



**HAL**  
open science

# Data Exfiltration and Anonymization of Medical Images based on Generative Models

Huiyu Li

► **To cite this version:**

Huiyu Li. Data Exfiltration and Anonymization of Medical Images based on Generative Models. Computer Science [cs]. Inria & Université Cote d'Azur, Sophia Antipolis, France, 2024. English. NNT: . tel-04875160

**HAL Id: tel-04875160**

**<https://inria.hal.science/tel-04875160v1>**

Submitted on 8 Jan 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



# THÈSE DE DOCTORAT

## Exfiltration et Anonymisation d'images Médicales à l'aide de Modèles Génératifs

Huiyu LI

CENTRE INRIA D'UNIVERSITÉ CÔTE D'AZUR, Équipe EPIONE

Thèse dirigée par Hervé DELINGETTE et co-dirigée par Nicholas AYACHE

Soutenue le 28 novembre 2024

Présentée en vue de l'obtention du grade de DOCTEUR EN AUTOMATIQUE, TRAITEMENT  
DU SIGNAL ET DES IMAGES d'UNIVERSITÉ CÔTE D'AZUR.

Devant le jury composé de :

Laurent AMSALEG	CNRS-IRISA	Rapporteur
Melek ÖNEN	EURECOM	Rapporteuse
Antoine BOUTET	INSA Lyon, Inria, CITI	Examineur
François BRÉMOND	Centre Inria d'Université Côte d'Azur	Présidente
Nicholas AYACHE	Centre Inria d'Université Côte d'Azur	Co-directeur de thèse
Hervé DELINGETTE	Centre Inria d'Université Côte d'Azur	Directeur de thèse



**Titre français**

Exfiltration et Anonymisation d'images Médicales à  
l'aide de Modèles Génératifs

**Titre anglais**

Data Exfiltration and Anonymization of Medical Images  
based on Generative Models

Jury:

Présidente

François BRÉMOND      Directeur de recherche      Centre Inria d'Université Côte d'Azur

Rapporteurs

Laurent AMSALEG      Directeur de recherche      CNRS-IRISA  
Melek ÖNEN      Professeure      EURECOM

Examineurs

Antoine BOUTET      Maître de conférence      INSA Lyon, Inria, CITI  
François BRÉMOND      Directeur de recherche      Centre Inria d'Université Côte d'Azur  
Nicholas AYACHE      Directeur de recherche      Centre Inria d'Université Côte d'Azur  
Hervé DELINGETTE      Directeur de recherche      Centre Inria d'Université Côte d'Azur



# Abstract

This thesis aims to address some specific safety and privacy issues when dealing with sensitive medical images within data lakes. This is done by first exploring potential data leakage when exporting machine learning models and then by developing an anonymization approach that protects data privacy.

Chapter 2 presents a novel data exfiltration attack, termed **Data Exfiltration by Compression (DEC)**, which leverages image compression techniques to exploit vulnerabilities in the model exporting process. This attack is performed when exporting a trained network from a remote data lake, and is applicable independently of the considered image processing task. By exploring both lossless and lossy compression methods, this chapter demonstrates how DEC can effectively be used to steal medical images and reconstruct them with high fidelity, using two public CT and MR datasets. This chapter also explores mitigation measures that a data owner can implement to prevent the attack. It first investigates the application of differential privacy measures, such as Gaussian noise addition, to mitigate this attack, and explores how attackers can create attacks resilient to differential privacy. Finally, an alternative model export strategy is proposed which involves model fine-tuning and code verification.

Chapter 3 introduces the **Generative Medical Image Anonymization** framework, a novel approach to balance the trade-off between preserving patient privacy while maintaining the utility of the generated images to solve downstream tasks. The framework separates the anonymization process into two key stages: first, it extracts identity and utility-related features from medical images using specially trained encoders; then, it optimizes the latent code to achieve the desired trade-off between anonymity and utility. We employ identity and utility encoders to verify patient identities and detect pathologies, and use a generative adversarial network-based auto-encoder to create realistic synthetic images from the latent space. During optimization, we incorporate these encoders into novel loss functions to produce images that remove identity-related features while maintaining their utility to solve a classification problem. The effectiveness of this approach is demonstrated through extensive experiments on the MIMIC-CXR chest X-ray dataset, where the generated images successfully support lung pathology detection.

Chapter 4 builds upon the work from Chapter 3 by utilizing generative adversarial networks (GANs) to create a more robust and scalable anonymization solution. The framework is structured into two distinct stages: first, we develop a streamlined encoder and a novel training scheme to map images into a latent space. In the second stage, we minimize the dual-loss functions proposed in Chapter 3 to optimize the latent representation of each image. This method ensures that the generated images effectively remove some identifiable features while retaining crucial diagnostic information. Extensive qualitative and quantitative experiments on the MIMIC-CXR dataset demonstrate that our approach produces high-quality anonymized images that maintain essential diagnostic details, making them well-suited for training machine learning models in lung pathology classification.

The conclusion chapter summarizes the scientific contributions of this work, and addresses remaining issues and challenges for producing secured and privacy preserving sensitive medical data.

**Keywords:** Data exfiltration attack, Privacy and security, Steganography, Medical image anonymization, Identity-utility extraction, Latent code optimization.

# Résumé

Cette thèse aborde certains problèmes spécifiques de sécurité et de confidentialité lors du traitement d'images médicales dans des lacs de données. Ainsi, on explore tout d'abord la fuite potentielle de données lors de l'exportation de modèles d'intelligence artificielle, puis on développe une approche d'anonymisation d'images médicales qui protège la confidentialité des données.

Le Chapitre 2 présente une nouvelle attaque d'exfiltration de données, appelée **Data Exfiltration by Compression (DEC)**, qui s'appuie sur les techniques de compression d'images pour exploiter les vulnérabilités du processus d'exportation du modèle. Cette attaque est effectuée lors de l'exportation d'un réseau de neurones entraîné au sein d'un lac de données distant et elle est applicable indépendamment de la tâche de traitement d'images considérée. En explorant à la fois les méthodes de compression sans perte et avec perte, ce chapitre montre comment l'attaque DEC peut être utilisée efficacement pour voler des images médicales et les reconstruire avec une grande fidélité, grâce à l'utilisation de deux ensembles de données CT et IRM publics. Ce chapitre explore également les mesures d'atténuation qu'un propriétaire de données peut mettre en œuvre pour empêcher l'attaque. Il étudie d'abord l'application de mesures de confidentialité différentielle, telles que l'ajout de bruit gaussien, pour atténuer cette attaque, et explore comment les attaquants peuvent créer des attaques résilientes à la confidentialité différentielle. Enfin, une stratégie d'exportation de modèle alternative est proposée, qui implique un réglage fin du modèle et une vérification du code.

Le Chapitre 3 présente une méthode **d'anonymisation d'images médicales par approche générative**, une nouvelle approche pour trouver un compromis entre la préservation de la confidentialité des patients tout en maintenant l'utilité des images générées pour résoudre les tâches de traitement d'images. Cette méthode sépare le processus d'anonymisation en deux étapes clés: tout d'abord, il extrait les caractéristiques liées à l'identité et à l'utilité des images médicales à l'aide d'encodeurs spécialement entraînés; ensuite, il optimise le code latent pour atteindre le compromis souhaité entre l'anonymisation et l'utilité de l'image. Nous utilisons des encodeurs d'identité et d'utilité pour vérifier l'identité des patients et détecter les pathologies, et utilisons un encodeur automatique génératif basé sur un réseau antagoniste pour créer des images synthétiques réalistes à partir de l'espace latent. Lors de l'optimisation, nous incorporons ces encodeurs dans de nouvelles fonctions de perte pour produire des images qui suppriment les carac-



téristiques liées à l'identité tout en conservant leur utilité pour résoudre un problème de classification. L'efficacité de cette approche est démontrée par des expériences approfondies sur l'ensemble de données de radiographie thoracique MIMIC-CXR, où les images générées prennent en charge avec succès la détection de pathologies pulmonaires.

Le Chapitre 4 s'appuie sur les travaux du Chapitre 3 en utilisant des réseaux antagonistes génératifs (GAN) pour créer une solution d'anonymisation plus robuste et évolutive. Le cadre est structuré en deux étapes distinctes: tout d'abord, nous développons un encodeur simplifié et un nouvel algorithme d'entraînement pour plonger chaque image dans un espace latent. Dans la deuxième étape, nous minimisons les fonctions de perte proposées dans le Chapitre 3 pour optimiser la représentation latente de chaque image. Cette méthode garantit que les images générées suppriment efficacement certaines caractéristiques identifiables tout en conservant des informations diagnostiques cruciales. Des expériences qualitatives et quantitatives approfondies sur l'ensemble de données MIMIC-CXR démontrent que notre approche produit des images anonymisées de haute qualité qui conservent les détails diagnostiques essentiels, ce qui les rend bien adaptées à la formation de modèles d'apprentissage automatique dans la classification des pathologies pulmonaires.

Le chapitre de conclusion résume les contributions scientifiques de ce travail et aborde les problèmes et défis restants pour produire des données médicales sensibles, sécurisées et préservant leur confidentialité.

**Mots-clés:** Attaque d'exfiltration de données, Compression d'images, Confidentialité, Stéganographie, Anonymisation d'images médicales, Extraction d'identité-utilité, Optimisation du code latent.

# Acknowledgement

Completing this PhD journey has been one of the most challenging and rewarding experiences of my life. It would not have been possible without the support, guidance, and encouragement of many individuals to whom I am deeply grateful.

First and foremost, I would like to express my sincere gratitude to my supervisors, Hervé Delingette and Nicholas Ayache, for their unwavering support, insightful feedback, and constant encouragement throughout this process. Your attention to detail and dedication to my work have been invaluable in shaping this thesis, and your belief in my abilities has kept me motivated even during the most difficult moments.

I am also sincerely thankful to the members of my jury, Laurent Amsaleg, Melek Önen, Antoine Boutet, and François Brémond, for their time, insightful questions, and constructive criticism, which helped refine my work and broaden my perspective.

I am deeply grateful to all the members of the Epione team for their camaraderie and for making this experience truly rewarding. I would like to extend a special thanks to Marco Milanesio and Paul Tourniaire for their invaluable support of my research. I am also grateful to Jairo Rodriguez for his constant encouragement, as well as to Nathalie Nordmann, Alix De Langlais, and Hari Sreedhar for their support in preparing my defense. Additionally, I am thankful to everyone at INRIA for their warmth and support, which have made my time here genuinely enjoyable.

My deepest appreciation goes to my family for their unconditional love, patience, and understanding. Your constant support and belief in me have been my source of strength, and I cannot thank you enough for standing by me throughout this journey. A special thanks to my husband for the countless insightful discussions and for making this journey even more memorable.

Finally, I would like to acknowledge the funding resources provided by 3IA Côte d'Azur, without which this research would not have been possible.

To everyone who has been a part of this journey, whether mentioned here or not, thank you from the bottom of my heart. Your support and contributions have been instrumental in the completion of this thesis.



# Financial Support

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Privacy Attacks and Defense Mechanisms . . . . .	2
1.1.1	Attacks Targeting the Data and Corresponding Defenses for Data Protection . . . . .	2
1.1.2	Attacks Targeting the Algorithm and Corresponding Defenses for Algorithm Protection . . . . .	4
1.2	Privacy Protection with Anonymization . . . . .	4
1.2.1	Pseudonymization vs. Anonymization . . . . .	6
1.2.2	Two Traditional Families of Anonymization: Randomization and Generalization . . . . .	6
1.3	Generative Anonymization: A New Family of Anonymization . . . . .	7
1.4	How to Check the Effectiveness of Anonymization? . . . . .	8
1.5	Objectives and Organization of the Thesis . . . . .	8
1.6	Publications . . . . .	11
<b>2</b>	<b>Data Exfiltration by Compression Attack: Definition and Evaluation on Medical Image Data</b>	<b>13</b>
2.1	Introduction . . . . .	14
2.2	Related Works & Background . . . . .	19
2.2.1	Network Pruning vs. Training a Smaller Network . . . . .	19
2.2.2	Steganography . . . . .	20
2.2.3	Privacy Defense . . . . .	21
2.3	Method . . . . .	21
2.3.1	Data Exfiltration by Compression Attack: Two Main Scenarios . . . . .	22
2.3.2	Lossy Compression and Utility Models . . . . .	24
2.3.3	Size Reduction of Compression Codes and Decoder Model . . . . .	25
2.3.4	Hiding Compression Codes in Exported Models . . . . .	26
2.3.5	Mitigation Measures . . . . .	27
2.3.6	DEC Attack Resisting to Differential Privacy . . . . .	28
2.4	Materials . . . . .	28
2.4.1	Dataset . . . . .	28
2.4.2	Pre and Post-processing . . . . .	29
2.4.3	Evaluation Metrics . . . . .	29
2.4.4	Compression and utility Models . . . . .	29

2.5	Results . . . . .	30
2.5.1	Attack Effectiveness . . . . .	30
2.5.2	EP Scenario vs. IT Scenario. . . . .	30
2.5.3	Type of Decoder: D vs D1. . . . .	30
2.5.4	Latent Channel Number Configurations. . . . .	32
2.5.5	Compression-fidelity Compromise. . . . .	32
2.5.6	Utility Task Performances . . . . .	33
2.5.7	Code Hiding Method . . . . .	34
2.5.8	Differential Privacy Mitigation . . . . .	36
2.6	Discussion and Limitations . . . . .	40
2.7	Conclusions . . . . .	43
<b>3</b>	<b>GMIA: Generative Medical Image Anonymization Based on Identity-utility Optimization</b>	<b>45</b>
3.1	Introduction . . . . .	46
3.2	Method . . . . .	49
3.2.1	Overview of GMIA . . . . .	49
3.2.2	Latent Code Mapping . . . . .	50
3.2.3	Identity Feature Extraction . . . . .	52
3.2.4	Utility Feature Extraction . . . . .	54
3.2.5	Latent Code Optimization . . . . .	55
3.3	Materials . . . . .	57
3.3.1	Dataset and Pre-processing . . . . .	57
3.3.2	Dataset Pre-processing . . . . .	57
3.3.3	Implementation Details . . . . .	59
3.3.4	Evaluation Metrics . . . . .	61
3.4	Results . . . . .	63
3.4.1	Network Pre-training . . . . .	63
3.4.2	Qualitative Results . . . . .	63
3.4.3	Utility Preservation . . . . .	64
3.4.4	Identity Elimination . . . . .	64
3.5	Discussion and Limitations . . . . .	66
3.5.1	Discussion . . . . .	67
3.5.2	Limitations . . . . .	67
3.6	Conclusions . . . . .	68
<b>4</b>	<b>GMIA2: Designing an Encoder for Generative Medical Image Anonymization</b>	<b>69</b>
4.1	Introduction . . . . .	70
4.2	Related Work . . . . .	71
4.2.1	StyleGAN . . . . .	71
4.2.2	StyleGAN Inversion . . . . .	72
4.3	Method . . . . .	73

4.3.1	Designing an Encoder for Image Reconstruction . . . . .	74
4.3.2	Latent Code Optimization for Image Anonymization . . . . .	78
4.4	Materials . . . . .	79
4.4.1	Evaluation Metrics . . . . .	79
4.4.2	Baseline and Ablation Study . . . . .	80
4.4.3	Implementation Details . . . . .	81
4.5	Results . . . . .	81
4.5.1	Reconstruction Results . . . . .	81
4.5.2	Anonymization Results . . . . .	82
4.5.3	Ablation Study . . . . .	85
4.5.4	Disentanglement Property of Latent Space . . . . .	91
4.6	Discussion and Limitations . . . . .	91
4.6.1	Discussion . . . . .	91
4.6.2	Limitations . . . . .	93
4.7	Conclusions . . . . .	95
<b>5</b>	<b>Conclusion</b>	<b>97</b>
5.1	Main Contributions . . . . .	97
5.2	Future research . . . . .	98
<b>A</b>	<b>Appendix of Chapter2</b>	<b>101</b>
<b>B</b>	<b>Appendix of Chapter3</b>	<b>105</b>
<b>C</b>	<b>Appendix of Chapter4</b>	<b>107</b>
	<b>Bibliography</b>	<b>113</b>



# Introduction

---

1.1	Privacy Attacks and Defense Mechanisms . . . . .	2
1.1.1	Attacks Targeting the Data and Corresponding Defenses for Data Protection . . . . .	2
1.1.2	Attacks Targeting the Algorithm and Corresponding Defenses for Algorithm Protection . . . . .	4
1.2	Privacy Protection with Anonymization . . . . .	4
1.2.1	Pseudonymization vs. Anonymization . . . . .	6
1.2.2	Two Traditional Families of Anonymization: Randomization and Generalization . . . . .	6
1.3	Generative Anonymization: A New Family of Anonymization . . . . .	7
1.4	How to Check the Effectiveness of Anonymization? . . . . .	8
1.5	Objectives and Organization of the Thesis . . . . .	8
1.6	Publications . . . . .	11

---

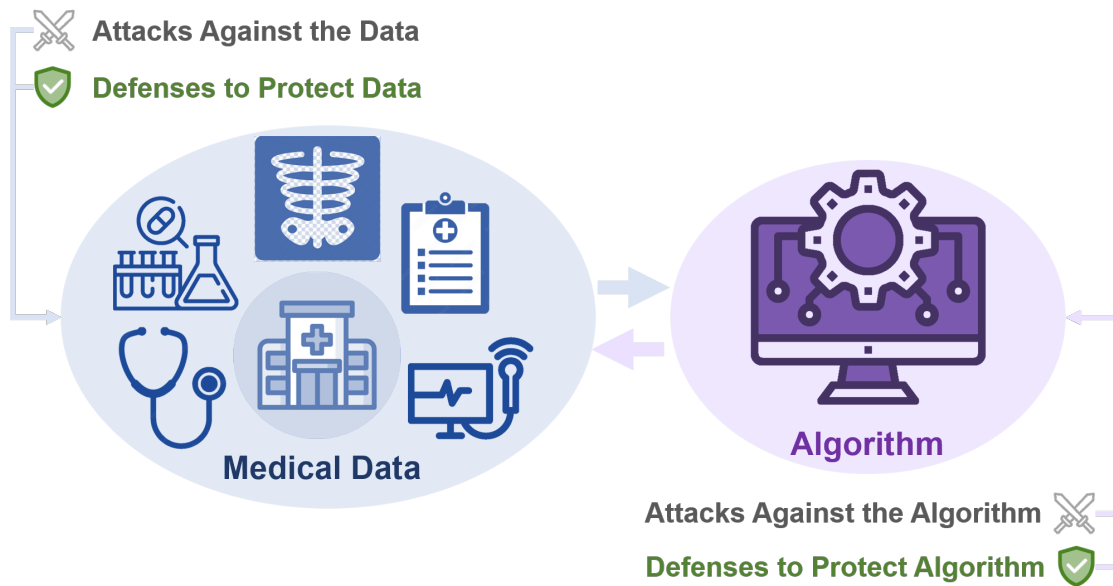
With advancements in artificial intelligence (AI), a wide range of AI-based methods are being applied to medicine, holding the potential to revolutionize the field. However, the sensitive nature of medical data, which includes personal patient information, combined with the vulnerabilities of AI-based algorithms, raises significant privacy and security concerns. As illustrated in Fig. 1.1, various types of attacks can target both the medical data and the algorithms, underscoring the need for robust defense strategies to protect both aspects.

In this thesis, we first examine vulnerabilities associated with medical data, particularly within a medical data lake [Gentner et al., 2023]<sup>1</sup>. Our focus then shifts to protect data privacy by exploring anonymization<sup>2</sup> techniques designed to safeguard sensitive information from the earliest stages of data collection and processing.

---

<sup>1</sup>A medical data lake is a centralized data repository that enables health organizations to store, process, and secure large amounts of structured, semi-structured, and unstructured data.

<sup>2</sup>Anonymized medical data refers to data from which the patient cannot be identified by the recipient of the information.



**Fig. 1.1.:** Schematic overview of the relationships and interactions between medical data and algorithms in the context of potential attack vectors and corresponding defense techniques.

## 1.1 Privacy Attacks and Defense Mechanisms

With the increasing digitization of healthcare records, medical imaging data is vulnerable to privacy attacks. These attacks can be broadly categorized into two main categories based on their target: attacks against the data and attacks against the algorithm [Kaissis et al., 2020], with corresponding defense mechanisms to against these attacks. A glossary of the terms presented in this section is available in Table 1.1 and Table 1.2.

### 1.1.1 Attacks Targeting the Data and Corresponding Defenses for Data Protection

Attacks against the data involve unauthorized access, manipulation, or disclosure of medical imaging data. These attacks aim to compromise the confidentiality, integrity, or availability of the dataset itself. Examples of attacks against the data include **Property Inference Attacks** [Zhang et al., 2022b], **Attribute Reconstruction Attack** (aka feature re-derivation, attribute inference) [Chen et al., 2022; Kaissis et al., 2020], **Membership Inference Attack** (aka tracing attack) [Liu et al., 2023a], **Re-identification Attack** [Chen et al., 2023; Ding et al., 2021; El Emam et al., 2011], **Model Inversion Attack** (MIA) [Zhu et al., 2023], **Gradient Inversion Attack** (GIA) [Liang et al., 2023], and **Data Exfiltration Attack** (DEA) [Amit et al., 2023]. Mitigating these attacks involves employing techniques such as **Pseudonymization** [Neubauer et al., 2011] and **Anonymization** [Chevrier et al., 2019].

**Tab. 1.1.:** Glossary of attack vectors, accompanied by conceptual descriptions.

Method	Description
<b>Attacks Against the Data</b>	
Property Inference Attack [Zhang et al., 2022b]	Inferring properties of target dataset that are not explicitly correlated to the learning task, using only the parameters of the target model as prior knowledge.
Attribute Reconstruction Attack [Chen et al., 2022]	Inferring sensitive attributes, such as a person’s gender, age, or other private information, of a target data given the access to the target model, its output, and the information about the non-sensitive attributes (synonyms: feature re-derivation, attribute inference).
Membership Inference Attack [Liu et al., 2023a]	Determining whether a particular data sample is present in the target dataset to train the target model (synonym: tracing attack).
Re-identification Attack [Chen et al., 2023]	Determining an individual’s identity by linking the information from a pseudo-anonymized personal data to another dataset in which the same individual is contained.
Model Inversion Attack [Zhu et al., 2023]	Reconstructing the target data from target model parameters and its output (synonym: Attribute Inference Attacks).
Gradient Inversion Attack [Liang et al., 2023]	Reconstructing target data by exploiting gradient updates in machine learning models.
Data Exfiltration Attack [Amit et al., 2023]	Obtaining the complete target data through either intentional or unintentional memorization by a model.
<b>Attacks Against the Algorithm</b>	
Adversarial Attack [Bortsova et al., 2021]	Manipulating input data to cause incorrect predictions or other unintended behavior of the machine learning model.
Model Poisoning Attack [Zhou et al., 2021]	Manipulating the model’s parameters to cause it to behave in an undesirable way.
Model Extraction Attack [Zhou et al., 2021]	Reconstructing or approximating the target model by querying the model with carefully chosen inputs and observing its outputs (synonyms: model stealing attack).

**Tab. 1.2.:** Glossary of defense techniques, accompanied by conceptual descriptions.

Method	Description
Defenses to Protect Data	
Pseudonymization [Kaissis et al., 2020]	Replacing direct identifiers in a dataset with artificial identifiers or pseudonyms.
Anonymization [Kaissis et al., 2020]	Removal or alteration of identifying information from a dataset to prevent individuals from being directly or indirectly identified.
Defenses to Protect Algorithm	
Federated Learning [Pfitzner et al., 2021]	A machine learning approach that enables training of models across decentralized devices or servers holding local data samples, without exchanging raw data.
Secure Multi-party Computation [Li et al., 2020a]	A cryptographic technique that allows multiple parties to jointly compute a function over their private inputs without revealing those inputs to each other.
Homomorphic Encryption [Kocabaş et al., 2014]	A form of encryption that allows computations to be performed on encrypted data without decrypting it first.

### 1.1.2 Attacks Targeting the Algorithm and Corresponding Defenses for Algorithm Protection

Attacks against the algorithm target the machine learning models or algorithms utilized for processing medical imaging data. Their objective is to exploit vulnerabilities in the algorithms to compromise privacy or security. Examples of such attacks include **Adversarial Attack** [Bortsova et al., 2021], **Model Poisoning Attack** [Zhou et al., 2021], and **Model Extraction Attack** [Zhou et al., 2021]. Protecting against these attacks requires robust defenses such as **Federated Learning** [Pfitzner et al., 2021], **Secure Multi-party Computation** [Li et al., 2020a], and **Homomorphic Encryption** [Kocabaş et al., 2014].

## 1.2 Privacy Protection with Anonymization

Anonymization is as a crucial defense mechanism against data attacks, utilizing techniques designed to make it practically impossible to re-identify individuals from the anonymized data. However, a common pitfall in exploring anonymization techniques is to consider pseudonymized data to be equivalent to anonymized data. It is essential to

recognize that pseudonymization, while it obscures identities, does not provide the same level of privacy protection as true anonymization.

In this section, we will explain the distinction between pseudonymized and anonymization, while also delving into various anonymization techniques. A glossary of the terms presented in this section is provided in Table 1.3.

**Tab. 1.3.:** Glossary of pseudonymization and anonymization techniques, accompanied by conceptual descriptions. \* is [Machanavajjhala et al., 2007].

Method	Description	
Pseudonymization [Kaissis et al., 2020]	Replacing direct identifiers with artificial identifiers or pseudonyms.	
Anonymization [Kaissis et al., 2020]	Removing or altering identifying information to prevent individuals from being directly or indirectly identified.	
Randomization	Noise Addition [Mivule, 2013]	Introducing random perturbations or noise into the data to protect privacy while preserving the overall statistical properties of the dataset.
	Permutation [Li et al., 2016]	Rearranging the order of elements in a dataset, particularly categorical variables, to prevent the disclosure of sensitive information while preserving data integrity.
	Differential Privacy [Dwork, 2006]	A rigorous privacy framework aimed at ensuring that the inclusion or exclusion of an individual's data does not significantly impact the outcome of a computation or analysis. This is achieved by adding carefully calibrated noise to the output of queries or computations performed on the dataset.
Generalization	Aggregation [Party, 2014]	Combining multiple data points or records into summary statistics or groups, thereby reducing the granularity of the data while preserving its overall trends or patterns.
	$K$ -anonymity [Sweeney, 2002]	A privacy model ensures that each record in a dataset is indistinguishable from at least $K - 1$ other records with respect to certain attributes.
	$L$ -diversity *	An extension of $K$ -anonymity that addresses the limitation of $K$ -anonymity by ensuring that each group of records contains at least $L$ distinct values for a sensitive attribute.
	$T$ -closeness [Li et al., 2006]	A privacy model that builds upon $K$ -anonymity and $L$ -diversity by ensuring that the distribution of sensitive attribute values in each group of records is close to the distribution in the overall dataset.

