag{2.8}$$

The definition of the privacy budget loss L_B poses two main challenges. First, the privacy budget-related annotations, Y_B, are not always available. Secondly, it is not sufficient to merely suppress the success rate of one f_b model. The author proposed to define a privacy prediction function family $\mathcal{P} : f_b(X) \rightarrow Y_B$. The ideal active degradation function f_d must suppress every possible model f_b from $\mathcal{P}$. Based on this idea, the problem becomes as follows:

$$\min_{f_T, f_d} \; L_T(f_T(f_f(X)), Y_T) \; + \; \gamma \max_{f_b \in \mathcal{P}} \; L_B(f_b(f_d(X)), Y_B) \tag{2.9}$$

For the solved f_d, the two goals should be simultaneously satisfied. Thus, there exists at least one f_T function that can predict Y_T from $f_d(X)$, and no function $f_b \in \mathcal{P}$ can predict Y_B from

$f_d(X)$. In this method, all three modules, f_d, f_T and f_b are learnable and implemented by neural networks. The entire system is trained under the combination of L_T and L_B in an adversarial manner.

By providing a solid foundation, this framework not only facilitates the development of novel privacy-preserving methods, but also serves as a point of reference for researchers and practitioners in the field.

2.3.3 Adversarial Learning Based Privacy-preserving Method

In the field of image recognition, Xu *et al.* (2019) proposed a method that added carefully-designed noise to gradients during the learning procedure in order to train a differentially-private GAN. In (Raval, Machanavajjhala & Cox, 2017), the authors presented a perturbation mechanism leveraging adversarial learning that optimises both privacy and utility objectives. During the training process, an unsupervised utility loss minimised based on the assumption that the encoded representations can be generated by removing sensitive attributes from an image while minimising alterations to the remaining components. These encoded representations are then used to learn a downstream task. Due to the fact that the encoding is performed independently of the downstream task, it may not be optimal for this task. Several studies used adversarial learning to simultaneously optimise privacy preservation and utility (Pittaluga *et al.*, 2019; Wu *et al.*, 2018; Yang *et al.*, 2018; Roy & Boddeti, 2019). Standard classification models are employed as mapping functions for adversarial and task-specific components in these studies. Notably, throughout the optimisation process, the number of classes within these models remains unchanged. *Chen et al.* introduced a model that integrates a Variational Autoencoder (VAE) and a Generative Adversarial Network (GAN) to generate identity-invariant representations for facial images (Chen *et al.*, 2018a). The model includes a discriminator which is used to predict the subject's identity within a synthesised image in order to explicitly regulate the preservation of specific features. Due to the requirement of a one-to-one correspondence between the number of possible classes and the number of subjects to be identified, the applicability of this method to large-scale scenarios, such as the one in this thesis, is limited.

In (Li & Choi, 2021), the author introduced DeepBlur, a straightforward yet efficacious approach to image obfuscation through the application of blurring within the latent space of a unconditionally pre-trained generative model capable of synthesizing facial images with photo-realistic attributes. In (Wu, Lim, Davis & Goldstein, 2020), to mount adversarial attacks on object detectors, the author employed conventional detection datasets to train patterns designed to attenuate the objectness scores generated by a spectrum of widely utilized detectors and detector ensembles. In (Li & Lin, 2019), the author introduces a novel framework named AnonymousNet, aimed at systematically mitigating concerns pertaining to the de-identification of face images, with a concerted emphasis on achieving a harmonious balance between usability and augmenting privacy in a discernible and intrinsic manner. The framework comprises four distinct stages: facial attribute estimation, face obfuscation guided by privacy metrics, targeted synthesis of natural images, and the introduction of adversarial perturbations.

In order to address the issue arising from a variable number of classes, Oleszkiewicz *et al.* (2018) introduced a Siamese architecture for the discriminator component. This architecture is designed to predict if two encoded images originate from the same subject or not. By employing this method, the researchers expected to circumvent the limitation posed by variable number of classes. Nevertheless, the objectives pursued by this method is significantly different from those of our own study. This approach's primary objective is to identify the minimal transformation required to remove identity-related information from an image, thereby enabling its subsequent use by non-specific applications. In contrast, the objective of our research in this thesis is to implement a highly robust image obfuscation transformation technique. The primary objective is to ensure that the subject's identity remains unrecoverable while facilitating the use of the encoded image for training image analysis tasks, particularly in the segmentation domain.

2.4 Subject re-identification from image:

In this thesis, we adopt MS-SSIM score and similarity score produced by a Siamese network as the metrices to measure the similarity between image. Hence, we proceed the subject re-identification experiment based on the similarity score. In the literature, there are also feature based method to compute the similarity between two MRI images.

In (Chauvin *et al.*, 2020), the authors presented a novel pairwise brain similarity measure by leveraging a distinct keypoint signature —namely, a collection of unique, localized patterns identified automatically in each image through a generic saliency operator. The quantification of the similarity between a pair of images is accomplished by assessing the proportion of keypoints they share, employing a novel Jaccard-like measure indicative of set overlap. Experimental validations demonstrated the notable efficiency and accuracy of the keypoint method, involving a dataset of 7536 T1-weighted Brain MRIs amalgamated from four publicly available neuroimaging repositories, encompassing subjects such as twins, non-twin siblings, and 3334 distinct individuals.

In (Toews, Wachinger, Estepar & Wells, 2015), the authors proposed an inference methodology particularly suitable for extensive collections of medical images. This approach is rooted in a framework where distinctive 3D scale-invariant features are efficiently indexed, enabling the identification of approximate nearest-neighbor (NN) feature matches with a computational complexity of $O(logN)$, where N represents the number of images. Consequently, this method demonstrates scalability to large datasets, a notable departure from approaches reliant on pairwise image registration or feature matching, which incur $O(N)$ complexity. A key innovation lies in the incorporation of a density estimator founded on a generative model that extends beyond conventional kernel density estimation and K-nearest neighbor (KNN) methods. The efficacy of the proposed method is substantiated through validation on an extensive multi-site dataset comprising 95,000,000 features extracted from 19,000 lung CT scans.

2.5 Summary

In this chapter, we reviewed the domain of privacy-preserving deep learning, examining a variety of innovative techniques designed to protect sensitive data and protect individual privacy. Federated learning, homomorphic encryption, and adversarial strategies were among the prominent strategies discussed.

Federated learning emerged as a revolutionary paradigm that permits multiple parties to train a machine learning model collaboratively without sharing their raw data. This paradigm ensures data privacy by distributing the learning process across decentralized devices, allowing the model to learn from diverse sources and capture a more comprehensive representation of real-world scenarios. Nonetheless, the implementation of federated learning (FL) in cloud-based solutions is restricted by the need for hardware resources to facilitate training at each individual site.

Homomorphic encryption, another potent technique, permits computations on encrypted data without the need for decryption, preserving the privacy of this data throughout the entirety of the computation process. This method provides a robust solution for secure deep learning, as it enables the analysis of sensitive data without compromising confidentiality, thereby minimizing the risk of exposing personal information. However, the computational burden associated with this approach imposes constraints on its wide-ranging applicability.

Adversarial techniques were also studied for the purpose of preserving privacy in deep learning. The objective of such techniques is to mitigate the risks associated with the disclosure of sensitive information by training an encoder so that a discriminator cannot recover private information. However, the application of this method to image segmentation remains challenging. Currently, the use of adversarial method is mainly restricted to situations in which the downstream task is classification.

This chapter shed light on recent advancements in privacy-preserving deep learning. By employing federated learning, homomorphic encryption and adversarial approaches, researchers and practitioners try to strike a balance between data utility and privacy protection, thereby opening up new horizons for the safe and responsible use of deep learning technologies across a variety of domains.

CHAPTER 3

PRIVACY-NET: AN ADVERSARIAL APPROACH FOR IDENTITY-OBFUSCATED SEGMENTATION OF MEDICAL IMAGES

Ngoc Bach, Kim[1] , Jose Dolz[1] , Pierre-Marc Jodoin[2] , Christian Desrosiers[1]

[1] Department of Software and IT Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Department of Computer Science, Université de Sherbrooke,
2500 Boul de L'université, Sherbrooke, Québec, Canada J1K 0A5

Article published in Journal *«IEEE Trans. Medical Imaging 40(7)»* 2020.

This chapter presents the article "Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images" by Kim, Dolz, Jodoin, and Desrosiers accepted by the *IEEE Trans. Medical Imaging 40(7)* journal for publication in 2020. The objective of this research is to propose a new method for privacy-preserving in medical image segmentation. Our approach is illustrated on the problem of segmenting brain MRI from the large-scale Parkinson Progression Marker Initiative (PPMI) dataset. we show that the discriminator learns to heavily distort input images while allowing for highly accurate segmentation results. Our results also demonstrate that an encoder trained on the PPMI dataset can be used for segmenting other datasets, without the need for retraining.

3.1 Introduction

Machine learning models like deep convolutional neural networks (CNNs) have achieved outstanding performances in complex medical imaging tasks such as segmentation, registration, and disease detection (Zhou, Greenspan & Shen, 2017; Litjens *et al.*, 2017). However, privacy restrictions on medical data including images impede the development of centralized cloud-based image analysis systems, a solution that has its share of benefits: no on-site specialized hardware, immediate trouble shooting or easy software and hardware updates, among others.

While server-to-client encryption can prevent attacks from outside the system, it cannot prevent cybercriminals within the system from gaining access to private medical data. Another approach to obfuscate the identity of a patient is to anonymize its data. In case of images, this is done by removing the patient-related DICOM tags or by converting it into a tag-free format such as PNG or NIFTI. However, as shown by Kumar et al. (Kumar, Toews, Chauvin, Colliot & Desrosiers, 2018) and further illustrated in this paper, the raw content of an image can be easily used to recover the identity of a person with up to 97% of accuracy.

A recent solution for decentralized training on multi-centric data is federated learning (McMahan *et al.*, 2016). The idea behind this strategy is to transfer the training gradients of the data instead of the data itself. While such approach is appealing to train a neural network with data hosted in different hospitals, it does not allow the use of a centralized cloud-based model for making predictions at test time without transmitting patient data.

Another solution for privacy protection is homomorphic-encryption (HE) (Dowlin *et al.*, 2016; Hesamifard *et al.*, 2017; Nandakumar *et al.*, 2019). Although it ensures absolute data protection, one can also train a neural network with both encrypted and non-encrypted data. Unfortunately, since the HE operations are limited to multiplication and addition, the non-linear operations of a CNN have to be approximated by polynomial functions which makes neural networks prohibitively slow. For example, (Nandakumar *et al.*, 2019) reports computation times above 30 minutes to process a single 28×28 image using an optimized network with only 954 nodes. Thus, homomorphic neural networks so far proposed have been relatively simplistic (Hardy *et al.*, 2017) and it is not clear how state-of-the-art medical image analysis CNNs like U-Net (Ronneberger *et al.*, 2015) could be implemented in such framework. Furthermore, HE imposes important communication overhead (Rouhani, Riazi & Koushanfar, 2018) and its use within a distributed learning framework is still cumbersome (Hardy *et al.*, 2017).

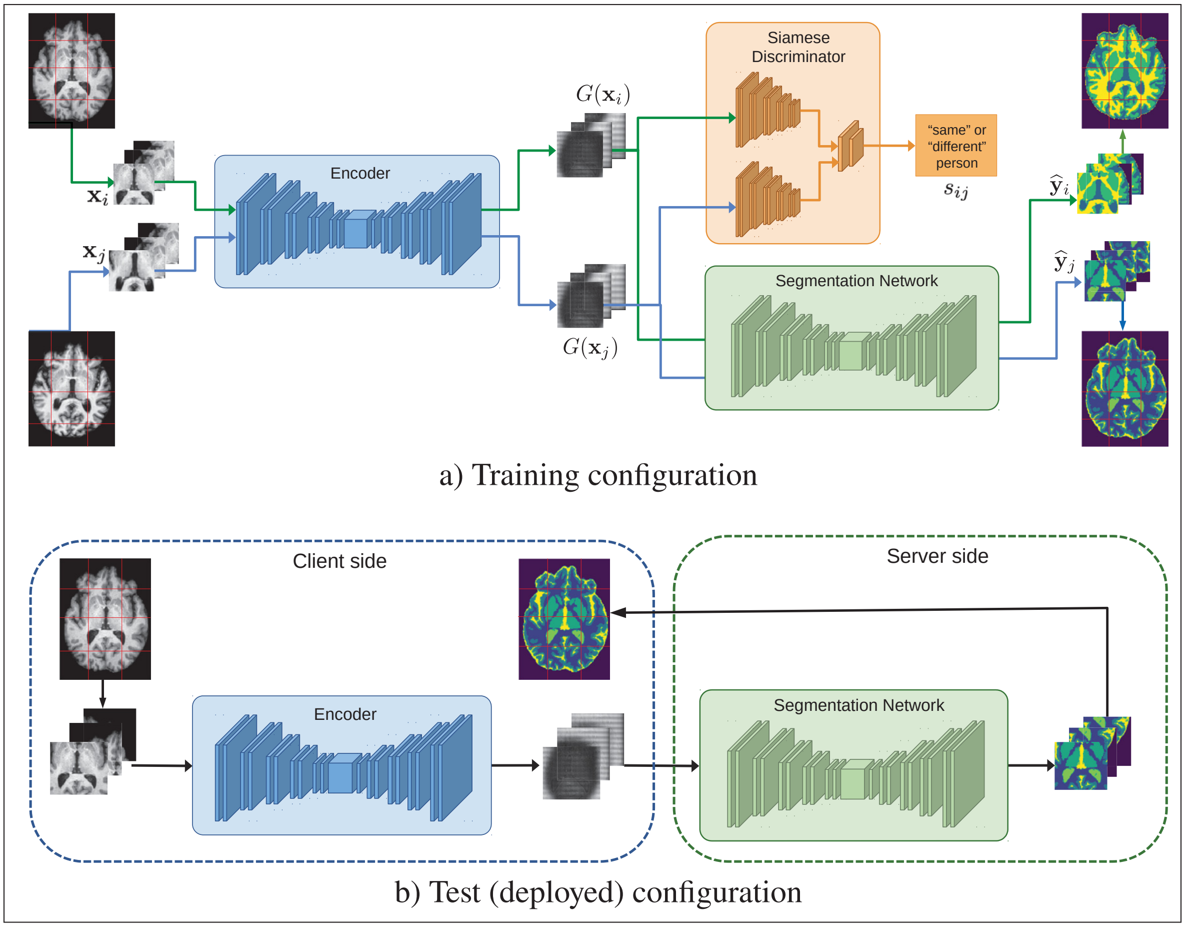


Figure 3.1 (a) Training configuration of our proposed system: 1) a client-side encoder network G converts input images $\mathbf{x}_i$ and $\mathbf{x}_j$ into two feature maps $G(\mathbf{x}_i)$ and $G(\mathbf{x}_j)$, 2) the discriminator network D tries to determine if its input data comes from the same patient ($s_{ij} = 1$) or not ($s_{ij} = 0$), and 3) a server-side segmentation network S segments the encoded images. (b) At test time, the discriminator is removed from the system and images are processed one at a time by the encoder on the client side and the segmentation network deployed on the server. The segmentation result is sent back to the client. Our networks' input size is $64 \times 64 \times 64$, thus the input images are cropped in to $64 \times 64 \times 64$ patches and shuffled before sending to server

In this paper, we propose a client-server system which allows for the analysis of multi-centric medical images while preserving patient identity. A high-level view of the proposed system is given in Fig. 3.1. On the client side is an encoder that converts patient-specific data into an identity-obfuscated signal containing enough semantic information to analyse its content. The encoded data is then sent to the server where it is analyzed and the results of this analysis are sent back to the client. Since each hospital has the same encoder, the server can keep on updating its system without having access to patient-specific information.

We achieve this with an adversarial learning approach inspired by generative adversarial networks (GAN) (Luc, Couprie, Chintala & Verbeek, 2016; Goodfellow *et al.*, 2014; Ganin *et al.*, 2016) but with two main differences. As illustrated in Fig. 3.1, instead of being a two-network configuration, our system involves three networks: 1) an image encoder, 2) a discriminator and 3) a medical image analysis network (a segmentation CNN in our case). Whereas the encoder's objective is to obfuscate the content of a raw input image, the goal of the discriminator is to determine whether two encoded images come from the same patient or not. The third network is a CNN which analyzes the content of the encoded image. As such, while the encoder tries to fool the discriminator, it must preserve enough information to allow the third network to successfully analyze its content. At test time, the encoder network residing on the client side converts a raw image $\mathbf{x}$ into an encoded (and yet secure) feature map $G(\mathbf{x})$. Thereafter, $G(\mathbf{x})$ is transferred to the cloud-based server where the segmentation network is deployed. The resulting segmentation map $\widehat{\mathbf{y}}$ is then sent back to the client.

The major contributions of this work are as follows:

1. We present the first client-server system for semantic medical image segmentation which allows for identity-preserving distributed learning. Obfuscating identity while preserving task-specific information is particularly challenging for segmentation, which requires to assign a label for each image pixel.
2. Our model proposes a novel architecture combining two CNNs, for the encoder and segmentation network, with a Siamese CNN for the discriminator. This Siamese discriminator learns identity-discriminative features from image pairs instead of a single image, allowing us to have a variable number of classes (i.e., subject IDs). Unlike the work in (Oleszkiewicz *et al.*, 2018), our model is trained using both an adversarial Siamese loss and a task-specific loss,

thereby providing encoded images that obfuscate identity while preserving the information required for the target task.

3. We provide a theoretical analysis showing that the proposed model minimizes the mutual information between pairs of encoded images and a variable indicating if these images are from the same subject. This analysis motivates our approach from a information theoretic perspective.
4. We demonstrate that the privacy-preserving encoder learned with a given dataset can be used to encode images from another dataset, and that these encoded images are useful to update the segmentation network.

3.2 Related Works

3.2.1 Privacy preserving in visual tasks:

Traditional methods to preserve privacy rely on cryptographic approaches (Ziad, Alanwar, Alzantot & Srivastava, 2016; Wang, Vong, Yang & Wong, 2017) which create local homomorphic encryptions of visual data. Although these methods perform well in some applications, homomorphic cryptosystems typically incur high computational costs (Paillier, 1999) and are mostly restricted to simple linear classifiers. This limits their usability in scenarios requiring more complex models like deep neural networks. Another solution consists in extracting feature descriptors from raw images, which are then transferred to the encrypted dataset server (Hsu, Lu & Pei, 2011). Nevertheless, sensitive information from original images can be still recovered from standard features, making these systems vulnerable to cyberattacks. An alternative strategy is to employ low-resolution images (Dai, Saghafi, Wu, Konrad & Ishwar, 2015; Chen, Wu, Richter, Konrad & Ishwar, 2016a) or image filtering techniques (Butler, Huang, Roesner & Cakmak, 2015; Jalal, Uddin & Kim, 2012) to degrade sensitive information. However, since these approaches also reduce the quality of the visual content, they are limited to a reduced set of tasks such as action or face expression recognition. More recently, McClure et al. (McClure *et al.*, 2018) proposed using continual learning to circumvent the issue of privacy preservation in the context of multi-center brain tumor segmentation. Nevertheless, unlike our method, their approach is not directly optimized to obfuscate identity from visual data.

3.2.2 Federated learning:

Federate learning has recently emerged as a solution to build machine learning models based on distributed data sets while preventing data leakage (Xie *et al.*, 2014; Konecný *et al.*, 2016; McMahan *et al.*, 2016; Vepakomma, Swedish, Raskar, Gupta & Dubey, 2018; Yang *et al.*, 2019). With this approach, the learning process involves collaboration from all the data owners without exposing their data to others. This can typically be achieved by sharing the architecture and parameters between the client and server during training, along with intermediate representations of the model that may include the gradients, activations and weight updates. Thus, the client downloads the model from the server and updates the weights based on its local data. Yet, a drawback of these strategies is their huge requirements for network bandwidth, memory and computational power, which strongly limits their scalability. More importantly, federated learning does not prevent, at test time, from having to send private data from the client to the server in a scenario such as ours where the server holds the model and processes the data. Also, while HE can be combined to federated learning, its communication protocol is cumbersome and imposes important communication overhead (Hardy *et al.*, 2017; Rouhani *et al.*, 2018).

3.2.3 Privacy preserving with adversarial learning:

The recent success of adversarial learning has led to the increased adoption of this technique for the protection of sensitive information, particularly in visual data. Xu et *al.* (Xu *et al.*, 2019) proposed to add carefully-designed noise to gradients during the learning procedure to train a differentially-private GAN in the context of image recognition. An unsupervised utility loss is employed for training in (Raval *et al.*, 2017), based on the assumption that removing private characteristics from an image while minimizing changes to the rest of the image yields encoded representations that can be used to learn a target task. However, since the encoding is performed independently of the task, it is potentially sub-optimal for this task. Other works (Pittaluga *et al.*, 2019; Wu *et al.*, 2018; Yang *et al.*, 2018; Roy & Boddeti, 2019) have leveraged adversarial training to jointly optimize privacy and utility objectives. In these works, the mapping functions for the adversarial and task-specific terms are standard classification models where the number of classes is fixed. In (Chen *et al.*, 2018a), a model which integrates a Variational Autoencoder (VAE) and a GAN is proposed to create an identity-invariant representation of face images. To explicitly control the features to be preserved, they include a discriminator which must predict the identity of the subject in a generated image. As the number of possible labels corresponds to the number of subjects to identify, this approach is not suitable for large-scale applications as the one considered in our work.