## 1.2.1 Pseudonymization vs. Anonymization

**Anonymization** [Kaissis et al., 2020] involves the removal or alteration of identifying information from a dataset to prevent individuals from being directly or indirectly identified. This typically includes removing personal identifiers such as names, addresses, social security numbers, and any other information that could be used to identify individuals. Anonymized data should not be able to be linked back to specific individuals, even with additional information or analysis.

**Pseudonymization** [Kaissis et al., 2020] replaces direct identifiers in a dataset with artificial identifiers or pseudonyms. These pseudonyms are unique to each individual but do not directly reveal their identity. Pseudonymized data can still be useful for analysis and processing, but the original identities of individuals are protected. However, if a separate key or lookup table is used to link pseudonyms back to individuals, additional security measures must be in place to protect the key or table from unauthorized access.

**Drawbacks of Pseudonymization.** Pseudonymization, while initially reversible, carries inherent risks of re-identification. This is because pseudonymized datasets can potentially be linked to other data sources, facilitating the identification of individuals. For instance, corresponding lists for identifiers and their pseudonyms or two-way encryption algorithms can inadvertently reveal individuals' identities when combined with external information sources. As a result, even though pseudonymization provides a layer of privacy protection, it may not offer sufficient safeguards against re-identification when linked with other datasets.

## 1.2.2 Two Traditional Families of Anonymization: Randomization and Generalization

In the domain of data anonymization, two main families of techniques are commonly employed: randomization and generalization [Party, 2014].

**Randomization** techniques involve introducing randomness or noise into the data to obscure individual identities while preserving the overall statistical properties of the dataset. This can include techniques such as **Noise Addition** [Mivule, 2013], which involves adding random noise to data, or **Permutation** [Li et al., 2016], which randomly permutes categorical variables. Another approach is **Differential Privacy** [Dwork, 2006], which adds noise to query responses to protect individual privacy. By injecting randomness, randomization methods aim to prevent the re-identification of individuals while still allowing for meaningful analysis of the data.

On the other hand, **generalization** techniques involve the aggregation or summarization of data to a higher level of abstraction, thereby reducing the level of detail and granularity in the dataset. This often entails grouping similar data points together or replacing specific values with broader categories or ranges. Examples of generalization techniques include **Aggregation**, where data points are combined into summary statistics,  **$K$ -anonymity** [Sweeney, 2002], which ensures that each individual in the dataset is indistinguishable from at least  $K - 1$  other individuals,  **$L$ -diversity** [Machanavajjhala et al., 2007], which ensures that each group of records contains at least  $L$  different values for a sensitive attribute, and  **$T$ -closeness** [Li et al., 2006], which ensures that the distribution of a sensitive attribute in each group of records is close to the distribution in the overall dataset. These techniques aim to anonymize data by removing fine-grained details while retaining the overall trends or patterns present in the dataset.

## 1.3 Generative Anonymization: A New Family of Anonymization

With the advent of generative models like Generative Adversarial Networks (GANs), several approaches have emerged that leverage the generative power of these networks to tackle anonymization by producing synthetic data. Unlike traditional anonymization methods, which fall into two primary categories—randomization, where controlled noise or alterations are introduced into the original data, and generalization, where data is aggregated or summarized at a higher level of abstraction—generative anonymization creates entirely new, privacy-safe data. This approach ensures that while the synthetic data retains key statistical properties and utility for analysis, it no longer corresponds to any real individual, offering a robust solution for protecting sensitive information.

With the growing use of facial recognition technologies, face anonymization has become increasingly important, prompting the development of numerous innovative methods [Barattin et al., 2023; Seo et al., 2024; Kuang et al., 2024]. These approaches have also served as inspiration for anonymizing medical images. A notable example is  $k$ -SALSA [Jeon et al., 2022], which employs GANs to generate retinal fundus images by using a local style alignment strategy to preserve the visual patterns of the original data. This method builds on the  $k$ -same framework [Meden et al., 2018], traditionally used for face deidentification. Expanding on  $k$ -SALSA, the PLAN [Pennisi et al., 2023] framework introduced a latent space navigation strategy, allowing for the generation of diverse synthetic samples suitable for downstream tasks. More recently, EchoNet-Synthetic [Reynaud et al., 2024] has developed a generation method based on diffusion models, establishing a protocol for anonymizing medical video datasets. These advancements demonstrate the cross-domain influence between face and medical image anonymization, highlighting the potential of generative models in preserving privacy without compromising data utility.

## 1.4 How to Check the Effectiveness of Anonymization?

To define anonymization, the European Data Protection Board (EDPB) outlined three risks in its opinion of 05/2014 [Party, 2014]. Consequently, a solution that effectively addresses these risks would mitigate the threat of re-identification attacks, rendering the data strictly anonymous. The three primary risks are outlined as follows:

- *Singling out*: the possibility to isolate some or all records which identify an individual in the dataset.
- *Linkability*: the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases).
- *Inference*: the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

The table below (1.4) provides an overview of the strengths and weaknesses of the techniques considered, evaluated against the three key requirements outlined by the EDPB.

**Tab. 1.4.:** Strengths and weakness of the pseudonymization and anonymization techniques in terms of privacy protection levels.

Technique		Singling Out Risk	Linkability Risk	Inference Risk
Pseudonymization		Yes	Yes	Yes
Anonymization	Noise addition	Yes	May Not	May Not
	Permutation	Yes	May Not	May Not
	Differential privacy	May Not	May Not	May Not
	Aggregation / K-anonymity	No	Yes	Yes
	L-diversity / T-closeness	No	Yes	May Not

## 1.5 Objectives and Organization of the Thesis

This thesis delves into the intricate relationship between medical data and the algorithms, as depicted in Fig 1.1, with a primary focus on the security and privacy aspects of medical imaging.



For the security of medical imaging, we investigate potential privacy breaches arising from the interactions between medical data and the algorithms. In a practical setting, a medical data lake (e.g., a hospital) holds the dataset, while a separate entity (e.g. a remote user or research institution) owns the algorithms. This separation of ownership between the dataset and the algorithms may initially appear as an effective way to preserve data privacy. However, a significant concern arises with algorithms, particularly deep neural networks, which can inadvertently encode sensitive information from the data. Such inadvertent leakage can lead to the unintended exposure of the entire original dataset. Thus, our research investigates the potential to reconstruct the original dataset using leaked information embedded within the algorithms. Our objective is to strike a delicate balance between malicious exploitation and preserving the utility of the stolen data.

Among various approaches to addressing privacy concerns in medical imaging—either protecting the algorithms or safeguarding the data—we concentrate on ensuring data privacy right from the outset of the dataset. Specifically, our focus lies on the anonymization of medical images. The primary objective of anonymization is to achieve a balance between the inherent trade-off of privacy and utility. To accomplish this, we explicitly address this trade-off by first identifying both identity and utility-related features within the data. We then focus on preserving the utility information critical for diagnostic purposes, while selectively removing or obfuscating the identity features to ensure privacy protection without compromising the overall data utility.

The manuscript is organized as follows, in accordance with the aforementioned research objectives:

In Chapter 2, we introduce a novel data exfiltration attack leveraging image compression techniques during the export of neural networks. This attack embeds information within the exported network, enabling the reconstruction of images originally stored in a data lake, outside of that secure environment. Specifically, we demonstrate that a network can be trained to perform both lossy and lossless image compression while simultaneously addressing utility tasks, such as image segmentation. The attack unfolds by exporting the compression decoder network along with specific image codes, facilitating the reconstruction of images outside the data lake. We validate the feasibility of this attack using CT and MR image databases, showing that it can produce perceptually meaningful reconstructions of the stolen dataset, which can then be used for a variety of tasks. Extensive experiments underscore that such data exfiltration attacks pose a significant threat to sensitive imaging data sources. This chapter also explores the application of differential privacy measures, like Gaussian noise addition, to mitigate the attack and examines how attackers might adapt to these defenses. Additionally, an alternative prevention strategy through model fine-tuning is proposed. This work was published at

the DeCaF workshop of the MICCAI conference in 2021 [Li et al., 2022a] and submitted to a journal [Li et al., ].

In Chapter 3, we introduce a new Generative Medical Image Anonymization framework, which is designed to address the identity-utility trade-off. Our key insight is to split the process of data anonymization into the following two distinct stages: an extraction of both identity and utility semantic features in images and then an optimization step of latent code to reach the desired trade-off. First, two identity and utility encoders are trained as components of classification networks that are aiming to verify patient identities and to detect pathologies in medical images. In addition, we train a generative adversarial network based auto-encoder to generate realistic images from a latent space. In a second stage, we use those two encoders in novel loss functions in order to generate synthetic images where the identity-related features have been removed while preserving their utility-related features. Through a comprehensive set of qualitative and quantitative experiments, we showcase the effectiveness of our approach on the MIMIC-CXR chest X-ray dataset by generating synthetic images that can serve as training set for detecting lung pathologies.

In Chapter 4, we extend the previous work by developing a streamlined encoder and co-training scheme designed to project input images into a more disentangled latent space. This encoder harnesses the power of StyleGAN2 [Karras et al., 2020], offering a more robust and scalable solution for generative medical image anonymization. As in the previous chapter, our framework operates in two stages: encoding and optimization. During the encoding stage, we accurately map input images into a latent space using our newly designed encoder. In the optimization stage, we implement a dual-loss approach—comprising identity removal loss and utility preservation loss—to refine the latent code, ensuring that anonymized images obscure identifiable features while maintaining critical diagnostic information. This carefully balanced strategy addresses the trade-off between identity removal and utility preservation. Through extensive qualitative and quantitative experiments on the MIMIC-CXR dataset, we demonstrate that our framework produces high-quality anonymized images while preserving essential diagnostic features, making them suitable for training machine learning models in lung pathology classification. This work will be published at the ISBI conference in 2025 [Li et al., 2025].

In Chapter 5, we summarize the contributions of this thesis, and discuss its potential outcomes, as well as the new perspectives and the future challenges that are left to explore.

## 1.6 Publications

The described contributions led to the following peer-reviewed publications in both conferences and journals.

### Journal Articles

- [Li et al., ] Li H, Ayache N, Delingette H. Data Exfiltration by Compression Attack: Definition and Evaluation on Medical Image Data. *Submitted to a journal.*

### Conference Papers

- [Li et al., 2022a] Li H, Ayache N, Delingette H. Data Stealing Attack on Medical Images: Is it Safe to Export Networks from Data Lakes? *International Workshop on Distributed, Collaborative, and Federated Learning. Cham: Springer Nature Switzerland, 2022: 28-36.*
- [Li et al., 2025] Li H, Ayache N, Delingette H. Generative Medical Image Anonymization Based on Latent Code Projection and Optimization. *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI). IEEE, 2025*

### Award

- Won the Best Paper Award at *International Workshop on Distributed, Collaborative, and Federated Learning. 2022.*



# Data Exfiltration by Compression Attack: Definition and Evaluation on Medical Image Data

---

2.1	Introduction . . . . .	14
2.2	Related Works & Background . . . . .	19
2.2.1	Network Pruning vs. Training a Smaller Network . . . . .	19
2.2.2	Steganography . . . . .	20
2.2.3	Privacy Defense . . . . .	21
2.3	Method . . . . .	21
2.3.1	Data Exfiltration by Compression Attack: Two Main Scenarios . . . . .	22
2.3.2	Lossy Compression and Utility Models . . . . .	24
2.3.3	Size Reduction of Compression Codes and Decoder Model . . . . .	25
2.3.4	Hiding Compression Codes in Exported Models . . . . .	26
2.3.5	Mitigation Measures . . . . .	27
2.3.6	DEC Attack Resisting to Differential Privacy . . . . .	28
2.4	Materials . . . . .	28
2.4.1	Dataset . . . . .	28
2.4.2	Pre and Post-processing . . . . .	29
2.4.3	Evaluation Metrics . . . . .	29
2.4.4	Compression and utility Models . . . . .	29
2.5	Results . . . . .	30
2.5.1	Attack Effectiveness . . . . .	30
2.5.2	EP Scenario vs. IT Scenario. . . . .	30
2.5.3	Type of Decoder: D vs D1. . . . .	30
2.5.4	Latent Channel Number Configurations. . . . .	32
2.5.5	Compression-fidelity Compromise. . . . .	32
2.5.6	Utility Task Performances . . . . .	33
2.5.7	Code Hiding Method . . . . .	34
2.5.8	Differential Privacy Mitigation . . . . .	36
2.6	Discussion and Limitations . . . . .	40
2.7	Conclusions . . . . .	43

---

**Abstract** With the rapid expansion of data lakes storing health data and hosting AI algorithms, a prominent concern arises: how safe is it to export machine learning models from these data lakes? In particular, deep network models, widely used for health data processing, encode information from their training dataset, potentially leading to the leakage of sensitive information upon export. This chapter thoroughly examines this issue in the context of medical imaging data and introduces a novel data exfiltration attack based on image compression techniques. This attack, termed *Data Exfiltration by Compression*, requires only access to a data lake and is based on lossless or lossy image compression methods. Unlike previous data exfiltration attacks, it is compatible with any image processing task and depends solely on an exported network model without requiring any additional information collected during the training process. We explore various scenarios, and techniques to limit the size of the exported model and conceals the compression codes within the network. Using two public datasets of CT and MR images, we demonstrate that this attack can effectively steal medical images and reconstruct them outside the data lake with high fidelity, achieving an optimal balance between compression and reconstruction quality. Additionally, we investigate the impact of basic differential privacy measures, such as adding Gaussian noise to the model parameters, to prevent the data exfiltration by compression attack. We also show how the attacker can make its attack resilient to differential privacy at the expense of decreasing the number of stolen images. Lastly, we propose an alternative prevention strategy by fine-tuning the model to be exported. This chapter was published as a conference paper in [Li et al., 2022a] and submitted to a journal [Li et al., ].

## 2.1 Introduction

The increasing use of AI technology on medical data for clinical research has stimulated the establishment of medical data warehouses or data lakes, where structured or unstructured data produced by major hospitals and health organizations, are stored and analysed. Of course, the access to these data lakes is heavily restricted and regulated, typically only allowing a remote access to external users. In general, remote data scientists are given access to a secured space where data curation and the training of machine learning algorithm can be performed.

The leakage of privacy-sensitive medical data from these data spaces poses a serious threat to the reputation of the health organizations managing a data lake. Cybercriminals could exploit the data leaks to harm third parties or ransom the health organization. To protect the privacy and security of these data lakes, it is crucial for data owners to

understand their potential vulnerabilities, prevent privacy attacks [Kaviani et al., 2022; Zhang et al., 2022a] and to develop effective mitigation measures.

Privacy attacks can be categorized into two main groups based on their objectives, as shown in Table 2.1. All attacks consider a *target dataset*, as the dataset that the attacker wants to steal or compromise, and a *target model*, as the machine learning model that the attacker aims to compromise or exploit. The first group of attacks focuses on the recovery of specific properties of the target dataset. Specifically, **Property Inference Attacks** [Zhang et al., 2022b; Wang et al., 2022b; Wang et al., 2022a; Zhang et al., 2021; Melis et al., 2019; Ganju et al., 2018; Ateniese et al., 2015] aim to infer additional properties of the target dataset, i.e. that are not explicitly correlated to the learning task, using only the parameters of the target model as prior knowledge. For instance, this includes the inference of the gender distribution in a patient dataset when the learning task is tumor segmentation. The **Attribute Reconstruction Attacks** (aka feature re-derivation, attribute inference) [Chen et al., 2022; Kaissis et al., 2020] aim to infer sensitive attributes of a target data from the knowledge of the target model, its output, and other non-sensitive attributes. In a **Membership Inference Attacks** (aka tracing attack) [Liu et al., 2023a; Liu et al., 2023b; Liu et al., 2023d; Zari et al., 2022; Hu et al., 2022], the attacker attempts to infer whether a particular data sample was used in the target dataset to train the target model. **Re-identification Attacks** [Chen et al., 2023; Ding et al., 2021; El Emam et al., 2011] determine an individual’s identity by linking the information from an pseudo-anonymised personal data to another dataset in which the same individual is contained.

Our work is primarily related to the second group of attacks, which aim at recovering the entire target dataset. These attacks exploit the output data generated by machine learning models to estimate the input target data. More precisely, **Gradient Inversion Attacks** (GIA) [Liang et al., 2023; Hatamizadeh et al., 2023; Hatamizadeh et al., 2022; Huang et al., 2021] leverage the gradients of the loss function with respect to the weights of the target model in order to reconstruct the training data. They exploit the gradient information to understand the relationship between the model’s output and input data, ultimately reconstructing the input data after benefiting from prior knowledge of the data domain. Different from GIA, **Model Inversion Attacks** (MIA) [Zhu et al., 2023; Nguyen et al., 2024; Chen et al., 2021; Fredrikson et al., 2015; Zhang et al., 2023], aka **Attribute Inference Attacks**, aim to reconstruct the target data from the target model’s output. Algorithmically, MIAs are designed as solving an optimization problem by estimating the input data that maximizes the likelihood of the output data.

**Data Exfiltration Attacks** [Ullah et al., 2018] (DEA) encompass various methods to leak data outside a data warehouse by an insider or an outsider of the data management organization. When dealing with machine learning algorithms, data exfiltration attacks aim at recovering the training data following the export of the model. Since a neural

**Tab. 2.1.1:** Privacy attacks categorized based on their target objectives: (1) targeting the recovery of specific properties (row 1-4) and (2) targeting the entire target data (row 5-7). These attacks depend on the attacker’s knowledge about the target model or the output of the model on the target data.

Attack	Adversary Knowledge				Adversary Output
	Target Model		Target Output	Additional	
	Architecture	Parameters			
Property Inference Attacks [Zhang et al., 2022b]	×	✓	×	×	Properties of target data
Attribute Reconstruction Attacks [Chen et al., 2022]	✓	✓	✓	×	Individual features
Membership Inference Attacks [Liu et al., 2023a]	✓	✓	✓	×	Membership of target data
Re-identification Attacks [Chen et al., 2023]	×	×	×	×	Individual identity
Model Inversion Attacks [Zhu et al., 2023]	✓	✓	✓	×	Reconstructed target data
Gradient Inversion Attacks [Liang et al., 2023]	✓	×	×	Gradients	Reconstructed target data
Data Exfiltration Attacks [Li et al., 2022a]	×	×	×	Compression Codes	Reconstructed target data

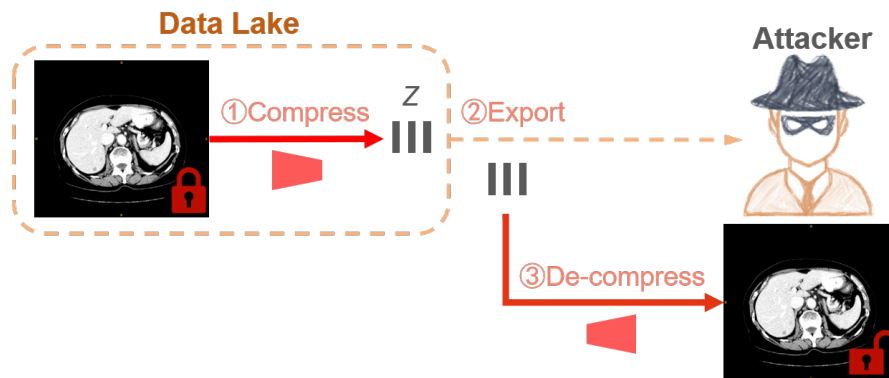


network, e.g. an autoencoder, somewhat memorizes its training data within its weights and as a result it is expected that the export of that network can be maliciously utilized to steal data. For instance, the technique of steganography was used to hide data or even malwares [Liu et al., 2020] inside the least significant bits of a neural network. Also, model inversion attacks solely based on the network weights was proposed in [Haim et al., 2022]. Under certain hypotheses about the network architecture (lack of skip connections) and training method (gradient based), the authors show that it is possible to reconstruct the training data given the trained model. However, the quality of the reconstructed training data is fairly limited and the proposed method is not suitable for U-net type image segmentation networks that include skip connections. Deep generative image models such as GANs or diffusion models are also likely to release their training data by using specific membership inference attacks [Carlini et al., 2023]. The authors show that GANs are less likely to generate their training data than diffusion models and that only a subset (typically 20%) of the original training data can be recovered. While the quality of recovered training images can be high, generative image models are by essence likely to lead to data theft and are therefore unlikely to be exported outside any data lake. Finally, transpose attack [Amit et al., 2023], was recently introduced to recover training image data by running a classification network backward. This approach was successfully tested on small size images and avoids to use multiple heads as in multi-task learning to hide a malicious model.

The GIA, MIA, and DEA aim at reconstructing the input training data, but they suffer from a number of shortcomings that limit their effectiveness in the context of medical data lakes. First, GIA and MIA require the attacker to have access to the gradient or the output of the target model but without accessing the input / target data. This situation is fairly uncommon except during the training stage of a network using federated learning [Li et al., 2020c]. Second, the quality of the reconstructed images is intrinsically limited for model inversion attacks or degrades rapidly with the number of memorized images increases for the transpose attack. Third, the attacks require that network to solve a specific task. For instance, model inversion and transpose attacks require the export of classification networks. Image generative models can exfiltrate high quality images but do not solve classical image processing tasks and can be easily detected. Finally, none of the previous attacks have been tested or evaluated on medical image datasets, which are typically large, often three-dimensional, and highly sensitive.

In this chapter, we introduce a novel attack coined as "**Data Exfiltration by Compression Attack**" which is widely applicable within medical data lakes, and that can lead to catastrophic data leakage. The Data Exfiltration by Compression (DEC) attack is a data exfiltration attack which is based on learned deep image compression networks (Fig. 2.1). The attacker compresses the target data into codes, and embeds them within the exported neural network, enabling the reconstruction of the target dataset outside the data lake. Despite the straightforward principle, the DEC attacker must cope with

several contradictory objectives: efficiently solving an image processing task, maximizing the number of stolen images, minimizing the size of the exported network, hiding the compression codes inside the network, and ensuring the compression codes are resistant to noise in the model.



**Fig. 2.1.:** Overview of Overview of Data Exfiltration by Compression Attack.

In the DEC attack, as with most other data exfiltration attacks, the exported network is intentionally designed to leak data whereas data leakage occurs unintentionally for the MIA and GIA. But unlike existing data exfiltration attacks, it can generate high quality and high resolution volumetric images and it is independent from the network architecture, and the task achieved by the model. Similarly to the transpose attack, in the external pre-training scenario, the exported network architecture is single headed and therefore cannot be discovered easily by the data owner. Besides, it is also resilient to the addition of Gaussian noise up to a certain level.

The DEC attack is evaluated in this chapter in the context of medical image analysis with the performance of image segmentation tasks. Please note that stealing medical images from data lakes is fairly challenging due to the size of the datasets (several tens of megabytes) and their subtle content. But the DEC attack principles could be extended to electronic health records, biosignals (e.g. ECG data), or biological data.

In this chapter, we have studied the different constraints associated with the DEC attack and proposed several mitigation plans. Our contributions can be summarized as follows:

1. We introduce a novel attack based on data compression which is agnostic to the architecture and task of the exported network. We introduce two distinct scenarios where the DEC attack is applicable, carefully balancing their inherent pros and cons. We explore how various hypotheses can influence the design of the attack. These hypotheses include the compression methodology (lossy vs. lossless), the ability to import a pretrained model into the data lake, the method of inserting

compression codes into the model, the addition of noise to the model, and the application of common mitigation measures.

2. We demonstrate that learned image compression methods, such as those in [Ballé et al., 2018], are well suited for: i) generating high-quality medical images with small compression codes, and ii) creating compression codes that are resilient to the addition of Gaussian noise in the network.
3. We conducted extensive experiments in real-world scenarios on two public datasets (LiTS and BraTS), focusing on stealing CT or MR images embedded within a segmentation network. To the best of our knowledge, this is the first time an exfiltration attack has been evaluated on a medical image dataset. We monitored the number and quality of stolen images, the efficacy of the segmentation network, and the size of the exported network. Our results show that the DEC attack poses a significant threat to medical data lakes, enabling attackers to efficiently balance a high number of stolen images with strong network performance in solving image segmentation tasks.

## 2.2 Related Works & Background

This section discusses prior work related to improving attack efficiency, specifically focusing on techniques to reduce network size and utilizing steganography to conceal codes within the exported network. Additionally, we briefly introduce privacy defense related to countering such attacks.

### 2.2.1 Network Pruning vs. Training a Smaller Network

Network pruning is a popular technique employed to reduce the computational cost of deep models in low-resource scenarios. Typically, pruning involves a three-stage pipeline: 1) training a large model, 2) pruning redundant weights based on a specific criterion while retaining important ones to maintain accuracy [Liu et al., 2018], and 3) fine-tuning the pruned model. When there are constraints on training resources, network pruning is often recommended to discover efficient architectures.

In our case, we explored various pruning techniques, including Structured pruning [Anwar et al., 2017], Unstructured pruning [Cheng et al., 2017], Quantization [Yang et al., 2019], and Knowledge Distillation (KD) [Gou et al., 2021]. However, Structured pruning, Unstructured pruning, and Quantization failed to identify efficient architectures due to the complexity of our model compared to baseline models typically used for parameter pruning. Similarly, KD did not yield did not produce satisfactory results in our

case, potentially due to the specific requirements of the reconstruction task. KD-based approaches are typically applied to classification tasks with a softmax function, which restricts their applicability. In our situation, using the teacher model's output as guidance for training the student model introduced lower-quality guidance, potentially misleading the student model in achieving high-quality image reconstruction.

Given the limitations of network pruning in our case, we opted to train a smaller model from scratch. Surprisingly, we made surprising observations that contradict common beliefs. First, training a smaller model from scratch achieved comparable or even higher accuracy. Additionally, the reduced model size facilitated training with less GPU memory and potentially faster compared to training the original large model. These unexpected findings open up new possibilities and challenge the conventional wisdom surrounding network pruning, showing the advantages of training smaller models from scratch in certain contexts.

## 2.2.2 Steganography

Steganography, as described by Morkel et al. [Ingemar et al., 2008], is the practice of concealing communication by hiding information within other data. By utilizing steganography, hidden messages can be embedded within carrier data, enabling covert communication that is difficult for observers to detect or decipher [Cheddad et al., 2010].

The process of steganography involves manipulating the carrier data in a manner that the alterations remain imperceptible to human senses. This is achieved by modifying specific features of the carrier data, such as subtly adjusting pixel values in an image or manipulating the least significant bits. As a result, the carrier data appears unaltered to casual observers while concealing the hidden message within.

Inspired by the imperceptible nature of steganography, we leverage it as a means to hide the compression codes within the exported network. By subtly adjusting the least significant bits of the network parameters, we can conceal the compression codes without modifying the size of the exported network. This approach allows for the secure transmission of compression codes out of the data lake without attracting attention or arousing suspicion from the data owner.

### 2.2.3 Privacy Defense

By understanding and addressing various privacy attacks as shown in Table. 2.1, we can develop robust defenses and preventive measures to ensure the privacy and confidentiality of medical data in data lakes.

In terms of privacy protection, a significant focus has been placed on Differential Privacy (DP) [Dwork, 2006]. DP offers privacy guarantees by introducing precisely calibrated noise to the input, model, or output of the data analysis process. Our experiments indicate that training models with DP is an effective defense mechanism in reducing the leakage of private information.

It is important to note, however, that the application of DP comes with a trade-off. While it effectively enhances privacy, the test accuracy of the target model is significantly impacted and may experience a noticeable decline. This degradation in accuracy is a challenge that needs to be addressed when considering the deployment of DP techniques. Balancing privacy protection and maintaining a satisfactory level of accuracy remains an ongoing area of research and development in the field of DP.

## 2.3 Method

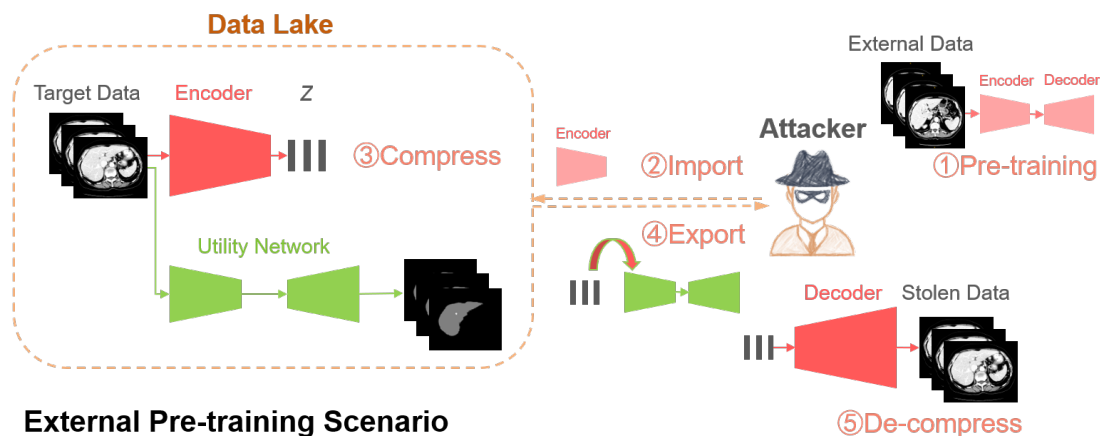
Data Exfiltration by Compression attack leads to the unauthorized leakage of sensitive data from a remote data lake. The attacker who carries out the attack, has access to the data lake, but with malicious intentions. The harmful objectives of the attacker may be for instance to steal data in order to create an external high-value dataset, to damage the data owner's reputation by publishing the stolen data, or to ransom the data owner. The attacker has access to the data lake either after getting an authorization by the data owner or after stealing the identity of an honest data lake user. Each data lake user has access to a secured demilitarized zone where data curation, model learning and testing take places without any possibility to export data. When a machine learning model reaches a certain performance level, the user can ask the data owner to export the trained model in order to test it on other proprietary or public data. The data owner must then decide what action should be taken to limit the risk of data leakage outside of the data lake following the export of the trained model. This chapter aims at evaluating the risk of data leakage of medical images from an healthcare organisation and at providing mitigation actions for the data owner.

The attack is based on data compression and takes place in three stages when exporting a trained model as seen in Fig. 2.1. First, the attacker encodes the target data into compression codes  $Z$ , and then the compression codes and the network are exported

outside the data lake. Finally, the attacker uses a decoder to decompress the compression codes, reconstructing the data into its original resolution and format.

Furthermore, to be realistic, the exported machine learning model should have a significant performance in order to convince the data owner that the original objective for accessing the data lake has been reached. The exported model solves a *utility task* which can be for instance an image segmentation, registration, detection or classification algorithm when processing medical images. A main advantage of the proposed attack is that the success of the attack does not depend on the chosen task or the network architecture.

We first present in Section 2.3.1, the two main scenarios of the attack depending whether the attacker is able to import a compression encoder or not. We then detail several key aspects such as the lossy compression network (in Section 2.3.2), the reduction of the exported model size (in Section 2.3.3), the hiding of compression codes and network models (in Section 2.3.4), the mitigation strategies (in Section 2.3.5), and the method to produce noise resilient attacks (in Section 2.3.6).



**Fig. 2.2.:** The pipeline of lossy compression based attack in External Pre-training Scenario, where the data owner has access to an encoder-decoder pair outside the data lake. Initially, an encoder-decoder pair is trained on an external dataset. The trained encoder is then imported into the data lake. Next, the encoder compresses the target data into compression codes  $Z$  while a utility network is simultaneously trained to conceal the attack behavior. Subsequently, the attacker exports the compressed codes and trained utility network from the data lake. Finally, with the exported  $Z$  and the trained decoder outside the data lake, the attacker can de-compress the stolen dataset.

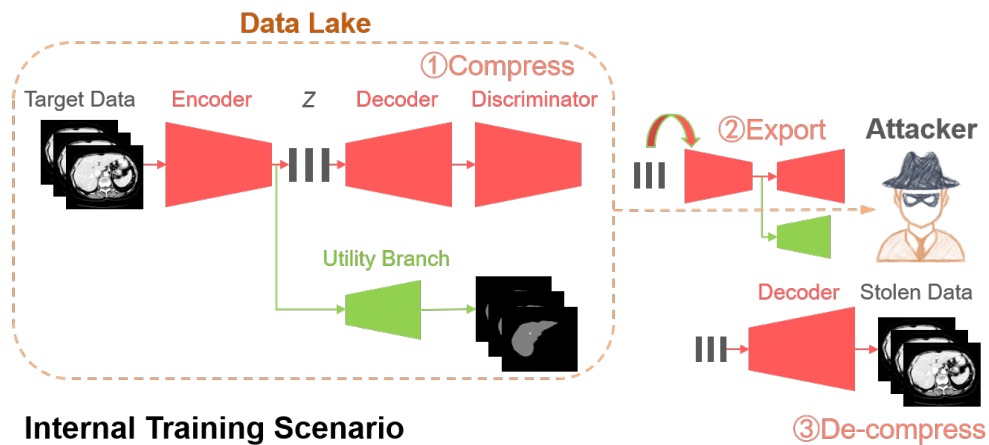
### 2.3.1 Data Exfiltration by Compression Attack: Two Main Scenarios

**External Pre-training (EP) Scenario.** In this first scenario, depicted in Fig. 2.2, the attacker creates an encoder-decoder pair outside the data lake, eventually by training

the pair on an external dataset. Then, the data owner allows the attacker to import the encoder inside the data lake, and the target data is then compressed by the encoder. To export the compression codes, the attacker trains a utility network that reaches the necessary performance and uses the network to hide the compression codes (see section 2.3.4 for more details) before the export. The attacker can then use the decoder outside the data lake to reconstruct the target data.

Thus, in this scenario, the attacker does not need to train and to export the decoder network outside the data lake. On the downside, the quality of the data reconstruction may suffer from a potential domain shift between the target and the external dataset from which the encoder was trained. Another issue for the attacker is to hide to the data owner, the real nature of the encoder network. The attacker may argue to import existing backbone or foundation models to boost the model performance or can resort to more sophisticated techniques such as hiding a model inside another model [Guo et al., 2020].

Lossless compression-based attack represents a specific configuration within this EP Scenario. In such case, the encoder-decoder pair can be a standard compression tool, such as ZIP / UNZIP. Note however that lossless image compression methods have inherent limitations in terms of compression ratio which restricts the amount of data that can be efficiently compressed. Nonetheless, in the remainder, we use the lossless compression-based attack as a baseline for comparison with lossy compression approaches.



**Fig. 2.3.:** The pipeline of lossy compression based attack in Internal Training Scenario, without access to an external encoder-decoder pair. First, an attack model trained inside the data lake compresses the target data into compression codes  $Z$ , utilizing a shared encoder with a utility decoder branch for hiding the attack behavior. Then, the attacker exports the utility network, attack decoder, and  $Z$  from the data lake. Finally, the stolen data can be de-compressed outside the data lake using the exported  $Z$  and the decoder.

**Internal Training (IT) Scenario.** In this second scenario (Fig. 2.3), we consider that the attacker does not have any access to a pretrained encoder-decoder pair unlike in

the previous scenario. Instead, the attacker trains the encoder-decoder pair inside the data lake, potentially utilizing adversarial autoencoders like VAE-GAN [Gao et al., 2020], deep compression networks [Agustsson et al., 2019], or a combination of both [Mentzer et al., 2020]. The attacker can then use the locally trained encoder to compress the target data. At this stage, the attacker must achieve 3 objectives: i) train a convincing utility model solving a specific task, ii) export the compressed data with the decoder, and iii) export the utility model. We propose an approach based on multi-task learning which solves the 3 objectives as can be seen in Fig. 2.3. More precisely, the utility model consists of the trained encoder as the backbone model combined with a dedicated *utility branch* to solve the utility task. The second branch including the decoder has the unique purpose to reconstruct the target data outside the data lake. An alternative design would be to create a utility network distinct from the encoder-decoder pair but this would have the disadvantage of potentially increasing a lot the size of the exported model. In [Amit et al., 2023], the authors argue that multi-headed networks can be easily detected as suspicious by the data owner. However, there are various ways to hide the network structure as discussed in [Guo et al., 2020] for instance. After training the utility branch, the attacker can hide the compression codes inside the two-headed network and ask the export of the resulting network to the data owner. Outside the data lake, the attacker can reconstruct the target data by leveraging the decoder branch and the compression codes.

This internal training scenario has the advantage of training a data specific encoder-decoder pair which is not subject to any domain shift. Please note that in this scenario, creating models that overfit the training data is a desirable property. However, this scenario requires the export of the decoder network which increases the risk of theft detection due to the increased size of the network and the two-heads in the network architecture. In Section 2.3.3, we explore various approaches to reduce the size of the decoder and the compression codes.

## 2.3.2 Lossy Compression and Utility Models

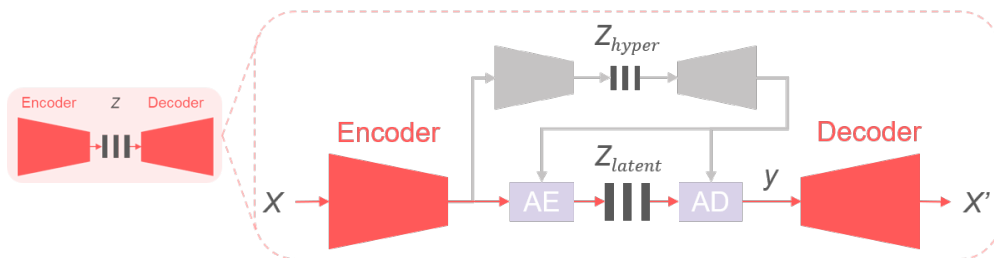
We detail below the lossy image compression model and utility branch / model implemented in this chapter. Motivated by the success of learned image compression techniques [Agustsson et al., 2019], we use the High Fidelity Compression (HiFiC) model proposed in [Mentzer et al., 2020] which combines GAN with learning based compression. The model includes an encoder that transforms an image  $x$  into a latent code  $y = E(x)$  and a decoder or generator which converts the latent code  $y$  into an approximation of the target image,  $x' = D(y) \approx x$ . Relying on adversary training, a discriminator  $\hat{D}(x')$  is used to estimate if the generated image is real or fake.



Thus the resulting architecture based on an encoder  $E$ , decoder  $D$  (aka the generator) and discriminator  $\hat{D}$ , is in some ways a mixture of a GAN (with  $D$  and  $\hat{D}$ ) with an auto-encoder (with  $E$  and  $D$ ). This architecture is suitable for learned image compression with high quality data reconstruction and it differs from other attack models [Zhu et al., 2023] (such as MIA) that are based on a GAN model with random noise as input.

The HiFiC model uses an hyperprior entropy model [Ballé et al., 2018] which helps to predict the distribution of the latent code (see Fig. 2.4). In fact, the (decoded) latent code  $y$  is generated by an arithmetic decoder  $AD$  which takes as input the latent variable  $Z_{\text{latent}}$  and is controlled by the transformed hyperlatent variables  $Z_{\text{hyper}}$ . As a result it is sufficient to store the pair of latent and hyperlatent variables  $Z = [Z_{\text{latent}}, Z_{\text{hyper}}]$  in order to generate the code  $y$  which then generates an image.

For the utility task, we primarily focus on image segmentation unlike most existing attacks that require solve classification tasks. Regarding the utility network in EP Scenario (Fig. 2.2), we employ a lightweight U-Net with 4 levels as the utility network, which takes target images as input and produces segmentation masks. In the IT Scenario (Fig. 2.3), the attacker uses the encoder of the generative compression model as the feature extraction network and trains a dedicated decoding branch specifically designed for the segmentation task. In this case the utility branch is the decoder part of a lightweight U-Net without any skip connections. This branch takes as input the output of the encoder, before the arithmetic encoding.



**Fig. 2.4.:** Operational diagrams of the learned image compression method using an hyperprior entropy model [Ballé et al., 2018]. The upper section corresponds to an hyperprior auto-encoder while the lower one shows an image auto-encoder architecture where AE, AD represent respectively arithmetic encoder and arithmetic decoder.

### 2.3.3 Size Reduction of Compression Codes and Decoder Model

To limit the risk of theft detection by the data owner, it is important to minimize the size of the compression codes in both scenarios, and the size of the decoder in the IT Scenario. To reduce the size of the compression codes  $Z$ , we can decrease the number of channels in both the hyperlatent  $Z_{\text{hyper}}$  and latent  $Z_{\text{latent}}$ . This reduction of the dimension of the

latent code directly decreases the overall size of the compression codes but also may potentially impact the quality of the reconstructed images.

To reduce the size of the decoder network  $D$ , we explored various classical model pruning techniques, including Structured pruning [Anwar et al., 2017], Unstructured pruning [Cheng et al., 2017], Quantization [Yang et al., 2019], and Knowledge Distillation (KD) [Gou et al., 2021]. However, Structured pruning, Unstructured pruning, and Quantization failed to create effective networks due to the complexity of our model compared to the ones typically used for parameter pruning. Similarly, KD did not produce any satisfactory results, potentially due to the fact that we were distilling a generative model instead of a classification model as commonly done.

Given the limitations of network pruning in our case, we opted for a more direct strategy. The original decoder  $D$  in the HiFiC model [Mentzer et al., 2020], is quite large with a size of 627 MB. A careful analysis of the architecture of  $D$  led to the observation that its size is mainly determined by 9 repeating ResNet blocks. We propose to replace the original decoder  $D$  with a smaller one  $D_1$ , where the 9 ResNet blocks have been removed, resulting in a much reduced model size of 29 MB.

### 2.3.4 Hiding Compression Codes in Exported Models

The data exfiltration by compression attack assumes that compression codes are hidden inside exported network weight files. More precisely, depending on the size of exported network, the compression codes can be hidden either in the least significant bits of the parameters or as an extra key-value pair in the parameter dictionary.

Indeed, if the attack model is large enough, it is possible to apply steganography [Ingemar et al., 2008] by storing some chunks of the codes in the least significant bits of the network weight file. After exporting the network outside the data lake, the compression codes can be easily extracted from the least significant bits of the checkpoint files, allowing the decoder to reconstruct the images. Using steganography has two major advantages. First, it does not affect the exported network size, and second it makes it very difficult for the data owner to detect the hidden codes.

When the exported model is small, steganography does not allow to hide enough data. In such case, the compression codes are stored in the checkpoint file as additional entries in a dictionary with dedicated keys, assuming for instance an HDF5 file format. This approach has the drawback of increasing the size of the exported network due to the added compression codes which may raise the suspicion from the data owner if the total size becomes excessive. Also, it is possible that the added dictionary entries may be detected since they are not connected to other nodes or layers in the exported model.

## 2.3.5 Mitigation Measures

To limit the risk of data exfiltration by compression attack or to detect the attack, data owners can apply a number of measures listed below.

**DEC Attack Prevention.** An obvious way to prevent the DEC attack is to forbid the export of any machine learning model outside the data lake. This drastic decision is however often not acceptable for the remote user who have developed the model. Alternatively, the data owner can implement several actions that limit the risk of DEC attack without eliminating it. First, the data owner should verify the integrity and honesty of the users that are granted access to the data lake. In addition, multi-factor authentication, adding an extra layer of security by requiring multiple forms of verification, should limit the risk for an attacker to steal the identity of an honest user. Third, data owners should closely monitor the machine learning models that are imported inside the data lake. Indeed the external pretraining scenario shows that importing a data compression encoder in the data lake makes the DEC attack very efficient (no decoder to export) and difficult to detect. Data owner should verify the origin of the model to be imported and its nature. Fourth, the data owner may control the last iterations of the model training (fine-tuning) before exporting the binary files outside the data lake. This allows the owner to verify that all weights of the model have been modified during training and that no additional data has been maliciously inserted into the model. This has the advantage of being relatively simple to implement by the data owner as it can be largely automated. A drawback is that the remote user has no control over the performance of the exported network.

Finally, the data owner can apply Differential Privacy (DP) [Dwork, 2006] by systematically adding noise to the exported model. The added noise aims to degrade or destroy the compression codes potentially stored in the exported model. Indeed, the addition of noise to lossless compression files makes the decompression impossible. Similarly, the noise added to the compression codes  $Z$  is also expected to have a significant effect since  $Z_{\text{latent}}$  is the input of an arithmetic encoder which is very noise sensitive. It is important to note however, that the application of DP must be calibrated since it effectively enhances privacy at the cost of degrading the performance of the exported utility model. In Section 2.3.6, we discuss how the attacker can adapt its methodology to cope with the addition of noise in the exported model.

**DEC Attack Detection.** The data owner can implement actions in order to detect that a remote user is performing a DEC attack. A first method consists in checking the source codes to find any manipulation of the neural networks such as the application of steganography or the addition of data in the dictionary file. This is obviously difficult to implement due to the time and expertise required. The second way to detect the attack

is to filter the exported networks depending on their size, and to carefully check the content of large neural networks that are exiting the data lake. Indeed, small networks cannot embed data of large size such as medical images.

### 2.3.6 DEC Attack Resisting to Differential Privacy

Knowing that the data owner is applying DP to the exported model, the attacker might opt for a compression method that is resilient to the addition of white noise, even if it means increasing the code size.

As discussed previously, the compression codes  $Z = [Z_{\text{hyper}}, Z_{\text{latent}}]$  are sensitive to noise addition since it can disrupt the precise intervals used for arithmetic decoding, potentially leading to the impossibility to produce any decoded data. However, the decoded latent code  $y$  following the arithmetic decoder as seen in Fig. 2.4, is by design much less sensitive to noise since it was shown in [Ballé et al., 2018] that  $y$  distribution closely resembles a Gaussian distribution. As a result, an effective strategy for the attacker to make network model resilient to Gaussian noise is to encode each image using the decoded latent variable  $y$  instead of  $Z$ , although its size is much larger than that of  $Z$ . In this case, the attacker sacrifices the quantity of the stolen images in favor of increasing their quality.

## 2.4 Materials

### 2.4.1 Dataset

To evaluate the feasibility of the proposed attack, we focus on a use case involving the storage of CT or MRI images in a medical data lake, with image segmentation as the designated utility task. Abdominal CT images are fairly large volumetric images (typically  $512 \times 512 \times 100$ ) encoded in 2 bytes per pixel and therefore are particularly challenging to exfiltrate.

The MICCAI 2017 Liver Tumor Segmentation (LiTS) Challenge dataset [Bilic et al., 2023] contains 130 abdominal CT cases for training and 70 CTs for testing. In this dataset, the utility task is to segment the liver in a supervised manner. The Multimodal Brain Tumor Segmentation (BraTS) Challenge 2021 dataset [Baid et al., 2021] includes 1251 skull stripped brain MR sequences for training and 219 cases for validation with the segmentation of the whole tumor from FLAIR MR sequences as the utility task. On the LiTS (resp. BraTS) dataset, we randomly partition the training set into 104/13/13 (resp. 1000/126/125) images that are used for training, validation, and testing of the utility

task. Also, for testing the attack model, we use the 70 (resp. 219) test images in the LiTS (resp. BraTS) dataset.

In the EP Scenario, two external datasets are required to pre-train the attack model outside the data lake which should closely resemble the two target dataset in order to minimize domain shift between the target and external datasets. As the external CT dataset, we chose the 2021 Fast and Low-resource semi-supervised Abdominal oRgan sEgmentation (FLARE) [Ma et al., 2022] dataset containing 361 abdominal CT cases for training and 50 abdominal CTs for testing. Similarly for MR, we used the 2022 Brain Tumor Sequence Registration (BraTS-Reg) Challenge [Baheti et al., 2021], having 120 cases for training, 20 cases for validation and 20 cases for testing. We used the same training and validation / test sets as those in the original FLARE and BraTS-Reg datasets.

## 2.4.2 Pre and Post-processing

In the LiTS and FLARE datasets, the network input size is  $512 \times 512$  pixels which is the same as the slice resolution. In the BraTS and BraTS-Reg dataset, edge padding is applied since the slice resolution is only  $240 \times 240$  pixels. Finally, a min-max intensity normalization is applied on the whole image. Since the HiFiC compression model takes RGB images as input, each slice is surrounded by its upper and lower slices to fill the three channels. For post-processing and display, all decompressed images are mapped back to their original minimum and maximum range.

## 2.4.3 Evaluation Metrics

To assess compression efficiency of the attack model, we used metrics such as BPP (bits per pixel) and a self-defined  $P_{\text{ratio}}$  which represents the ratio between the disk size of the compression codes generated by lossy compression and the disk size of the baseline lossless compression (as given by *gzip 3.6.0*). The metrics  $P_{\text{ratio}}$  provides an indication of the effectiveness of lossy compression compared to lossless compression. Additionally, to assess the reconstruction fidelity of the decompressed images compared to the original ones, we make use of the PSNR (Peak Signal-to-Noise Ratio) and MS\_SSIM (Multi-Scale Structural Similarity) [Wang et al., 2003] metrics. Finally, we evaluate the image segmentation performance of the utility model with the Dice score.

## 2.4.4 Compression and utility Models

The HiFiC image compression model is borrowed from [Mentzer et al., 2020] while the utility model is a simplified version of the network in [Li et al., 2020b]. The same

HiFiC model architecture is used in both the external pretraining and internal training scenarios. All models are optimized with the Adam [Kingma et al., 2014] algorithm and training continues till validation loss has converged.

## 2.5 Results

### 2.5.1 Attack Effectiveness

We assess the potential trade-offs faced by the attacker when crafting its attack model. These trade-offs encompass maximizing the quality of the compressed images, minimizing the size of the compression codes, minimizing the overall model size, and ensuring the utility task is effectively fulfilled.

To optimize those metrics, the attacker can modify the architecture of the encoder-decoder model, and select a specific exfiltration scenario. Regarding the model architecture, the attacker may either choose the full size original decoder  $D$  (627 MB) or its reduced version  $D1$  (29 MB) (see Section 2.3.3). In addition, the number of latent and hyperlatent channels may be decreased to reduce the compression code size. We write as  $C_{|Z_{\text{latent}}|, |Z_{\text{hyper}}|}$  the network configuration where the number of latent (resp. hyperlatent) channels is  $|Z_{\text{latent}}|$  (resp.  $|Z_{\text{hyper}}|$ ).

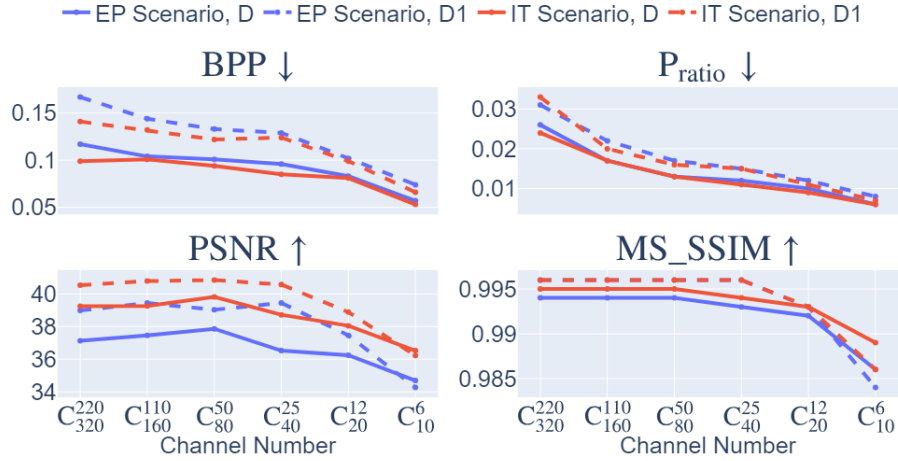
In Fig. 2.5, the attack performance is evaluated in terms of compression efficiency (BPP,  $P_{\text{ratio}}$ ) and reconstruction fidelity (PSNR, SSIM) across different network architectures and scenarios. The key results are summarized below.