To alleviate the problem of a non-fixed number of classes, (Oleszkiewicz *et al.*, 2018) uses a Siamese architecture for the discriminator which predicts whether two encoded images come from the same subject or not. This paper focuses on biometrical data (e.g., fingerprint), which are dissimilar in nature from the medical images used in our method, and seeks a very different goal: finding the smallest possible transformation to an image which removes identity information and such that images can later be used by non-specific applications. In contrast, we 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 an image analysis (i.e., segmentation) task. This translates into important methodological differences. First, while the model in (Oleszkiewicz *et al.*, 2018) has a generator and a discriminator, our architecture is composed of three separate networks, i.e., an encoder, a Siamese discriminator and a segmentation network. Second, the final objective is different since we aim at maximizing the same-subject classification error, *as well as* optimizing a task-specific loss related to segmentation. In summary, both the structure and objectives between (Oleszkiewicz *et al.*, 2018) and our work are different.

The work in (Wang, Ding & Fu, 2018) tackles a task opposite to privacy-preserving image analysis, where faces in input images are rejuvenated while preserving, and not removing as in our work, information related to kinship. This is done by minimizing a discriminative sparse metric learning loss encouraging generated images for members of the same family to be nearby in a low-dimensional subspace. In (Xia & Ding, 2020), Xia et *al.* also employed adversarial training to develop a novel Generative cross-domain learning method via Structure-Preserving (GSP). The method attempts to transform target data into the source domain in order to take advantage of source supervision.

3.3 Methodology

3.3.1 Proposed system

As shown in Fig. 3.1, our system implements a zero-sum game involving three separate CNN networks. At the input of our system is a raw image $\mathbf{x} \in \mathbb{R}^{H \times W \times D}$ (in our case a 3D T1 magnetic resonance image (MRI)). During training, images come in pairs $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{X}^2, i \neq j$. Each image pair is associated to the corresponding ground-truth segmentation maps $(\mathbf{y}_i, \mathbf{y}_j)$ and binary target s_{ij} which equals 1 when $\mathbf{x}_i$ and $\mathbf{x}_j$ come from the same patient and 0 otherwise. As mentioned in Section 3.4.1, pairs of images from the same patient are not identical as they were acquired during different acquisition sessions, often months apart.

The first network of our system is an encoder network G parameterized by θ_G. The output of the encoder is a feature map $G(\mathbf{x}) \in \mathbb{R}^{H \times W \times D}$ which can be seen as an encoded version of the input image. While the encoder could return feature maps of any size, we chose maps with the same size as the input image $\mathbf{x}$ for the following important reasons. First, it allows preserving the information and spatial resolution of the input image. In contrast, using a compressed representation could lead to loss of details. This is why, for example, state-of-art segmentation networks employ skip connections that concatenate detailed features from downsampling layers with low-resolution features from upsampling layers (Dolz, Desrosiers & Ayed, 2018; Ronneberger *et al.*, 2015). Second, despite the high spatial resolution of encoding $G(\mathbf{x})$, it is still more compact than convolutional features of standard networks like VGG which have a lower spatial resolution but a larger number of channels (e.g., $14 \times 14 \times 512$ = 100,352 features at the last convolutional layer of VGG compared to $224 \times 224 \times 1$ = 50,176 features for our encoding, in the case of 224×224 images). Third, it enables a fair comparison of segmentation performance with the model using non-encoded images. Last, preserving the same shape as the input image allows processing the encoding image in sub-regions (i.e., 3D patches), which provides additional protection when these sub-regions are sent in a random order to the server for segmentation. While training the system, the encoder is fed with a pair of images $(\mathbf{x}_i, \mathbf{x}_j)$ and returns two encoded images $G(\mathbf{x}_i)$ and $G(\mathbf{x}_j)$. Here, $\mathbf{x}_i$ and $\mathbf{x}_j$ are processed individually and not concatenated together.

The second network is the Siamese discriminator network D with parameters θ_D, which is fed with a pair of encoded images. The goal of this network is to determine whether the two images come from the same patient or not. By fooling D (i.e., maximizing its loss), the encoder transforms the images and makes it difficult to identify the patient. Last, the third CNN is the segmentation network S with parameters θ_S, whose goal is to recover the correct segmentation map $\mathbf{y}$ given the encoded image $G(\mathbf{x})$. During training, both $G(\mathbf{x}_i)$ and $G(\mathbf{x}_j)$ are segmented. For this network, we used the widely-adopted U-Net (Ronneberger *et al.*, 2015), which is very effective at segmenting medical images.

3.3.2 Training losses

As in most adversarial models, our system is trained with two losses that steer the model in opposite directions. In our case, the training procedure involves a segmentation loss and an adversarial discriminator loss:

$$\begin{aligned}\min_{\theta_G,\theta_S}\max_{\theta_D}\ \mathcal{L}(\theta_G,\theta_S,\theta_D) \;=\; & \mathbb{E}_{\mathbf{x},\mathbf{y}\sim P(\mathbf{x},\mathbf{y})}\left[\ell_S(S(G(\mathbf{x})),\mathbf{y})\right] \\ & -\lambda\,\mathbb{E}_{\mathbf{x}_i,\mathbf{x}_j\sim P(\mathbf{x})}\Big[\ell_D\big(D(G(\mathbf{x}_i),G(\mathbf{x}_j)),s_{ij}\big)\Big]\end{aligned} \tag{3.1}$$

where ℓ_S is attached to the segmentation network, ℓ_D is attached to the discriminator, and $s_{ij} = \mathbf{1}_{\mathrm{id}(\mathbf{x}_i)=\mathrm{id}(\mathbf{x}_j)}$ is a binary indicator function indicating whether two encoded images come from the same patient or not.

Using $\widehat{\mathbf{y}} = S(G(\mathbf{x}))$ as shorthand notation for the predicted segmentation map, we employ the generalized Dice loss (Sudre, Li, Vercauteren, Ourselin & Cardoso, 2017a) to train the segmentation network, i.e.

$$\ell_S(\widehat{\mathbf{y}},\mathbf{y}) \;=\; 1 - \frac{2\sum_p y_p\,\widehat{y}_p}{\sum_p y_p \,+\, \sum_p \widehat{y}_p} \tag{3.2}$$

For the adversarial loss, we want the discriminator to differentiate subject identity in pairs of encoded images $G(\mathbf{x}_i)$, $G(\mathbf{x}_j)$. Here, we define discriminator's classification loss ℓ_D using binary cross entropy:

$$\ell_D(\widehat{s},s) \;=\; -s\log\widehat{s} \;-\; (1-s)\log(1-\widehat{s}) \tag{3.3}$$

Like most adversarial models, the parameters of our system cannot be updated all at once through a gradient step. Instead, we first update the encoder and segmentation parameters θ_D, θ_G by taking the following gradient descent step:

$$(\theta_S^{t+1}, \theta_G^{t+1}) \leftarrow (\theta_S^t, \theta_G^t) - \eta \nabla \widetilde{\mathcal{L}}(\theta_G^t, \theta_S^t) \tag{3.4}$$

The gradient is estimated using random batches of image pairs $\mathcal{B} \subset |\mathcal{X}| \bowtie |\mathcal{X}|$, as follows:

$$\nabla \widetilde{\mathcal{L}}(\theta_G, \theta_S) = \frac{1}{|\mathcal{B}|} \sum_{(i,j)\in\mathcal{B}} \nabla_{\theta_G, \theta_S} \Big[\ell_S(\widehat{\mathbf{y}}_i, \mathbf{y}_i) + \ell_S(\widehat{\mathbf{y}}_j, \mathbf{y}_j) - \lambda \ell_D(D(G(\mathbf{x}_i), G(\mathbf{x}_j)), s_{ij}) \Big] \tag{3.5}$$

We then update the discriminator parameters by taking a gradient ascent step

$$\theta_D^{t+1} \leftarrow \theta_D^t + \eta \nabla \widetilde{\mathcal{L}}(\theta_D^t) \tag{3.6}$$

with the batch gradient computed as

$$\nabla \widetilde{\mathcal{L}}(\theta_D) = -\frac{\lambda}{|\mathcal{B}|} \sum_{(i,j)\in\mathcal{B}} \nabla_{\theta_D} \ell_D(D(G(\mathbf{x}_i), G(\mathbf{x}_j)), s_{ij}) \tag{3.7}$$

Details of our training method are provided in Algo. 3.1.

3.3.3 Link to mutual information minimization

The idea of using adversarial learning to obfuscate identity is well-grounded on the principles of information theory. Hence, it can be shown that training a subject-ID classifier as discriminator in an adversarial learning model implicitly minimizes the mutual information between the encoded image and the corresponding subject ID. However, as mentioned before, this strategy is ill-suited to our problem since the number of classes (i.e. the number of subject ID) is not fixed and instead increases as new subjects are added to the system. This poses two major problems: 1) the output size of D varies over time, and 2) the classification task is hard to learn due to the large number of classes compared to the very low number of samples per classes (i.e., 1–4

Algorithm 3.1 Privacy-preserving network learning

```
Input: Images $\mathcal{X}$ and ground-truth masks $\mathcal{Y}$
Output: Network parameters $\theta_G, \theta_D, \theta_S$
/* Initialization */
Initialize network parameters $\theta_G, \theta_D, \theta_S$;
/* Main loop */
for epoch = $1, \ldots, E_{\max}$ do
    for iter = $1, \ldots, T_{\max}$ do
        Randomly select batch $\mathcal{B} \subset |\mathcal{X} \bowtie \mathcal{X}|$;
        Update encoder and segmentation network parameters $(\theta_S, \theta_G)$ using Eq. (3.4)
          and (3.5);
        Update discriminator parameters $(\theta_D)$ using Eq. (3.6) and (3.7);
    end for
end for
return $\theta_G, \theta_D, \theta_S$;
```

images per subject). This motivates our approach, based on a Siamese discriminator, where the identification task is to determine if two encoded images are from the same subject. This approach can naturally incorporate new subjects/classes over time and is easier to learn since it corresponds to a binary classification problem and training samples are more abundant (i.e., image pairs instead of images).

As a theoretical contribution of this work, we show that our proposed privacy-preserving learning approach based on a Siamese discriminator also relates to mutual information minimization. This is done in the following theorem.

Theorem 3.3.1. *Let* $\mathbf{x}$, $\mathbf{x}'$ *be two images, and* $G(\mathbf{x})$, $G(\mathbf{x}')$ *be their encoded version obtained by the generator* G. *Denoting as* $z = \mathrm{id}(\mathbf{x})$ *the subject ID of image* $\mathbf{x}$ *and* $s = \mathbf{1}_{z=z'}$ *the random variable indicating whether images* $\mathbf{x}$ *and* $\mathbf{x}'$ *are from the same subject, optimizing the problem defined in Eq. (4.2) corresponds to minimizing mutual information* $I(G(\mathbf{x}), G(\mathbf{x}');\, s)$ *between encoded images* $G(\mathbf{x})$, $G(\mathbf{x}')$ *and random variable* s.

Proof. We proceed by defining mutual information and then bounding it using a variational approach. The mutual information $I(G(\mathbf{x}), G(\mathbf{x}');\ s)$ between encoded images $G(\mathbf{x})$, $G(\mathbf{x}')$ and random variable s can be defined as

$$\begin{aligned} I(G(\mathbf{x}), G(\mathbf{x}'); s) &= H(s) - H(s \mid G(\mathbf{x}), G(\mathbf{x}')) && (3.8)\\ &= H(s) + \mathbb{E}_{\mathbf{x},\, \mathbf{x}' \sim P(\mathbf{x},\mathbf{x}')}\Big[\mathbb{E}_{s' \sim P(s \mid G(\mathbf{x}),G(\mathbf{x}'))}[\log P(s' \mid G(\mathbf{x}), G(\mathbf{x}'))]\Big] && (3.9)\\ &= H(s) + \mathbb{E}_{s \sim P(s),\, \mathbf{x},\, \mathbf{x}' \sim P(\mathbf{x},\mathbf{x}' \mid s)}\Big[\mathbb{E}_{s' \sim P(s \mid G(\mathbf{x}),G(\mathbf{x}'))}[\log P(s' \mid G(\mathbf{x}), G(\mathbf{x}'))]\Big] && (3.10) \end{aligned}$$

with $H(x)$ being the Shannon entropy of a random variable x and using the fact that G is a deterministic function. To deal with the intractable computation of $P(s \mid G(\mathbf{x}), G(\mathbf{x}'))$, we derive a lower bound using variational distribution $Q(s \mid G(\mathbf{x}), G(\mathbf{x}'))$:

$$\begin{aligned} (3.10) &= H(s) + \mathbb{E}_{s \sim P(s),\, \mathbf{x},\, \mathbf{x}' \sim P(\mathbf{x},\mathbf{x}' \mid s)}\Big[\mathbb{E}_{s' \sim P(s \mid G(\mathbf{x}),G(\mathbf{x}'))}[\\ &\qquad \underbrace{D_{\mathrm{KL}}(P(s' \mid G(\mathbf{x}), G(\mathbf{x}')) \parallel Q(s' \mid G(\mathbf{x}), G(\mathbf{x}')))}_{\geq 0} + \log Q(s' \mid G(\mathbf{x}), G(\mathbf{x}'))]\Big] && (3.11)\\ &\geq H(s) + \mathbb{E}_{s \sim P(s),\, \mathbf{x},\, \mathbf{x}' \sim P(\mathbf{x},\mathbf{x}' \mid s)}\Big[\mathbb{E}_{s' \sim P(s \mid G(\mathbf{x}),G(\mathbf{x}'))}[\log Q(s' \mid G(\mathbf{x}), G(\mathbf{x}'))]\Big] && (3.12) \end{aligned}$$

Next, we use the fact that, for random variables X, Y and function $f(x, y)$, $\mathbb{E}_{x \sim X,\, y \sim Y \mid x}[f(x, y)] = \mathbb{E}_{x \sim X,\, y \sim Y \mid x,\, x' \sim X \mid y}[f(x', y)]$ (see Appendix A.1 of (Chen *et al.*, 2016b) for proof) to get:

$$\begin{aligned} (3.12) &= H(s) + \mathbb{E}_{s \sim P(s),\, \mathbf{x},\, \mathbf{x}' \sim P(\mathbf{x},\mathbf{x}' \mid s)}[\log Q(s \mid G(\mathbf{x}), G(\mathbf{x}'))] && (3.13)\\ &= H(s) + \mathbb{E}_{s \sim P(s),\, z,z' \sim P(z,z' \mid s),\, \mathbf{x} \sim P(\mathbf{x} \mid z),\, \mathbf{x}' \sim P(\mathbf{x}' \mid z')}[\log Q(s \mid G(\mathbf{x}), G(\mathbf{x}'))] && (3.14) \end{aligned}$$

Last, we equate (3.14) with Eq. (4.2) using the following: 1) $H(s)$ can be treated as a constant, and 2) the variational distribution Q is modeled using our Siamese discriminator D, and 3) $\log Q(s \mid G(\mathbf{x}), G(\mathbf{x}'))$ is equal to minus the cross-entropy loss ℓ_D of Eq. (3.3). Maximizing the lower bound in (3.14) thus increases its tightness to $I(G(\mathbf{x}), G(\mathbf{x}');\ s)$, the two becoming equal when $P(\cdot \mid G(\mathbf{x}), G(\mathbf{x}')) = Q(\cdot \mid G(\mathbf{x}), G(\mathbf{x}'))$. Consequently, optimizing the loss function of (4.2) minimizes a maximally-tight bound to mutual information. □

By minimizing mutual information, which is a symmetric measure of co-dependence between two variables, we ensure that subject identity cannot be established by matching an encoded image with those previously seen in the system. Moreover, a powerful property of mutual information is that it is invariant to any monotone and uniquely invertible transformation of the variables (Kraskov, Stögbauer & Grassberger, 2004). Consequently, it provides a certain robustness to small transformations applied to the (encoded) images, such as translation and rotation. Similarly, it avoids the trivial and non-obfuscating solution where the discriminator is forced to systematically flip its predictions to $s_{\text{flip}} = 1 - s$, since this does not change the mutual information, i.e. $I(G(\mathbf{x}), G(\mathbf{x}');\ s) = I(G(\mathbf{x}), G(\mathbf{x}');\ s_{flip})$.

3.3.4 Implementation details

In this study, we used a U-Net architecture (Ronneberger *et al.*, 2015) but with 3D convolution kernels both for the encoder and the segmentation network. The discriminator is a Siamese network as in (Koch, Zemel & Salakhutdinov, 2015). We used a DenseNet architecture (Huang, Liu & Weinberger, 2017) with 3D convolution kernels for the CNN backbone. The CNN Siamese backbone (i.e. the left-most CNN inside the discriminator box in Fig. 3.1) is used to extract the features of input images. The last layer of the discriminator contains two fully-connected layers to predict if two encoded images are from the same patient.

The system was implemented with Pytorch. We used the Adam optimizer with a learning rate of 10^{-4} for the whole training process. The PC used for training is an Intel(R) Core(TM) i7-6700K 4.0GHz CPU, equipped with a NVIDIA GeForce GTX 1080Ti GPU with 12 GB of memory. Training our framework takes roughly 30 minutes per epoch, and around 2 days for the fully-trained system.

Since our networks employ 3D convolutions, and due to the large size of MRI volumes, dense training cannot be applied to the whole volume. Instead, volumes are split into smaller patches of size $64 \times 64 \times 64$, which allows dense training in our hardware setting. During training, the patches are randomly cropped from the MRI volume. In testing, the volume to segment is instead divided in evenly-spaced 3D patches, which are then segmented separately. Individual patch outputs are then combined to obtain full-size segmentation maps.

An important advantage of segmenting patches separately is that they can be sent in a random order to the server once the image has been encoded on the client side. This makes obtaining the identity of a subject even more challenging, since a potential attacker must either reorder patches to recover the full-size encoded image and segmentation map, or match small-size patches with previously seen ones. In Section 3.4.3.7, we illustrate this advantage in our experiments by performing a subject-ID retrieval analysis on output patches instead of full-size segmentation maps.

To help the learning process in early training stages, the encoder is pre-trained using an auto-encoder loss. Hence, when the real training starts, the encoder generates encoded images which are almost identical to input ones. Likewise, both the segmentation and discriminator networks were pre-trained on the original images from the the PPMI dataset.

3.4 Experimental results

3.4.1 Datasets

3.4.1.1 PPMI Dataset:

We experiment on brain tissue segmentation of 5 classes: white matter (WM), gray matter (GM), nuclei, internal cerebrospinal fluid (CSF int.) and external cerebrospinal fluid (CSF ext.). We used the T1 images of the publicly-available Parkinson's Progression Marker Initiative (PPMI) dataset (Marek *et al.*, 2011). We took images from 350 subjects, most of which with a recently diagnosed Parkinson disease. Each subject underwent one or two baseline acquisitions and one or two acquisitions 12 months later for a total of 773 images. PPMI MR images were acquired on Siemens Tim Trio and Siemens Verio 3 Tesla machines from 32 different sites. The images have been registered onto a common MNI space and resized to $144 \times 192 \times 160$ with a 1 mm^3 resolution. More information on the MRI acquisition and processing can be found online: www.ppmi-info.org.

The dataset was divided into a training and a testing set as shown in Table 3.1. We split the data in a stratified manner so that images from the same subject are not included in both the training and testing sets. In order to keep a good balance between the pairs of images, during training and testing, we randomly sampled an equal number of negative and positive samples. Due to the burden of manually annotating volumetric images, we resort to Freesurfer to obtain the segmentation ground-truth, similar to recently-published approaches on large-scale datasets (Dolz *et al.*, 2018; Roy, Conjeti, Navab, Wachinger *et al.*, 2019). We use cross-validation to measure the performance of our approach and properly set the hyper-parameters. The training set was randomly divided into 5 stratified subsets, each containing around 54 subjects. We then trained our system for 5 rounds, each time using a different group of 4 subsets for training and

Table 3.1 PPMI data used for training and testing our method

	Training	Testing	Total
Num. of subjects	269	81	350
Num. of images	592	181	773
Num. of positive pairs	509	148	657

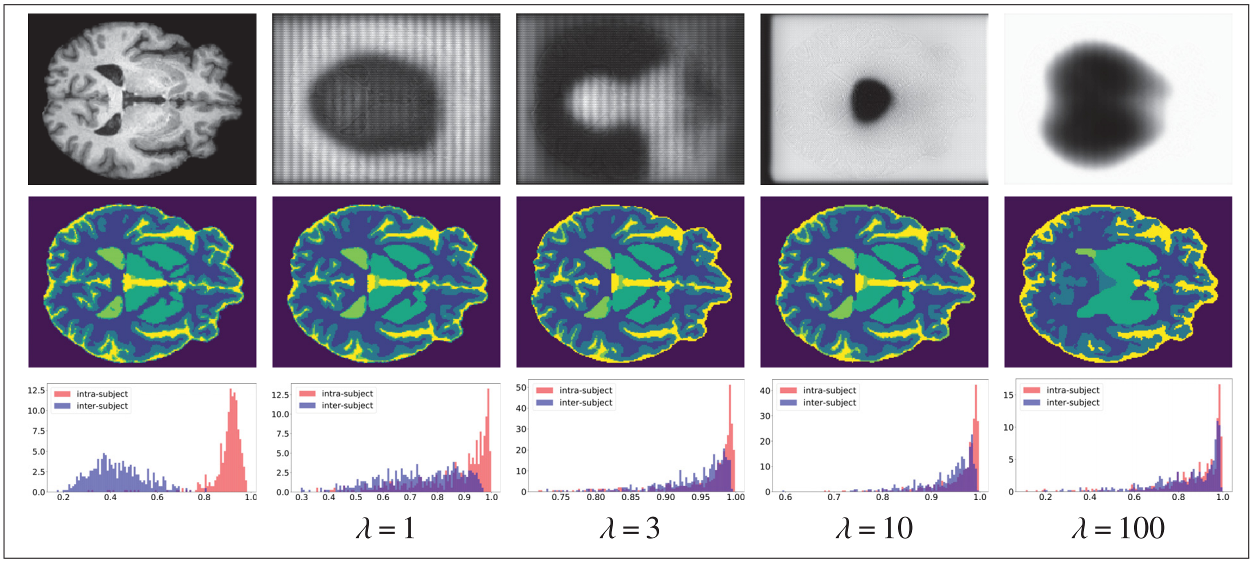


Figure 3.2 Impact of discriminator loss (λ). [**First column**] (Top row): input MRI image $\mathbf{x}$, (Second row): ground truth segmentation map $\mathbf{y}$, (Third row): distribution of inter-subject and intra-subject MS-SSIM score on the PPMI dataset. [**Remaining columns**], (Top row): encoded image $G(\mathbf{x})$, (Second row): predicted segmentation $\widehat{\mathbf{y}}$ and (Third row): distribution of MS-SSIM values between encoded images $G(\mathbf{x}_i)$ and $G(\mathbf{x}_j)$

the remaining one for validation. After validation, we retrained the system on the entire training set and reported results from the independent test set.