### 2.5.2 EP Scenario vs. IT Scenario.

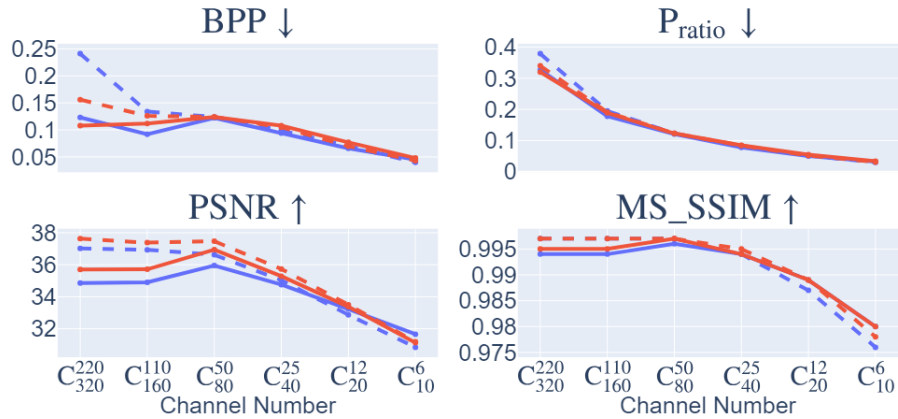
The internal training (IT) scenario leads to improved image quality compared to the external pretraining (EP) scenario. This can be explained by the potential domain shift between the external dataset and the target dataset. In terms of model size, however, the two scenarios are equivalent.

### 2.5.3 Type of Decoder: D vs D1.

In Fig. 2.5, we observe that the reduced decoder D1 (29 MB) surprisingly outperforms the original large decoder D (627 MB) both in terms of quality of reconstruction, and compression efficiency, and also for both scenarios. Thus, stripping the ResNet blocks from decoder D leads to improved image quality, which can be explained by a better balance in terms of model size among the encoder, decoder, and discriminator components.



(a) CT images.

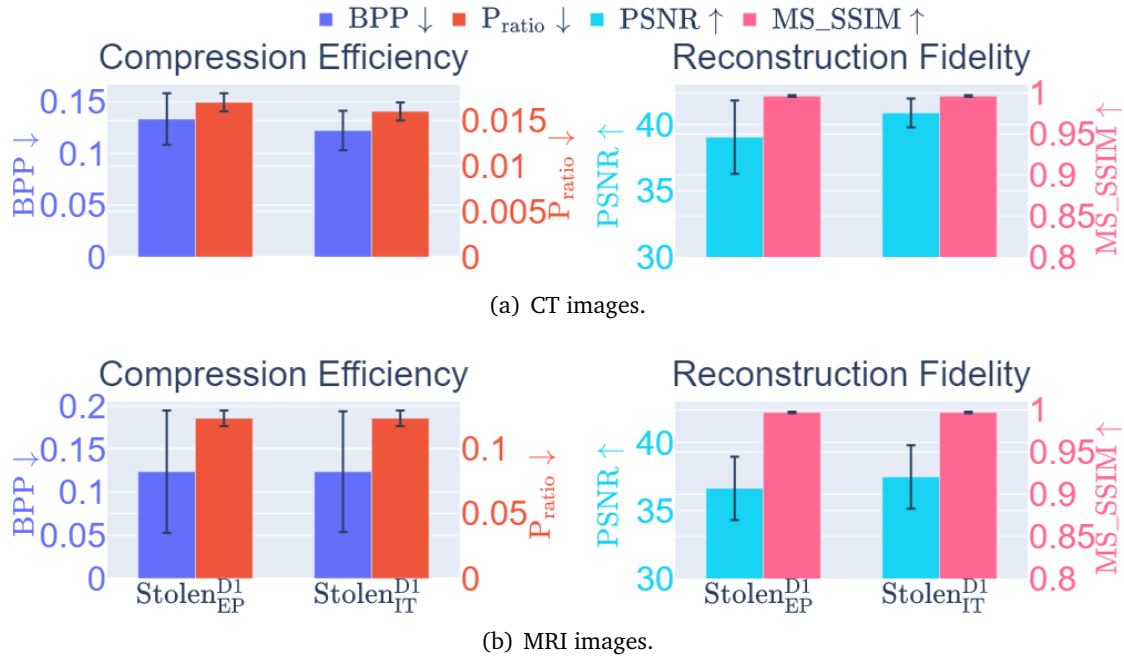


(b) MRI images.

**Fig. 2.5.:** Lossy compression based attack on CT images (a) and MRI images (b) with various channel numbers  $C_{\text{hyper}}^{\text{latent}}$  in x axis. The compression efficiency is measured using the criteria of BPP and  $P_{\text{ratio}}$  in the first row, while the reconstruction fidelity is assessed using the criteria of PSNR and MS\_SSIM in the second row.

## 2.5.4 Latent Channel Number Configurations.

As the number of channels of the latent/hyperlatent variables decreases, we observe an increase in the compression efficiency of the attack model but also a reduction of the reconstruction fidelity. The quality of reconstructed images decreases more sharply when the channel number is equal or below  $C_{40}^{25}$ . Thus, we found that setting the channel numbers at  $C_{80}^{50}$  resulted in a good trade-off between compression efficiency and reconstruction fidelity.



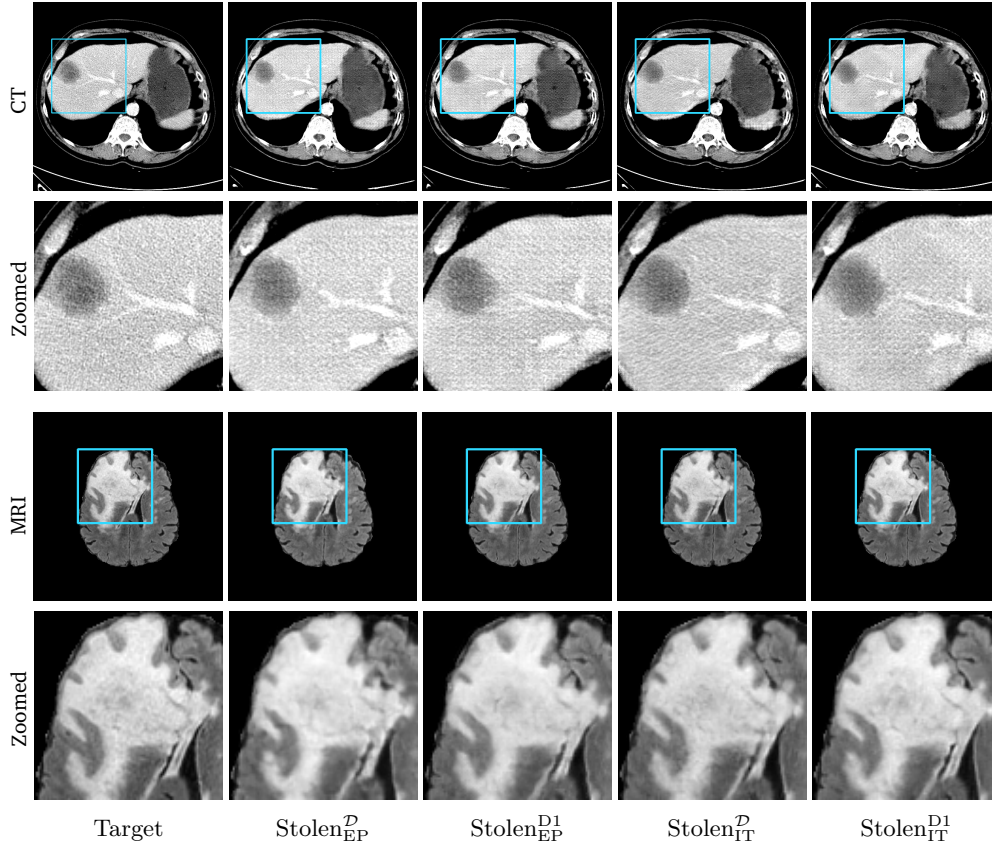
**Fig. 2.6.:** Lossy compression based attack on CT images (a) and MRI images (b) with a specific channel numbers ( $C_{80}^{50}$ ) for the EP and IT Scenarios.  $Stolen_{EP}^{D1}$  denotes the decompressed images in the EP Scenario with a reduced D1 decoder.

## 2.5.5 Compression-fidelity Compromise.

Based on the previous results, we have selected the decoder  $D1$  and the configuration of the latent and hyperlatent variables  $C_{80}^{50}$ , as the optimal architecture of the HiFiC encoder-decoder pair. In Fig. 2.6, we display more specifically the compression and reconstruction quality performances. In terms of reconstruction quality, we obtain a PSNR of approximately 40 for CT images and around 38 for MRI images while the MS\_SSIM values are close to 1. This indicates an excellent perceptual quality of the reconstructed images that are hardly discernable from the original ones. In terms of compression efficiency, the  $P_{ratio}$  for CT images is approximately 0.015, indicating that the lossy image compression-based attack generated compressed images are 67 times smaller than those produced by the lossless zipped image compression-based attack. For MRI images, the  $P_{ratio}$  is around 0.12, 10 times higher than that of CT images, which can



be attributed to the presence of a large uniform background in the skull-stripped original MR images.



**Fig. 2.7.:** Lossy image reconstructions on CT (row 1, 2) and MRI (row 3, 4) images, where the row 2, 4 provide a zoomed-in view of the bounding box region of the row 1, 3. The leftmost column represents the target images, while the subsequent four columns show the stolen images reconstructed by the decoder  $D$  or  $D1$  in two scenarios.

A visual comparison between target and stolen images is available in Fig. 2.7. We observe that the stolen images from IT Scenario closely resemble the input ones, particularly in the tumor regions, whereas stolen images from EP Scenario exhibit blurring artifacts in finer details. In both cases, the stolen images reconstructed by  $D1$  demonstrate a comparable quality to those reconstructed by  $D$ , thus further confirming the effectiveness of the reduced decoder  $D1$ .

## 2.5.6 Utility Task Performances

To solve the utility task, in EP Scenario, we employed a whole dedicated U-Net network (UN as Utility Network), whereas in the IT Scenario, we utilized a Utility Branch (UB) network. These models were trained either on the target dataset from the data lake or on the stolen dataset reconstructed by the attacker. Subsequently, they were tested on an

unseen test dataset (a subset of the target dataset). The UN model, trained on the target dataset, also serves as a baseline model for comparison.



**Fig. 2.8.:** Utility task results on CT and MRI images. The UN and the UB models trained on the target dataset ( $UN_{EP}^{Target}$ ,  $UB_{IT}^{Target}$ ) are utilized to execute the utility task within the data lake. The UN model trained on the stolen dataset ( $UN_{EP}^{Stolen}$ ,  $UN_{IT}^{Stolen}$ ) is to evaluate the practical utility of the stolen dataset in solving the same utility task.

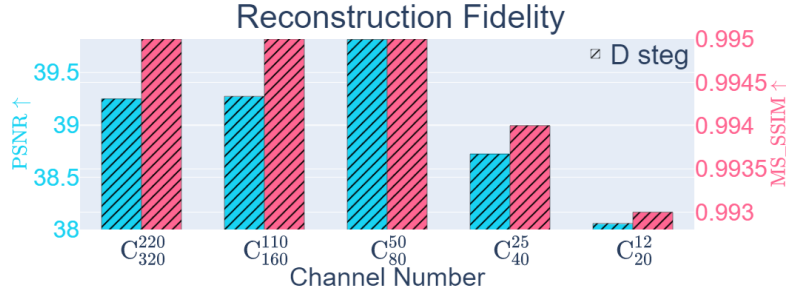
As shown in Fig. 2.8, both the UN and UB models have good results on both the target and stolen dataset. This means that in both scenarios the attacker can provide convincing evidences to the data owner that the model to be exported is effective. This also indicates that the stolen image data is indeed useful for training a model for solving the same utility task.

## 2.5.7 Code Hiding Method

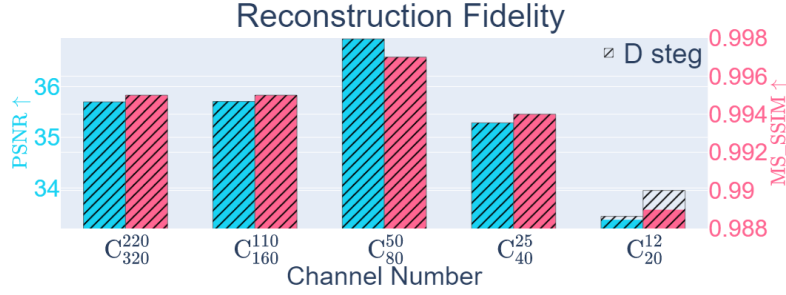
To hide the compression codes within the exported network, the attacker can either use steganography or add entries in the file dictionary (see Section 2.3.4). The choice may depend on the size of the decoder: steganography is probably preferred if the decoder is large enough and otherwise the codes are stored as additional dictionary entries.

For the steganography method, the values of each parameter are converted from float32 into binary (32 bits), and we decided to allocate the least significant 16 bits to store the compression codes. This approach does not increase the size of the exported network, but it requires that the size of the exported network be at least twice as large as that of the compression codes. This is why it is well suited in the case where a full-size decoder (D) is used for the IT Scenario. However, steganography modifies the parameter values of the model and it may impact the model accuracy if a large number of bits are allocated. It is therefore essential to verify that the performance of the exported model is not significantly affected by steganography.

We evaluated the performance of D based attack models before and after applying steganography to conceal the code. As shown in Fig. 2.9, there is no discernible impairment in the compression and reconstruction performance of the attack models after using steganography. This is further supported by the visualizations in Fig. A.1, which



(a) CT images.



(b) MRI images.

**Fig. 2.9.:** Comparison of the lossy compression-based attack on CT and MRI images before and after applying steganography. The bars colored in solid indicate the results before steganography, while the bars with slashes represent the results after steganography.

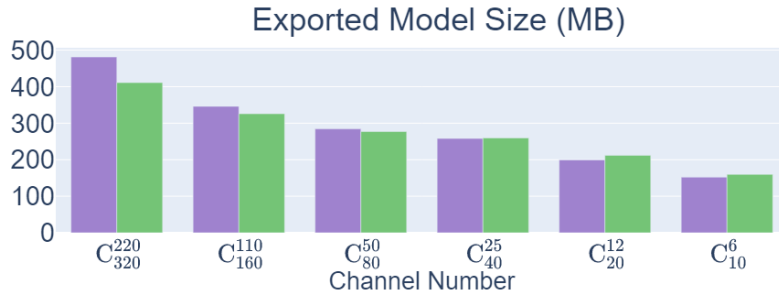
display the reconstructed images after applying steganography, showing minimal impact on image quality.

**Tab. 2.2.:** The size (MB) of encoded compression codes of decoder D in IT Scenario when stealing 100 images.

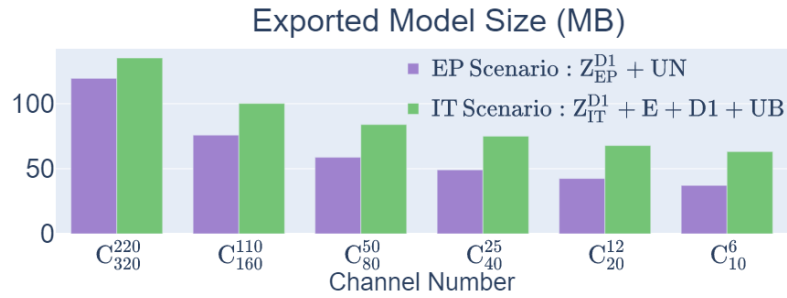
Data	Model					
	$C_{320}^{220}$	$C_{160}^{110}$	$C_{80}^{50}$	$C_{40}^{25}$	$C_{20}^{12}$	$C_{10}^6$
CT	230.504	179.866	153.824	132.260	121.610	81.728
MRI	44.153	28.425	21.078	15.771	10.513	7.318

Table 2.2 shows that the size of the compression codes when stealing 100 images, is consistently smaller than half the size of D (627 MB), confirming that the full decoder  $D$  has sufficient capacity for steganography. Moreover, we found no noticeable degradation of the compression and reconstruction performance of the HiFiC model when allocating 16 bits for data exfiltration. But when allocating more than 16 bits, image artifacts start to appear in stolen images, confirming that it corresponds to a good compromise.

In the dictionary method, the compression codes are stored as a key-value pairs within the checkpoint of the exported network. This does not alter the network parameters, but increases the size of the exported network. In Fig. 2.10, the size of the exported model when exfiltrating 100 CT or MR images are displayed as a function of the number of latent channels and scenario type. In the EP Scenario, one needs to export the codes and utility network whereas in the IT scenario one has to also export the encoder, the



(a) CT images.



(b) MRI images.

**Fig. 2.10.:** The size of the exported network when stealing 100 images with the dictionary approach. In the EP Scenario, the exported network includes the compression codes  $Z$  and the utility network UN, while in the IT Scenario, it includes the compression codes  $Z$ , the decoder  $D1$ , and the encoder  $E$  and utility branch UB from the data lake.

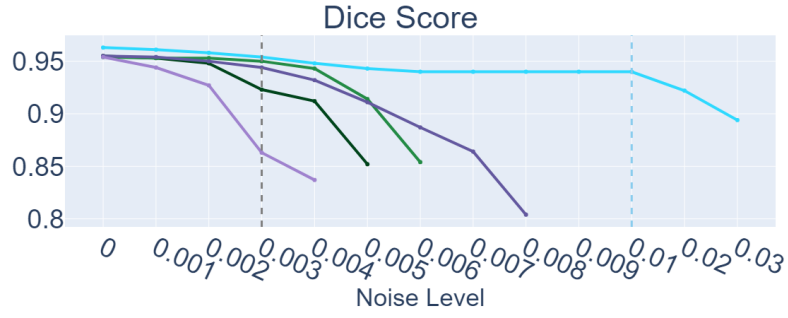
reduced decoder  $D1$ . The exported network is always below 500 MB, and even less than 140MB for MR images. Given that commonly used backbone models often exceed 500 MB in disk size (e.g., VGG16 is 576 MB), the size of the exported model is unlikely to raise suspicion from the data owner.

## 2.5.8 Differential Privacy Mitigation

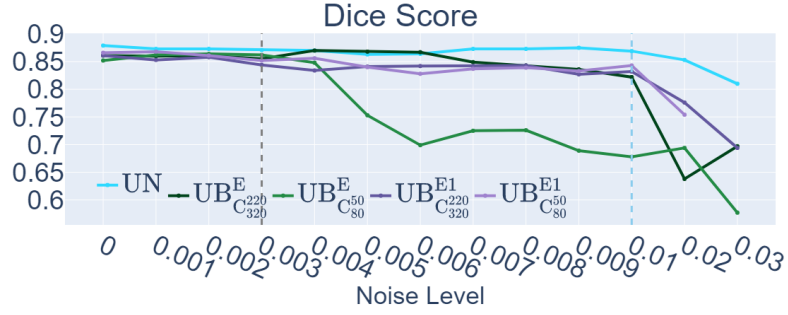
We study below the impact of differential privacy on the DEC attack with and without decoded latent variables. We consider a basic DP protocol that adds calibrated Gaussian noise with a zero mean and specified standard deviation to the exported model.

### Noise Level Calibration.

It is crucial to carefully calibrate the noise standard deviation, as model performance is expected to decline with increasing noise levels. In Fig. 2.11, we visualize the mean Dice score of the utility network (UN) or utility branch (UB) as a function of the noise level (i.e. standard deviation for Gaussian noise), the network architecture and the type of image to encode. We observe that a noise level less than 0.01 preserves the model performance in EP Scenario whereas a level of 0.002 (resp. 0.003) is a limit for the UB model of the IT Scenario trained on CT (resp. MR) images. The difference of noise



(a) CT images.



(b) MRI images.

**Fig. 2.11.:** Utility task performances on CT and MRI images with varying noise levels applied to the utility model (UN for EP Scenario and UB for IT Scenario). For the UB model, several encoders and channel numbers are considered.

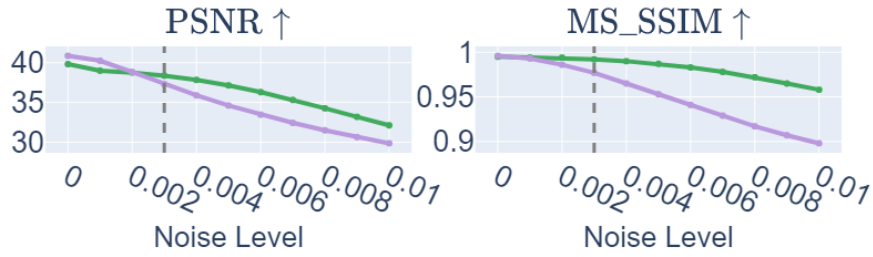
sensitivity between the two scenarios could be explained by the fact that the encoder is not jointly trained with the UB in the IT scenario.

### Resilience of the DEC Attack to DP.

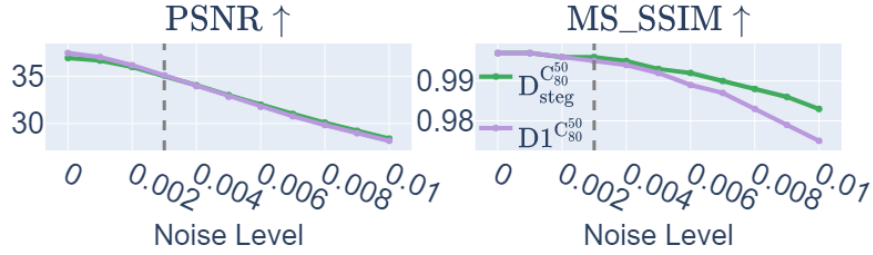
As detailed in the Section 2.3.6, the attacker knowing that the data owner applies DP, is likely to adopt a different strategy by exporting the decoded latent variables instead of the encoded latent and hyperlatent variables, at the expense of stealing less images. We have performed several experiments to assess the resilience of the attack both in terms of quality and quantity of the stolen images.

In the EP Scenario, the performance of the attack model seems unaffected by the application of noise levels less than 1%. We have consistently obtained a PSNR of 39.0 (resp. 36.9) and MS\_SSIM of 0.996 (resp. 0.997) for CT (resp. MR) images. This resilience can be attributed to the inherent nature of the GAN-based HiFiC model where the decoded latent codes follows a Gaussian distribution with a standard deviation larger than 1%.

In IT Scenario, the noise addition perturbs both the compression codes and the decoder. From Fig. 2.12, we see that the image quality of the stolen images remains acceptable when the noise level is below 0.003 (PSNR > 35 and MS\_SSIM > 0.98) for both image



(a) CT images.



(b) MRI images.

**Fig. 2.12.:** Lossy compression-based attack on CT and MRI images before and after applying Differential Privacy in IT Scenario, with varying noise levels applied to the attack decoder and the compression codes.

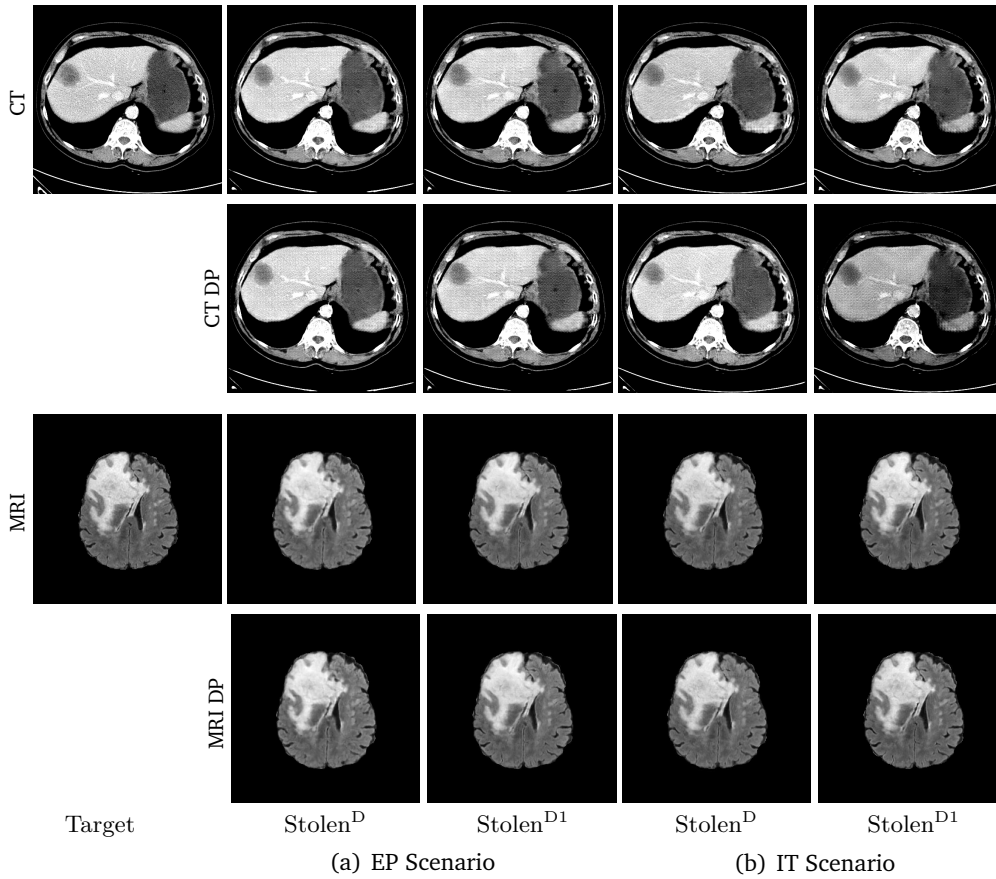
modalities. Besides, the architecture with the reduced decoder  $D1$  is more sensitive to noise than the original full-size one. This is due to the dilution of noise created by the large ResNet blocks.

It should be noted that this threshold of 0.003 is roughly the same as the noise level acceptable for the utility branch as displayed in Fig. 2.11. This implies that the data owner is likely to apply a noise level of 0.003 or lower in order to preserve the performance of the exported network. But with this level of noise, the data owner will also preserve the quality of the exfiltrated images.