3.4.1.2 MRBrainS Dataset:

To further validate the proposed method and investigate its generalization ability, we also tested it on segmenting MRI scans from the MRBrainS 2013 challenge dataset (Mendrik *et al.*, 2015). These images were acquired on a 3.0T Philips Achieva MR scanner and come with expert-annotated segmentation masks including three classes: WM, GM and CSF. We employed a single modality (i.e., MR-T1) in our experiments. Bias correction was performed as a pre-processing step. Original images had a resolution of $0.96 \times 0.96 \times 3\,\text{mm}^3$ and were registered onto the MNI space using ANTs (Avants *et al.*, 2011).

3.4.2 Evaluation metrics

To gauge the performance of our system, we use the classification accuracy for measuring the discriminator's ability to identify images from the same person, and employ the Dice score for evaluating segmentation results. We also use the multiscale structural-similarity (MS-SSIM) score to measure image-to-image distance as a proxy of perceived image quality (Wang, Simoncelli & Bovik, 2003).

Since the number of inter-subject samples is much larger than the number of inter-subject samples, we balanced the dataset by using every intra-subject examples but randomly selected inter-subject examples. We tested the discriminator on the dataset for 5 times and reported the mean ± std. dev. accuracy.

Last, to determine if an encoded image can be used to recover the subject in a top-k retrieval setting (Kumar *et al.*, 2018), we use mean average precision (mAP). Given an image i, we rank other images in the dataset by their similarity to image i. The similarity between two images is the cosine similarity between the feature vectors of each image extracted by the CNN backbone of the Siamese discriminator. Let T_i be the set of images of the same subject as image i, and denote as S_i^k the set containing the k images most similar to i (i.e., the k nearest neighbors of i). For a given value of k, we evaluate the retrieval performance using the measure of top-k

precision (also known as precision-at-k):

$$(\text{precision@}k)_i = \frac{T_i \cap S_i^k}{k} \tag{3.15}$$

Considering each encoded test image $G(\mathbf{x}_i)$ as a separate retrieval task where one must find other encoded images from the same person, the average precision (AP) for $G(\mathbf{x}_i)$ is given by

$$\text{AP}_i \;=\; \frac{1}{\sum_{j \neq i} s_{ij}} \sum_{k=1}^{|\mathcal{X}|} (\text{precision@}k)_i \cdot s_{ik} \tag{3.16}$$

where precision@k is the precision at cut-off k, i.e. the ratio of k encoded images most similar to $G(\mathbf{x}_i)$ which belong to the same person. mAP is then the mean of AP values computed over all test examples.

3.4.3 Results

3.4.3.1 Results on non-encoded images

We first processed the dataset without the adversarial component, i.e, by independently training the segmentation and the discriminator networks without the encoder. We call this setting *non-encoded* in our results. In the first row of Table 3.2, we see that the discriminator obtains a testing accuracy of 95.3%. This underlines how easy it is for a neural network to recognize a patient based on the content of a brain MRI. More surprising is the 97% classification accuracy that we obtain by simply thresholding the image-to-image MS-SSIM score. This can be explained by the inter-subject and intra-subject MS-SSIM distribution plots shown in the third row of the first column of Fig. 3.2. As can be seen, when considering non-encoded images, the intra-subject MS-SSIM scores (red curve) are significantly larger than that of the inter-subjects (blue curve).

Table 3.2 Intra-subject and inter-subject prediction accuracy on test examples obtained by thresholding MS-SSIM scores, using the adversarial discriminator (D_{adv}), or training a separate discriminator on the encoded image (D_{new}). The mAP column is the mean average precision of a top-k retrieval analysis using the Siamese discriminator's embedding as representation. Results are reported for non-encoded images or encoded images for different λ values

		Accuracy			mAP
		MS-SSIM	D_{adv}	D_{new}	
Non-encoded		0.963 ± 0.010	—	0.949 ± 0.021	0.850
Encoded	$\lambda = 1$	0.561 ± 0.013	0.545 ± 0.031	0.608 ± 0.028	0.189
	$\lambda = 3$	0.543 ± 0.016	0.536 ± 0.027	0.597 ± 0.026	0.152
	$\lambda = 10$	0.514 ± 0.015	0.525 ± 0.029	0.575 ± 0.021	0.141
	$\lambda = 100$	0.509 ± 0.012	0.518 ± 0.032	0.513 ± 0.024	0.087

This again illustrates the ease of recognizing the identity of a person based on the content of a medical image.

The PPMI segmentation Dice scores on non-encoded images for the five brain regions are in the first row of Table 3.3. We also report the overall Dice computed as the mean of Dice scores in all regions, weighted by the regions' size. These results correspond roughly to those obtained in recent publications for the same architecture (Dolz *et al.*, 2019). Note that the nuclei and the internal CSF have a lower Dice due to the smaller size of these regions.

Table 3.3 Segmentation Dice score on the PPMI test set for different values of λ. Non-enc refers to the model trained with non-encoded images

	GM	WM	Nuclei	CSF int.	CSF ext.	Overall
Non-enc	0.941	0.853	0.657	0.665	0.825	0.848
$\lambda = 1$	0.925	0.824	0.580	0.598	0.752	0.812
$\lambda = 3$	0.899	0.793	0.549	0.550	0.693	0.778
$\lambda = 10$	0.881	0.796	0.555	0.531	0.685	0.771
$\lambda = 100$	0.847	0.692	0.454	0.405	0.513	0.684

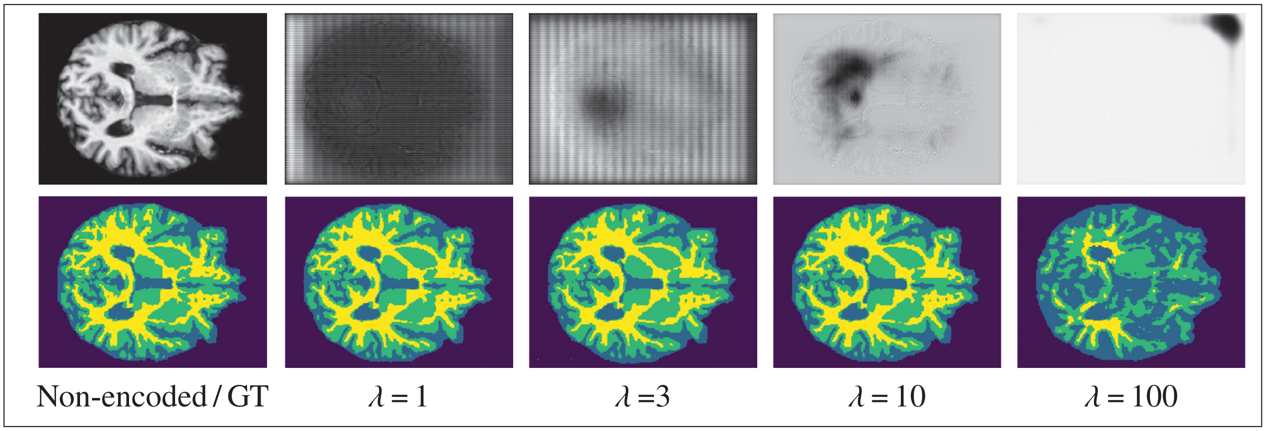


Figure 3.3 Test results on the MRBrainS dataset. [**First column**] (Top row): input MRI image $\mathbf{x}$, (Bottom row): ground truth segmentation map $\mathbf{y}$. [**Remaining columns**], (Top row): encoded image $G(\mathbf{x})$, (Bottom row): predicted segmentation $\widehat{\mathbf{y}}$ with re-trained segmentation networks on MRBrainS

3.4.3.2 Adversarial results

We next report results of our adversarial approach obtained with different values of parameter λ, which controls the trade-off between segmentation accuracy and identity obfuscation. The first row of Fig. 3.2 shows encoded images $G(\mathbf{x})$ with the corresponding raw input MRI $\mathbf{x}$. As can be seen, the larger the λ value is, the more distorted the encoded image gets. Nonetheless, except for extreme cases (e.g., $\lambda = 100$) the encoded images contain enough information for the segmentation network to recover a good segmentation map (c.f., the second row of Fig. 3.2). The obfuscating power of our method is also illustrated by the MS-SSIM plots (c.f., third row of Fig. 3.2). As λ increases, the distribution of inter-subject MS-SSIM between encoded images $G(\mathbf{x}_i)$ and $G(\mathbf{x}_j)$ becomes more and more similar to that of intra-subjects.

The encoder's ability to obfuscate identity is evaluated quantitatively in Table 3.2. Four different techniques are used to measure this property. First, based on the observation that the distribution of MS-SSIM values differs between images from the same patient and images from different patients (c.f., last row of Fig. 3.2), we compute the accuracy obtained by the best possible tresholding of MS-SSIM values (i.e., values below or equal to the threshold are predicted as same-patient images, and those above as different-patient images). Second, we report the classification accuracy of the discriminator used for training the encoder, denoted as D_{adv} in Table 3.2. Third, since the encoder was trained to fool D_{adv}, we also trained a new ResNet discriminator (D_{new}) as in (He *et al.*, 2016) on the fixed encoded images to measure how good the encoder is with respect to an independent network that was not involved in training our system. Last, to assess whether an encoded image can be used to find the corresponding subject with a retrieval approach, we considered the embedding of Siamese discriminator D_{new} as representation of each encoded test image and used Euclidean distance to find most similar encoded images. We employ mAP to measure retrieval performance.

Results in Table 3.2 show the same trend for all four obfuscation measures. When images are not encoded, identifying the subject's identity either by comparing two images or using a retrieval-based approach is fairly easy. However, this becomes much harder for encoded images, with accuracy and mAP rates dropping as λ increases. Moreover, as shown in column D_{new}, employing a discriminator trained independently from the encoder does not help re-identify the subject's ID. This demonstrates the robustness of our method to classification approaches.

The segmentation performance obtained with different privacy-segmentation trade-off, defined by the λ parameter, is given in Table 3.3. As expected, the Dice score degrades when increasing λ values, since a greater importance is then given to identity obfuscation compared to segmentation. Nevertheless, the segmentation performance on encoded images is still sufficient for many medical applications, especially when using $\lambda = 1$ or $\lambda = 3$. Although the definition of suitable performance is application-dependent, some authors have reported DSC values of 70% (Zijdenbos, Dawant, Margolin & Palmer, 1994; Zou *et al.*, 2004; Gambacorta *et al.*, 2013; Anders *et al.*, 2011) or 80% (Mattiucci *et al.*, 2013) as threshold for clinically-acceptable segmentations. For $\lambda = 1$, the overall difference in DSC compared to segmentation of non-encoded images is only 3.6%, an impressive result considering that subject identity information is largely removed in encoded images for this λ value.

Anatomical information capturing the shape of brain regions and cortical folds (i.e., sulci) can be predictive of both subject identity and segmentation contours. This can be observed in Fig. 3.2 (first row), where small anatomical details are visible in the encoded images, especially for $\lambda = 1$. To obfuscate identity, the encoder must therefore produce strong noise and artifacts that dominate this morphological information. The impact of this noise in encoded images can be seen in the last row of Fig. 3.2, in which the histogram of MS-SSIM scores between different-subject images (i.e., inter-subject) is pushed towards the one for same-subject images (i.e., intra-subject).

3.4.3.3 Generalization to new dataset

In previous experiments, we considered the scenario where the encoder and segmentation network are trained once with some available data, and then clients send encoded images to the server for segmentation. However, this approach may fail when trying to process images different from those seen in training, for instance, coming from another hospital or acquired with different parameters. In this section, we show that the privacy-preserving encoder learned with a given dataset can be used to encode images from another dataset, and that these encoded images are useful to update the segmentation network.

To test this configuration, we consider the same encoder as before, which was trained using the longitudinal data from the PPMI dataset, and use it as an identity obfuscation module for clients with other data. To simulate this other data source, we used images from the MRBrainS dataset which were acquired with a different acquisition protocol than PPMI and have three labels instead of five, i.e., WM, GM, and CSF. We first tested on MRBrainS images our model pre-trained with PPMI data. In order to match the three-class ground-truth, we merged the CSF int. and CSF ext. outputs into a single CSF class, and the GM and nuclei outputs into a single GM class. Results for different λ values are shown at the top of Table 3.4. As expected, these results are slightly lower than those on PPMI (see Table 3.3). Interestingly, we observe a small improvement in the overall Dice when using encoded images, compared to the baseline network trained with non-encoded PPMI images. This suggests that the system trained with adversarial loss generalizes better to new data than the baseline segmentation network, despite the heavy distortion of encoded images. This improved generalization of our method is due to optimizing the encoder for both segmentation accuracy and identity obfuscation, which possibly removes site-related variability (e.g., intensity distribution, tissue contrast, etc.) from encoded images.

As can be seen at the bottom of Table 3.4, segmentation accuracy improves when the segmentation network is retrained on MRBrainS data following a distributed learning schedule. This shows that the segmentation network of our system can be updated, even after being deployed onto a cloud server. Segmentation maps as well as encoded MRBrainS images are given in Fig 3.3. However, like other deep learning methods, when the discrepancy between domains becomes too large (i.e., different image modalities), the system would need to be retrained end-to-end.

3.4.3.4 Robustness analysis

To make sure that our system does not work only on high-quality images such as those of PPMI, we performed a robustness analysis where we trained our method on the original PPMI dataset (with $\lambda = 1$) and tested it on noisy versions of the PPMI test images or on images with a lower $2 \times 2 \times 2$ cm^3 resolution. Results are provided in Fig. 3.4 and Table 3.5. As can be seen, the encoding and segmentation is not much affected by noise. The segmentation accuracy is close to the one obtained on the original PPMI test set, with overall Dice scores around 80% for noisy images with an SNR of 15db or 10db. Reducing image resolution seems to induce more significant changes in the encoding, however the segmentation network appears robust to these changes.

Table 3.4 Segmentation Dice score on the MRBrainS dataset for different values of λ. (Top) the CNNs have not been retrained while (Bottom) the segmentation network has been retrained following a distributed learning approach

		GM	WM	CSF	Overall
No retrain	Non-enc	0.742	0.805	0.778	0.783
	$\lambda = 1$	0.768	0.822	0.804	0.796
	$\lambda = 3$	0.767	0.852	0.798	0.804
	$\lambda = 10$	0.757	0.798	0.768	0.772
	$\lambda = 100$	0.499	0.464	0.648	0.537
Retrain	Non-enc	0.832	0.866	0.840	0.845
	$\lambda = 1$	0.819	0.827	0.823	0.821
	$\lambda = 3$	0.794	0.807	0.831	0.814
	$\lambda = 10$	0.780	0.747	0.797	0.790
	$\lambda = 100$	0.605	0.360	0.572	0.586

Table 3.5 Segmentation Dice score on PPMI dataset with different levels of Rician noise (measured in dB) and low resolution setting

	GM	WM	Nuclei	CSF int.	CSF ext.	Overall
Noise (15 dB)	0.921	0.818	0.572	0.585	0.743	0.808
Noise (10 dB)	0.917	0.804	0.566	0.582	0.696	0.797
Noise (5 dB)	0.821	0.706	0.514	0.472	0.331	0.669
Low res.	0.881	0.781	0.552	0.516	0.683	0.759

The robustness of our model to noise and resolution can be explained as follows. Due to the adversarial optimization between the encoder and discriminator, the segmentation network sees a wide variety of distortion patterns in encoded images as the encoder tries to fool the discriminator. By forcing the segmentation network to produce the same output for these different distorted inputs, we regularize training and make this network more robust to noise and resolution. This principle is at the core of powerful regularization techniques for semi-supervised learning, such as Virtual Adversarial Training (VAT) (Miyato, Maeda, Koyama & Ishii, 2018).

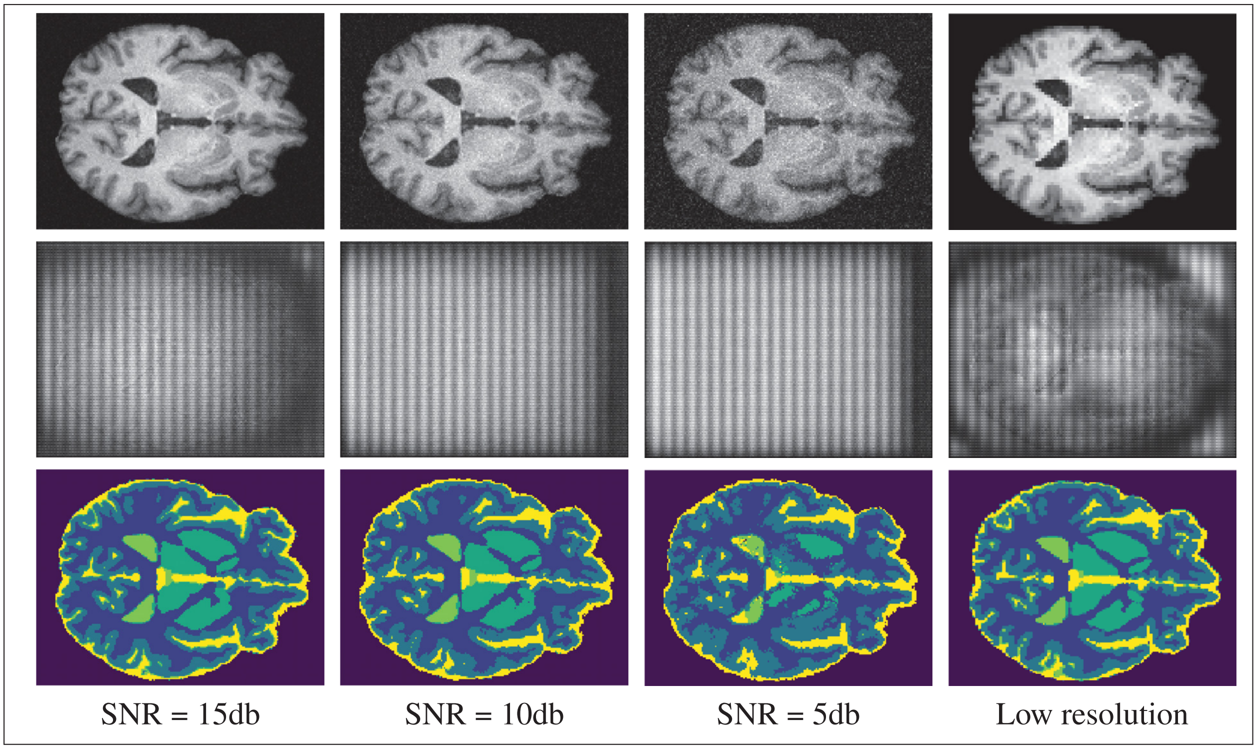


Figure 3.4 Segmentation with different noise level and a lower resolution setting. (Top row): Degraded images, (Middle row): Encoded images, (Bottom row): Segmentation Results

3.4.3.5 Dimension of encoded images

By default, our encoder outputs a one-channel feature map (c.f., image $G(x_i)$ in Fig. 3.1). Our motivation for using a single channel is that the encoder should preserve the amount of information while transforming the input. To validate this hypothesis, we repeated the same experiment but with a larger number of channels for the encoded images. The intra-subject and inter-subject prediction results are reported in Table 3.6 and the segmentation Dice scores are in Table 3.8. These results show that, even though the segmentation performance slightly increases when encoding images with multiple channels, privacy preservation is compromised, as shown by the significant increase in MS-SSIM scores, discriminator accuracy (D_{adv}) and mAP.

Table 3.6 Intra-subject and inter-subject prediction accuracy on test examples using different numbers of channels in encoded images

Num. of channels	MS-SSIM	D_{adv}	mAP
1	0.564	0.520	0.189
2	0.658	0.557	0.230
3	0.642	0.541	0.216

3.4.3.6 Advantage of a Siamese discriminator

To assess the benefit of using a Siamese discriminator in our model, we replaced it by a multi-class classifier with the same pre-trained DenseNet CNN backbone as the Siamese network. In this new model, each subject is given a different class ID and the multi-class discriminator has to predict the class ID of encoded images. Unlike for the Siamese network, training this classifier requires to have images from the same subject in both the training set tand validation set. Hence, we divided the original dataset into a new training set and validation set, and used this new split to retrain the whole system. Multi-class cross-entropy was used as loss function for the discriminator. Since the number of classes is not known in advance (i.e., new subjects can be added to the system after training) and the number of samples per class is very limited (1–4 images per subject), the classifier's output cannot be used directly to evaluate its ability to identify new subjects in testing. Instead, we considered the same out-of-sample strategy as with the Siamese discriminator, and used the features obtained from the last convolutional layer as an embedding for a nearest-neighbor retrieval analysis.

Results in Table 3.9 show that the multi-class classifier is worse than the Siamese network at recovering identity in non-encoded images, i.e., 0.442 vs 0.850 in terms of mAP. However, when used to train the entire network, the patient identity is more easily recovered when the images are encoded with the multi-class classifier (0.360) than by the Siamese network (0.189). This motivates the use of a Siamese network as discriminator in our model.

Table 3.7 Top-k retrieval analysis results (mAP) using segmentation maps

	$\lambda = 1$	$\lambda = 3$	$\lambda = 10$	$\lambda = 100$
Full-size segmentation maps	0.826	0.811	0.804	0.592
$64 \times 64 \times 64$ patches	0.716	0.683	0.697	0.398
$32 \times 32 \times 32$ patches	0.632	0.624	0.583	0.240

3.4.3.7 Top-k retrieval analysis on segmentation maps

The goal of our privacy-preserving method is to segment medical images on a server without having to provide sensitive information about the subject. For example, sending non-encoded images could reveal the gender and age of a subject, or if this subject suffers from a neurological disease/disorder like Alzheimer's. Our method achieves this by distorting the input image with noise patterns so that 1) visual interpretation is nearly impossible and 2) identity cannot be recovered easily. However, because we encode the input, but not the output segmentation, an attacker still has access to some information that can help determine the subject's identity and condition.

In the next experiment, we evaluate whether subject identity can be recovered from the segmentation network's output, using a top-k retrieval analysis similar to the one presented in Section 3.4.3.2. For this analysis, we suppose that an attacker compares the segmentation map of a test image against those of training images, and identifies the subject as the one corresponding to the most similar image. Since a direct pixel-to-pixel comparison is highly sensitive to transformations such as translation, scaling and rotation, as in the previous retrieval analyses, we instead use the representation of a Siamese discriminator as feature vector for matching.

Results of are reported in Table 3.7. The first row gives the mAP when using whole-image segmentation maps, for different values of λ. Parameter λ affects the retrieval indirectly since a higher value leads to a less accurate segmentation and, thus, a noisy representation for matching. For a small $\lambda = 1$, we get an mAP of 0.826 similar to the one of 0.850 obtained for non-encoded images (c.f., Table 3.2). This shows that the geometry of segmented brain structures is informative of subject identity. However, increasing λ to the high value of 100 leads to an important drop in mAP to 0.592 caused by the poor segmentation resulting from this setting.

As mentioned in Section 3.3.4, an important advantage of our method is that encoded images can be cut in small 3D patches which are sent to the server in a random order for processing. This is possible because the segmentation network requires to segment an image one patch at a time. Once the client receives the segmentation output for each patch from the server, it can recover the full-size segmentation map by assembling patches following the same random order.

We assume that reassembling the segmentation maps from randomly-permuted patches is challenging, and that potential attackers instead try to identify the subject's identity by matching patches against a database of previously seen patches. Based on this assumption, we repeat the same top-k retrieval analysis as before, except we now match the Siamese network representation of test patches with those of training patches. The second and third rows of Table 3.2 give retrieval mAP when employing patches of size $64 \times 64 \times 64$ and $32 \times 32 \times 32$, respectively. We observe that retrieval rates substantially degrade when sending encoded patches to the server, with respective mAP drops of 0.110 and 0.194 for patches of size $64 \times 64 \times 64$ and $32 \times 32 \times 32$, when using $\lambda = 1$. This indicates that the limited spatial context of patches, compared to whole segmentation maps, renders more difficult the identification of subjects. While mAP values stay relatively similar for $\lambda = 1$–10, they sharply decrease for $\lambda = 100$ due to the poor segmentation obtained with this setting.

Although the non-encrypted segmentation map can be seen as a security weakness, the results in able 3.7 show that the retrieval accuracy is greatly reduced when the client sends shuffled

Table 3.8 Segmentation Dice score on encoded images with different number of channels

Nb channels	GM	WM	Nuclei	CSF int.	CSF ext.	Overall
1	0.925	0.824	0.580	0.598	0.752	0.812
2	0.935	0.851	0.621	0.633	0.786	0.824
3	0.932	0.848	0.617	0.637	0.802	0.829

Table 3.9 Image-retrieval performance (mAP) of the multi-class classifier and Siamese discriminator

	Multi-class classifier			Siamese discr.	
	Train.	Valid.	Test	Train.	Test
Non-encoded	0.712	0.478	0.442	0.894	0.850
Encoded	0.404	0.324	0.360	0.153	0.189

patches for segmentation. Moreover, our method not only removes patient identity from the image, but also scrubs out most of its content. Hence, even if the identity of a patient was to be recovered via the inspection of the segmentation map, all an attacker would have access to are encoded image patches and not the original non-encoded MRI.

3.4.3.8 Runtime analysis

In terms of runtime, an average of 0.08 seconds is required to encode an image on a NVIDIA GTX 1080Ti, whereas the entire segmentation process requires around 0.1 seconds per 3D MRI image. This runtime is negligible compared to the 8 to 12 hours required by Freesurfer.

3.5 Discussion and conclusion

We presented a novel framework which integrates an encoder, a segmentation CNN and a Siamese network to preserve the privacy of medical imaging data. Experimental results on two independent datasets showed that the proposed method can preserve the identity of a patient while maintaining the performance on the target task. While this is an interesting application *per se*, it opens the door to appealing potential uses. For example, this approach can be integrated in a continual learning scenario trained on a decentralized dataset, where images have to be shared across institutions but privacy needs to be preserved. From a clinical perspective, obfuscating visual data in addition to current anonymization techniques may foster multi-centre collaborations, resulting in larger datasets as well as more complete and heterogeneous clinical studies.

Additionally, we have shown that the proposed privacy-preserving model generalizes well to novel datasets, unlike similar works (Chen *et al.*, 2018a) which cannot generalize to encoded images of subjects not seen in training. This facilitates the scalability of our approach to new datasets or tasks. As preliminary step towards preventing privacy leakage in medical imaging data, this study has however some limitations. For example, the domain shift between employed datasets is not significantly large, since both include MRI images of adult brains (even though the acquisition protocols and parameters across scanners differ). Although similar domain shift has resulted in a performance degradation in segmentation networks (Dolz *et al.*, 2018), results demonstrate the good generability of the proposed method in these cases. Future investigations will explore the generalization capabilities of the trained encoder on datasets where the domain shift is larger, for example, between infant and adult brains or even between different image modalities such as MRI and CT.

Finally, one disadvantage of our method is that unlike classification outputs, segmentation maps could still contain patient identifiable information. As discussed in the section VI.C.7, because our method does not encode the output segmentation map, if the attacker gains access to the full-size segmentation maps, the level of privacy leaking risk is on the similar level of directly sending non-encoded images. In order to minimize this risk, it is crucial to perform the segmentation in patches sent in random order so that the attacker does not gain access to the full-size segmentation maps. Because this random shuffling is not a complete solution, thus encoding the output segmentation map will be our next target.

CHAPTER 4

PRIVACY PRESERVING FOR MEDICAL IMAGE ANALYSIS VIA NON-LINEAR DEFORMATION PROXY

Ngoc Bach, Kim[1] , Jose Dolz[1] , Christian Desrosiers[1] Pierre-Marc Jodoin[2]

[1] Department of Software and IT Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Department of Computer Science, Université de Sherbrooke,
2500 Boul de L'université, Sherbrooke, Québec, Canada J1K 0A5

Article published in conference *«British Machine Vision Conference (BMVC)»* 2021.

This chapter presents the article "Privacy Preserving for Medical Image Analysis via Non-Linear Deformation Proxy" by Kim, Dolz, Desrosiers and Jodoin accepted to the conference *BMVC2021* for publication in 2021. The objective of this research is to develop a new method for 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.

4.1 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 (Wang *et al.*, 2016; Dolz *et al.*, 2018; Lee, Chun, Hong & et al., 2020; Litjens *et al.*, 2017). 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.

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 (Kumar *et al.*, 2018; Kim, Dolz, Jodoin & Desrosiers, 2021b). Results reported in Section 4.4 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 (Ziad *et al.*, 2016; Nandakumar *et al.*, 2019). Although these methods perform well in some applications, homomorphic

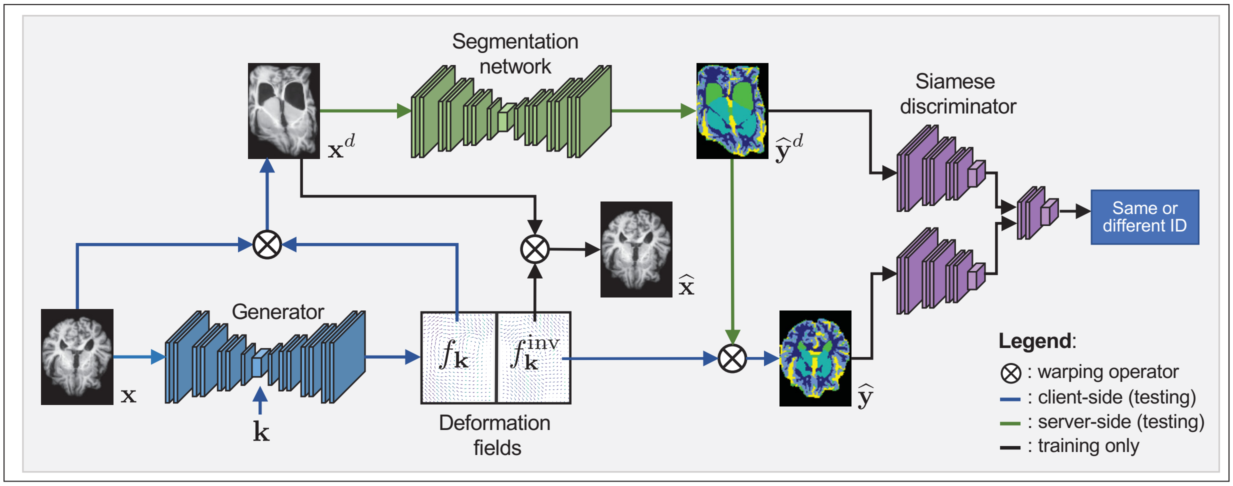


Figure 4.1 Diagram of the proposed system. Once deployed (testing), the client performs operations identified by blue lines, and the server operations corresponding to green lines

cryptosystems typically incur high computational costs (Hesamifard *et al.*, 2017; Nandakumar *et al.*, 2019; Paillier, 1999) 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. 4.1, 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.

4.2 Related works

4.2.1 Homomorphic encryption

One way of preserving privacy is via homomorphic-encryption (HE) (Dowlin *et al.*, 2016; Hesamifard *et al.*, 2017; Nandakumar *et al.*, 2019), 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 (Rouhani *et al.*, 2018). 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.* (2019) report processing rates of 30 min per image). Thus, homomorphic networks have been relatively simplistic (Hardy *et al.*, 2017) and it is not clear how state-of-the-art deep neural nets (Ronneberger *et al.*, 2015) can accommodate this approach.

4.2.2 Federated learning

Another solution for multi-centric deep learning data analysis is federated learning. (Xie *et al.*, 2014; Konecný *et al.*, 2016; McMahan *et al.*, 2016; Vepakomma *et al.*, 2018; Yang *et al.*, 2019). 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 (Wu *et al.*, 2019; Zhu, Liu & Han, 2019).

4.2.3 Privacy preserving with adversarial learning

A popular solution consists in training a generator to create perturbed images, from either a noise distribution (Xu *et al.*, 2019) or real images (Raval *et al.*, 2017). 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 (Pittaluga *et al.*, 2019; Wu *et al.*, 2018; Yang *et al.*, 2018; Roy & Boddeti, 2019; Ren, Jae Lee & Ryoo, 2018; Xiao, Tsai, Sohn, Chandraker & Yang, 2020). 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 (Oleszkiewicz *et al.*, 2018; Kim *et al.*, 2021b).

4.2.4 Differences with existing methods

In contrast with prior works, the proposed framework can easily scale-up to non-fixed classes scenarios. Furthermore, compared to (Oleszkiewicz *et al.*, 2018), our approach presents significant differences both in the objectives and methodology. First, privacy preserving is investigated in the context of biometrical data in (Oleszkiewicz *et al.*, 2018) (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 (Kim *et al.*, 2021b) 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.

4.3 Methods

4.3.1 Proposed architecture

As shown in Fig. 4.1, 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.

4.3.1.1 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 (Kim *et al.*, 2021b) 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 (Wolberg, 1999)). 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 $f^{\text{inv}}_{\mathbf{k}}$, and use a reconstruction loss (see Section 4.3.3.3) to impose that $f^{\text{inv}}_{\mathbf{k}} \circ f_{\mathbf{k}} = I$. 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 4.3.2. 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.

4.3.1.2 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 (Çiçek, Abdulkadir, Lienkamp, Brox & Ronneberger, 2016) which implements a convolutional encoder-decoder architecture with skip-connections between corresponding levels of the encoder and decoder.

4.3.1.3 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 (Kim *et al.*, 2021b) 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.

4.3.1.4 Test-time system

At testing, the system can be used for privacy-preserving segmentation as illustrated in Fig. 4.1. 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 $\widehat{\mathbf{y}} = f_{\mathbf{k}}^{\mathrm{inv}}(\widehat{\mathbf{y}}^d)$.

4.3.2 Transformation function

As in (Balakrishnan, Zhao, Sabuncu, Guttag & Dalca, 2019; Chaitanya *et al.*, 2019), 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 (Jaderberg, Simonyan, Zisserman & kavukcuoglu, 2015), this geometric deformation is based on grid sampling. Let **b** be the base-grid of size $(H, W, D, 3)$ containing the coordinates of image voxels, and **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 **d** using tri-linear interpolation:

$$x^d_{u,v,w} = \sum_{(u',v',w')\in\Omega} x_{u',v',w'} \cdot \max\left(0, 1-|d_u-b_{u'}|\right) \cdot \max\left(0, 1-|d_v-b_{v'}|\right) \cdot \max\left(0, 1-|d_w-b_{w'}|\right) \quad (4.1)$$

Since Eq. (4.1) is differentiable, we can back-propagate gradients during optimization.

4.3.3 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 :

$$\mathcal{L}_{\text{total}}(S, G, D) = \mathcal{L}_{\text{seg}}(S) + \lambda_1\mathcal{L}_{\text{adv}}(G, D) + \lambda_2\mathcal{L}_{\text{inv}}(G) + \lambda_3\mathcal{L}_{\text{smt}}(G) + \lambda_4\mathcal{L}_{\text{div}}(G) \quad (4.2)$$

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.

4.3.3.1 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 (Sudre, Li, Vercauteren, Ourselin & Cardoso, 2017b) to measure the difference between the reconstructed predicted segmentation and its corresponding ground-truth:

$$\mathcal{L}_{\text{seg}}(S) \;=\; \min_{S}\; \mathbb{E}_{(\mathbf{x},\mathbf{y}),\mathbf{k}}\big[\ell_{\text{Dice}}(\mathbf{y}, \widehat{\mathbf{y}})\big] \;=\; \min_{S}\; \mathbb{E}_{(\mathbf{x},\mathbf{y}),\mathbf{k}}\big[\ell_{\text{Dice}}(\mathbf{y}, (f_{\mathbf{k}}^{\text{inv}} \circ S \circ f_{\mathbf{k}})(\mathbf{x}))\big] \tag{4.3}$$

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

4.3.3.2 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

$$\begin{aligned}\mathcal{L}_{\text{adv}}(G, D) \;=\; \min_{G}\,\max_{D}\; & \mathbb{E}_{\mathbf{y}_i,\mathbf{y}_j}\big[s_{ij}\log D(\mathbf{y}_i, \mathbf{y}_j) \;+\; (1-s_{ij})\log\left(1 - D(\mathbf{y}_i, \mathbf{y}_j)\right)\big] \\ & + \mathbb{E}_{\mathbf{y},\mathbf{k}}\big[\,\log\left(1 - D(\widehat{\mathbf{y}}^{d}, \widehat{\mathbf{y}})\right)\big] \end{aligned} \tag{4.4}$$

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. (4.4) minimizes the mutual information between a pair $(\widehat{\mathbf{y}}^d, \widehat{\mathbf{y}})$ and the same-identity variable s_{ij} (Kim *et al.*, 2021b). Consequently, it impedes a potential attacker from retrieving subject identity for a given distorted image by matching it with a database of existing images.

4.3.3.3 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:

$$\mathcal{L}_{\text{inv}}(G) \, = \, \min_{G} \, \mathbb{E}_{(\mathbf{x},\mathbf{y}),\mathbf{k}}\Big[\ell_{\text{SSIM}}(\mathbf{x}, (f_{\mathbf{k}}^{\text{inv}} \circ f_{\mathbf{k}})(\mathbf{x})) \, + \, \ell_{\text{Dice}}(\mathbf{y}, (f_{\mathbf{k}}^{\text{inv}} \circ f_{\mathbf{k}})(\mathbf{y}))\Big] \tag{4.5}$$

where $\ell_{\text{SSIM}}(\mathbf{x}, \mathbf{y}) \in [0, 1]$ is the SSIM loss as in (Zhao, Gallo, Frosio & Kautz, 2017).

The global SSIM is generated at each voxel using a 11×11×11 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 (Wang, Simoncelli & Bovik, 2003).

4.3.3.4 Transformation smoothness loss

The transformation invertibility loss in Eq. (4.5) 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

Table 4.1 Segmentation and re-identification results on the PPMI dataset

	Segmentation DSC						Re-id. F1-score		Re-id. mAP	
Method	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 (Kim *et al.*, 2021b)	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

smoothness:

$$\mathcal{L}_{\text{smt}}(G) \;=\; \mathbb{E}_{\mathbf{x},\mathbf{k}}\left[\frac{1}{|\Omega|}\sum_{(u,v,w)\in\Omega}\|\nabla f_{u,v,w}\|_2\right] \tag{4.6}$$

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

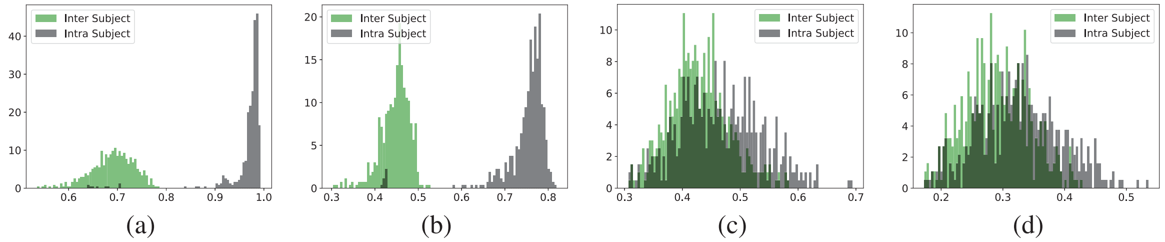
4.3.3.5 Transformation diversity loss

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 $\mathbf{k}'$:

$$\mathcal{L}_{\text{div}}(G) \;=\; \max_{G}\; \mathbb{E}_{(\mathbf{x},\mathbf{y}),\mathbf{k},\mathbf{k}'}\left[\ell_{\text{SSIM}}(f_{\mathbf{k}}(\mathbf{x}), f_{\mathbf{k}'}(\mathbf{x})) \;+\; \ell_{\text{Dice}}(f_{\mathbf{k}}(\mathbf{y}), f_{\mathbf{k}'}(\mathbf{y}))\right] \tag{4.7}$$

Table 4.2 Influence of the different terms of the loss function $\mathcal{L}_{\text{total}}$ in the reconstruction

		Segmentation DSC					
Method	MS-SSIM	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 4.2 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

4.4 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.

4.4.1 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 (Marek *et al.*, 2011) and MRBrainS13 Challenge (Mendrik *et al.*, 2015) 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.

4.4.2 Baseline

4.4.2.1 No-proxy Baseline

Re-identification Result: 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 4.1, 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. 4.2 (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 4.1, 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

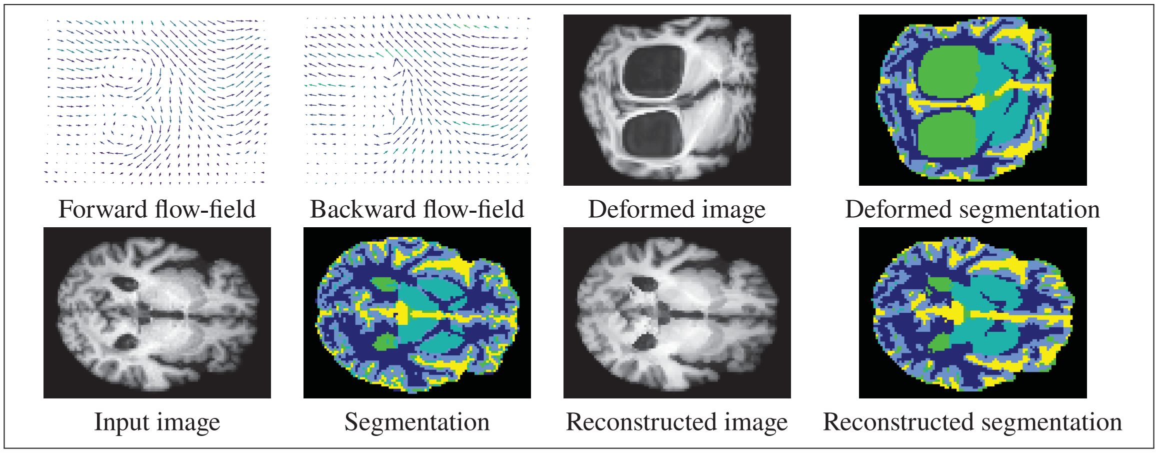


Figure 4.3 Visualization of forward f and backward f^{inv} deformation fields, input images with its associated ground truth map, deformed image and segmentation map and the reconstructed images and segmentation maps

to those obtained in a recent publication for a similar architecture (Dolz *et al.*, 2019). Note that the nuclei and the internal CSF have a lower Dice due to their smaller sizes.

4.4.2.2 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 4.1. 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.

4.4.2.3 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 4.1. 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.

4.4.3 Results on PPMI

4.4.3.1 Re-identification Result

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 4.1. 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. 4.2 (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 4.3 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.

4.4.3.2 Reconstruction Result

The first row of Table 4.2, 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.