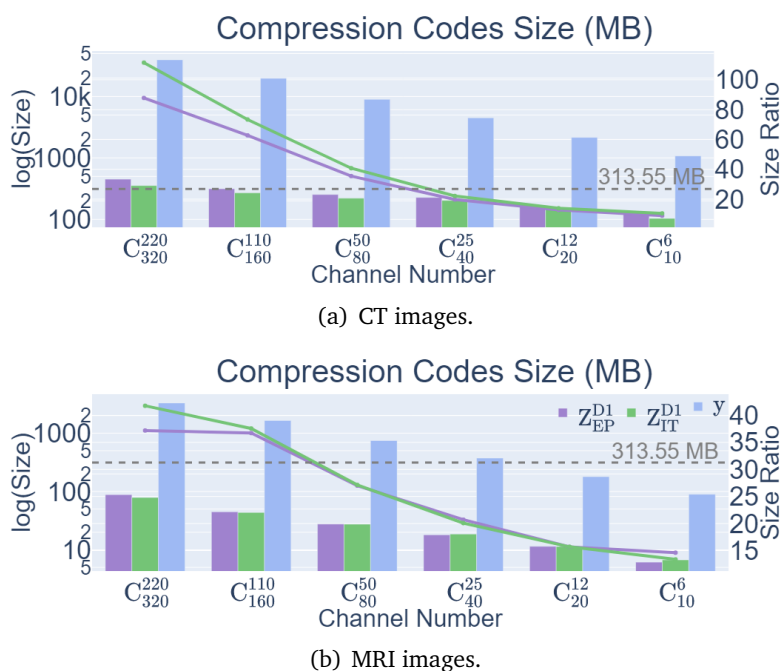
The visual comparison of exfiltrated images before and after applying DP is displayed in Fig. 2.13. In the EP Scenario, no noticeable differences are observed but in the IT Scenario, we observe slight variations of image intensity in the stolen CT images generated by the decoder  $D1$ . For MR images no discernible differences are present. Additional visual examples are provided in Fig. A.2 and Fig. A.3, offering further insights into the effects of steganography on the reconstructed images.

### The Size of the Exported Network.

Exporting decoded latent variables  $y$  instead of encoded variables  $Z$  results in an increase in the size of the compression codes. This phenomenon is visible in Fig. 2.14, where we show the size ratio  $|y|/|Z|$  is around 40 for CT images and 30 for MRI images for the specific channel numbers ( $C_{80}^{50}$ ). Specifically, the number of exfiltrated images is reduced by 40-fold for CT images and 30-fold for MRI images to ensure resilience against



**Fig. 2.13.:** Lossy image reconstructions of CT and MRI images before (row 1, 3) and after (row 2, 4) applying DP with calibrated noise levels. The first column displays the target images. The middle two columns show the results for the EP Scenario, while the last two columns is for the IT Scenario. For each scenario, we show the stolen images reconstructed by the decoder D or D1, respectively, utilizing a specific channel numbers ( $C_{80}^{50}$ ).



**Fig. 2.14.:** The size of the encoded and decoded compression codes. The bar chart with the left y-axis represents their sizes when stealing 100 images, while the line chart with the right y-axis illustrates the size ratio between  $y$  and  $Z$ .

differential privacy measures. In a realistic scenario, if the exported model has disk size around 627 MB (using the full-size decoder  $D$ ), then we can exfiltrate 313.55 MB of compression codes (displayed by the gray dashed line in Fig. 2.14), thus corresponding to the exfiltration of only 4 CT images or 42 MRI images after resorting to steganography.

## 2.6 Discussion and Limitations

### Discussion.

We have demonstrated that the DEC attack is realistic in the context of a medical data lake where the remote user is given access to CT or MR medical images with a utility task consisting in solving an image segmentation task. In both scenarios, an attacker can exfiltrate a large set of medical images when exporting a network solving a utility task (such as image detection, classification, registration, segmentation) outside the data lake. The attack feasibility is mainly a consequence of the two following facts : i) neural networks for image analysis tend to be large in disk size, for instance with backbone models being larger than 500MB, and ii) current deep compression networks allow to reach an excellent fidelity / compression compromise. We discuss below several key topics about the DEC attack.

**IT vs EP Scenarios.** The external pretraining scenario where a data encoder is imported inside the data lake is far simpler to handle for an attacker than the internal training

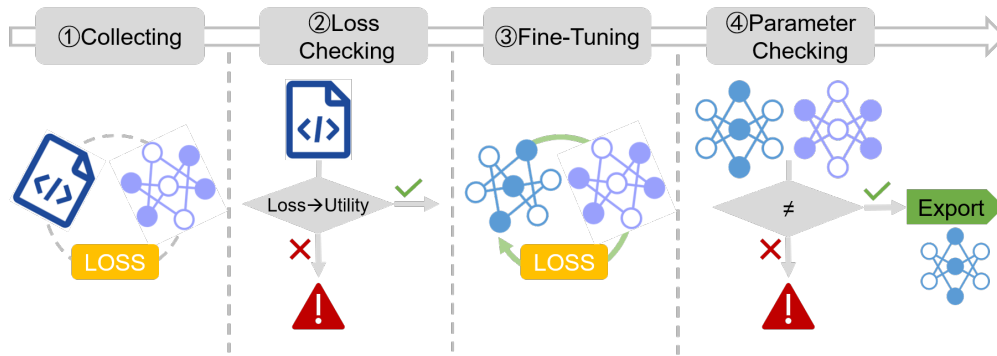


scenario. Indeed, in this case, the attacker has only to produce the compression codes and hide them inside the exported network. The risk of detection is mainly related to the import of the encoder that must be hidden to the data owner. In the alternative IT scenario, the risk of detection occurs during the export of the model due to the dual heads of the model. Both risks are related to the ability of the attacker to hide the real nature of the model.

**Nature of Utility Network.** While previously proposed attacks such as the model inversion [Zhu et al., 2023] or transpose [Amit et al., 2023] attacks were assuming that the exported model was a classifier (without skip connections), the DEC attack does not put any constraint on the architecture of the utility model or the utility branch. In the EP scenario, the network is only a recipient to hiding compression codes, while in the IT scenario, we have shown that the data encoder can serve as a backbone from which a task specific head (solving image segmentation in our case) can be connected. This greatly expands the applicability of the attack compared to previously proposed ones.

**Effective Compression Network.** Based on the HiFiC architecture, we have optimised the performance / size ratio of this deep compression network based on a reduced decoder  $D_1$ , and a limited number of channels of the latent and hyperlatent variables. With this configuration, our compression model was able to produce CT and MR images of high quality with compressed files that are around 30-50 times smaller than the size obtained with lossless compression (e.g. gzip algorithm) with a decoder weighting less than 30MB. The research on universal deep image compression [Tsubota et al., 2023] is very active, and it is expected that the trade-off between fidelity and compression will continue to improve favorably with higher quality images from smaller codes. For instance, one way to improve the HiFiC network would be to design a more efficient entropy coding [Hu et al., 202] module, for instance by incorporating a deep prior about the distribution of the latent variables  $Z_{\text{latent}}$ .

**Resilience to Differential Privacy.** Given the considered scenarios, we have studied the addition of noise on the model parameters rather than on the images or the gradient during training. The addition of Gaussian noise to the exported model has a drastic effect on the DEC attack, making it impossible to reconstruct any image, if the hidden compression codes  $Z_{\text{latent}}$  are the encoded latent variables. However, when exporting the decoded latent  $y$  instead of  $Z_{\text{latent}}$ , the attacker can reconstruct a limited number of stolen images with sufficient quality for a certain range of noise level. Besides, we found that the amount of noise necessary of prevent image reconstruction is higher than the noise level seriously decreasing the performance of the utility model. This leads the data owner with a difficult dilemma: either exporting a secured but seriously limited model, or exporting an effective model but with a risk of data exfiltration. Obviously, the compromise made by the data owner depends on the trustworthiness of the data user.



**Fig. 2.15.:** Flowchart for preventing data exfiltration through fine-tuning. The data owner implements this prevention strategy within the data lake in four steps. ①Collecting: Gather the source code for the model and the loss functions required for fine-tuning. ②Loss Checking: Review the source code and confirm whether the loss functions align with the utility task objectives. If they do, the process proceeds to the next step; otherwise, an alert is triggered. ③Fine-Tuning: Refine the model until all parameters have significantly deviated from their initial values. ④Parameter Checking: Check if all parameters have changed after fine-tuning. If any remain unchanged, an alert is raised. If all parameters are modified, the fine-tuned model is approved for export from the data lake.

**Prevention based on Model Fine-tuning.** To mitigate the risk of data leakage when exporting a machine learning model from a data lake, we propose an alternative to differential privacy based on model fine-tuning. The prevention process, as illustrated in Fig. 2.15, involves four key steps executed by the data owner within the data lake. First, the data owner collects the source code for the model intended for export and the associated loss functions required for fine-tuning on the training data. Second, the source code is reviewed to ensure that the loss functions align with the utility task objectives. If alignment is confirmed, the model is fine-tuned until all model parameters have been sufficiently modified. If some parameters have not been modified after a fixed number of epochs, then an alert is raised. If all stages have been cleared, the model is exported. Otherwise, the data owner does not proceed with the model export and start discussing with the data user about potential security issues in the model.

We believe that this export process enforces privacy with existing data exfiltration attacks while preserving the utility of the model. Indeed, similarly to differential privacy, the proposed process modifies the model parameters thus making it impossible to hide compression codes with it. Besides it allows to detect secondary branches in the utility model that are not modified during the fine-tuning but that can serve to decode compression codes. Finally, by checking the nature of the loss functions optimized during fine-tuning, the data owner can check the relevance of those objective functions with respect to the utility task and at this occasion, can detect most attacks including the transpose attack [Amit et al., 2023]. The advantage of this Fine-Tuning prevention is that data privacy is not obtained at the expense of the model performance, as limited fine-tuning over a few epochs should not substantially alter its effectiveness. Additionally, it is not

intrusive for the data user, but it does require the data owner to have the capability to understand the code related to the loss functions, fine-tune the model, and monitor any modifications to the model parameters.

### **Limitations.**

There are limitations associated with the Arithmetic Coding module. First, reducing the channel numbers of both hyperlatent  $Z_{hyper}$  and latent  $Z_{latent}$  may not be the most effective approach for reducing the size of compression codes  $Z$ . A more suitable strategy would involve designing a more efficient entropy coding approach to enhance compression efficiency. For instance, incorporating a deep prior could effectively estimate the probability of latent  $Z_{latent}$ , instead of relying on the simplified assumption that it follows a Gaussian distribution. Second, the Arithmetic Coding renders the attack approach susceptible to mitigation strategies such as DP. The addition of random noise can disrupt the precise intervals used for decoding, potentially rendering it impossible to produce any decoded data. Hence, it would be advisable to replace Arithmetic Coding with a more robust noise-resistant approach.

## **2.7 Conclusions**

In this chapter, we have proposed a data exfiltration attack on medical images consisting in using pretrained or learned compression / decompression algorithms. We have shown that this is a realistic and effective attack if a remote user acts as an attacker. We have demonstrated that by combining different methods (training smaller model from scratch, steganography, latent channel downsizing...) it is possible to decrease the model size, and the compression codes while preserving the quality of stolen images and exporting an effective utility model. For instance, with learned lossy image compression, it is possible to obtain compression codes that are 60 times smaller than the ones obtained with baseline lossless compression. The number of images that can be stolen is on the order of at least one hundred of images for an exported model size of 300 MB. Importing a pretrained model in the data lake does not have a large influence on the attack performance compared to the internal training scenario. Finally, the quality of the stolen images with this attack is high (MS\_SSIM around 0.996) and can be used to successfully train an utility model outside a data lake.

To prevent this attack, the data owner should use multi-factor authentication combined with some monitoring of the user activity inside the data lake, although it may be difficult to implement efficiently. The use of different privacy is effective when encoded latent and hyperlatent codes are exported within the model. But, we have shown that the attacker can make the attack resilient to a simple differential privacy protection by exporting decoded latent variables instead of encoded latent and hyperlatent variables. Finally, we

have introduced an export process that can be implemented by a data owner to prevent the data exfiltration by compression attack. This process is based on the fine-tuning of a model until all model parameters have significantly changed, and the inspection of the source codes providing the loss functions.

Future work will concentrate on the implementation and evaluation of fine-tuning based attack prevention, but also will explore other privacy preserving technique such as knowledge distillation [Li et al., 2022b], homomorphic encryption or the generation of anonymized imaging data from original ones.

# GMIA: Generative Medical Image Anonymization Based on Identity-utility Optimization

---

3.1	Introduction . . . . .	46
3.2	Method . . . . .	49
3.2.1	Overview of GMIA . . . . .	49
3.2.2	Latent Code Mapping . . . . .	50
3.2.3	Identity Feature Extraction . . . . .	52
3.2.4	Utility Feature Extraction . . . . .	54
3.2.5	Latent Code Optimization . . . . .	55
3.3	Materials . . . . .	57
3.3.1	Dataset and Pre-processing . . . . .	57
3.3.2	Dataset Pre-processing . . . . .	57
3.3.3	Implementation Details . . . . .	59
3.3.4	Evaluation Metrics . . . . .	61
3.4	Results . . . . .	63
3.4.1	Network Pre-training . . . . .	63
3.4.2	Qualitative Results . . . . .	63
3.4.3	Utility Preservation . . . . .	64
3.4.4	Identity Elimination . . . . .	64
3.5	Discussion and Limitations . . . . .	66
3.5.1	Discussion . . . . .	67
3.5.2	Limitations . . . . .	67
3.6	Conclusions . . . . .	68

---

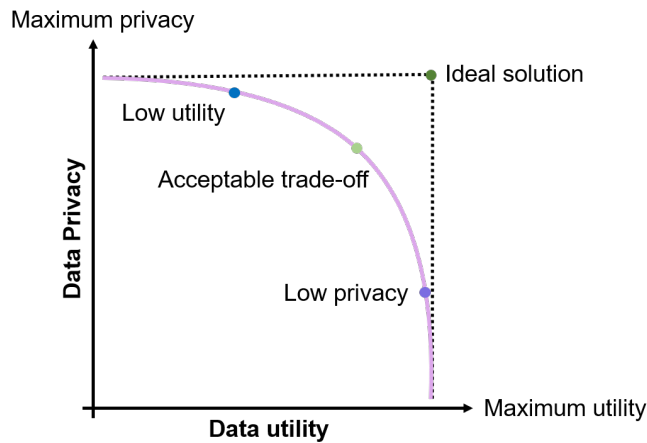
**Abstract** Medical image anonymization aims to remove the patient identity to protect privacy, while maintaining the data utility to solve downstream tasks such as image classification. This chapter introduces the Generative Medical Image Anonymization framework to address the identity-utility trade-off. Our key insight is to split the process of data anonymization into two distinct stages: first, extracting identity and utility features using specialized encoders, and second, optimizing latent codes to reach the desired trade-off. We train identity and utility encoders to verify patient identities and detect pathologies, and employ a generative adversarial network-based auto-encoder to produce realistic synthetic images from the latent space. In the optimization phase, we integrate these two encoders into innovative loss functions to generate images with removed identity-related features while retaining their diagnostic value. Through a comprehensive set of qualitative and quantitative experiments, we showcase the effectiveness of our approach on the MIMIC-CXR chest X-ray dataset by generating synthetic images that can serve as training set for detecting lung pathologies.

## 3.1 Introduction

The advancement of digital medical imaging has revolutionized the field of healthcare, enabling more accurate diagnostics, efficient treatment planning, and innovative research. However, the widespread sharing and utilization of medical images raise significant privacy concerns. Medical images often contain sensitive information that can be linked to individual patients, posing risks of privacy breaches and identity attack. Therefore, robust anonymization techniques are essential to safeguard patient confidentiality while preserving the clinical utility of the images.

Medical image anonymization involves the process of removing or obfuscating personally identifiable information (PII) from both the images and their associated metadata. This includes not only obvious identifiers such as patient names and dates of birth but also more subtle data like facial features, unique anatomical characteristics, and embedded metadata that could potentially re-identify individuals [Packhäuser et al., 2022]. The challenge lies in balancing the removal of identifiable information for privacy protection, while preserving diagnostically relevant features to maintain data utility. As illustrated in Fig. 3.1, this balance represents a trade-off between data privacy and utility. While the ideal scenario would involve maximizing the data privacy without compromising data utility, this is practically impossible to reach.

Existing standards and regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation



**Fig. 3.1.:** Trade-off between privacy level and utility level of data.

(GDPR) in Europe, impose rigorous requirements for the anonymization of medical data. These frameworks highlight the crucial role of anonymization in protecting patient privacy while enabling the legitimate use of medical data in areas like clinical practice, research, and telemedicine. As a result, a variety of anonymization techniques have emerged to address this growing challenge.

Traditional anonymization techniques, such as redaction [Orekondy et al., 2018], blurring [Du et al., 2019], pixelation [Gerstner et al., 2013], and masking [Dadoun et al., 2021] have been used to obscure identifiable regions of medical images. While these methods provide a straightforward way to protect patient identity, they may significantly degrade the overall quality and utility of the data, making it less valuable for clinical or research purposes. In response, researchers have begun focusing on more advanced approaches, including the generation of synthetic data using deep learning techniques such as generative adversarial networks (GANs).

GAN-based models offer the potential to generate realistic synthetic images that maintain privacy while retaining valuable information. For instance, Jeon et al. [Jeon et al., 2022] introduced  $k$ -SALSA, a framework utilizing a GAN-based approach to synthesize a  $K$ -anonymous dataset from private retinal images. However, the process of sample aggregation in the latent space, aimed to ensure privacy according to the  $K$ -anonymity [Sweeney, 2002] principle, comes with the drawback of reducing the dataset size by a factor of  $k$ . To address this issue, Pennisi et al. [Pennisi et al., 2023] introduced a latent space navigation strategy to generate a wide array of diverse images. This approach involves sampling a non-linear walk between two arbitrary latent points in the GAN latent space. However, sampling a synthetic image from two latent points within the GAN latent space is akin to blending two images corresponding to those latent points to a certain degree. Consequently, ensuring the interpretability of the synthetic image becomes challenging. As depicted in Pennisi et al. [Pennisi et al., 2023], discrepancies

between the ribs and lungs arise due to differences between the images from the latent points.

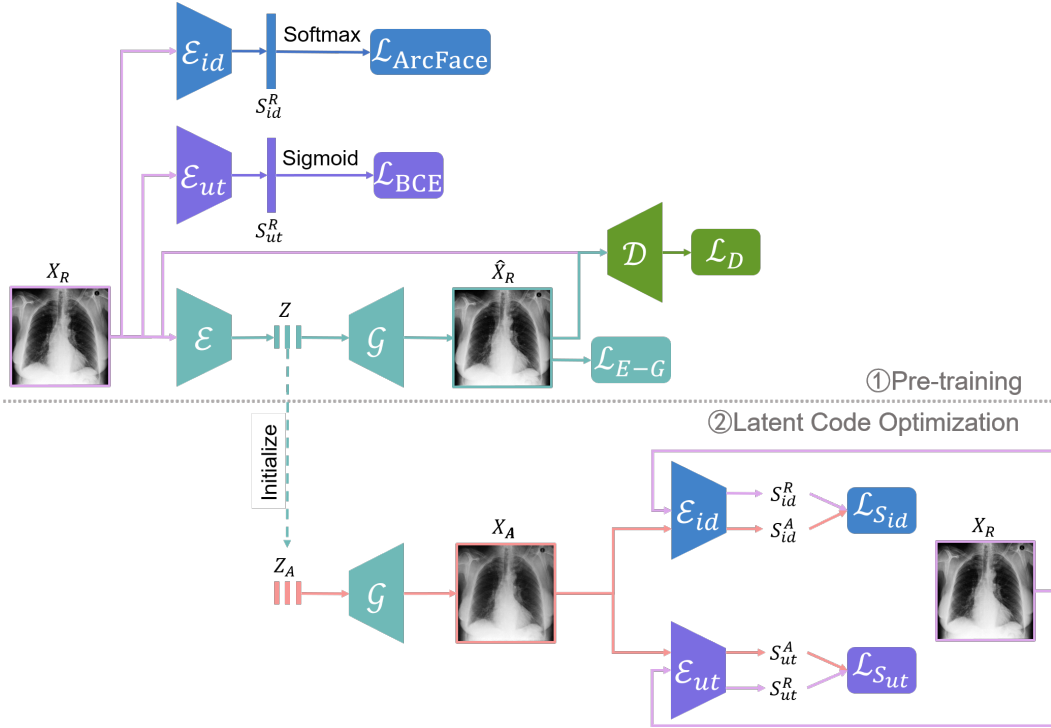
Several other GAN-based studies have focused on facial de-identification [Barattin et al., 2023; Seo et al., 2024; Kuang et al., 2024], offering solutions that could be adapted for medical image anonymization. One particularly influential work is [Barattin et al., 2023], which, for the first time, addresses the challenge of identity anonymization while explicitly preserving facial attributes essential for downstream tasks. This is achieved through two novel loss functions designed to balance the trade-off between anonymizing identity and preserving facial attributes. However, this approach heavily relies on pre-trained networks tailored for facial image processing, and the image anonymization initialization process is unnecessarily intricate.

In contrast to existing generative medical image anonymization methods, which typically rely on GAN inversion to discover the latent representation of an input image, we approach the learning of the latent code representation as an image reconstruction task. Specifically, we employ an Autoencoder-Generative Adversarial Network (AE-GAN) to learn a precise latent code representation for each input image. The Autoencoder (AE) is responsible for establishing a direct one-to-one mapping between the input image and its reconstructed counterpart, ensuring accurate replication of the input. Meanwhile, the Generative Adversarial Network (GAN) component, through its discriminator, enhances the realism of the reconstructed image by incorporating adversarial loss, driving the generator to produce more lifelike images and further improving reconstruction quality. This combined architecture enables a more robust and faithful latent code representation for the task of medical image anonymization. Building on the latent code optimization techniques discussed in [Barattin et al., 2023], we incorporate both an identity removal loss and a utility preservation loss. Specifically, the identity removal minimizes the similarity between the original and anonymized images in the identity feature space, while the utility preservation loss ensures alignment of high-level diagnostic features between the two images. This enables us to generate high-quality anonymized images that effectively remove identifiable information while retaining essential clinical attributes.

**Contributions.** (1) We propose a novel approach to the medical image anonymization problem by accurately identifying and removing identity information from images while explicitly retaining crucial utility attributes essential for downstream tasks. (2) We develop a novel framework for generative medical image anonymization, enhanced by novel loss functions designed to generate high-resolution anonymized datasets from large real datasets with numerous identities. (3) Extensive experiments on the MIMIC-CXR-JPG dataset confirm the effectiveness of our method in obscuring identity while notably enhancing downstream utility, particularly in detecting lung pathologies.



## 3.2 Method



**Fig. 3.2.:** Overview of the proposed Generative Medical Image Anonymization (GMIA) framework, consisting of two key stages: (1) Pre-training of three specialized networks—identity encoder, utility encoder, and AE-GAN network—for feature extraction and image reconstruction, and (2) Latent code optimization, which focuses on identity removal and utility preservation to generate anonymized images that retain critical diagnostic information.

### 3.2.1 Overview of GMIA

Given a private real image dataset, our objective is to generate an anonymized version that ensures patient privacy while preserving data utility. For each image  $X_R$  in the dataset, the corresponding anonymized image  $X_A$  must retain essential diagnostic or utility-related features, while effectively modifying identity-related attributes to ensure patient confidentiality.

As illustrated in Fig. 3.2, the proposed Generative Medical Image Anonymization (GMIA) framework operates in two main stages:

- 1. Pre-training specialized networks:** This involves training three networks, including identity encoder  $\mathcal{E}_{id}$  for extracting identity features ( $S_{id}$ ), utility encoder  $\mathcal{E}_{ut}$  for utility feature extraction ( $S_{ut}$ ), and an AE-GAN model comprising the encoder ( $\mathcal{E}$ ), generator ( $\mathcal{G}$ ), and discriminator ( $\mathcal{D}$ ) to map input images into a latent representation ( $Z$ ) for image reconstruction.

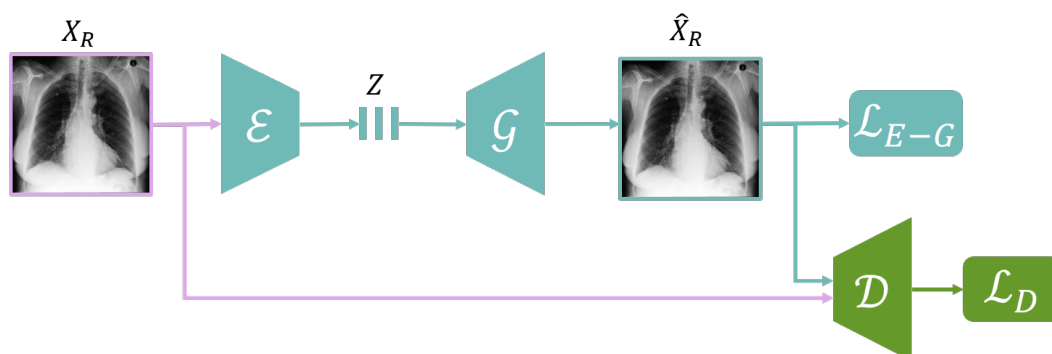
2. **Latent code optimization:** This stage builds upon the pre-trained networks from the previous step. Initially, the anonymized latent code  $Z_A$  is initialized from the original latent code  $Z$ . The optimization process then focuses on modifying  $Z_A$  to remove identity features using the pre-trained identity encoder  $\mathcal{E}_{id}$  while preserving key utility features with the pre-trained utility encoder  $\mathcal{E}_{ut}$ . The result is the anonymized image  $X_A$ , which effectively conceals identity attributes while maintaining the essential data utility.

Compared to the approach presented in [Barattin et al., 2023], the proposed GMIA framework offers a more streamlined solution. Our framework eliminates the reliance on an auxiliary dataset for latent code initialization and removes the need for latent code pairing with a kNN classifier. These simplifications reduce computational complexity while maintaining effective anonymization, making the GMIA framework more efficient and easier to implement.