4.4.3.3 Segmentation Result

The segmentation DSC achieved by our method is reported in the *All losses* row of Table 4.1. 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 (Zijdenbos *et al.*, 1994; Zou *et al.*, 2004; Gambacorta *et al.*, 2013; Anders *et al.*, 2011) while others, more conservative, suggest minimum DSC values of 0.80 (Mattiucci *et al.*, 2013). That said, if an application requires a larger Dice score, one can improve it by reducing the λ_4 Diversity coefficient (c.f. Eq.(4.2)). The segmentation results for different values of λ_4 are reported in the *All losses* rows of Table 4.1. 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 4.1.

4.4.3.4 Comparison to the state-of-art

We also compared our system to the recently-proposed Privacy-Net (Kim *et al.*, 2021b). 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 (Kim *et al.*, 2021b) and the proposed framework resort to UNet as backbone segmentation architecture. This demonstrates that *i)* our approach preserves the segmentation capabilities shown in (Kim *et al.*, 2021b), and also *ii)* it can drastically improve the obfuscation of identity.

4.4.4 Ablation study

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

4.4.4.1 Invertibility loss

We trained the whole system without the invertibility loss of Eq. (4.5). Although the segmentation loss in Eq. (4.3) implicitly handles the reconstruction of segmentation maps, it is not sufficient for learning a reversible transformation. As can be seen from Tables 4.1 and 4.2, 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.

4.4.4.2 Smoothness loss

We trained the system without the smoothness loss of Eq. (4.6) 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 4.1 and 4.2.

4.4.4.3 Diversity loss

As indicated in Tables 4.1 and 4.2, removing the transformation diversity loss of Eq. (4.7) 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 4.1. 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 4.1).

4.4.5 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 4.3 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.

Table 4.3 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

4.5 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.

CHAPTER 5

MIXUP-PRIVACY: A SIMPLE YET EFFECTIVE APPROACH FOR PRIVACY-PRESERVING SEGMENTATION

Ngoc Bach, Kim[1] , Jose Dolz[1] , Christian Desrosiers[1] Pierre-Marc Jodoin[2]

[1] Department of Software and IT Engineering, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Department of Computer Science, Université de Sherbrooke,
2500 Boul de L'université, Sherbrooke, Québec, Canada J1K 0A5

Article published in conference *«Information Processing In Medical Imaging (IPMI)»* 2023.

5.1 Present:

This chapter presents the article "Mixup-Privacy: A simple yet effective approach for privacy-preserving segmentation" by Kim, Dolz, Jodoin and Desrosiers accepted by the conference *IPMI2023* for publication in 2023. The objective of this article is to propose a client-server image segmentation system which allows for the analysis of multi-centric medical images while preserving patient privacy. In this approach, the client protects the to-be-segmented patient image by mixing it to a reference image. As shown in our work, it is challenging to separate the image mixture to exact original content, thus making the data unworkable and unrecognizable for an unauthorized person. This proxy image is sent to a server for processing. The server then returns the mixture of segmentation maps, which the client can revert to a correct target segmentation. Our system has two components: 1) a segmentation network on the server side which processes the image mixture, and 2) a segmentation unmixing network which recovers the correct segmentation map from the segmentation mixture. Furthermore, the whole system is trained end-to-end. 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 comparable to a system trained on raw images, and outperforms other privacy-preserving methods with little computational overhead.

5.2 Introduction

Neural networks are the *de facto* solution to numerous medical analysis tasks, from disease recognition, to anomaly detection, segmentation, tumor resurgence prediction, and many more (Dolz *et al.*, 2018; Litjens *et al.*, 2017). Despite their success, the widespread clinical deployment of neural nets has been hindered by legitimate privacy restrictions, which limit the amount of data the scientific community can pool together.

Researchers have explored a breadth of solutions to tap into massive amounts of data while complying with privacy restrictions. One such solution is federated learning (FL) (Konecný *et al.*, 2016; Yang *et al.*, 2019), for which training is done across a network of computers each holding its local data. While FL has been shown to be effective, it nonetheless suffers from some limitations when it comes to medical data. First, from a cybersecurity standpoint, communicating with computers located in a highly-secured environment such as a hospital, while complying with FDA/MarkCE cybersecurity regulation, is no easy feast. Second, having computers communicate with their local PACS server is also tricky. And third, since FL is a decentralized *training* solution, it requires a decentralized set of computers to process images at test time, making it ill-suited for software as a service (SAAS) cloud services. Another solution is to train a centralized network with homomorphic data encryption (Hardy *et al.*, 2017). While this ensures a rigorous data protection, as detailed in Section 5.3, the tremendous computational complexity of homomorphic networks prohibits their use in practice.

Recent studies have investigated centralized cloud-based solutions where data is encoded by a neural network prior being sent to the server (Kim *et al.*, 2021b). While the encoded data is unworkable for unauthorized parties, it nonetheless can be processed by a network that was trained to deal with such encoded data. In some methods, such as Privacy-Net (Kim *et al.*, 2021b), the data sent back to the client (e.g., predicted segmentation maps) is not encoded and may contain some private information about the patient (e.g., the patient's identity or condition). To ensure that the returned data is also unworkable for non-authorized users, Kim et al.(Kim, Dolz, Jodoin & Desrosiers, 2021a) proposed an encoding method based on reversible image warping, where the warping function is only known by the client.

In this paper, we propose a novel client-server cloud system that can effectively segment medical images while protecting subjects' data privacy. Our segmentation method, which relies on the hardness of blind source separation (BSS) as root problem (Jain & Rai, 2012; Cardoso, 1998; Nouri *et al.*, 2022; Davies & James, 2007), leverages a simple yet powerful technique based on mixup (Guo *et al.*, 2019). In the proposed approach, the client protects the to-be-segmented patient image by mixing it to a reference image only known to this client. This reference image can be thought as a private key needed to encode and decode the image and its segmentation map. The image mixture renders the data unworkable and unrecognizable for a non-authorized person, since recovering the original images requires to solve an intractable BSS problem. This proxy image is sent to a server for a processing task, which corresponds to semantic segmentation in this work. Instead of sending back the non-encoded segmentation map, as in (Kim *et al.*, 2021b), the server returns to the client a mixture of the target and reference segmentation maps. Finally, because the client knows the segmentation map for the reference image, as well as the mixing coefficients, it can easily recover the segmentation for the target.

Our work makes four contributions to privacy-preserving segmentation:

1. We introduce a simple yet effective method inspired by mixup, which encodes 3D patches of a target image by mixing them to reference patches with known ground-truth. Unlike FL approaches, which require a bulky training setup, or homomorphic networks which are

computationally prohibitive, our method works in a normal training setup and has a low computational overhead.

2. We also propose a learning approach for recovering the target segmentation maps from mixed ones, which improves the noisy results of directly reversing the mixing function.
3. Results are further improved with a test-time augmentation strategy that mixes a target image with different references and then ensembles the segmentation predictions to achieve a higher accuracy.
4. We conduct extensive experiments on two challenging 3D brain MRI benchmarks, and show our method to largely outperform state-of-art approaches for privacy-preserving segmentation, while being simpler and faster than these approaches and yet offering a similar level of privacy.

5.3 Related works

Most privacy-preserving approaches for image analysis fall in two categories: those based on homomorphic encryption and the ones using adversarial learning.

5.3.1 Homomorphic encryption (HE)

This type of encryption (Dowlin *et al.*, 2016; Hesamifard *et al.*, 2017; Nandakumar *et al.*, 2019) enables to compute a given function on encrypted data without having to decrypt it first or having access to the private key. Although HE offers strong guarantees on the security of the encrypted data, this approach suffers from two important limitations: 1) it has a prohibitive computational/communication overhead (Rouhani *et al.*, 2018); 2) it is limited to multiplications and additions, and non-linear activation functions have to be approximated by polynomial functions. As a result, homomorphic networks have been relatively simplistic (Hardy *et al.*, 2017), and even computing the output of a simple CNN is prohibitively slow (e.g., 30 minutes for a single image (Nandakumar *et al.*, 2019)).

5.3.2 Adversarial learning (AL)

This type of approach uses a neural net to encode images so that private information is discarded, yet the encoded image still holds the necessary information to perform a given image analysis task (Xu *et al.*, 2019; Raval *et al.*, 2017). The encoder is trained jointly with two downstream networks taking the encoded image as input, the first one seeking to perform the target task and the other one (the discriminator) trying to recover the private information. The parameters of the encoder are updated to minimize the task-specific utility loss while maximizing the loss of the discriminator. In medical imaging tasks, where patient identity should be protected, the discriminator cannot be modeled as a standard classifier since the number of classes (e.g., patient IDs) is not fixed. To alleviate this problem, the method in (Kim *et al.*, 2021b) uses a Siamese discriminator which receives two encoded images as input and predicts if the images are from the same patient or not. While input images are encoded, the method produces non-encoded segmentation maps which may still be used to identify the patient. The authors of (Kim *et al.*, 2021a) overcome this limitation by transforming input images with a reversible non-linear warping which depends on a private key. When receiving a deformed segmentation map from the server, the client can recover the true segmentation by reversing the transformation. However, as the method in (Kim *et al.*, 2021b), this approach requires multiple scans of the same patient to train the Siamese discriminator, which may not be available in practice. Furthermore, the learned encoder is highly sensitive to the distribution of input images and fails to obfuscate identity when this distribution shifts. In contrast, our method does not require multiple scans per patient. It is also simpler to train and, because it relies on the general principle of BSS, is less sensitive to the input image distribution.

5.4 Methodology

We first introduce the principles of blind source separation and mixup on which our work is based, and then present the details of our Mixup-Privacy method.

5.4.1 Blind source separation

Blind source separation (BSS) is a well-known problem of signal processing which seeks to recover a set of unknown source signals from a set of mixed ones, without information about the mixing process. Formally, let $x(t) = [x_1(t), \ldots, x_n(t)]^T$ be a set of n source signals which are mixed into a set of m signals, $y(t) = [y_1(t), \ldots, y_m(t)]^T$, using matrix $A \in \mathbb{R}^{m \times n}$ as follows: $y(t) = A \cdot x(t)$. BSS can be defined as recovering $x(t)$ when given only $y(t)$. While efficient methods exist for cases where $m = n$, the problem is much harder to solve when $m < n$ as the system of equations then becomes under-determined (Jain & Rai, 2012). For the extreme case of single channel separation ($n = 1$), (Davies & James, 2007) showed that traditional approaches such as Independent Component Analysis (ICA) fail when the sources have substantially overlapping spectra. Recently, the authors of (Jayaram & Thickstun, 2020) proposed a deep learning method for single channel separation, using the noise-annealed Langevin dynamics to sample from the posterior distribution of sources given a mixture. Although it achieves impressive results for the separation of RGB natural images, as we show in our experiments, this method does not work on low-contrast intensity images such as brain MRI. Leveraging the ill-posed nature of single source separation, we encode 3D patches of images to segment by mixing them with those of reference images.

5.4.2 Mixup training

Mixup is a data augmentation technique that generates new samples via linear interpolation between random pairs of images as well as their associated one-hot encoded labels (Zhang, Cisse, Dauphin & Lopez-Paz, 2018a). Let (x_i, y_i) and (x_j, y_j) be two examples drawn at random from the training data, and $\alpha \sim \text{Beta}(b, b)$ be a mixing coefficient sampled from the Beta distribution with hyperparameter b. Mixup generates virtual training examples $(\tilde{x}, \tilde{y})$ as follows:

$$\tilde{x} = \alpha x_i + (1-\alpha)x_j; \quad \tilde{y} = \alpha y_i + (1-\alpha)y_j \tag{5.1}$$

While Mixup training has been shown to bring performance gains in various problems, including image classification (Guo *et al.*, 2019) and semantic segmentation (Zhou, Qi & Shi, 2022), it has not been explored as a way to preserve privacy in medical image segmentation.

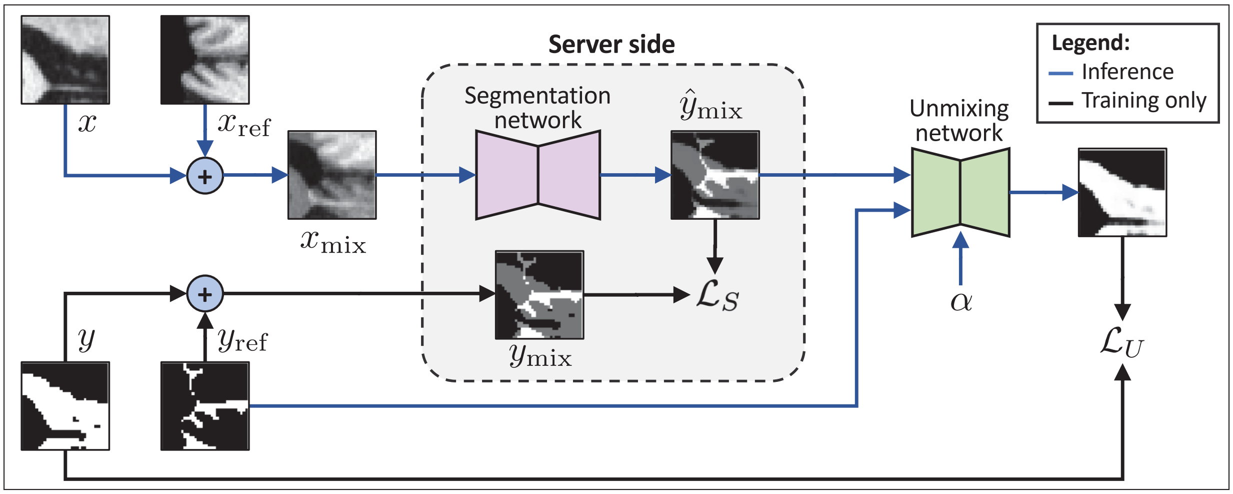


Figure 5.1 Training diagram of the proposed system with the client (left and right) and the server (middle). The client mixes the input image x and segmentation map y with a reference pair $(x_{\text{ref}}, y_{\text{ref}})$. The mixed data is then fed to a segmentation network located on a server and whose output is a mixed segmentation map. The resulting segmentation map is sent back to the client, which decodes it with a unmixing network and the reference map y_{ref}

5.4.3 Proposed system

As shown in Fig 5.1, our method involves a client which has an image x to segment and a server which has to perform segmentation without being able to recover private information from x. During training, the client mixes an image x and its associated segmentation map y with a reference data pair x_{ref} and y_{ref}. The mixed data $(x_{\text{mix}}, y_{\text{mix}})$ is then sent to the server. Since unmixing images requires to solve an under-determined BSS problem, x cannot be recovered from x_{mix} without x_{ref}. This renders x_{mix} unusable if intercepted by an unauthorized user. During inference, the server network returns the mixed segmentation maps $\hat{y}_{\text{mix}}$ to the client, which then recovers the true segmentation maps y by reversing the mixing process. The individual steps of our method, which is trained end-to-end, are detailed below.

5.4.3.1 Data mixing

Since 3D MR images are memory heavy, our segmentation method processes images in a patch-wise manner. Each patch $x \in \mathbb{R}^{H \times W \times D}$ is mixed with a reference patch of the same size:

$$x_{\text{mix}} \;=\; \alpha x_{\text{target}} \;+\; (1-\alpha) x_{\text{ref}} \tag{5.2}$$

where $\alpha \in [0, 1]$ is a mixing weight drawn randomly from the uniform distribution[1]. During training, the one-hot encoded segmentation ground-truths $y \in [0, 1]^{C \times H \times W \times H}$ are also mixed using the same process:

$$y_{\text{mix}} \;=\; \alpha y_{\text{target}} \;+\; (1-\alpha) y_{\text{ref}} \tag{5.3}$$

and are sent to the server with the corresponding mixed image patches x_{mix}.

5.4.3.2 Segmentation and unmixing process

The server-side segmentation network $S(\cdot)$ receives a mixed image patch x_{mix}, predicts the mixed segmentation maps $\hat{y}_{\text{mix}} = S(x_{\text{mix}})$ as in standard Mixup training, and then sends $\hat{y}_{\text{mix}}$ back to the client. Since the client knows the ground-truth segmentation of the reference patch, y_{ref}, it

[1] Unlike Mixup which uses the Beta distribution to have a mixing weight close to 0 or 1, we use the uniform distribution to have a broader range of values.

can easily recover the target segmentation map by reversing the mixing process as follows:

$$\hat{y}_{\text{target}} \;=\; \frac{1}{\alpha}(\hat{y}_{\text{mix}} \;-\; (1-\alpha)y_{\text{ref}}) \tag{5.4}$$

However, since segmenting a mixed image is more challenging than segmenting the ones used for mixing, the naive unmixing approach of Eq. (5.4) is often noisy. To address this problem, we use a shallow network $D(\cdot)$ on the client side to perform this operation. Specifically, this unmixing network receives as input the mixed segmentation $\hat{y}_{\text{mix}}$, the reference segmentation y_{ref}, and the mixing coefficient α, and predicts the target segmentation as $\hat{y}_{\text{target}} = D(\hat{y}_{\text{mix}}, y_{\text{ref}}, \alpha)$.

5.4.4 Test-time augmentation

Test-time augmentation (TTA) is a simple but powerful technique to improve performance during inference (Wang *et al.*, 2019a). Typical TTA approaches generate multiple augmented versions of an example x using a given set of transformations, and then combine the predictions for these augmented examples based on an ensembling strategy. In this work, we propose a novel TTA approach which augments a target patch x_{target} by mixing it with different reference patches $\{x^k_{\text{ref}}\}^K_{k=1}$:

$$x^k_{\text{mix}} \;=\; \alpha x_{\text{target}} \;+\; (1-\alpha)x^k_{\text{ref}} \tag{5.5}$$

The final prediction for the target segmentation is then obtained by averaging the predictions of individual mixed patches:

$$\hat{y}_{\text{target}} \;=\; \frac{1}{K}\sum_{k=1}^{K} D(\hat{y}^k_{\text{mix}},\; y^k_{\text{ref}},\; \alpha) \tag{5.6}$$

As we will show in experiments, segmentation accuracy can be significantly boosted using only a few augmentations.

5.5 Experimental setup

5.5.1 Datasets

We evaluate our method on the privacy-preserving segmentation of brain MRI from two public benchmarks, the Parkinson's Progression Marker Initiative (PPMI) dataset (Marek *et al.*, 2011) and the Brain Tumor Segmentation (BraTS) 2021 Challenge dataset. For the PPMI dataset, we used T1 images from 350 subjects for segmenting brain images into three tissue classes: white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF). Each subject underwent one or two baseline acquisitions and one or two acquisitions 12 months later for a total of 773 images. The images were registered onto a common MNI space and resized to $144 \times 192 \times 160$ with a 1mm^3 isotropic resolution. We divided the dataset into training and testing sets containing 592 and 181 images, respectively, so that images from the same subject are not included in both the training and testing sets. Since PPMI has no ground-truth annotations, as in (Kim *et al.*, 2021b,a), we employed Freesurfer to obtain a pseudo ground-truth for training. We included the PPMI dataset in our experiments because it has multiple scans per patient, which is required for some of the compared baselines (Kim *et al.*, 2021b,a).

BraTS 2021 is the largest publicly-available and fully-annotated dataset for brain tumor segmentation. It contains 1,251 multi-modal MRIs of size $240 \times 240 \times 155$. Each image was manually annotated with four labels: necrose (NCR), edema (ED), enhance tumor (ET), and background. We excluded the T1, T2 and FLAIR modalities and only use T1CE. From the 1,251 scans, 251 scans were used for testing, while the remaining constituted the training set.

5.5.2 Evaluation metrics

Our study uses the 3D Dice similarity coefficient (DSC) to evaluate the segmentation performance of tested methods. For measuring the ability to recover source images, we measure the Multiscale Structural Similarity (MS-SSIM) (Wang *et al.*, 2003) between the original source image and the one recovered from a BSS algorithm (Jayaram & Thickstun, 2020). Last, to evaluate the privacy-preserving ability of our system, we model the task of recovering a patient's identity as a retrieval problem and measure performance using the standard F1-score and mean average precision (mAP) metrics.

5.5.3 Implementation details

We used patches of size $32\times32\times32$ for PPMI and $64\times64\times64$ for BraTS. Larger patches were considered for BraTS to capture the whole tumor. We adopted architectures based on U-Net (Ronneberger *et al.*, 2015) for both the segmentation and unmixing networks. For the more complex segmentation task, we used the U-Net++ architecture described in (Zhou, Rahman Siddiquee, Tajbakhsh & Liang, 2018), whereas a small U-Net with four convolutional blocks was employed for the unmixing network. For the latter, batch normalization layers were replaced by adaptive instance normalization layers (Huang & Belongie, 2017) which are conditioned on the mixing coefficient α. Both the segmentation and unmixing networks are trained using combination of multi-class cross entropy loss and 3D Dice loss (Milletari, Navab & Ahmadi, 2016). End-to-end training was performed for 200,000 iterations on a NVIDIA A6000 GPU, using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 4.

5.5.4 Compared methods

We evaluate different variants of our Mixup-Privacy method for privacy-preserving segmentation. For the segmentation unmixing process, two approaches were considered: a *Naive* approach which reverses the mixing process using Eq. (5.4), and a *Learned* one using the unmixing network $D(\cdot)$. Both approaches were tested with and without the TTA strategy described in Section 5.4.4, giving rise to four different variants. We compared these variants against a segmentation *Baseline* using non-encoded images and two recent approaches for cloud-based

Table 5.1 Main results of the proposed approach across different tasks - including segmentation, blind source separation and test-retest reliability - and two datasets (PPMI and BraTS2021)

	PPMI				BraTS2021			
	GM	WM	CSF	Avg	NCR	ED	ET	Avg
SEGMENTATION (DICE SCORE)								
Baseline	0.930	0.881	0.876	0.896	0.846	0.802	0.894	0.847
Privacy-Net (Kim *et al.*, 2021b)	0.905	0.804	0.732	0.813	—	—	—	—
Deformation-Proxy (Kim *et al.*, 2021a)	0.889	0.825	0.757	0.823	—	—	—	—
Ours (*Naive*)	0.758	0.687	0.634	0.693	0.656	0.635	0.692	0.661
Ours (*Naive* + *TTA*)	0.852	0.829	0.793	0.825	0.775	0.737	0.804	0.772
Ours (*Learned*)	0.893	0.833	0.795	0.840	0.805	0.763	0.842	0.803
Ours (*Learned* + *TTA*)	**0.925**	**0.879**	**0.863**	**0.889**	**0.841**	**0.808**	**0.872**	**0.840**
BLIND SOURCE SEPARATION (MS-SSIM)								
Separation Accuracy	0.602 ± 0.104				0.588 ± 0.127			
TEST-RETEST RELIABILITY (ICC VALUE)								
ICC	0.845	0.812	0.803	—	0.842	0.812	0.803	—
Upper bound	0.881	0.856	0.844	—	0.878	0.855	0.839	—
Lower bound	0.798	0.783	0.771	—	0.805	0.777	0.768	—

privacy-preserving segmentation: *Privacy-Net* (Kim *et al.*, 2021b) and *Deformation-Proxy* (Kim *et al.*, 2021a). The hyperparameters of all compared methods were selected using 3-fold cross-validation on the training set.

5.5.5 Results

5.5.5.1 Segmentation performance.

The top section of Table 5.1 reports the segmentation performance of the compared models. Since Privacy-Net and Deformation-Proxy require longitudinal data to train the Siamese discriminator, we only report their results for PPMI, which has such data. Comparing the naive and learned approaches for segmentation unmixing, we see that using an unmixing network brings a large boost in accuracy. Without TTA, the learned unmixing yields an overall Dice improvement of

14.7% for PPMI and of 14.2% for BraTS2021. As shown in Fig. 5.2, the naive approach directly reversing the mixing process leads to a noisy segmentation which severely affects accuracy.

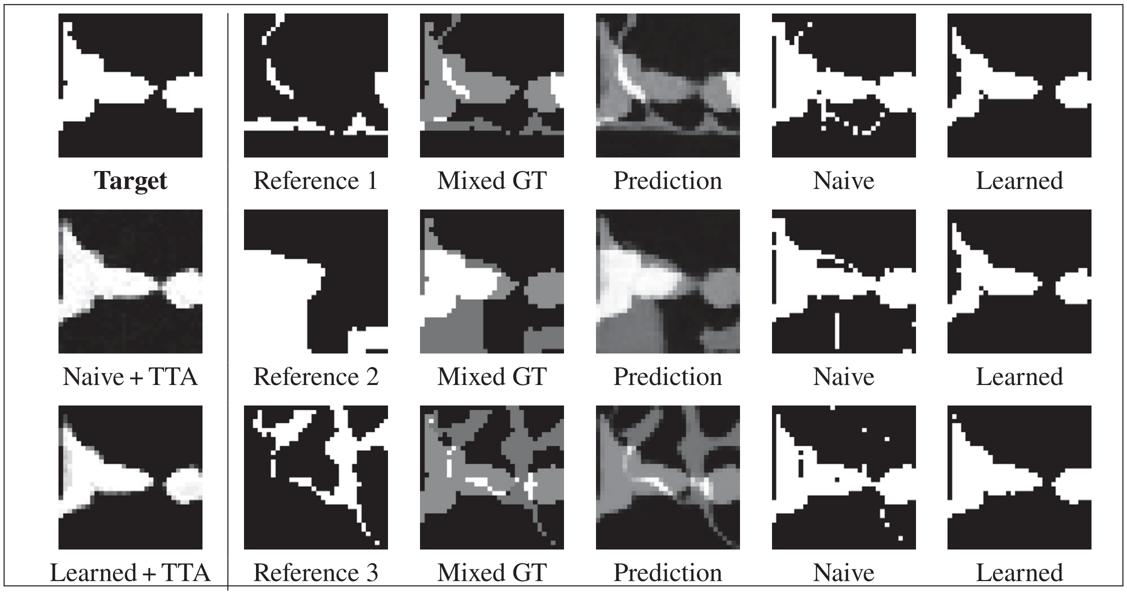


Figure 5.2 Examples of segmented patches obtained by the naive and learned unmixing approaches from the same target and three different references. Naive + TTA and Learned + TTA show the mean prediction of these approaches for 30 augmentations (each one using a different reference)

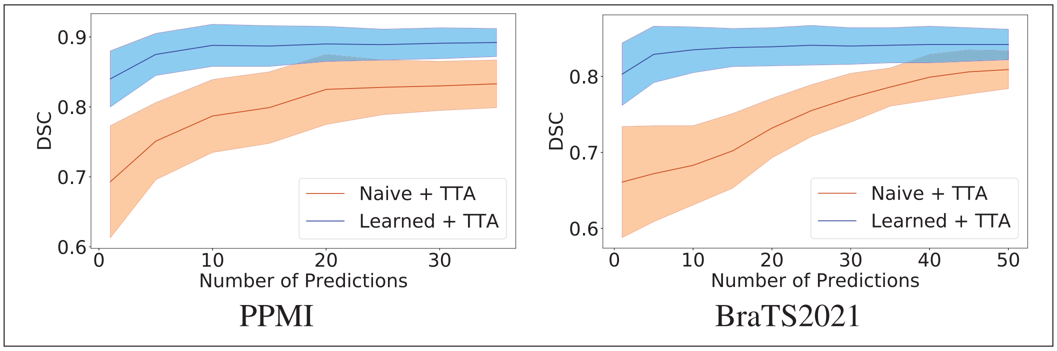


Figure 5.3 Segmentation accuracy (DSC) against the number of TTA predictions

Results in Table 5.1 also demonstrate the positive impact of our TTA strategy on segmentation performance. Thus, adding this strategy to the naive unmixing approach increases the overall Dice by 13.2% for PPMI and by 11.1% for BraTS2021. Likewise, combining it with the learned unmixing approach boosts the overall Dice by 4.9% for PPMI and by 3.7% in the case of

BraTS2021. Looking at the predictions for different reference patches in Fig. 5.2, we see a high variability, in particular for the naive unmixing approach. As can be seen in the first column of the figure (Naive + TTA and Learned + TTA), averaging multiple predictions in our TTA strategy reduces this variability and yields a final prediction very close to the ground-truth. As in other TTA-based approaches, our TTA strategy incurs additional computations since a segmentation prediction must be made for each augmented example (note that these predictions can be made in a single forward pass of the segmentation network). It is therefore important to analyze the gain in segmentation performance for different numbers of TTA augmentations. As shown in Fig. 5.3, increasing the number of predictions for augmented examples leads to a higher Dice, both for the naive and learned unmixing approaches. Interestingly, when using the learned unmixing (i.e., Learned + TTA), the highest accuracy is reached with only 10-15 augmentations. In summary, our TTA strategy brings considerable improvements with limited computational overhead.

5.5.5.2 Blind source separation

To assess whether our mixing-based image encoding effectively prevents an authorized person to recover the source image, we try to solve this BSS problem using the Deep Generative Priors algorithm introduced in (Jayaram & Thickstun, 2020). This algorithm uses a Noise Conditional Score Network (NCSN) (Song & Ermon, 2019) to compute the gradient of the log density function with respect to the image at a given noise level σ, $\nabla_x \log p_\sigma(x)$. An iterative process based on noise-annealed Langevin dynamics is then employed to sample from the posterior distribution of sources given a mixture. We use the U-Net++ as model for the NCSN, and train this model from scratch for each dataset with a Denoising Score Matching loss. Training is performed for 100,000 iterations on NVIDIA A6000 GPU, using the Adam optimizer with a learning rate of 5×10^{-4} and a batch size of 16.

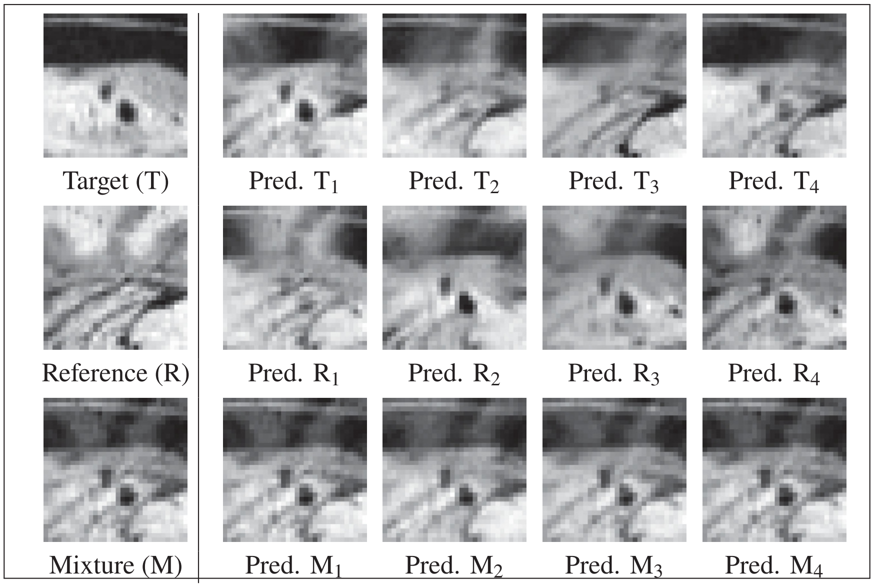


Figure 5.4 Examples of blind source separation (BSS) results for the mixture of given target and reference patches. Columns 2-5 correspond to results for different random initializations of the BSS algorithm

The second section of Table 5.1 gives the mean (± stdev) of MS-SSIM scores (ranging from 0 to 1) between original target images and those recovered from the BSS algorithm: 0.602 ± 0.104 for PPMI and 0.588 ± 0.127 for BraTS2021. These low values indicate that the target image cannot effectively be recovered from the mixed one. This is confirmed in Fig. 5.4 which shows the poor separation results of the BSS algorithm for different random initializations.

5.5.5.3 Test-retest reliability

One source of variability in our method (without TTA) is the choice of the reference image used for mixing. To evaluate the stability of our method with respect to this factor, we perform a test-retest reliability analysis measuring the intra-class correlation coefficient (ICC) (Mcgraw & Wong, 1996) of the test DSC for two predictions using different references. A higher ICC (ranging from 0 to 1) corresponds to a greater level of consistency. The third section of Table 5.1 reports the ICC score obtained for each segmentation class, as well as the upper and lower bounds at 95% confidence. We see that all ICC values are above 0.75, indicating a good reliability.

Table 5.2 Subject re-identification analysis on the PPMI dataset

Method	F1-score	mAP
No Proxy	0.988	0.998
Privacy-Net (Kim *et al.*, 2021b)	0.092	0.202
Deformation-Proxy (Kim *et al.*, 2021a)	0.122	0.147
Ours	0.284	0.352

5.5.5.4 Subject re-identification

To measure how well our method protects the identity of patients, we carry out a patient re-identification analysis using the PPMI dataset which has multiple scans for the same patient. In this analysis, we encode each image in the dataset by mixing it with a randomly chosen reference. For an encoded image x_{mix}, we predict the patient identity as the identity of the other encoded image x'_{mix} most similar to x_{mix} based on the MS-SSIM score. Table 5.2 compares the F1-score and mAP performance of our method to a baseline with no image encoding (No Proxy), Privacy-Net and Deformation-Proxy. As can be seen, the re-identification of patients is quite easy when no encoding is used (mAP of 0.998), and all encoding-based methods significantly reduce the ability to recover patient identity using such retrieval approach. While our mixing based method does not perform as well as the more complex Privacy-Net and Deformation-Proxy approaches, it still offers a considerable protection while largely improving segmentation accuracy (see Table 5.1).

5.6 Conclusion

We introduced an efficient method for privacy-preserving segmentation of medical images, which encodes 3D patches of a target image by mixing them to reference patches with known ground-truth. Two approaches were investigated for recovering the target segmentation maps from the mixed output of the segmentation network: a naive approach reversing the mixing process directly, or using a learned unmixing model. We also proposed a novel test-time augmentation (TTA) strategy to improve performance, where the image to segment is mixed by different references and the predictions for these mixed augmentations are averaged to generate the final prediction.

We validated our method on the segmentation of brain MRI from the PPMI and BraTS2021 datasets. Results showed that using a learned unmixing instead of the naive approach improves DSC accuracy by more than 14% for both datasets. Our TTA strategy, which alleviates the problem of prediction variability, can also boost DSC performance by 3.7%–13.2% when added on top of its single-prediction counterpart. Compared to state-of-art approaches such as Privacy-Net and Deformation-Proxy, our method combining learned unmixing and TTA achieves a significantly better segmentation, while also offering a good level of privacy.

In the future, we plan to validate our method on other segmentation tasks involving different imaging modalities. While we encoded a target image by mixing it to a reference one, other strategies could be also explored, for example, mixing more than two images. This could make the BSS more difficult, hence increasing the security of the method, at the cost of a reduced segmentation accuracy. The prediction variance of our TTA strategy could also be used as a measure of uncertainty in semi-supervised segmentation settings or to suggest annotations in an active learning system.

CONCLUSION AND DISCUSSION

6.1 Conclusion

This thesis' introduction chapter outlined the privacy concerns associated with the application of deep learning to medical image processing research. Deep learning has demonstrated enormous potential for revolutionizing medical image analysis and diagnosis, resulting in substantial advances in healthcare. The use of sensitive medical data in deep learning models, however, raises ethical and privacy concerns. Medical images contain patient-specific information, such as anatomical details and identifiers, that can be used to retrieve the identity of a patient. To maintain patients' confidence and abide by legal and ethical obligations, protecting patient privacy and assuring the secure handling of medical image data is of paramount importance. This thesis seeks to propose effective privacy-preserving solutions that enable the responsible use of medical image data.

In the background chapter, a thorough examination of the fundamental concepts of deep learning and the techniques utilized within the scope of this thesis was conducted. CNNs were used as the primary architecture for image segmentation modules, image transformation, and the acquisition of sensitive patient data. Adversarial learning was adopted as a key technique for generating image transformations that effectively obfuscate sensitive patient data, thereby preserving their privacy. In addition, contrastive learning techniques were utilized to extract sensitive information from patients data. In summary, the chapter described the main deep learning techniques used in the thesis and their applicability in addressing the challenges associated with medical image analysis and privacy protection.

In the literature review chapter, a comprehensive analysis of the strengths and limitations of various privacy-preserving methods was conducted. Specifically, three prominent techniques, federated learning, homomorphic encryption, and adversarial learning, were extensively reviewed. Federated learning enables the training of models on decentralized data sources while maintaining data confidentiality, thereby facilitating collaborative learning without data sharing. Homomorphic encryption, on the other hand, enables computations on encrypted data, ensuring the security of sensitive information throughout the training process. Finally, adversarial learning approaches use adversarial examples or generative models to obscure sensitive information within the data, thereby enhancing privacy. The literature review chapter provided a critical analysis of these methods, underscoring their potential benefits and limitations in machine learning applications that protect privacy. This comprehensive evaluation serves as the basis for the ensuing research, guiding the selection and application of appropriate privacy-protecting techniques within the context of the thesis.

In the methodological chapters (chapters 3, 4, and 5), we proposed various techniques to address the privacy-preserving problem in medical imaging, resulting in three distinct contributions.

In chapter 3 we presented a client/server model that protects privacy in the context of multicentric medical image analysis. Our method is based on adversarial learning, which encodes images to conceal the patient's identity while retaining sufficient data for the downstream task. Our innovative architecture comprises three components: 1) an encoder network that extracts identity-specific features from input medical images, 2) a discriminator network that attempts to identify the subject from the input images, and 3) a medical image analysis network that analyzes the content of the encoded images (in our case, segmentation). By simultaneously fooling the discriminator and training the medical analysis network, the encoder learns to eliminate privacy-specific features while retaining the performance of the target task. The problem of segmenting brain MRI from the Parkinson Progression Marker Initiative (PPMI) dataset is used to illustrate our method. Using longitudinal PPMI data, we demonstrated that the encoder learns to severely distort input images while still allowing for producing highly accurate segmentation results. Additionally, our results demonstrate that an encoder trained on the PPMI dataset can be used to segment other datasets, MRBrainS dataset in this case, without retraining.

In chapter 4, we introduced a client-server system that enables the analysis of multi-centric medical images while maintaining patient privacy. In our method, the client protects the identity of the patient by applying a pseudo-random non-linear deformation to the input image. This generates a proxy image that is then sent to the server for processing. The server then returns the deformed processed image, which the client reconstructs into its original state. Our system consists of three elements: 1) a flow-field generator that produces a pseudo-random deformation function, 2) a Siamese discriminator that learns the identity of the patient from the processed image, and 3) a medical image processing network that analyzes the content of the proxy images. The entire system is trained in an adversarial manner. By fooling the discriminator, the flow-field generator learns to generate a bi-directional non-linear deformation that enables the subject's identity to be removed from the input image and recovered from the output image. The flow-field generator is deployed on the client side and the segmentation network is deployed on the server side following end-to-end training. The proposed technique for MRI brain segmentation is validated using images from two distinct datasets. Results indicate that the segmentation accuracy of our method is comparable to that of a system trained on non-encoded images, while the ability to recover subject identity is drastically diminished.

In chapter 5, we proposed a simple and efficient client-server system for privacy-preserving image segmentation. In this method, the client protects the image of the patient to be segmented by combining it with a reference image. As demonstrated by our research, it is challenging to separate the image mixture into its exact original components, rendering the data unusable and unidentifiable to an unauthorized individual. This proxy image is transmitted for processing to a server. The server then returns a mixture of segmentation maps, which the client can use to determine the correct segmentation for the target. Our system contains two components: 1) a server-side segmentation network that processes the image mixture, and 2) a segmentation unmixing network that recovers the correct segmentation map from the segmentation mixture. The entire system is trained end to end. The proposed technique for MRI brain segmentation is validated using images from two datasets, PPMI and BraTS 2021. Our method's segmentation accuracy is comparable to that of a system trained on raw images, and it outperforms other privacy-preserving methods with minimal computational overhead.

6.2 Discussion

In this section, we critique the merits and drawbacks of each contribution. We then present some directions for future research that could build on these contributions:

1. Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images:

   While the proposed method offered a more effective solution for the cloud-based segmentation of medical images, this solution is incomplete and may not be suitable for real-world applications. The non-encoding of the segmentation map output is one of the primary causes for concern. This poses a substantial risk, as the subject's identity could potentially be recovered. Privacy is of the utmost importance in the realm of medical data, and any system or methodology must prioritize the preservation of patient information. Until the issue of non-encoded segmentation maps is resolved, implementing this solution in real-world scenarios could compromise patient confidentiality, thereby limiting the purpose it is intended to serve.

   Another minor issue associated with this method is the subtle encoding of the segmentation boundary within the encrypted images, akin to a digital watermark. Despite the imperceptibility of this information to the human eye, as depicted in Fig 3.2, and the incapacity of the discriminator to recover the subject's identity from the encrypted images, it still poses a risk to the efficacy of the method.

Therefore, the adoption of such approach for medical image analysis requires additional research and development to assure the existence of robust privacy-preserving techniques, thus leading to the second method of this thesis.

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

   This method is a considerable improvement over the one presented in Chapter 3. Significant progress has been made with the encoding of the segmentation map output through a structure transformation. The transformation generated by the proposed method can be seen as an approximation of a diffeomorphism, wherein the transformation is invertible; and both the transformation and its inverse are continuous differentiable. As discuss in the Chapter 4, Eq 4.5 ensures the invertibility of the transformation, while Eq 4.6 guarantees the continuously differentiable properties for both the transformation and its inverse.

   It is noteworthy that our method is capable of generating diverse transformations contingent upon distinct input vectors denoted as k. The input image and the produced transformation are kept at the client's side. The input image and the resultant transformation remain securely stored at the client's side. Consequently, the risk of reverse engineering for image reconstruction, segmentation recovery, or subject's identity retrieval by an external entity lacking prior knowledge of the transformation is notably limited. However, it is imperative to acknowledge that this method does not alter the image intensity, thereby permitting the potential inference of other privacy-related information, such as the device used for image capture, race, age, etc., through statistical analyses. This aspect represents a vulnerability inherent in the method.

   Another disadvantage of this approach is that this method still requires a longitudinal dataset for training purposes. Despite the fact that longitudinal datasets provide valuable insights and facilitate the examination of changes over time, their availability and acquisition can be limited, posing a challenge for the implementation on a larger scale. Consequently, this lead to the third methodology employed within this thesis.

3. Mixup-Privacy: A simple yet effective approach for privacy-preserving segmentation:

   This novel approach allows for the encoding of both the input image and the output segmentation maps, providing a comprehensive solution for protecting privacy in medical image analysis. By encapsulating both components, information contained within the input image and segmentation maps is protected, ensuring patient confidentiality. Another benefit of this method is that it does not rely on longitudinal datasets for training, making it more accessible and applicable to a broader spectrum of medical imaging scenarios.

   Unlike the previous approaches, this method avoids using adversarial learning, a technique that has demonstrated some instability in certain applications. Adversarial learning involves simultaneously training a generator and a discriminator, resulting in a delicate balance that can be difficult to maintain. By averting adversarial learning, this technique circumvents the associated instability concerns. Instead, it employs alternative strategies to preserve privacy, ensuring the approach's robustness and dependability in real-world applications. By incorporating the ability to encode both input images and segmentation maps, eliminating the need for longitudinal datasets, and circumventing the instability concerns associated with adversarial learning, this method represents a significant advance in medical image analysis that protects patient privacy. It paves the way for broader adoption and application of this method, thereby enhancing patient privacy and facilitating more accurate and dependable analysis of medical images in a variety of healthcare settings.

4. Here we propose recommendation for future works:

   First of all, as mentioned earlier, the first approach alters the intensity of images, whereas the second approach solely distorts the structural attributes of the images. Consequently, a potential direction for future research will be integrating both approaches to concurrently distort both the structure and intensity of images. The combination method will enhance privacy protection.