### 3.2.2 Latent Code Mapping

Latent code mapping involves mapping the real image  $X_R$  to its latent code  $Z$  with high accuracy, as this latent code  $Z$  is critical for optimization-based anonymization.

Unlike previous approaches [Jeon et al., 2022; Pennisi et al., 2023; Barattin et al., 2023], which rely on encoder-based GAN inversion for latent code mapping, our method frames latent code mapping as an image reconstruction task. We employ an AE-GAN (Autoencoder-Generative Adversarial Network) framework (see Fig. 3.3) to achieve this. This network comprises an encoder  $\mathcal{E}$ , which transforms a real image  $X_R$  into its latent code  $Z = \mathcal{E}(X_R)$ , and a generator which transforms the latent code  $Z$  into an approximation of the real image,  $\hat{X}_R = \mathcal{G}(Z)$ . Additionally, a discriminator  $\mathcal{D}(X_R)$  is employed to determine whether the generated image is real or fake.



**Fig. 3.3.:** Overview of the generative reconstruction network: The encoder  $\mathcal{E}$  processes a real image  $X_R$  to extract its latent representation  $Z$ . This latent code  $Z$  is then passed to a generator  $\mathcal{G}$ , which reconstructs the image as  $\hat{X}_R$ . Finally, a discriminator  $\mathcal{D}$  assesses whether the reconstructed image  $\hat{X}_R$  is real or fake.

We train the networks  $\mathcal{E}$  (encoder),  $\mathcal{G}$  (generator), and  $\mathcal{D}$  (discriminator) in an adversarial framework using two distinct loss functions. In each training step, we first optimize the encoder  $\mathcal{E}$  and generator  $\mathcal{G}$  jointly using a composite loss function  $\mathcal{L}_{\mathcal{E},\mathcal{G}}$ , followed by optimizing the discriminator  $\mathcal{D}$  with its corresponding loss  $\mathcal{L}_{\mathcal{D}}$ .

During the first step, the encoder  $\mathcal{E}$  and generator  $\mathcal{G}$  are trained with a composite loss that integrates three key components: pixel-wise similarity loss, perceptual similarity loss, and adversarial (generator) loss. The pixel-wise and perceptual similarity losses ensure that the reconstructed image  $\hat{X}_R$  accurately resembles the original input image  $X_R$  both at the pixel level and in terms of high-level perceptual features extracted from pre-trained networks. This combination helps preserve fine-grained details while maintaining the overall structure and semantics of the image, ensuring both low-level accuracy and high-level visual coherence. Simultaneously, the adversarial loss drives the generator  $\mathcal{G}$  to produce images that are indistinguishable from real images. By optimizing this loss, the generator learns to fool the discriminator, thereby enhancing the realism and authenticity of the reconstructed images.

To stabilize the training process and enhance GAN performance, we employ the non-saturating GAN loss [Goodfellow et al., 2020]. In the standard GAN setup, when the discriminator becomes highly effective at distinguishing real data from generated data, the generator’s loss term  $\log(1 - D(G(z)))$  can quickly approach zero. This leads to very small gradients for the generator, a phenomenon known as saturation, which significantly slows down learning and hampers the generator’s ability to improve.

To address this, Goodfellow [Goodfellow et al., 2020] proposed a non-saturating generator loss to mitigate gradient saturation. Rather than minimizing  $\log(1 - D(G(z)))$ , the generator instead maximizes  $\log(D(G(z)))$ , thereby encouraging more stable gradient flow. This modification incentivizes the generator to produce more realistic samples by directly rewarding high discriminator outputs for generated data, rather than merely minimizing the probability of being classified as fake.

The loss function to optimize the generative reconstruction network are defined as follows

$$\begin{aligned} \mathcal{L}_{\mathcal{E}-\mathcal{G}}(X_R, \hat{X}_R) = & \lambda_{\text{pixel}} \|X_R - \hat{X}_R\|_2 \\ & + \lambda_{\text{perceptual}} \|\phi(X_R) - \phi(\hat{X}_R)\|_2 \\ & - \lambda_{\mathcal{G}} \mathbb{E}_{\hat{X}_R} [\log \mathcal{D}(\hat{X}_R)] \end{aligned} \quad (3.1)$$

where  $\lambda_{\text{pixel}}$ ,  $\lambda_{\text{perceptual}}$ ,  $\lambda_{\mathcal{G}}$  are hyper-parameters used to balance the contributions of the respective loss components. In this formulation, perceptual loss  $\|\phi(X_R) - \phi(\hat{X}_R)\|_2$  plays a crucial role in improving the quality of generated images by comparing high-level semantic features rather than relying solely on pixel-level differences. By leveraging feature maps from pre-trained networks, such as the 19-layer VGG network  $\phi$  [Simonyan

et al., 2014], perceptual loss encourages the preservation of fine textures and structural details, resulting in sharper, more realistic images that are better aligned with human visual perception.

In the second step, we optimize the discriminator  $\mathcal{D}$  using the discriminator loss  $\mathcal{L}_{\mathcal{D}}$ . This loss measures the discriminator’s ability to correctly differentiate between real images and the fake images generated by the generator. As a result, the discriminator becomes more efficient at distinguishing between authentic and generated images, which in turn challenges the generator to create even more convincing reconstructions in future iterations.

The discriminator loss can be expressed as:

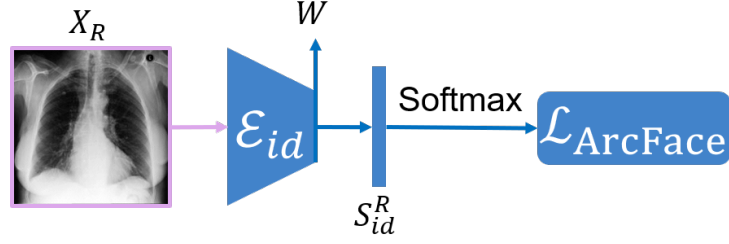
$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}_{X_R}[-\log \mathcal{D}(X_R)] + \mathbb{E}_{\hat{X}_R}[-\log(1 - \mathcal{D}(\hat{X}_R))] \quad (3.2)$$

### 3.2.3 Identity Feature Extraction

Although biometric markers in medical images may not be as prominent as those in facial images, they still contain identifiable information that can lead to patient re-identification [Boutet et al., 2023]. For example, a study by [Packhäuser et al., 2022] demonstrated that a well-trained deep learning model was able to re-identify patients from chest X-rays with an impressive AUC of 0.9940. This finding highlights the presence of subtle but significant biometric patterns in medical images, which a sufficiently trained deep learning system can effectively recognize, underscoring the need for robust anonymization techniques.

For identity feature extraction, we train a dedicated identity network for multi-class classification (see Fig. 3.4), where each patient is assigned a unique identity ID (a number) from a set of mutually exclusive labels.

To ensure robust identity classification, we adopt an open-set setting as outlined in [Liu et al., 2017], where the identities in the test set are distinct from those in the training set. Unlike the closed-set setting, where testing identities are predefined during training, the open-set setting is more challenging and reflective of real-world scenarios. Since open-set classification is inherently a metric learning problem, the goal is to learn highly discriminative, large-margin features. To achieve this, we integrate Additive Angular Margin Loss (ArcFace) [Deng et al., 2019], which enhances the discriminative power of the extracted identity features by introducing angular margins between identity classes. This ensures that the features are not only more distinct but also better separated, enabling more precise identity removal during the anonymization process.



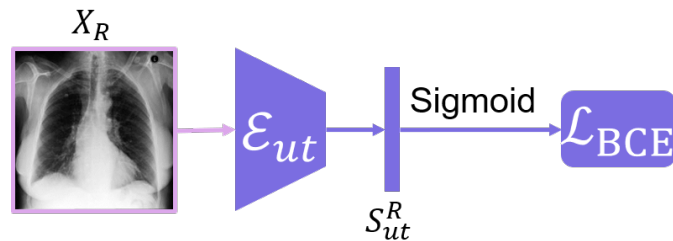
**Fig. 3.4.:** Overview of the identity extraction network: The encoder  $\mathcal{E}_{id}$  processes a real image  $X_R$  to extract its identity semantic feature  $S_{id}^R \in \mathbb{R}^{512}$ . These semantic feature  $S_{id}$  is then passed through a softmax layer to generate the output probability. Finally, a Arcface loss [Deng et al., 2019] is used to perform the multi-class identity classification.  $W$  is the weight matrix from the last fully connected layer of  $\mathcal{E}_{id}$ .

The ArcFace loss [Deng et al., 2019] is presented as follows:

$$\mathcal{L}_{\text{ArcFace}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{e^{s(\cos(\theta_{y_i} + m))}}{e^{s(\cos(\theta_{y_i} + m))} + \sum_{j=1, j \neq y_i}^C e^{s \cos(\theta_j)}} \quad (3.3)$$

where  $N$  is the number of training samples.  $C$  is the number of classes. As demonstrated in [Liu et al., 2017; Liu et al., 2016], the weight  $W$  from the last fully connected layer of a classification network trained with softmax loss provided a kind of center for each class. Using the normalized feature  $S_{id}^R$  and the normalized weight  $W$ , the  $\cos \theta_j$  for each class is calculated as  $W_j^T S_{id}^R$ . Next, the angle  $\theta_{y_i}$  between the feature and the target class weight  $W_{y_i}$  is calculated as  $\arccos(W_{y_i}^T S_{id}^R)$ . An angular margin penalty  $m$  is then added to the target angle  $\theta_{y_i}$ . Following this, the calculated logits are multiplied with a feature scale  $s$ , passed through the softmax function, and finally contribute to the cross-entropy loss.

The identity network architecture preceding the classifier head (softmax layer) functions as the identity encoder  $\mathcal{E}_{id}$ , extracting identity semantic features  $S_{id}^R \in \mathbb{R}^{512}$  from the real image  $X_R$ .



**Fig. 3.5.:** Overview of the utility extraction network: The encoder  $\mathcal{E}_{ut}$  processes a real image  $X_R$  to extract its utility semantic feature  $S_{ut}^R \in \mathbb{R}^{512}$ . These semantic feature  $S_{ut}$  is then passed through a sigmoid layer to generate the output probability. Finally, a Binary Cross-Entropy (BCE) loss is used to perform the multi-label pathology classification.

### 3.2.4 Utility Feature Extraction

For utility feature extraction, we train a utility network for multi-label pathology classification (see Fig. 3.5), where each instance can be assigned multiple labels simultaneously.

In this multi-label pathology classification task, the training set consists of image-label pairs  $(X_R, y)$ . A significant portion of the dataset includes uncertain labels, meaning each input image  $X_R$  is associated with a label  $y \in \{0, 1, -1\}$ , where 0, 1, and -1 correspond to negative, positive, and uncertain, respectively.

To effectively utilize the large number of uncertain labels, we employ a label smoothing strategy similar to [Pham et al., 2021]. This approach prevents the model from making overconfident predictions on training examples that might contain mislabeled data.

The strategy for replacing the uncertainty label  $y$  with a smoothed version  $\tilde{y}$  is presented as follows:

$$\tilde{y} = \begin{cases} u, & \text{if } y = -1 \\ y, & \text{otherwise} \end{cases} \quad (3.4)$$

where  $u \sim U(a_1, b_1)$  is a uniformly distributed random variable between  $a_1$  and  $b_1$  (the hyper-parameters). The loss function is then given by

$$\mathcal{L}_{\text{BCE}} = -(\tilde{y} \log(\hat{y}) + (1 - \tilde{y}) \log(1 - \hat{y})) \quad (3.5)$$

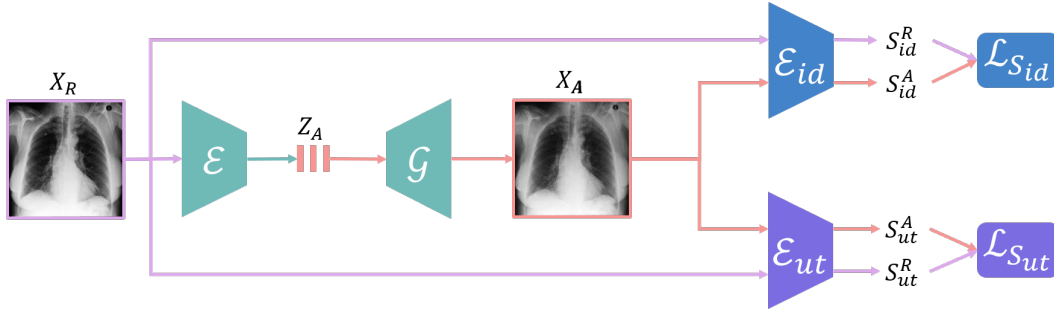
where  $\tilde{y}$  is the smoothed binary label and  $\hat{y}$  is the predicted probability for the positive class.

The utility network utilizes the architecture preceding the classifier head (sigmoid layer) as its encoder to get the utility semantic features  $S_{ut}^R \in \mathbb{R}^{512}$  from the real image  $X_R$ .

It is important to note that the type of utility task is highly dependent on the specific dataset being used. For example, in this chapter, since the dataset focuses on lung pathology, the utility task is lung pathology classification. While the proposed anonymization framework leverages utility features to preserve attributes necessary for solving downstream tasks, it is not limited to any specific task. The framework is adaptable to a wide range of downstream tasks depending on the original dataset's purpose, whether it be classification, segmentation, detection, or other objectives.

### 3.2.5 Latent Code Optimization

The generative medical image anonymization process is based on latent code optimization. It begins by using the pre-trained generative image reconstruction network (as described in Section 3.2.2) to initialize the anonymized image  $X_A$  as the reconstructed version of  $\hat{X}_R$ , with the anonymized latent code initialized as the corresponding latent representation  $Z$  of the reconstructed image. The identity and utility extraction networks (see Section 3.2.3, 3.2.4) are utilized as loss functions for optimizing the latent code  $Z_A$ .



**Fig. 3.6.:** Overview of the latent code optimization process: the encoder  $\mathcal{E}$  projects a real image  $X_R$  into its latent space  $Z$ , providing the initial value for  $Z_A$ . This latent code is then optimized using two deep network-based loss functions:  $\mathcal{L}_{S_{id}}(X_R, X_A)$ , which removes the identity information, and  $\mathcal{L}_{S_{ut}}(X_R, X_A)$ , which preserves the utility information. Finally, a generator  $\mathcal{G}$  projects the optimized latent code  $Z_A$  back into the image space, producing the anonymized image  $X_A$  corresponding to the original image  $X_R$ .

To enable the latent code optimization, we set  $Z_A$  as a trainable parameter. During the training process, we optimize  $Z_A$  to minimize the sum of two loss functions.

It should be noted that, although the latent code optimization ideas are adapted from the paper [Barattin et al., 2023], we use a different initialization approach for the latent code. Consequently, we do not require the creation of an auxiliary dataset generated from random noise. Additionally, we train the AE-GAN model concurrently, creating a seamless connection between the input image and its latent representation. This approach enables us to find a more precise latent code for each input image compared to the method in [Barattin et al., 2023], which employs a pre-trained GAN inversion approach to approximate the latent code. Moreover, due to the lack of usable pre-trained networks for medical images, we have to train all the networks from scratch. We hope this can eliminate the need for other researchers to undergo the costly process of training the image reconstruction network, as well as the identity and utility extraction networks.

The identity removal loss  $\mathcal{L}_{S_{id}}(X_R, X_A)$  utilizes the identity extraction networks described in Section 3.2.3 to separately extract the identity semantic features  $S_{id}^R$  from the real image  $X_R$  and  $S_{id}^A$  from its anonymized version  $X_A$ . The  $\mathcal{L}_{S_{id}}(X_R, X_A)$  ensures that

$X_A$  has a different identity from  $X_R$ , up to a desired margin. This loss <sup>1</sup> is defined as follows:

$$\mathcal{L}_{S_{id}}(X_R, X_A) = \max(0, \cos(\mathcal{E}_{id}(X_R), \mathcal{E}_{id}(X_A)) - m) \quad (3.6)$$

where  $\cos(a, b) = \frac{a \cdot b}{\max(\|a\|_2 \cdot \|b\|_2, \epsilon)}$  denotes the cosine similarity, and  $\epsilon$  is a small value to avoid division by zero.  $\mathcal{E}_{id}$  is the identity encoder, and  $m$  is a hyperparameter that controls the dissimilarity between the real and the anonymized images, constrained to be equal to or greater than  $\arccos(m)$ .

The utility preservation loss,  $\mathcal{L}_{S_{ut}}(X_R, X_A)$ , employs the utility extraction networks outlined in Section 3.2.4 to separately extract the utility semantic features  $S_{ut}^R$  from the real image  $X_R$  and  $S_{ut}^A$  from its anonymized version  $X_A$ . The  $\mathcal{L}_{S_{ut}}(X_R, X_A)$  enforces that utility attributes of  $X_R$  are preserved in  $X_A$ . The utility preservation loss is defined as follows:

$$\mathcal{L}_{S_{ut}}(X_R, X_A) = \|\mathcal{E}_{ut}(X_R) - \mathcal{E}_{ut}(X_A)\|_2 \quad (3.7)$$

where  $\mathcal{E}_{ut}$  denotes the utility encoder.

The overall loss is represented as

$$\mathcal{L} = \lambda_{id}\mathcal{L}_{S_{id}}(X_R, X_A) + \lambda_{ut}\mathcal{L}_{S_{ut}}(X_R, X_A) \quad (3.8)$$

where the hyperparameters  $\lambda_{id}$  and  $\lambda_{ut}$  are used to balance between utility preserving and privacy protection. The identity removal loss  $\mathcal{L}_{S_{id}}(X_R, X_A)$  ensures that the identity semantic features of the real and anonymized images differ by at least  $90^\circ$ , controlled by a hyper-parameter margin  $m \leq 0.0$ . The utility-preserving loss  $\mathcal{L}_{S_{ut}}(X_R, X_A)$  enforces that the utility semantic features of the real and anonymized images are as similar as possible.

Gradient-based optimization algorithms, such as Adam [Kingma et al., 2014], are used to iteratively adjust the latent code  $Z_A$ . During each iteration, the algorithm updates  $Z_A$  to minimize the total loss:

$$Z_A^{new} = Z_A^{old} - \eta \nabla_Z \mathcal{L} \quad (3.9)$$

where  $\eta$  is the learning rate and  $\nabla_Z \mathcal{L}$  is the gradient of the total loss with respect to the latent code.

<sup>1</sup><https://pytorch.org/docs/stable/generated/torch.nn.CosineEmbeddingLoss.html>



The optimized latent code  $Z_A$  is then used to generate the anonymized image. This image should have identifiable features effectively obfuscated or removed, while preserving essential visual and diagnostic information.

## 3.3 Materials

### 3.3.1 Dataset and Pre-processing

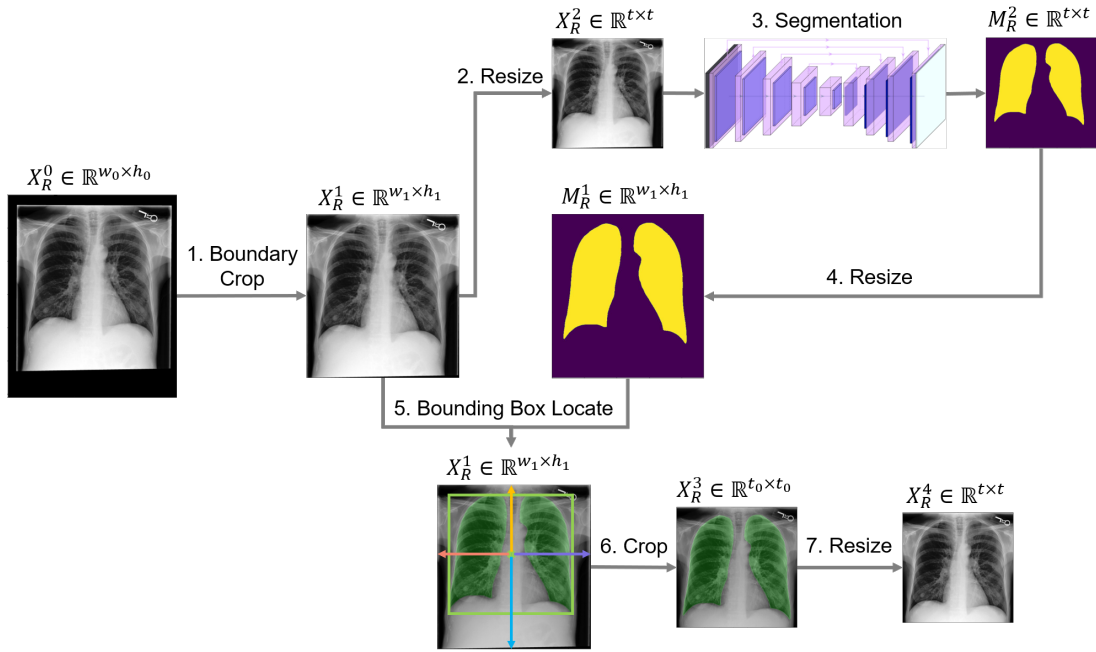
Our experiments are conducted on MIMIC-CXR-JPG dataset [Johnson et al., 2019], which is a large publicly available dataset of chest radiographs in JPG format including several images for a given patient. The dataset contains 377 110 JPG format images and structured labels derived from the 227 827 free-text radiology reports associated with these images. Each label column contains one of four values: 1.0, -1.0, 0.0, or missing. These labels have the following interpretation:

- 1.0 - The label was positively mentioned in the associated study, and is present in one or more of the corresponding images
- 0.0 - The label was negatively mentioned in the associated study, and therefore should not be present in any of the corresponding images
- -1.0 - The label was either: (1) mentioned with uncertainty in the report, and therefore may or may not be present to some degree in the corresponding image, or (2) mentioned with ambiguous language in the report and it is unclear if the pathology exists or not.
- Missing (empty element) - No mention of the label was made in the report

We adopt the dataset partition provided by the competition, focusing solely on frontal images with a minimum of 20 images per patient. This results in 50 875 training images, 522 validation images, and 1 641 testing images, corresponding to 1 538 patients in the training set, 14 patients in the validation set, and 47 patients in the test set.

### 3.3.2 Dataset Pre-processing

The pre-processing involves seven steps (see Fig. 3.7) aimed at obtaining a square image centered on the lungs from the raw dataset, since the original images may vary in size and may not be centered either. Example images are shown in Fig. 3.8.



**Fig. 3.7.:** Overview of data pre-processing step.

Starting with the raw image  $X_R^0 \in \mathbb{R}^{w_0 \times h_0}$ , the objective is to process it into a square image of size  $t$ . The preprocessing steps are as follows:

- **Step 1: Boundary Crop**  
This step involves cropping the black boundaries surrounding the image  $X_R^0 \in \mathbb{R}^{w_0 \times h_0}$  to retain the main content and exclude irrelevant information.
- **Step 2: Image Resize**  
The image  $X_R^1 \in \mathbb{R}^{w_1 \times h_1}$  is resized to  $X_R^2 \in \mathbb{R}^{t \times t}$  to fit the input size of a U-net segmentation network.
- **Step 3: Segmentation**  
A pre-trained lung segmentation algorithm from [Ovcharenko, 2019] is used to obtain the lung segmentation mask.
- **Step 4: Mask Resize**  
The predicted lung mask is resized to match the dimensions of  $X_R^1 \in \mathbb{R}^{w_1 \times h_1}$ .
- **Step 5: Bounding Box Localization**  
This step involves locating the main lung region with a refined bounding box. Initially, a bounding box (depicted in green in Fig. 3.7) is placed around the lung mask to fully encompass it within the image. Subsequently, we determine the center point of this bounding box. The distance between this center point and both the bounding box's edges and the image boundary is then compared. Finally, a

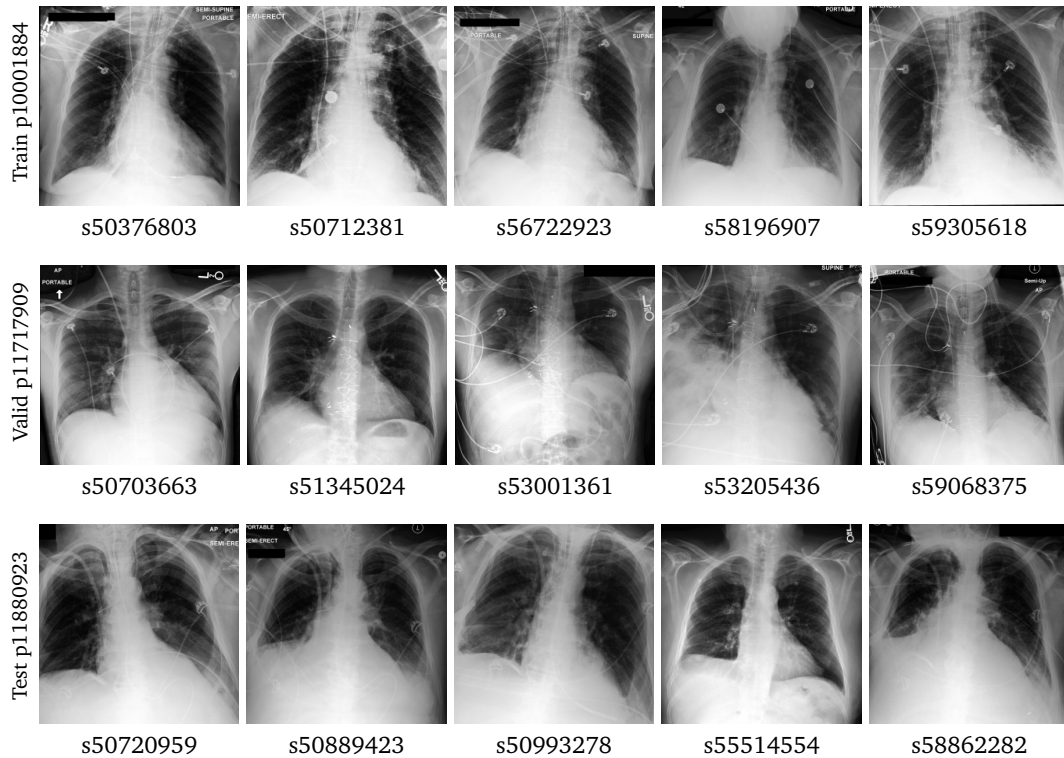
refined bounding box centered on the same center point is determined, with its dimensions set to  $t_0$ , representing the length of the shortest lines identified.