Secondly, in the future, as this method continues to evolve, it will be necessary to validate its efficacy across a wider variety of data types. Different medical imaging modalities pose unique challenges and characteristics that must be considered, despite the fact that the current validation has demonstrated its efficacy. Using diverse data modalities, such as magnetic resonance imaging (MRI), computed tomography (CT), and ultrasound, its versatility and adaptability can be comprehensively evaluated. This validation process will enable researchers and clinicians to evaluate the method's robustness and reliability across various imaging technologies, ensuring its applicability in clinical settings.

Thirdly, future studies could also concentrate on investigating different mixing schemes. Mixing schemes play a crucial role in the protection of privacy and the method's overall performance. By experimenting with various combining techniques, these studies will aid in refining the method and extending its applicability to additional scenarios.

Last but not least, many publicly available labeled medical images datasets frequently exhibit a relatively limited size when comparing with datasets in the broader field of computer vision. This discrepancy in size leads to an inadequacy of training data. Consequently, it becomes imperative to implement learning mechanisms capable of aggregating and leveraging data from diverse origins. In pursuit of this objective, the mix-up method proves instrumental in formulating a distributed learning algorithm, wherein data is distributed across different organizations.

BIBLIOGRAPHY

AboulAtta, M., Ossadnik, M. & Ahmadi, S. (2019). Stabilizing Inputs to Approximated Nonlinear Functions for Inference with Homomorphic Encryption in Deep Neural Networks. *CoRR*, abs/1902.01870. Retrieved from: http://arxiv.org/abs/1902.01870.

Aggarwal, N., Gupta, C. & Sharma, I. (2014). Fully Homomorphic symmetric scheme without bootstrapping. *Proceedings of 2014 International Conference on Cloud Computing and Internet of Things*, pp. 14-17. doi: 10.1109/CCIOT.2014.7062497.

Al Badawi, A., Jin, C., Lin, J., Mun, C., Jie, S., Tan, B., Nan, X., Khin, A. & Chandrasekhar, V. (2020). Towards the AlexNet Moment for Homomorphic Encryption: HCNN, the First Homomorphic CNN on Encrypted Data with GPUs. *IEEE Transactions on Emerging Topics in Computing*, PP, 1-1. doi: 10.1109/TETC.2020.3014636.

Albelwi, S. & Mahmood, A. (2017). A Framework for Designing the Architectures of Deep Convolutional Neural Networks. *Entropy*, 19(6). Retrieved from: https://www.mdpi.com/1099-4300/19/6/242.

Aledhari, M., Razzak, R., Parizi, R. M. & Saeed, F. (2020). Federated Learning: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Access*, 8, 140699-140725. doi: 10.1109/ACCESS.2020.3013541.

Anders, L., Stieler, F., Siebenlist, K., Schäfer, J., Lohr, F. & Wenz, F. (2011). 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. doi: 10.1016/j.radonc.2011.08.043.

Arjovsky, M., Chintala, S. & Bottou, L. (2017, 06–11 Aug). Wasserstein Generative Adversarial Networks. *Proceedings of the 34th International Conference on Machine Learning*, 70(Proceedings of Machine Learning Research), 214–223. Retrieved from: https://proceedings.mlr.press/v70/arjovsky17a.html.

Avants, B., Tustison, N., Song, G., Cook, P., Klein, A. & Gee, J. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *Neuroimage*, 54(3), 2033-2044.

Badrinarayanan, V., Kendall, A. & Cipolla, R. (2017). SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2481-2495. doi: 10.1109/TPAMI.2016.2644615.

Bakshi, M. & Last, M. (2020). CryptoRNN - Privacy-Preserving Recurrent Neural Networks Using Homomorphic Encryption. *Cyber Security Cryptography and Machine Learning*, pp. 245–253.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. V. & Dalca, A. V. (2019). VoxelMorph: A Learning Framework for Deformable Medical Image Registration. *IEEE Transactions on Medical Imaging*, PP, 1-1. doi: 10.1109/TMI.2019.2897538.

Baldi, P. (2012, 02 Jul). Autoencoders, Unsupervised Learning, and Deep Architectures. *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, 27(Proceedings of Machine Learning Research), 37–49. Retrieved from: https://proceedings.mlr.press/v27/baldi12a.html.

Banabilah, S., Aloqaily, M., Alsayed, E., Malik, N. & Jararweh, Y. (2022). Federated learning review: Fundamentals, enabling technologies, and future applications. *Information Processing & Management*, 59(6), 103061. doi: https://doi.org/10.1016/j.ipm.2022.103061.

Bang, D. & Shim, H. (2021, October). MGGAN: Solving Mode Collapse Using Manifold-Guided Training. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2347-2356.

Bau, D., Zhu, J.-Y., Wulff, J., Peebles, W., Strobelt, H., Zhou, B. & Torralba, A. (2019, October). Seeing What a GAN Cannot Generate. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).*

Bhagoji, A. N., Chakraborty, S., Mittal, P. & Calo, S. (2019, 09–15 Jun). Analyzing Federated Learning through an Adversarial Lens. *Proceedings of the 36th International Conference on Machine Learning*, 97(Proceedings of Machine Learning Research), 634–643. Retrieved from: https://proceedings.mlr.press/v97/bhagoji19a.html.

Bhagyashree, Kushwaha, V. & Nandi, G. C. (2020). Study of Prevention of Mode Collapse in Generative Adversarial Network (GAN). *2020 IEEE 4th Conference on Information & Communication Technology (CICT)*, pp. 1-6. doi: 10.1109/CICT51604.2020.9312049.

Boneh, D., Goh, E.-J. & Nissim, K. (2005). Evaluating 2-DNF Formulas on Ciphertexts. *Theory of Cryptography*, pp. 325–341.

Bos, J. W., Lauter, K., Loftus, J. & Naehrig, M. (2013). Improved Security for a Ring-Based Fully Homomorphic Encryption Scheme. *Cryptography and Coding*, pp. 45–64.

Bourse, F., Minelli, M., Minihold, M. & Paillier, P. (2018). Fast homomorphic evaluation of deep discretized neural networks. *Advances in Cryptology–CRYPTO 2018: 38th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19–23, 2018, Proceedings, Part III 38*, pp. 483–512.

Boykov, Y. & Jolly, M.-P. (2000). Interactive Organ Segmentation Using Graph Cuts. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Retrieved from: https://api.semanticscholar.org/CorpusID:16519743.

Brakerski, Z. (2012). Fully Homomorphic Encryption without Modulus Switching from Classical GapSVP. *Advances in Cryptology – CRYPTO 2012*, pp. 868–886.

Brakerski, Z. & Vaikuntanathan, V. (2011a). Fully Homomorphic Encryption from Ring-LWE and Security for Key Dependent Messages. *Advances in Cryptology – CRYPTO 2011*, pp. 505–524.

Brakerski, Z. & Vaikuntanathan, V. (2011b). Efficient Fully Homomorphic Encryption from (Standard) LWE. *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*, pp. 97-106. doi: 10.1109/FOCS.2011.12.

Brakerski, Z., Gentry, C. & Vaikuntanathan, V. (2012). (Leveled) Fully Homomorphic Encryption without Bootstrapping. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, (ITCS '12), 309–325. doi: 10.1145/2090236.2090262.

Bromley, J., Bentz, J., Bottou, L., Guyon, I., Lecun, Y., Moore, C., Sackinger, E. & Shah, R. (1993). Signature Verification using a "Siamese" Time Delay Neural Network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7, 25. doi: 10.1142/S0218001493000339.

Brutzkus, A., Gilad-Bachrach, R. & Elisha, O. (2019). Low latency privacy preserving inference. *International Conference on Machine Learning*, pp. 812–821.

Butler, D. J., Huang, J., Roesner, F. & Cakmak, M. (2015). The privacy-utility tradeoff for remotely teleoperated robots. *Proc of ACM/IEEE ICHRI*, pp. 27–34.

Cardoso, J.-F. (1998). Blind signal separation: statistical principles. *Proceedings of the IEEE*, 86(10), 2009-2025. doi: 10.1109/5.720250.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660.

Chabanne, H., De Wargny, A., Milgram, J., Morel, C. & Prouff, E. (2017). Privacy-preserving classification on deep neural network. *Cryptology ePrint Archive*.

Chaitanya, K., Karani, N., Baumgartner, C. F., Donati, O., Becker, A. S. & Konukoglu, E. (2019). Semi-Supervised and Task-Driven Data Augmentation. *CoRR*, abs/1902.05396. Retrieved from: http://arxiv.org/abs/1902.05396.

Chang, Y.-T. et al. (2020). Mixup-CAM: Weakly-supervised Semantic Segmentation via Uncertainty Regularization. *BMVC*.

Chauvin, L., Kumar, K., Wachinger, C., Vangel, M., de Guise, J., Desrosiers, C., Wells, W. & Toews, M. (2020). Neuroimage signature from salient keypoints is highly specific to individuals and shared by close relatives. *NeuroImage*, 204, 116208. doi: https://doi.org/10.1016/j.neuroimage.2019.116208.

Chen, J., Wu, J., Richter, K., Konrad, J. & Ishwar, P. (2016a). Estimating head pose orientation using extremely low resolution images. *proc of IEEE SSIAI*, pp. 65–68.

Chen, J., Konrad, J. & Ishwar, P. (2018a). VGAN-based image representation learning for privacy-preserving facial expression recognition. *proc of CVPR-W*, pp. 1570–1579.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S. & Miao, Y. (2021). Review of Image Classification Algorithms Based on Convolutional Neural Networks. *Remote Sensing*, 13(22). doi: 10.3390/rs13224712.

Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. (2020a). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, pp. 1597–1607.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I. & Abbeel, P. (2016b). InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, pp. 2172–2180.

Chen, Y., Sun, X. & Jin, Y. (2020b). Communication-Efficient Federated Deep Learning With Layerwise Asynchronous Model Update and Temporally Weighted Aggregation. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10), 4229-4238. doi: 10.1109/TNNLS.2019.2953131.

Chen, Y., Luo, F., Li, T., Xiang, T., Liu, Z. & Li, J. (2020c). A training-integrity privacy-preserving federated learning scheme with trusted execution environment. *Information Sciences*, 522, 69-79. doi: https://doi.org/10.1016/j.ins.2020.02.037.

Chen, Z., Yeo, C. K., Lee, B. S. & Lau, C. T. (2018b). Autoencoder-based network anomaly detection. *2018 Wireless Telecommunications Symposium (WTS)*, pp. 1-5. doi: 10.1109/WTS.2018.8363930.

Cheon, J. H., Coron, J.-S., Kim, J., Lee, M. S., Lepoint, T., Tibouchi, M. & Yun, A. (2013). Batch Fully Homomorphic Encryption over the Integers. *Advances in Cryptology – EUROCRYPT 2013*, pp. 315–335.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. (2016). 3D U-Net: learning dense volumetric segmentation from sparse annotation. *International conference on medical image computing and computer-assisted intervention*, pp. 424–432.

Coron, J.-S., Lepoint, T. & Tibouchi, M. (2014). Scale-Invariant Fully Homomorphic Encryption over the Integers. *Public-Key Cryptography – PKC 2014*, pp. 311–328.

Dai, J., Saghafi, B., Wu, J., Konrad, J. & Ishwar, P. (2015). Towards privacy-preserving recognition of human activities. *proc of ICIP*, pp. 4238–4242.

Davies, M. & James, C. (2007). Source separation using single channel ICA. *Signal Processing*, 87(8), 1819-1832.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255. doi: 10.1109/CVPR.2009.5206848.

Diffie, W. & Hellman, M. (1976). New directions in cryptography. *IEEE Transactions on Information Theory*, 22(6), 644-654. doi: 10.1109/TIT.1976.1055638.

Dolz, J., Gopinath, K., Yuan, J., Lombaert, H., Desrosiers, C. & Ben Ayed, I. (2019). HyperDense-Net: A Hyper-Densely Connected CNN for Multi-Modal Image Segmentation. *IEEE TMI*, 38(5), 1116-1126.

Dolz, J., Desrosiers, C. & Ayed, I. B. (2018). 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. *NeuroImage*, 170, 456–470.

Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M. & Wernsing, J. (2016). CryptoNets: Applying Neural Networks to Encrypted Data with High Throughput and Accuracy. *proc of ICML.*

Duan, M., Liu, D., Chen, X., Tan, Y., Ren, J., Qiao, L. & Liang, L. (2019). Astraea: Self-Balancing Federated Learning for Improving Classification Accuracy of Mobile Deep Learning Applications. *2019 IEEE 37th International Conference on Computer Design (ICCD)*, pp. 246-254. doi: 10.1109/ICCD46524.2019.00038.

Elgamal, T. (1985). A public key cryptosystem and a signature scheme based on discrete logarithms. *IEEE Transactions on Information Theory*, 31(4), 469-472. doi: 10.1109/TIT.1985.1057074.

Esmeral, L. C. M. & Uhl, A. (2022). Low-Effort Re-identification Techniques Based on Medical Imagery Threaten Patient Privacy. *Medical Image Understanding and Analysis*, pp. 719–733.

Fousse, L., Lafourcade, P. & Alnuaimi, M. (2011, 07). Benaloh's dense probabilistic encryption revisited. *AFRICACRYPT'11 - 4th international conference on Progress in cryptology in Africa*, (Progress In Cryptology Africacrypt 2011 4th International Conference On Cryptology In Africa Dakar Senegal July 5 7 2011 Proceedings), 348-362. Retrieved from: https://hal.science/hal-00769449.

Freymann, J., Kirby, J., Perry, J. & Clunie, D. (2012). Image Data Sharing for Biomedical Research—Meeting HIPAA Requirements for De-identification. *Journal of digital imaging : the official journal of the Society for Computer Applications in Radiology*, 25, 14-24. doi: 10.1007/s10278-011-9422-x.

Galbraith, S. (2001). Elliptic Curve Paillier Schemes. *Journal of Cryptology*, 15. doi: 10.1007/s00145-001-0015-6.

Galbraith, S. D., Gebregiyorgis, S. W. & Murphy, S. (2016). Algorithms for the approximate common divisor problem. *LMS Journal of Computation and Mathematics*, 19(A), 58–72. doi: 10.1112/S1461157016000218.

Gambacorta, M. A., Valentini, C., Dinapoli, N., Boldrini, L., Caria, N., Barba, M. C., Mattiucci, G. C., Pasini, D., Minsky, B. & Valentini, V. (2013). 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.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M. & Lempitsky, V. (2016). Domain-adversarial training of neural networks. *JMLR*, 17(1), 2096–2030.

Gao, L., Fu, H., Li, L., Chen, Y., Xu, M. & Xu, C.-Z. (2022). FedDC: Federated Learning with Non-IID Data via Local Drift Decoupling and Correction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10102-10111. doi: 10.1109/CVPR52688.2022.00987.

Gentry, C. (2009). *A fully homomorphic encryption scheme*. (Ph.D. thesis, Stanford University). Retrieved from: crypto.stanford.edu/craig.

Gentry, C., Halevi, S. & Smart, N. P. (2012). Homomorphic Evaluation of the AES Circuit. *Advances in Cryptology – CRYPTO 2012*, pp. 850–867.

Gentry, C., Sahai, A. & Waters, B. (2013). Homomorphic Encryption from Learning with Errors: Conceptually-Simpler, Asymptotically-Faster, Attribute-Based. *Advances in Cryptology – CRYPTO 2013*, pp. 75–92.

Gjøsteen, K. (2005). Symmetric Subgroup Membership Problems. *Public Key Cryptography - PKC 2005*, pp. 104–119.

Goldwasser, S. & Micali, S. (1982). Probabilistic Encryption & How to Play Mental Poker Keeping Secret All Partial Information. *Proceedings of the Fourteenth Annual ACM Symposium on Theory of Computing*, (STOC '82), 365–377. doi: 10.1145/800070.802212.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. *proc NIPS*, pp. 2672–2680.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M. et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33, 21271–21284.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V. & Courville, A. C. (2017). Improved Training of Wasserstein GANs. *Advances in Neural Information Processing Systems*, 30. Retrieved from: https://proceedings.neurips.cc/paper_files/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

Guo, H. et al. (2019). Mixup as locally linear out-of-manifold regularization. *proc of AAAI*, 33(01), 3714–3722.

Halevi, S. & Shoup, V. (2018). Faster Homomorphic Linear Transformations in HElib. *Advances in Cryptology - CRYPTO 2018 - 38th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2018, Proceedings, Part I*, 10991(Lecture Notes in Computer Science), 93–120. doi: 10.1007/978-3-319-96884-1_4.

Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G. & Thorne, B. (2017). Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *ArXiv*, abs/1711.10677.

Hatamizadeh, A., Yin, H., Molchanov, P., Myronenko, A., Li, W., Dogra, P., Feng, A., Flores, M., Kautz, J., Xu, D. & Roth, H. (2023). Do Gradient Inversion Attacks Make Federated Learning Unsafe? *IEEE Transactions on Medical Imaging*, PP, 1-1. doi: 10.1109/TMI.2023.3239391.

He, K., Zhang, X., Ren, S. & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770-778. doi: 10.1109/CVPR.2016.90.

Hemalatha, R., Thamizhvani, T., Dhivya, A. J. A., Joseph, J. E., Babu, B. & Chandrasekaran, R. (2018). Active Contour Based Segmentation Techniques for Medical Image Analysis. In Koprowski, R. (Ed.), *Medical and Biological Image Analysis* (ch. 2). Rijeka: IntechOpen. doi: 10.5772/intechopen.74576.

Hesamifard, E., Takabi, H. & Ghasemi, M. (2017). CryptoDL: Deep neural networks over encrypted data. *arXiv preprint arXiv:1711.05189*.

Hinton, G. & Salakhutdinov, R. (2006). Reducing the Dimensionality of Data with Neural Networks. *Science (New York, N.Y.)*, 313, 504-7. doi: 10.1126/science.1127647.

Hinton, G. E., Osindero, S. & Teh, Y.-W. (2006). A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.*, 18(7), 1527–1554. doi: 10.1162/neco.2006.18.7.1527.

Hoffstein, J., Pipher, J. & Silverman, J. H. (1998). NTRU: A ring-based public key cryptosystem. *Algorithmic Number Theory*, pp. 267–288.

Hsu, C.-Y., Lu, C.-S. & Pei, S.-C. (2011). Homomorphic encryption-based secure SIFT for privacy-preserving feature extraction. *proc of MWSF-III*, 7880.

Huang, G., Liu, Z. & Weinberger, K. Q. (2017). Densely Connected Convolutional Networks. *proc CVPR*.

Huang, X., Ding, Y., Jiang, Z. L., Qi, S., Wang, X. & Liao, Q. (2020). DP-FL: a novel differentially private federated learning framework for the unbalanced data. *World Wide Web*, 23, 2529 - 2545.

Huang, X. & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510.

Ishai, Y. & Paskin, A. (2007). Evaluating Branching Programs on Encrypted Data. *Theory of Cryptography*, pp. 575–594.

Islam, M., Vibashan, V. S., Jose, V. J. M., Wijethilake, N., Utkarsh, U. & Ren, H. (2020). Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet. *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pp. 262–272.

Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. (2017, July). Image-To-Image Translation With Conditional Adversarial Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jaderberg, M., Simonyan, K., Zisserman, A. & kavukcuoglu, k. (2015). Spatial Transformer Networks. *Advances in Neural Information Processing Systems*, 28. Retrieved from: https://proceedings.neurips.cc/paper_files/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

Jain, S. & Rai, D. (2012). Blind source separation and ICA techniques: a review. *IJEST*, 4, 1490-1503.

Jalal, A., Uddin, M. Z. & Kim, T.-S. (2012). Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home. *IEEE TCE*, 58(3), 863–871.

Jayaram, V. & Thickstun, J. (2020). Source separation with deep generative priors. *International Conference on Machine Learning (ICML)*, pp. 4724–4735.

Jiang, X., Zhang, R. & Nie, S. (2012). Image Segmentation Based on Level Set Method. *Physics Procedia*, 33, 840-845. doi: 10.1016/j.phpro.2012.05.143.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawit, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Qi, H., Ramage, D., Raskar, R., Raykova, M., Song, D., Song, W., Stich, S. U., Sun, Z., Theertha Suresh, A., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H. & Zhao, S. (2021). *Advances and Open Problems in Federated Learning*. Now Foundations and Trends.

Kaliski, B. (2011). Quadratic Residuosity Problem. In van Tilborg, H. C. A. & Jajodia, S. (Eds.), *Encyclopedia of Cryptography and Security* (pp. 1003–1003). Boston, MA: Springer US. doi: 10.1007/978-1-4419-5906-5_429.

Kang, J., Xiong, Z., Niyato, D., Zou, Y., Zhang, Y. & Guizani, M. (2020). Reliable Federated Learning for Mobile Networks. *IEEE Wireless Communications*, 27(2), 72-80. doi: 10.1109/MWC.001.1900119.

Karras, T., Aila, T., Laine, S. & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=Hk99zCeAb.

Kass, M., Witkin, A. & Terzopoulos, D. (1988). Snakes: Active contour models. *International Journal of Computer Vision*, 1(2), 321-331. doi: 10.1007/BF00133570.

Kawachi, A., Tanaka, K. & Xagawa, K. (2007). Multi-bit Cryptosystems Based on Lattice Problems. *Public Key Cryptography – PKC 2007*, pp. 315–329.

Kim, B. N., Dolz, J., Jodoin, P.-M. & Desrosiers, C. (2021a). Privacy Preserving for Medical Image Analysis via Non-Linear Deformation Proxy. *British Machine and Vision Conference (BMVC).*

Kim, B. N., Dolz, J., Jodoin, P.-M. & Desrosiers, C. (2021b). Privacy-Net: An Adversarial Approach for Identity-Obfuscated Segmentation of Medical Images. *IEEE Transactions on Medical Imaging*, 40(7), 1737-1749. doi: 10.1109/TMI.2021.3065727.

Koch, G., Zemel, R. & Salakhutdinov, R. (2015). Siamese Neural Networks for One-shot Image Recognition. *proc of ICML.*

Konecný, J., McMahan, H. B., Ramage, D. & Richtárik, P. (2016). Federated Optimization: Distributed Machine Learning for On-Device Intelligence. *CoRR*, abs/1610.02527.

Kong, J., Kim, J. & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33, 17022–17033.

Kraskov, A., Stögbauer, H. & Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6), 066138.

Kumar, K., Desrosiers, C., Siddiqi, K., Colliot, O. & Toews, M. (2017). Fiberprint: A subject fingerprint based on sparse code pooling for white matter fiber analysis. *NeuroImage*, 158, 242-259. doi: https://doi.org/10.1016/j.neuroimage.2017.06.083.

Kumar, K., Toews, M., Chauvin, L., Colliot, O. & Desrosiers, C. (2018). Multi-modal brain fingerprinting: a manifold approximation based framework. *NeuroImage*, 183, 212–226.

Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324. doi: 10.1109/5.726791.

Lee, B., Chun, S., Hong, J. & et al. (2020). DeepBTS: Prediction of Recurrence-free Survival of Non-small Cell Lung Cancer Using a Time-binned Deep Neural Network. *Scientific Report*, 1952, 456–470.

Li, T. & Choi, M. S. (2021). DeepBlur: A Simple and Effective Method for Natural Image Obfuscation. *CoRR*, abs/2104.02655. Retrieved from: https://arxiv.org/abs/2104.02655.

Li, T. & Lin, L. (2019, June). AnonymousNet: Natural Face De-Identification With Measurable Privacy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Li, W., Fan, L., Wang, Z., Ma, C. & Cui, X. (2021). Tackling mode collapse in multi-generator GANs with orthogonal vectors. *Pattern Recognition*, 110, 107646. doi: https://doi.org/10.1016/j.patcog.2020.107646.

Lin, P., Zheng, C., Yang, Y. & Gu, J. (2004). Medical Image Segmentation by Level Set Method Incorporating Region and Boundary Statistical Information. *Progress in Pattern Recognition, Image Analysis and Applications*, pp. 654–660.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B. & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *MedIA*, 42, 60–88.

Liu, T., Li, Z., Yu, C. & Qin, Y. (2017). NIRS feature extraction based on deep auto-encoder neural network. *Infrared Physics & Technology*, 87, 124-128. doi: https://doi.org/10.1016/j.infrared.2017.07.015.

Liu, X., Li, H., Xu, G., Lu, R. & He, M. (2020a). Adaptive privacy-preserving federated learning. *Peer-to-Peer Networking and Applications*, 13. doi: 10.1007/s12083-019-00869-2.

Liu, Y., Zhang, L., Ge, N. & Li, G. (2020b). A Systematic Literature Review on Federated Learning: From A Model Quality Perspective. *CoRR*, abs/2012.01973. Retrieved from: https://arxiv.org/abs/2012.01973.

López-Alt, A., Tromer, E. & Vaikuntanathan, V. (2012). On-the-Fly Multiparty Computation on the Cloud via Multikey Fully Homomorphic Encryption. *Proceedings of the Forty-Fourth Annual ACM Symposium on Theory of Computing*, (STOC ’12), 1219–1234. doi: 10.1145/2213977.2214086.

Lu, X., Liao, Y., Lio, P. & Hui, P. (2020). Privacy-Preserving Asynchronous Federated Learning Mechanism for Edge Network Computing. *IEEE Access*, 8, 48970-48981. doi: 10.1109/ACCESS.2020.2978082.

Luc, P., Couprie, C., Chintala, S. & Verbeek, J. (2016). Semantic Segmentation using Adversarial Networks. *NIPS Workshop on Adversarial Training*.

Lyubashevsky, V., Peikert, C. & Regev, O. (2010). On Ideal Lattices and Learning with Errors over Rings. *Advances in Cryptology – EUROCRYPT 2010*, pp. 1–23.

Marek, K., Jennings, D., Lasch, S., Siderowf, A., Tanner, C., Simuni, T., Coffey, C., Kieburtz, K., Flagg, E., Chowdhury, S. et al. (2011). The Parkinson Progression Marker Initiative (PPMI). *Progress in neurobiology*, 95(4), 629–635.

Mattiucci, G. C., Boldrini, L., Chiloiro, G., D'Agostino, G., Chiesa, S., de Rose, F., Azario, L., Pasini, D., Gambacorta, M., Balducci, M. & Valentini, V. (2013). Automatic delineation for replanning in nasopharynx radiotherapy: What is the agreement among experts to be considered as benchmark? *Acta Oncologica*, 52, 1417 - 1422.

McClure, P., Zheng, C. Y., Kaczmarzyk, J., Rogers-Lee, J., Ghosh, S., Nielson, D., Bandettini, P. A. & Pereira, F. (2018). Distributed weight consolidation: A brain segmentation case study. *proc of NIPS*, pp. 4093–4103.

Mcgraw, K. & Wong, S. (1996). Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods*, 1, 30-46.

McMahan, B., Moore, E., Ramage, D., Hampson, S. & Arcas, B. A. y. (2017, 20–22 Apr). Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54(Proceedings of Machine Learning Research), 1273–1282. Retrieved from: https://proceedings.mlr.press/v54/mcmahan17a.html.

McMahan, H. B., Moore, E., Ramage, D. & y Arcas, B. A. (2016). Federated Learning of Deep Networks using Model Averaging. *CoRR*, abs/1602.05629.

Mendrik, A. M., Vincken, K. L., Kuijf, H. J., Breeuwer, M., Bouvy, W. H., De Bresser, J., Alansary, A., De Bruijne, M., Carass, A., El-Baz, A. et al. (2015). MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Comp. Intel. and Neuro.*, 2015, 1.

Milletari, F., Navab, N. & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571.

Minister Responsible for Access to Information and the Protection of Personal Information. (2021). Bill 64, an Act to modernize legislative provisions as regards the protection of personal information - National Assembly of Québec. Minister Responsible for Democratic Institutions, Electoral Reform and Access to Information. Retrieved from: https://m.assnat.qc.ca/en/travaux-parlementaires/projets-loi/projet-loi-64-42-1.html.

Miyato, T., Maeda, S.-i., Koyama, M. & Ishii, S. (2018). Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8), 1979–1993.

Mohri, M., Sivek, G. & Suresh, A. T. (2019, 09–15 Jun). Agnostic Federated Learning. *Proceedings of the 36th International Conference on Machine Learning*, 97(Proceedings of Machine Learning Research), 4615–4625. Retrieved from: https://proceedings.mlr.press/v97/mohri19a.html.

Mowla, N. I., Tran, N. H., Doh, I. & Chae, K. (2020). Federated Learning-Based Cognitive Detection of Jamming Attack in Flying Ad-Hoc Network. *IEEE Access*, 8, 4338-4350. doi: 10.1109/ACCESS.2019.2962873.

Munjal, K. & Bhatia, R. (2022). A systematic review of homomorphic encryption and its contributions in healthcare industry. *Complex & Intelligent Systems*, 1-28. doi: 10.1007/s40747-022-00756-z.

Nandakumar, K., Ratha, N., Pankanti, S. & Halevi, S. (2019). Towards Deep Neural Network Training on Encrypted Data. *proc of CVPR-W*, pp. 0–0.

Nguyen, T. V., Dakka, M. A., Diakiw, S. M., VerMilyea, M. D., Perugini, M., Hall, J. M. M. & Perugini, D. (2022). A novel decentralized federated learning approach to train on globally distributed, poor quality, and protected private medical datag. *Scientific Reports*.

Nouri, A. et al. (2022). A new approach to feature extraction in MI-based BCI systems. In *Artificial Intelligence-Based Brain-Computer Interface* (pp. 75–98).

Oleszkiewicz, W., Kairouz, P., Piczak, K., Rajagopal, R. & Trzciński, T. (2018). Siamese generative adversarial privatizer for biometric data. *proc of ACCV*, pp. 482–497.

Oord, A. v. d., Li, Y. & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Packhäuser, K., Gündel, S., Münster, N., Syben, C., Christlein, V. & Maier, A. (2022). Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data. *Scientific Reports*, 12, 14851. doi: 10.1038/s41598-022-19045-3.

Paillier, P. (1999). Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. *Advances in Cryptology — EUROCRYPT '99*, pp. 223–238.

Pisa, P. S., Abdalla, M. & Duarte, O. C. M. B. (2012). Somewhat homomorphic encryption scheme for arithmetic operations on large integers. *2012 Global Information Infrastructure and Networking Symposium (GIIS)*, pp. 1-8. doi: 10.1109/GIIS.2012.6466769.

Pittaluga, F., Koppal, S. & Chakrabarti, A. (2019). Learning privacy preserving encodings through adversarial training. *proc of IEEE WACV*, pp. 791–799.

Prayitno, Shyu, C.-R., Putra, K. T., Chen, H.-C., Tsai, Y.-Y., Hossain, K. S. M. T., Jiang, W. & Shae, Z.-Y. (2021). A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. *Applied Sciences*, 11(23). doi: 10.3390/app112311191.

Pulido-Gaytan, L., Tchernykh, A., Cortés-Mendoza, J., Babenko, M., Radchenko, G., Avetisyan, A. & Drozdov, A. (2021). Privacy-preserving neural networks with Homomorphic encryption: Challenges and opportunities. *Peer-to-Peer Networking and Applications*, 14. doi: 10.1007/s12083-021-01076-8.

Qian, X., Wang, J., Guo, S. & Li, Q. (2013). An active contour model for medical image segmentation with application to brain CT image. *Medical physics*, 40(2), 021911.

Qin, X., Huang, R. & Fan, H. (2021). An Effective NTRU-Based Fully Homomorphic Encryption Scheme. *Mathematical Problems in Engineering*, 2021, 1-9. doi: 10.1155/2021/9914961.

Ramaiah, Y. G. & Kumari, G. V. (2012). Efficient public key Homomorphic Encryption over integer plaintexts. *2012 International Conference on Information Security and Intelligent Control*, pp. 123-128. doi: 10.1109/ISIC.2012.6449723.

Raval, N., Machanavajjhala, A. & Cox, L. P. (2017). Protecting visual secrets using adversarial nets. *proc of CVPR-W*, pp. 1329–1332.

Regev, O. (2009). On Lattices, Learning with Errors, Random Linear Codes, and Cryptography. *J. ACM*, 56(6). doi: 10.1145/1568318.1568324.

Ren, Z., Jae Lee, Y. & Ryoo, M. S. (2018). Learning to anonymize faces for privacy preserving action detection. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 620–636.

Rivest, R., Adleman, L. & Dertouzos, M. (1978). On data banks and privacy homomorphisms. *Foundations on Secure Computation, Academia Press*, pp. 169-179.

Ronneberger, O., Fischer, P. & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241.

Rouhani, B., Riazi, S. & Koushanfar, F. (2018). DeepSecure: Scalable Provably-Secure Deep Learning. *in proc of Design Auto. Conf. (DAC)*.

Roy, A. G., Conjeti, S., Navab, N., Wachinger, C. et al. (2019). QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. *NeuroImage*, 186, 713–727.

Roy, P. C. & Boddeti, V. N. (2019). Mitigating information leakage in image representations: A maximum entropy approach. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2586–2594.

Rumelhart, D. E. & McClelland, J. L. (1987). Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations* (pp. 318-362).

Rutherford, M., Mun, S., Levine, B., Bennett, W., Smith, K., Farmer, P., Jarosz, Q., Wagner, U., Freyman, J., Blake, G., Tarbox, L., Farahani, K. & Prior, F. (2021). A DICOM dataset for evaluation of medical image de-identification. *Scientific Data*, 8. doi: 10.1038/s41597-021-00967-y.

Sander, T., Young, A., Yung, M. & Inc, C. (2001). Non-Interactive CryptoComputing for NC1. *Foundations of Computer Science, 1975., 16th Annual Symposium on*. doi: 10.1109/SFFCS.1999.814630.

Shamir, L. (2013). MRI-based knee image for personal identification. *International Journal of Biometrics*, 5(2), 113–125.

Shokri, R. & Shmatikov, V. (2015). Privacy-preserving deep learning. *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321.

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR 2015)*, pp. 1-14.

Song, Y. & Ermon, S. (2019). Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32.

Stehlé, D. & Steinfeld, R. (2011). Making NTRU as Secure as Worst-Case Problems over Ideal Lattices. *Advances in Cryptology – EUROCRYPT 2011*, pp. 27–47.

Su, H., Deng, J. & Fei-Fei, L. (2012). Crowdsourcing annotations for visual object detection. *Workshops at the 26th AAAI Conference on Artificial Intelligence*.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. (2017a). 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* (pp. 240–248).

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S. & Cardoso, M. J. (2017b). 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* (pp. 240–248). Springer.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015). Going deeper with convolutions. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1-9. doi: 10.1109/CVPR.2015.7298594.

Takabi, H., Hesamifard, E. & Ghasemi, M. (2016). Privacy preserving multi-party machine learning with homomorphic encryption. *29th Annual Conference on Neural Information Processing Systems (NIPS).*

Toews, M., Wachinger, C., Estepar, R. & Wells, W. (2015, 07). A Feature-Based Approach to Big Data Analysis of Medical Images. *International Conference on Information Processing in Medical Imaging*, 24, 339-50. doi: 10.1007/978-3-319-19992-4_26.

Tran, N.-T., Bui, T.-A. & Cheung, N.-M. (2018, September). Dist-GAN: An Improved GAN using Distance Constraints. *Proceedings of the European Conference on Computer Vision (ECCV).*

Truex, S., Liu, L., Chow, K.-H., Gursoy, M. E. & Wei, W. (2020). LDP-Fed: Federated Learning with Local Differential Privacy. *Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking*, (EdgeSys '20), 61–66. doi: 10.1145/3378679.3394533.

Vachhani, B. B., Bhat, C., Das, B. & Kopparapu, S. K. (2017). Deep Autoencoder Based Speech Features for Improved Dysarthric Speech Recognition. *Interspeech*. Retrieved from: https://api.semanticscholar.org/CorpusID:19830245.

Van Dijk, M., Gentry, C., Halevi, S. & Vaikuntanathan, V. (2010). Fully Homomorphic Encryption over the Integers. *Advances in Cryptology – EUROCRYPT 2010*, pp. 24–43.

Vepakomma, P., Swedish, T., Raskar, R., Gupta, O. & Dubey, A. (2018). No Peek: A Survey of private distributed deep learning. *CoRR*, abs/1812.03288.

Wachinger, C., Golland, P., Kremen, W., Fischl, B. & Reuter, M. (2015). BrainPrint: A discriminative characterization of brain morphology. *NeuroImage*, 109, 232-248. doi: https://doi.org/10.1016/j.neuroimage.2015.01.032.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S. & Vercauteren, T. (2019a). Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338, 34–45.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D. & Khazaeni, Y. (2020). Federated Learning with Matched Averaging. *International Conference on Learning Representations*. Retrieved from: https://openreview.net/forum?id=BkluqlSFDS.

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

Wang, L., Wang, W. & Li, B. (2019b). CMFL: Mitigating Communication Overhead for Federated Learning. *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*, pp. 954-964. doi: 10.1109/ICDCS.2019.00099.

Wang, S., Ding, Z. & Fu, Y. (2018). Cross-generation kinship verification with sparse discriminative metric. *IEEE transactions on pattern analysis and machine intelligence*, 41(11), 2783–2790.

Wang, W., Huang, Y., Wang, Y. & Wang, L. (2014, June). Generalized Autoencoder: A Neural Network Framework for Dimensionality Reduction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Wang, W., Vong, C.-M., Yang, Y. & Wong, P.-K. (2017). Encrypted image classification based on multilayer extreme learning machine. *MSSP*, 28(3), 851–865.

Wang, Z., Simoncelli, E. P. & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. *proc of IEEE ACSSC*, pp. 1398-1402.

Wang, Z., Song, M., Zhang, Z., Song, Y., Wang, Q. & Qi, H. (2019c). Beyond Inferring Class Representatives: User-Level Privacy Leakage From Federated Learning. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pp. 2512-2520. doi: 10.1109/INFOCOM.2019.8737416.

Wang, Z., Simoncelli, E. P. & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, 2, 1398–1402.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., Jin, S., Quek, T. Q. S. & Vincent Poor, H. (2020). Federated Learning With Differential Privacy: Algorithms and Performance Analysis. *IEEE Transactions on Information Forensics and Security*, 15, 3454-3469. doi: 10.1109/TIFS.2020.2988575.

Wolberg, G. (1999). Image Morphing: A Survey. *The Visual Computer*, 14. doi: 10.1007/s003710050148.

Wu, B., Zhao, S., Sun, G., Zhang, X., Su, Z., Zeng, C. & Liu, Z. (2019). P3SGD: Patient privacy preserving SGD for regularizing deep CNNs in pathological image classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2099–2108.

Wu, Z., Wang, Z., Wang, Z. & Jin, H. (2018). Towards privacy-preserving visual recognition via adversarial training: A pilot study. *proc of ECCV*, pp. 606–624.

Wu, Z., Lim, S.-N., Davis, L. S. & Goldstein, T. (2020). Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors. *Computer Vision – ECCV 2020*, pp. 1–17.

Xia, H. & Ding, Z. (2020, June). Structure Preserving Generative Cross-Domain Learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiao, T., Tsai, Y.-H., Sohn, K., Chandraker, M. & Yang, M.-H. (2020). Adversarial Learning of Privacy-Preserving and Task-Oriented Representations. *AAAI*.

Xie, P., Bilenko, M., Finley, T., Gilad-Bachrach, R., Lauter, K. E. & Naehrig, M. (2014). Crypto-Nets: Neural Networks over Encrypted Data. *CoRR*, abs/1412.6181.

Xu, C., Ren, J., Zhang, D., Zhang, Y., Qin, Z. & Ren, K. (2019). GANobfuscator: Mitigating information leakage under GAN via differential privacy. *IEEE TIFS*, 14(9), 2358–2371.

Xu, G., Li, H., Liu, S., Yang, K. & Lin, X. (2020). VerifyNet: Secure and Verifiable Federated Learning. *IEEE Transactions on Information Forensics and Security*, 15, 911-926. doi: 10.1109/TIFS.2019.2929409.

Yang, Q., Liu, Y., Chen, T. & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2), 12.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T. & Yu, H. (2020a). Federated Transfer Learning. In *Federated Learning* (pp. 83–93). Cham: Springer International Publishing. doi: 10.1007/978-3-031-01585-4_6.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T. & Yu, H. (2020b). Horizontal Federated Learning. In *Federated Learning* (pp. 49–67). Cham: Springer International Publishing. doi: 10.1007/978-3-031-01585-4_4.

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T. & Yu, H. (2020c). Vertical Federated Learning. In *Federated Learning* (pp. 69–81). Cham: Springer International Publishing. doi: 10.1007/978-3-031-01585-4_5.

Yang, T.-Y., Brinton, C., Mittal, P., Chiang, M. & Lan, A. (2018). Learning Informative and Private Representations via Generative Adversarial Networks. *proc of ICBD*, pp. 1534–1543.

Yao, A. C. (1982). Protocols for secure computations. *23rd Annual Symposium on Foundations of Computer Science (sfcs 1982)*, pp. 160-164. doi: 10.1109/SFCS.1982.38.

Yao, X., Huang, C. & Sun, L. (2018). Two-Stream Federated Learning: Reduce the Communication Costs. *2018 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1-4. doi: 10.1109/VCIP.2018.8698609.

Yao, X., Huang, T., Wu, C., Zhang, R. & Sun, L. (2019). Towards Faster and Better Federated Learning: A Feature Fusion Approach. *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 175-179. doi: 10.1109/ICIP.2019.8803001.

Yi, F. & Moon, I. (2012). Image segmentation: A survey of graph-cut methods. *2012 International Conference on Systems and Informatics (ICSAI2012)*, pp. 1936-1941. doi: 10.1109/ICSAI.2012.6223428.

Yi, X., Paulet, R. & Bertino, E. (2014). Homomorphic Encryption. In *Homomorphic Encryption and Applications* (pp. 27–46). Cham: Springer International Publishing. doi: 10.1007/978-3-319-12229-8_2.

Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. (2018a). mixup: Beyond Empirical Risk Minimization. *International Conference on Learning Representations*.

Zhang, Z., Li, M. & Yu, J. (2018b). On the Convergence and Mode Collapse of GAN. *SIGGRAPH Asia 2018 Technical Briefs*, (SA ’18). doi: 10.1145/3283254.3283282.

Zhao, H., Gallo, O., Frosio, I. & Kautz, J. (2017). Loss Functions for Image Restoration With Neural Networks. *IEEE Transactions on Computational Imaging*, 3(1), 47-57. doi: 10.1109/TCI.2016.2644865.

Zhao, H., Shi, J., Qi, X., Wang, X. & Jia, J. (2017, jul). Pyramid Scene Parsing Network. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6230-6239. doi: 10.1109/CVPR.2017.660.

Zhou, S., Greenspan, H. & Shen, D. (2017). *Deep Learning for Medical Image Analysis*. Elsevier Science.

Zhou, Z., Qi, L. & Shi, Y. (2022). Generalizable Medical Image Segmentation via Random Amplitude Mixup and Domain-Specific Image Restoration. *ECCV*, pp. 420–436.

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *proc. of DLMIA ML-CDS*, pp. 3–11.

Zhu, L., Liu, Z. & Han, S. (2019). Deep leakage from gradients. *Advances in Neural Information Processing Systems*, pp. 14774–14784.

Ziad, M. T. I., Alanwar, A., Alzantot, M. & Srivastava, M. (2016). CryptoImg: Privacy preserving processing over encrypted images. *proc of IEEE CNS*, pp. 570–575.

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

Zou, K., Warfield, S., Bharatha, A., Tempany, C., Kaus, M., Haker, S., Wells, W., Jolesz, F. & Kikinis, R. (2004). Statistical Validation of Image Segmentation Quality Based on a Spatial Overlap Index. *Academic radiology*, 11, 178-89. doi: 10.1016/S1076-6332(03)00671-8.