- **Step 6: Image Crop**

The image  $X_R^1 \in \mathbb{R}^{w_1 \times h_1}$  is cropped to a square image  $X_R^3 \in \mathbb{R}^{t_0 \times t_0}$  with the main lung region preserved and centered.

- **Step 7: Image Resize**

The image  $X_R^3 \in \mathbb{R}^{t_0 \times t_0}$  is resized to the target size  $t$ . In this thesis,  $t$  is set to 512.



**Fig. 3.8.:** Example of pre-processed images in MIMIC-CXR-JPG dataset. For example, 'Test p11880923' indicates that the image originates from the test dataset and is associated with patient ID '11880923'. Images from different studies are displayed in separate columns, with the corresponding study IDs listed below the images.

### 3.3.3 Implementation Details

#### Reconstruction Network.

The reconstruction network, adapted from [Mentzer et al., 2020], has been modified to exclude the compression component. The network architecture replaces the 9 ResNet blocks with 9 DenseNet blocks, resulting in a smaller network size while maintaining nearly the same performance. To accelerate training, we initially train this network using an auto-encoder structure with a perceptual distortion loss [Mentzer et al., 2020]. Subse-


quently, we integrate the discriminator and incorporate an adversarial loss [Goodfellow et al., 2014].

The reconstruction network processes input and output images at a resolution of  $512 \times 512$  pixels, with a latent representation  $Z \in \mathbb{R}^{256 \times 16 \times 16}$ . We set the weights for the pixel-wise loss  $\lambda_{\text{pixel}}$ , perceptual loss  $\lambda_{\text{perceptual}}$ , and generator loss  $\lambda_G$  to 1, 1, and 0.15, respectively. The model is trained using the Adamw [Loshchilov et al., 2017] optimizer with a learning rate of 3e-5 and a weight decay of 1e-2 for 100 epochs.

### Identity and Utility Network.

We utilize iResNet50 [Duta et al., 2021] for multi-class identity classification. The identity network takes input images with a resolution of  $256 \times 256$ . To learn discriminative identity features across a large number of identities (1 599 identities), we employ the ArcFace loss [Deng et al., 2019], setting the additive angular margin penalty  $m$  to 0.5. The network is trained using the SGD optimizer [Sutskever et al., 2013] with a learning rate of 0.1, a momentum of 0.9, and a weight decay of  $5 \times 10^{-4}$  for 100 epochs.

For multi-label pathology classification, we utilize DenseNet121 [Iandola et al., 2014]. This utility network is designed to predict four selected labels: Lung Opacity, Atelectasis, Pleural Effusion, and Support Devices. To stabilize the training process, we initialize the network with pre-trained weights from ImageNet27 [Russakovsky et al., 2015]. The utility network also takes input images of resolution  $256 \times 256$  and is trained using the Adam optimizer [Kingma et al., 2014] with a learning rate of 0.0005 and no weight decay for 100 epochs. An example of images from the MIMIC-CXR-JPG dataset, along with their corresponding classification labels, is presented in Fig. 3.9. The accompanying table outlines the specific pathology labels that the utility classifier is designed to predict.



Atelectasis	0	1	0	0	0	1	-1	-1
Lung Opacity	0	0	1	0	0	1	1	-1
Pleural Effusion	0	0	0	1	0	1	1	-1
Support Devices	0	0	0	0	1	1	1	1

**Fig. 3.9.:** Example of images from the MIMIC-CXR-JPG dataset alongside their corresponding classification labels. The accompanying table indicates the specific labels for the lung pathologies that the utility classifier is designed to predict. In the table, 0 indicates negative, 1 represents positive, and -1.0 indicates an uncertain label.

### Latent Code Optimization.

For latent code optimization, both the input and the output anonymized images have a resolution of  $512 \times 512$ , with a batch size of 1. The margin  $m$  in  $\mathcal{L}_{S_{id}}(X_R, X_A)$  is set to

0. We set  $\lambda_{id}$  and  $\lambda_{ut}$  to 0.5. The Adam optimizer [Kingma et al., 2014] is used with a learning rate of  $1 \times 10^{-3}$  and a weight decay of  $1 \times 10^{-2}$ , and training is conducted for a maximum of 500 epochs.

### 3.3.4 Evaluation Metrics

We evaluate our method by quantifying image reconstruction quality, utility preservation and identity elimination using the following metrics:

#### **Image Reconstruction Quality.**

To evaluate image quality, we utilize metrics including PSNR (Peak Signal-to-Noise Ratio), SSIM (Structural Similarity) [Wang et al., 2004], MS\_SSIM (Multi-Scale Structural Similarity) [Wang et al., 2003], and LPIPS (Learned Perceptual Image Patch Similarity) [Zhang et al., 2018] to assess the fidelity of the reconstruction network.

#### **Utility Preservation.**

To evaluate the preservation of utility attributes in the anonymized images, we conduct a standard multi-label classification task. We train the classifier on the anonymized training set and evaluate its performance on the real test set. To assess utility preservation, we report metrics such as Accuracy (Acc), Average Precision (AP), Area Under the Receiver Operating Characteristic Curve (AUROC), and F1 Score (F1).

#### **Identity Elimination.**

To measure the privacy-preserving properties of our approach, we first evaluate it using privacy metrics. Then, we introduce a protocol to assess linkability risk<sup>2</sup>. Lastly, we perform membership inference attacks [Shokri et al., 2017].

**Privacy Metrics and Singling-out Risk.** To quantify identity elimination, we initially employ two general privacy metrics: DCR (distance to closest record) and NNDR (nearest neighbor distance ratio) [Zhao et al., 2021]. The DCR is computed as the median distance between each anonymized image and its closest real images, while NNDR is the median of all ratios between the distance to the closest and the distance to the second closest real neighbor for each anonymized image.

Additionally, we consider two metrics specific to assess the singling-out risk<sup>3</sup>: LC (local cloaking) and HR (hidden rate) [Guillaudoux et al., 2023]. LC is calculated as the median number of anonymized images that look more like an individual than the anonymized image generated. HR is defined as the percentage of individuals in the real dataset whose

---

<sup>2</sup>Linkability: the ability to link, at least, two records concerning the same data subject or a group of data subjects (either in the same database or in two different databases).

<sup>3</sup>Singling out: the possibility to isolate some or all records which identify an individual in the dataset.

anonymized images is not the most similar to them. This metric evaluates the probability of an attack being wrong when linking an anonymized image with the individual to whom it is most similar. For DCR and LC, higher values indicate a better privacy level. NNDR and HR, being bounded in the range  $[0, 1]$ , also show a higher privacy level when closer to 1.

In practice, we compute privacy metrics in both the identity semantic feature space  $S_{id}$  and the latent space  $Z$ , utilizing L2 Euclidean distance for comparison. The identity semantic feature space is rich in identity-specific information, making it a strong indicator of how well identity features are removed. On the other hand, the  $Z$  latent space is where the anonymization process occurs, offering insight into the efficacy of transformations applied for anonymization. This dual evaluation offers a comprehensive understanding of how well our method preserves privacy by reducing the likelihood of re-identification while retaining diagnostically relevant features. Notably, privacy metrics are not computed directly in the image space, as such metrics are better suited for assessing feature-level changes rather than the images themselves.

**Linkability Risk.** To evaluate linkability risk, we introduce a protocol that assesses this risk from two perspectives:

- **Inner risk:** Verifying whether two anonymized images belong to the same patient.
- **Outer risk:** Verifying whether a real and its anonymized image belong to the same patient.

Specifically, we frame the evaluation as a standard identity verification task, where the identity classifier is trained to determine whether two images originate from the same patient. For the inner risk, we assess its performance on image pairs from the anonymized test sets, where an equal number of positive pairs (consisting of two images from the same patient) and negative pairs (consisting of two images from different patients) are included. For the outer risk, we evaluate its performance on negative pairs, where each pair consists of a real image from the real test set and its anonymized image from the anonymized test set. We employ metrics including Accuracy (Acc), True Accept Rate (TAR), False Accept Rate (FAR), and F1 Score (F1).

**Mitigation of Membership Inference Attacks.** To evaluate the robustness of the anonymized dataset, we conduct membership inference attacks from two perspectives:

- Predicting whether an anonymized image was part of the real training set.
- Predicting whether a real image was part of the anonymized training set.

In the first case, the attacker is trained on the real dataset and evaluated on the anonymized dataset (target dataset), while in the second case, the attacker is trained on the anonymized dataset and evaluated on the real dataset (target dataset).

In both cases, we excluded the validation dataset used to train the target utility classifier due to its unclear membership status. The attacker is trained on 80% of the training set and the test set, with 10% of each dataset for validation and testing, separately. Subsequently, we evaluate the attack on the entire target dataset.

## 3.4 Results

### 3.4.1 Network Pre-training

As illustrated in Fig. 3.6, the generative reconstruction network, identity and utility encoders serve as core components in the latent code optimization process. Therefore, we initially evaluate the effectiveness of each component independently to ensure their efficacy before integrating them to contribute for the subsequent steps of latent code optimization.

#### Image Reconstruction.

As presented in Table. 3.1, high fidelity reconstruction is achieved with a PSNR of 32, MS\_SSIM [Wang et al., 2003] of 0.994 and LPIPS [Zhang et al., 2018] of 0.011. Furthermore, the visual comparison between the original and reconstructed images in Fig. 3.10 further validate the effectiveness of the image reconstruction network.

**Tab. 3.1.:** The evaluation results of the generative reconstruction network pre-training, which conducted on the test dataset with 1 641 real images.

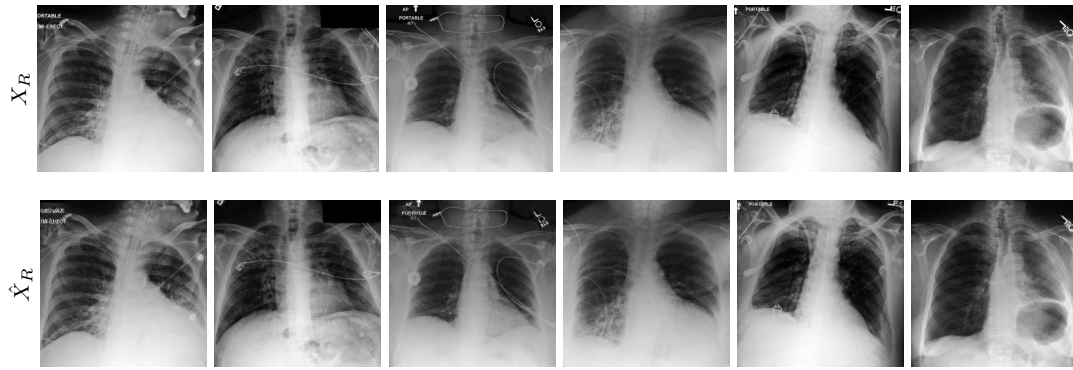
PSNR $\uparrow$	SSIM $\uparrow$	MS_SSIM $\uparrow$	LPIPS $\downarrow$
32.080 $\pm$ 1.139	0.947 $\pm$ 0.008	0.994 $\pm$ 0.001	0.011 $\pm$ 0.002

#### Identity and Utility Extraction.

The pre-training results for the utility networks are presented in Table 3.2 (row 1), while those for the identity networks are shown in Table 3.4 (column 1). These results show an accuracy of 0.983 for identity classification and a global accuracy of 0.752 for pathology classification.

### 3.4.2 Qualitative Results

The visualization of anonymized images is presented in Fig. 3.11. As observed, with increasing epochs during the optimization process, the anonymized images progressively



**Fig. 3.10.:** Reconstruction results of the latent code generation process. The first row displays the real images  $X_R$ , while the second row displays their reconstructed counterparts  $\hat{X}_R$ .

become visually blurry compared to their real counterparts. This outcome is expected and reflects the fundamental principle of our approach, which involves removing identity information from real images up to the preservation of its utility.

Early anonymized images may appear quite similar to their real counterparts, but this similarity is only apparent from a human perspective and may not be reliable in some cases, such as in adversarial attacks. Additionally, this similarity is partly due to the faithful mapping of the reconstruction network and is related to the underlying concept of our approach, which is to remove identity information within a real image rather than replacing it with new information, such as through structural deformations [Jeon et al., 2022; Pennisi et al., 2023]. The visualizations of the difference maps, represented as  $|X_R - X_A|$ , are displayed in Fig. B.1, highlighting the discrepancies between the real and anonymized images.

### 3.4.3 Utility Preservation

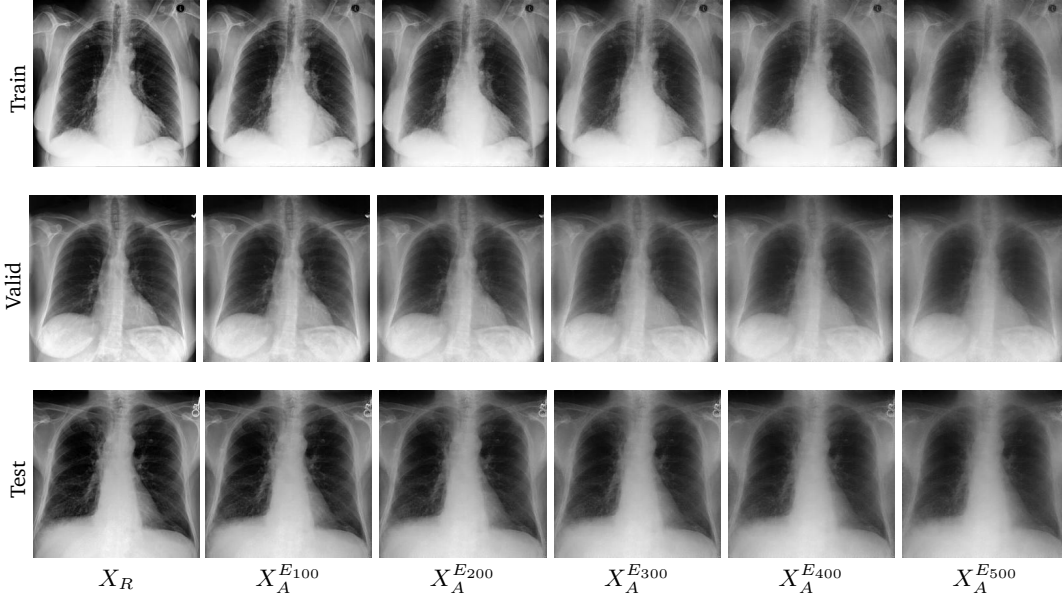
From Table 3.2, it is evident that the utility classifier’s performance with anonymized images (row 2-6) is comparable to that achieved with the real images (row 1) across various recognized metrics. This indicates the ability of our method in preserving utility attributes essential for downstream tasks.

### 3.4.4 Identity Elimination

#### Privacy Metrics and Singling-out Risk.

As illustrated in Table 3.3, the anonymized images exhibit robust performance across various privacy metrics, notably excelling in those associated with singling-out risk.





**Fig. 3.11.:** Anonymization results. Three real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The anonymized images  $X_A$  at different optimization epochs  $E$  are shown in the last five columns. For instance,  $X_A^{E_{100}}$  denotes the anonymized images  $X_A$  at 100th optimization epoch.

**Tab. 3.2.:** Evaluation results of utility network for pre-training and utility preservation on the test set with 1641 real images. 'Tr': training set, 'Acc': accuracy, 'AP': average precision, 'AUROC': area under the ROC curve, 'F1': F1 score. For instance,  $X_A^{E_{100}}$  denotes the anonymized images  $X_A$  at 100th optimization epoch.

	Tr	Acc $\uparrow$	AP $\uparrow$	AUROC $\uparrow$	F1 $\uparrow$
Pre-training	$X_R$	$0.752 \pm 0.072$	$0.619 \pm 0.100$	$0.726 \pm 0.110$	$0.448 \pm 0.104$
Preservation	$X_A^{E_{100}}$	$0.741 \pm 0.067$	$0.614 \pm 0.090$	$0.711 \pm 0.106$	$0.453 \pm 0.097$
	$X_A^{E_{200}}$	$0.735 \pm 0.069$	$0.612 \pm 0.096$	$0.711 \pm 0.106$	$0.456 \pm 0.094$
	$X_A^{E_{300}}$	$0.732 \pm 0.069$	$0.602 \pm 0.102$	$0.701 \pm 0.105$	$0.468 \pm 0.101$
	$X_A^{E_{400}}$	$0.720 \pm 0.063$	$0.591 \pm 0.091$	$0.696 \pm 0.110$	$0.401 \pm 0.094$
	$X_A^{E_{500}}$	$0.718 \pm 0.070$	$0.575 \pm 0.123$	$0.680 \pm 0.109$	$0.403 \pm 0.099$

**Tab. 3.3.:** Results based on privacy metrics. ' $N_R$ ': median number of real images between a real image and its anonymization, ' $N_A$ ' median number of anonymized images between a real image and its anonymization. ' $S_{id}$ ': the identity semantic space. ' $Z$ ': the latent space. For instance,  $E_{100}$  denotes the corresponding features at 100th optimization epoch.

Space	DCR $\uparrow$	NNDR $\uparrow$	HR $\uparrow$	LC $\uparrow$	
				$N_R$	$N_A$
$S_{id}^{E_{100}}$	19.330	0.987	0.998	158	307
$S_{id}^{E_{500}}$	17.277	0.977	0.999	201	1105
$Z^{E_{100}}$	35.907	0.977	0.764	20	11
$Z^{E_{500}}$	28.667	0.980	0.998	67	685

## Linkability Risk.

Table 3.4 shows that though the anonymized images exhibit high inner linkability risk, they demonstrate extremely low outer risk. This suggests that while the anonymized images hold distinct identities from their real counterparts, they still retain some shared similarities among images from the same patient.

**Tab. 3.4.:** Results of identity network for pre-training and linkability risk. 'Pre': pre-training, 'Te': test set, 'Acc': accuracy, 'TAR': true accept rate, 'FAR': false accept rate, 'F1': F1 score.

	Pre	Inner risk				
Te	$X_R$	$X_A^{E_{100}}$	$X_A^{E_{200}}$	$X_A^{E_{300}}$	$X_A^{E_{400}}$	$X_A^{E_{500}}$
Acc↑	0.983	0.850	0.845	0.821	0.816	0.801
TAR↑	0.966	0.724	0.700	0.661	0.671	0.710
FAR↓	0.034	0.276	0.300	0.339	0.329	0.290
F1↑	0.983	0.828	0.819	0.787	0.785	0.781
	Pre	Outer risk				
Te	$X_R$	$X_A^{E_{100}}$	$X_A^{E_{200}}$	$X_A^{E_{300}}$	$X_A^{E_{400}}$	$X_A^{E_{500}}$
Acc↑	0.983	0.000	0.010	0.001	0.000	0.000
TAR↑	0.966	0.000	0.010	0.001	0.000	0.000
FAR↓	0.034	1.000	0.990	0.999	1.000	1.000
F1↑	0.983	0.000	0.019	0.002	0.000	0.000

## Mitigation of Membership Inference Attacks.

The attacker demonstrated good performance on its test set (Acc: 0.968) but performed extremely poorly (Acc: 0) on the target dataset in both cases (see Table 3.5), indicating its inability to identify samples used for training. These results are consistent with the HR close to 1 in Table 3.4, indicating a high probability of the attacker incorrectly associating an anonymized image with its real counterparts.

**Tab. 3.5.:** Results of MIA. 'Te': test set, ' $X_R$ ': real images, ' $X_A$ ': anonymized images. ' $E_{100}$ ': the corresponding anonymized images  $X_A$  at 100th optimization epoch.

	Acc↑	AP↑	AUROC↓	F1↑
$X_R^{Te}$	0.968±0.176	0.978±0.121	0.968±0.176	0.968±0.176
$X_A^{E_{100}}$	0.000±0.000	0.312±0.000	0.000±0.000	0.000±0.000
$X_A^{Te}$	0.968±0.176	0.978±0.121	0.968±0.176	0.968±0.176
$X_R$	0.000±0.000	0.312±0.000	0.000±0.000	0.000±0.000

## 3.5 Discussion and Limitations

The proposed GMIA framework first captures the latent representation of the real image using a generative reconstruction network. An anonymized image is then generated by optimizing the latent code through identity-removing and utility-preserving loss functions. While the framework demonstrates a strong effort in balancing privacy protection and data utility, several key considerations and limitations must be acknowledged.

### 3.5.1 Discussion

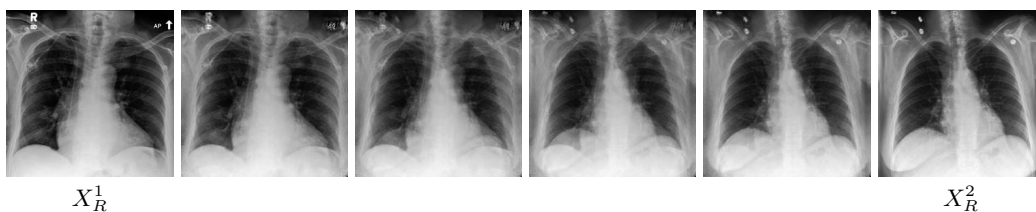
**Preservation of Diagnostic Information:** The primary advantage of using generative models for anonymization is the potential to produce high-fidelity synthetic images that retain essential diagnostic characteristics. This is crucial in medical settings where the accuracy of images directly impacts patient care and treatment outcomes. By training on diverse and representative datasets, the GAN can learn to replicate crucial medical features, ensuring that anonymized images remain clinically useful.

**Scalability and Efficiency:** The proposed generative image anonymization approach, once trained, can anonymize large volumes of images rapidly, which is a significant benefit for healthcare facilities handling extensive imaging datasets. This scalability makes the approach viable for integration into clinical workflows where time and resources are critical.

**Regulatory Compliance:** The use of the proposed generative image anonymization approach aligns with privacy metrics such as the singling out risk and the linkability risk by effectively removing personally identifiable information. This compliance is essential for legal and ethical reasons, as it protects patient confidentiality while allowing data sharing for research and collaboration.

### 3.5.2 Limitations

**Entangled Latent Space** To achieve the goal of removing identity-related information while preserving utility-based features, the disentanglement property of the latent space is crucial.



**Fig. 3.12.:** Linear interpolation between the two arbitrary latent points. The leftmost and rightmost images correspond to the two latent points, while the images in the middle represent the linear interpolation between these latent points with a new latent code of  $(1 - \lambda)Z_{X_R^1} + \lambda Z_{X_R^2}$ .

In Fig. 3.12, we analyze the entanglement property of the latent code by performing a linear interpolation between two arbitrary points in the latent space. The resulting interpolations resemble overlapping images rather than clinically meaningful ones, suggesting that the latent space is highly entangled in our case.

Furthermore, since the goal of the generative network is not only to faithfully reconstruct the real image but also to effectively anonymize it through latent code optimization, we face the reconstruction-editing trade-off discussed in previous works [Katsumata et al., 2024]. This trade-off refers to the balance between faithfully reconstructing an image and enabling modifications, such as attribute editing. If the model prioritizes reconstruction, it may struggle to effectively alter key features for image editing. Conversely, if the model emphasizes editing flexibility, the quality and fidelity of the reconstructed images may deteriorate. In our case, the high-fidelity reconstruction indicates a compromise in the optimization of the latent code, referred to as editing in previous studies. This also explains why the identity and utility information exhibit a highly global nature in the latent space in our experiments.

Therefore, to further enhance anonymization performance, it is crucial to first construct a more disentangled latent space. In a disentangled latent space, each dimension (or group of dimensions) corresponds to a distinct and independent feature or factor of the data. This structure allows for more precise and targeted control over the underlying variations, making it particularly useful for manipulating specific data characteristics while maintaining privacy.

## 3.6 Conclusions

We present GMIA, a novel framework for generative medical image anonymization, which involves discerning and removing identity information from real images. Experimental results demonstrate that GMIA can effectively anonymize large datasets with numerous identities while preserving both privacy and downstream utility. In future work, we aim to further improve the performance of the current method. First, we plan to relax the strong constraint imposed by the faithful mapping of the AE-GAN model. Second, we intend to locate specific identity channels in the latent space and conduct fine optimization exclusively on these channels. Lastly, we intend to extend the application of our approach to other imaging modalities, such as MR or CT.

# GMIA2: Designing an Encoder for Generative Medical Image Anonymization

---

4.1	Introduction . . . . .	70
4.2	Related Work . . . . .	71
4.2.1	StyleGAN . . . . .	71
4.2.2	StyleGAN Inversion . . . . .	72
4.3	Method . . . . .	73
4.3.1	Designing an Encoder for Image Reconstruction . . . . .	74
4.3.2	Latent Code Optimization for Image Anonymization . . . . .	78
4.4	Materials . . . . .	79
4.4.1	Evaluation Metrics . . . . .	79
4.4.2	Baseline and Ablation Study . . . . .	80
4.4.3	Implementation Details . . . . .	81
4.5	Results . . . . .	81
4.5.1	Reconstruction Results . . . . .	81
4.5.2	Anonymization Results . . . . .	82
4.5.3	Ablation Study . . . . .	85
4.5.4	Disentanglement Property of Latent Space . . . . .	91
4.6	Discussion and Limitations . . . . .	91
4.6.1	Discussion . . . . .	91
4.6.2	Limitations . . . . .	93
4.7	Conclusions . . . . .	95

---

**Abstract** In this chapter, we build upon the work introduced previously by designing a streamlined encoder and a co-training scheme to learn a more disentangled latent space. Our encoder leverages the power of StyleGAN2 to create a more robust and scalable solution. As in the previous chapter, the generative medical image anonymization framework consists of two primary stages: encoding and optimization. In the encoding stage, we develop an effective encoder to accurately map input images into a latent space. During the optimization stage, we introduce an identity removal loss and a utility preservation loss to refine the latent code, ensuring that the anonymized images obscure identifiable features while retaining essential diagnostic information. This dual-loss strategy carefully balances the trade-off between identity removal and utility preservation. We demonstrate the effectiveness of our framework through extensive qualitative and quantitative experiments on the MIMIC-CXR dataset. The results show that our method not only produces high-quality anonymized images but also preserves the integrity of critical diagnostic features, making the images suitable for training machine learning models for lung pathology classification. This chapter will be published as a conference paper in [Li et al., 2025].

## 4.1 Introduction

The disentanglement property of a latent space plays a critical role in latent code optimization for image anonymization. A disentangled latent space ensures that different dimensions (or sets of dimensions) represent independent, semantically meaningful features of the image.

In the context of image anonymization, this separation is crucial. If the latent space is not well-disentangled, modifications aimed at altering identity attributes may inadvertently affect other important features, such as those critical for medical diagnosis. Disentanglement allows for fine-grained control during latent code optimization, enabling precise editing to identity features while preserving utility features intact. This capability ensures that anonymized images retain their diagnostic value while effectively protecting patient privacy.

StyleGAN2 [Karras et al., 2020] is known for its high-quality image generation and well-structured latent space, making it a strong candidate for tasks like image editing and anonymization. The latent space of StyleGAN2 is hierarchical [Katzir et al., 2022], with different layers controlling different levels of detail in the image. For example, early layers in the latent space correspond to high-level features (e.g., pose or shape), while later layers influence fine details (e.g., texture or color). This hierarchical structure

facilitates disentanglement, allowing certain aspects of an image (such as identity) to be modified without affecting others (such as utility-related features).

To leverage the latent space of StyleGAN2 for image anonymization, it's essential to first project real images into this latent space. This process, known as StyleGAN inversion [Wei et al., 2022], involves designing an effective encoder that maps a given real image  $X_R$  to its corresponding latent code  $W$  in the StyleGAN2 latent space. Unlike conventional encoder-based StyleGAN inversion approaches [Richardson et al., 2021; Tov et al., 2021], which typically train the encoder while keeping the StyleGAN generator fixed, we design an effective encoder and introduce a co-training scheme that jointly optimizes the encoder and the StyleGAN2 network. The goal is to minimize the reconstruction loss, defined as the difference between the original image  $X_R$  and the reconstructed image  $G(E(X_R))$ . By co-training the encoder and generator, the encoder becomes more closely aligned with the StyleGAN2 generator, leading to more accurate latent space projections. This, in turn, enables a more precise latent code optimization process where identity features can be selectively modified without affecting utility-related features. As a result, the framework generates anonymized images that maintain diagnostic value while ensuring robust privacy protection.

## 4.2 Related Work

### 4.2.1 StyleGAN

Generative Adversarial Networks (GANs) [Goodfellow et al., 2014] have revolutionized the field of generative modeling. The core idea involves training two neural networks, a generator and a discriminator, in a minimax game, where the generator aims to produce realistic data samples, and the discriminator endeavors to distinguish between real and generated samples. This pioneering work set the foundation for numerous advancements and variations in GAN architectures.

One of the most significant evolutions from the original GAN is the Deep Convolutional GAN (DCGAN) [Radford et al., 2015]. DCGANs integrated convolutional layers into GANs, which significantly improved the quality of generated images and enabled more stable training. This was a critical step towards applying GANs to high-resolution image synthesis.

Further advancements were achieved with Progressive Growing of GANs (ProGANs) [Karras et al., 2017]. ProGANs incrementally increase the resolution of both the generator and discriminator during training, which facilitates the generation of high-resolution

images. This technique effectively addressed the instability issues associated with training GANs on high-resolution datasets and laid the groundwork for StyleGAN.

StyleGAN [Karras et al., 2019] brought a novel approach to high-resolution image synthesis by disentangling the generation process into a mapping network and a synthesis network. This architecture introduced style-based generator control, enabling intuitive manipulation of image attributes by adjusting the latent space of the GAN. This design not only improved image quality but also provided unprecedented control over generated images, marking a significant milestone in generative modeling.

In subsequent work, StyleGAN2 [Karras et al., 2020] further refined this architecture by addressing certain artifacts and enhancing the fidelity of generated images. Key improvements included reconfiguring the normalization techniques and the introduction of weight demodulation, which led to even higher quality outputs. StyleGAN3 [Karras et al., 2021], continuing this trend, aimed to eliminate aliasing artifacts and improve the generation of images that are more coherent and free of pixel-level imperfections.

Additionally, other variations and extensions of GANs have contributed to the field. Conditional GANs (cGANs) [Mirza et al., 2014] condition the generation process on auxiliary information, allowing for controlled generation of images based on class labels or other input data. Pix2Pix [Isola et al., 2017] and CycleGAN [Zhu et al., 2017] are notable examples that have shown success in tasks like image-to-image translation.

Finally, Variational Autoencoders (VAEs) [Kingma et al., 2013] and their hybrids with GANs, such as VAE-GANs [Larsen et al., 2016], provide alternative approaches to generative modeling. VAEs introduce a probabilistic framework for generating data, complementing the deterministic nature of GANs and offering unique advantages in terms of latent space structure and interpretability.

## 4.2.2 StyleGAN Inversion

StyleGAN inversion is a significant area of research that focuses on mapping real images back into the latent space of a pre-trained StyleGAN model. This task is crucial for enabling various image manipulation applications, such as editing, attribute transfer, and reconstruction. The process of inversion involves several challenges and has been addressed through diverse approaches in the literature.

The concept of GAN inversion dates back to early works on Generative Adversarial Networks (GANs). Zhu et al. [Zhu et al., 2017] proposed an initial framework for GAN inversion by optimizing the latent code to reconstruct real images using the generator network. This optimization-based approach, although effective, was computationally



intensive and often resulted in suboptimal reconstructions due to the non-convex nature of the latent space.

Building upon these foundations, early approaches to StyleGAN inversion focused on improving the accuracy and efficiency of the inversion process. Richardson et al. introduced pSp (pixel2style2pixel) [Richardson et al., 2021], a versatile framework for image-to-image translation tasks, including StyleGAN inversion. pSp employed an encoder network to predict latent codes directly, bypassing the need for iterative optimization and enabling faster and more accurate reconstructions.

Subsequent research has sought to refine the quality of inversions and extend the capabilities of inversion frameworks. Tov et al. presented e4e (encoder4editing) [Tov et al., 2021], which improved upon pSp by using a more sophisticated encoder architecture that better preserved high-frequency details and semantic consistency in the inverted images. This work highlighted the importance of encoder design in achieving high-fidelity reconstructions.

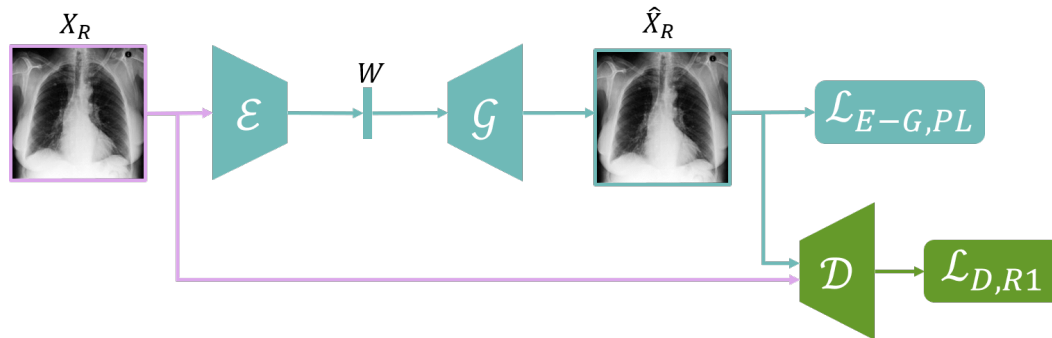
Another notable contribution to the field is the work on StyleGAN2-based inversion techniques. Image2StyleGAN [Abdal et al., 2019] introduced an efficient algorithm to embed a given image into the latent space of a pre-trained StyleGAN. This approach was further enhanced with Image2StyleGAN++ [Abdal et al., 2020], which incorporated an iterative refinement process to fine-tune the latent codes for improved reconstructions. This extension balanced the trade-off between computational efficiency and reconstruction quality, making it a practical solution for real-world applications.

Moreover, hybrid approaches that combine optimization-based and encoder-based methods have emerged to leverage the strengths of both paradigms. For instance, StyleALAE [Pidhorskyi et al., 2020] integrated an encoder network with an adversarial autoencoder framework, achieving high-quality inversions while maintaining computational efficiency. This hybrid strategy demonstrated the potential of combining different inversion techniques to achieve superior performance.

## 4.3 Method

As discussed in Chapter 3, the entanglement of the latent space  $Z$  poses a significant challenge in selectively anonymizing identifiable characteristics without inadvertently altering other crucial, diagnostically relevant features of the image. This entanglement can cause the optimization of latent code  $Z$  to have a broad, unintended impact on the anonymized image, often resulting in undesirable artifacts, such as blurring, which degrade overall image quality. To address this issue, we leverage a more disentangled

latent space, ensuring that modifications made during the anonymization process are both localized and controlled. Based on previous research [Katzir et al., 2022], the  $W$  space of StyleGAN2 [Karras et al., 2020] has demonstrated superior disentanglement properties, making it an ideal choice for our approach to achieve effective, targeted anonymization while preserving image utility.



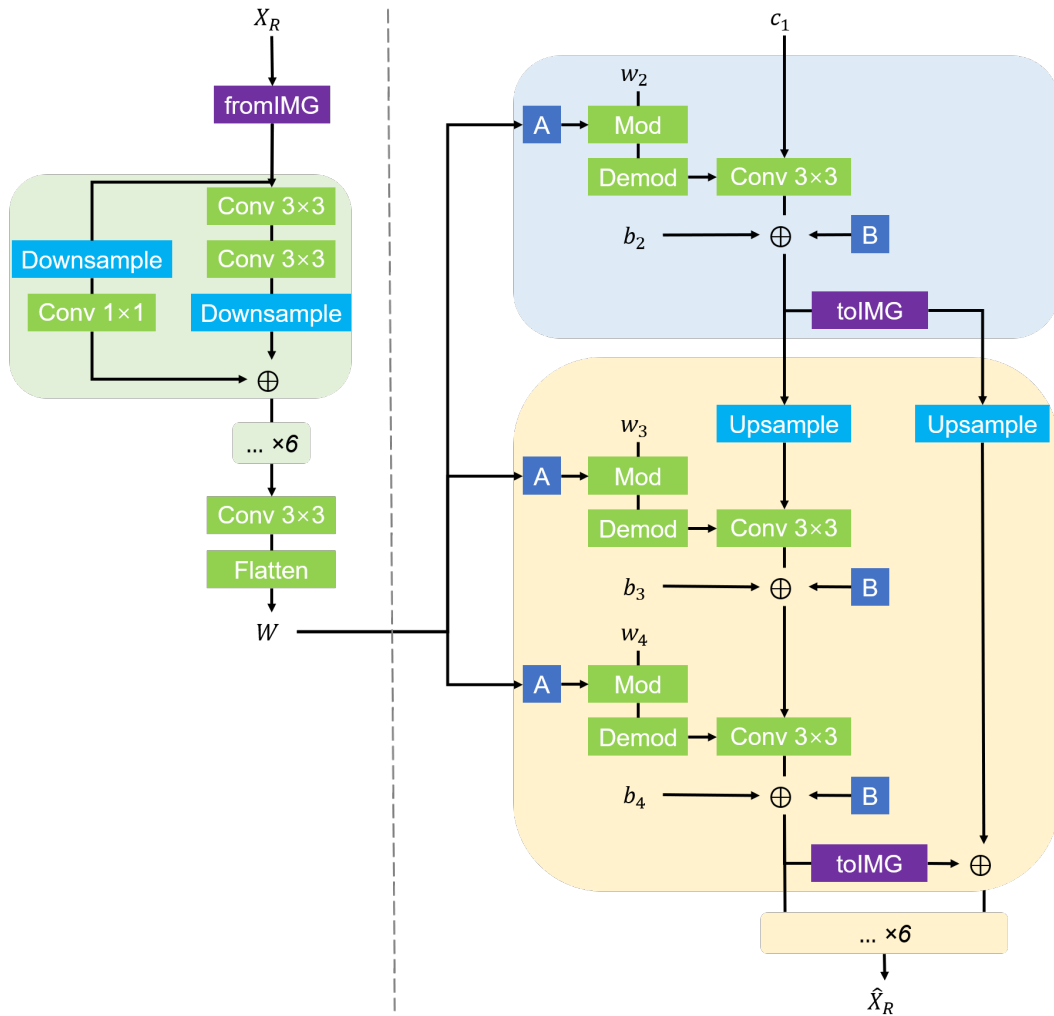
**Fig. 4.1.:** Overview of the generative reconstruction network: The encoder  $\mathcal{E}$  is a custom-designed model that processes a real image  $X_R$  to extract its latent representation  $W$ . This latent code  $W$  is then fed into the StyleGAN2 generator  $\mathcal{G}$  [Karras et al., 2020], which reconstructs the image as  $\hat{X}_R$ . To ensure the quality and realism of the reconstruction, the StyleGAN2 discriminator  $\mathcal{D}$  evaluates whether  $\hat{X}_R$  appears authentic or synthetic.

To achieve precise latent code mapping from the input image, we build on the approach introduced in Chapter 3, where the latent code mapping is framed as an image reconstruction task using an AE-GAN architecture (see Fig. 4.1). However, in this case, we replace the GAN model in Chapter 3 with the more advanced StyleGAN2 network. Consequently, we design a specialized encoder tailored to integrate seamlessly with the StyleGAN2 framework within the AE-GAN architecture. Additionally, we introduce a novel co-training strategy that jointly optimizes both the encoder and the StyleGAN2 generator, allowing for a more seamless integration between the two components.

### 4.3.1 Designing an Encoder for Image Reconstruction

The goal of designing an encoder for image reconstruction is to develop a model that accurately maps input images to a latent space, from which a corresponding decoder can reconstruct the images with high fidelity.

The core idea behind the custom encoder is to design a simple yet effective architecture that accurately captures the latent representation of a given input image while maximizing the benefits of latent space disentanglement. To accomplish this, we structure the encoder's architecture symmetrically to the generator's design. Symmetrically designing the encoder with respect to the generator allows for the reuse of StyleGAN2 components. This not only ensures better alignment between the two networks but also reduces the need to design completely new modules [Wei et al., 2022]. Moreover, inspired by the



**Fig. 4.2.:** The architecture consists of the encoder (left of the dashed line) and the generator (right of the dashed line). The `fromIMG` and `toIMG` blocks handle the conversion between feature maps and image data. `Upsample` and `Downsample` refer to bilinear upsampling and downsampling operations, respectively. `Conv 1×1` and `Conv 3×3` represent 1×1 and 3×3 convolution layers. The `Flatten` layer converts 2D feature maps into long linear vectors via a fully connected layer. The `A` operation is a learned affine transform from  $W$  that produces a style, while `B` denotes a noise broadcast operation. `Mod` and `Demod` refer to the modulation and demodulation operations as described in StyleGAN2 paper [Karras et al., 2020]. The variables  $c$ ,  $w$ , and  $b$  represent the constant input, learned weights, and biases, respectively. The notation  $\times 6$  indicates that the operations in the colored boxes are repeated six times.

success of symmetric architectures like AE, this approach leverages the symmetry between the encoder and generator networks, facilitating more effective feature preservation during inversion. As illustrated in Fig. 4.2, the process begins with a fromIMG operation that converts the input image into feature maps for subsequent convolutional processing. At each resolution, we implement a residual network architecture, consisting of two  $3 \times 3$  convolution layers followed by a bilinear downsampling operation to reduce the feature map size. The skip connection includes a bilinear downsampling operation and a  $1 \times 1$  convolution layer.

As noted by [Tov et al., 2021], the  $W$  latent space offers better editability and perceptual quality compared to its  $W^+$  counterpart. The  $W^+$  space is an extended version of  $W$ , where the  $W$  latent code is a 512-dimensional vector, while  $W^+$  is a concatenation of 16 different 512-dimensional  $W$  vectors, each controlling a different layer of the generator. The  $W^+$  allows for finer, layer-specific control over the generated image, but comes at the cost of increased complexity. Additionally, StyleGAN2 was originally trained in the  $W$  space, making it a natural choice for our approach. Accordingly, we first map the input image to its latent representation in the  $W$  space. We then use the broadcasting operation, as in StyleGAN2, to duplicate the  $W$  latent code into 16 copies, which subsequently control the generator through modulation and demodulation at each convolution layer.

**Training Scheme** To achieve a smoother and more accurate mapping from the input image to its  $W$  latent space, we propose a co-training scheme that simultaneously trains both the encoder and the StyleGAN2 model from scratch. Unlike traditional encoder-based GAN inversion methods, which typically train the encoder on a fixed, pre-trained GAN model, our co-training approach optimizes the encoder and StyleGAN2 generator together. This joint optimization leads to better integration between the two networks, resulting in improved performance and more accurate latent space projections.

This approach offers several key advantages. First, there is no available pre-trained StyleGAN2 model specifically for medical images, which makes our method particularly valuable. The co-training scheme eliminates the need to fine-tune a pre-trained StyleGAN2 model on a new medical image dataset, thus ensuring that the model is tailored from the outset to the specific characteristics of the medical images. Second, by training the encoder alongside the StyleGAN2 network in an image reconstruction context, we achieve a more accurate mapping of input images to their corresponding latent codes, enhancing the fidelity and effectiveness of the GAN inversion process.

In the co-training process, we incorporate reconstruction losses to achieve the objective of accurate image reconstruction. The training process is divided into four distinct phases:

**Phase One: Encoder and Generator Training.** In this initial phase, we train the encoder and the generator simultaneously using a combination of reconstruction loss and non-

saturating generator loss [Goodfellow et al., 2020]. The reconstruction loss ensures that the generated images closely match the original input images, while the generator loss drives the generator to produce realistic images.

$$\begin{aligned} \mathcal{L}_{\mathcal{E}-\mathcal{G}}(X_R, \hat{X}_R) = & \lambda_{\text{pixel}} \|X_R - \hat{X}_R\|_2 \\ & + \lambda_{\text{perceptual}} \|\phi(X_R) - \phi(\hat{X}_R)\|_2 \\ & - \lambda_{\mathcal{G}} \mathbb{E}_{\hat{X}_R} [\log \mathcal{D}(\hat{X}_R)] \end{aligned} \quad (4.1)$$

where  $\lambda_{\text{pixel}}$ ,  $\lambda_{\text{perceptual}}$ ,  $\lambda_{\mathcal{G}}$  are hyper-parameters used to balance the contributions of the respective loss components.  $\phi$  represents the feature extraction network, such as a pre-trained VGG network [Simonyan et al., 2014] with 19 layers.

**Phase Two: Path Length Regularization.** In the second phase, we apply path length regularization [Karras et al., 2020] to the generator. This regularization technique stabilizes the generator by enforcing smoothness in the latent space, ensuring that small changes in the latent code result in proportional changes in the generated images. This helps maintain the consistency and quality of the generated images over the training process.

The path length regularization loss can be expressed as:

$$L_{PL} = \mathbb{E}_{W, \tilde{X} \in \mathcal{N}(0, I)} \left( \left\| \nabla_{\mathbf{w}}(\mathcal{G}(W) \cdot \tilde{X}) \right\|_2 - a \right)^2 \quad (4.2)$$

where  $\tilde{X}$  are random images with normally distributed pixel intensities. The constant  $a$  is set dynamically during optimization as the exponential moving average of the path lengths  $\left\| \nabla_{\mathbf{w}}(\mathcal{G}(W) \cdot \tilde{X}) \right\|_2$ , allowing the optimization to find a suitable global scale by itself.

**Phase Three: Discriminator Training.** During the third phase, we focus on training the discriminator using the discriminator loss of the nonsaturating loss [Goodfellow et al., 2020]. This loss function helps the discriminator differentiate between real and generated images, thereby improving its ability to guide the generator towards producing more realistic outputs.

The discriminator loss can be expressed as:

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}_{X_R} [-\log \mathcal{D}(X_R)] + \mathbb{E}_{\hat{X}_R} [-\log(1 - \mathcal{D}(\hat{X}_R))] \quad (4.3)$$

where  $X_R$  is the real image while  $\hat{X}_R$  represents the reconstructed image.

**Phase Four:  $R_1$  Regularization.** In the final phase, we apply the  $R_1$  regularization [Mescheder et al., 2018] to the discriminator. This regularization term penalizes the

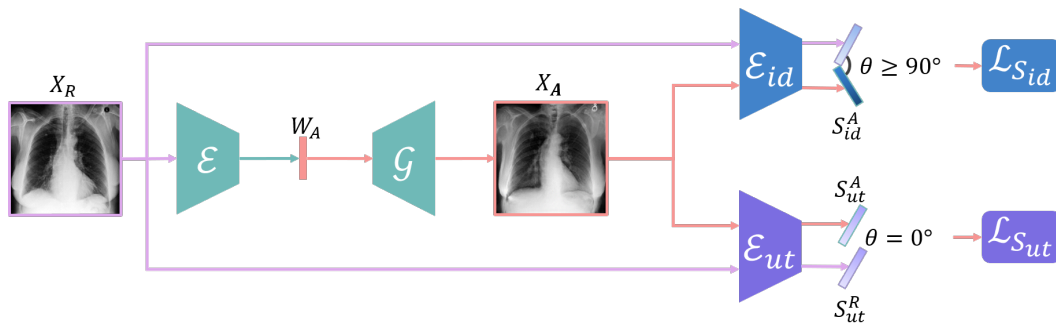
gradient of the discriminator’s output with respect to its input, enhancing the stability of the discriminator and preventing it from producing overly sharp gradients.

The  $R_1$  regularization loss can be expressed as:

$$L_{R_1} = \frac{1}{2} \mathbb{E}_{\mathbf{X}_R} \left[ \|\nabla_{\mathbf{X}_R} \mathcal{D}(\mathbf{X}_R)\|^2 \right] \quad (4.4)$$

We incorporate lazy regularization, as implemented in StyleGAN2 [Karras et al., 2020], to improve training efficiency. Specifically, lazy regularization is applied to both path length regularization and  $R_1$  regularization. Rather than calculating these regularization terms at every iteration, they are computed less frequently, reducing computational overhead while preserving the stability and overall effectiveness of the training process.

By systematically progressing through these phases, we effectively co-train the encoder and the StyleGAN2 components, leveraging reconstruction and regularization techniques to achieve high-fidelity image reconstructions.



**Fig. 4.3.:** Overview of the latent code optimization process: the encoder  $\mathcal{E}$  first projects a real image  $X_R$  into its latent space  $W$ , initializing the latent code  $W_A$ . This latent code is then optimized using two deep network-based loss functions:  $\mathcal{L}_{S_{id}}(X_R, X_A)$ , which removes the identity information, and  $\mathcal{L}_{S_{ut}}(X_R, X_A)$ , which preserves the utility information within the latent code  $W$ . Finally, a generator  $\mathcal{G}$  projects the optimized latent code  $W$  back into the image space, producing the anonymized version  $X_A$  of the real image  $X_R$ .

### 4.3.2 Latent Code Optimization for Image Anonymization

The process of latent code optimization follows the same approach as described in Section 3.2.5. However, instead of using the latent code  $Z$  from Chapter 3, we now employ the  $W$  latent code from StyleGAN2 [Karras et al., 2020], as illustrated in Fig. 4.3. The disentangled  $W$  space offers a powerful tool for controlling and manipulating various attributes of generated images. By optimizing the  $W$  latent code, we can selectively modify specific, identifiable features while preserving essential, diagnostically relevant characteristics of the image.

## 4.4 Materials

### 4.4.1 Evaluation Metrics

We evaluate our method by quantifying image reconstruction quality, utility preservation, and privacy preservation. Below, we briefly introduce the metrics used for these evaluations.

#### Image Reconstruction Quality

To assess the quality of reconstructed images, we use several image quality metrics, including those mentioned in Section 3.3.4. Additionally, we report the Information Content Weighted Structural Similarity Measure (IW\_SSIM) [Wang et al., 2010]. IW\_SSIM extends the Structural Similarity (SSIM) [Wang et al., 2004] index by incorporating the concept of information content weighted pooling, which provides a more comprehensive evaluation of image quality by considering the importance of different regions within the image.

#### Utility Preservation

We propose a protocol to quantify how well the anonymized dataset retains data utility. Specifically, the evaluation is conducted through a standard downstream task: classification. More specifically, we employ a DenseNet121 [Iandola et al., 2014] model to perform multi-label classification, training it with soft labels to handle uncertainty and ensure robustness. Unlike the approach in Section 3.3.4, we perform a bilateral evaluation, assessing:

1.  $X_A \rightarrow X_R$  **Utility Evaluation:** The accuracy of utility classifiers evaluated on real test sets (1 641 images) when trained on anonymized training sets (50 875 images).
2.  $X_R \rightarrow X_A$  **Utility Evaluation:** The accuracy of utility classifiers evaluated on anonymized test sets (1 641 images) when trained on real training sets (50 875 images).

#### Privacy Preservation

To evaluate privacy preservation, we measure how effectively identifiable features are removed or obfuscated in the anonymized images.

First, we use privacy metrics to evaluate the Singling-out Risk. We report these metrics in two spaces: the identity semantic feature space  $S_{id}$  of the identity classifier and the  $W$  latent space of StyleGAN2. In addition to the metrics listed in Section 3.3.4, we introduce a new metric, 'Dist', to evaluate the distance between each real image and its anonymized counterpart.

Then, we employ identity verification classifiers to quantify the reduction in identifiable information, assessing the Linkability Risk from two perspectives:

1. **Inner Risk Evaluation:** This assesses whether two anonymized images belong to the same patient.
2. **Outer Risk Evaluation:** This evaluates whether a real image and its corresponding anonymized image belong to the same patient.

For this evaluation, the identity classifier is trained on the real training set. The details of the evaluation data collection and methodology can be referred to in Section 3.4.4. By using these two evaluation metrics, we can effectively measure both the internal consistency of anonymized images and the degree to which anonymization masks the identity of individuals when compared to their real images.

Finally, we further evaluate privacy preservation using the success rates of Membership Inference Attacks (MIA). This evaluation is conducted in a bilateral manner:

1.  $X_R \rightarrow X_A$  **Attack Evaluation:** The attacker predicts the membership on anonymized test sets when trained on real training sets.
2.  $X_A \rightarrow X_R$  **Attack Evaluation:** The attacker predicts the membership on real test sets when trained on anonymized training sets.

In the MIA, the attacker is implemented as a simple fully connected classification network. The training/validation/test split structure follows the one outlined in Section 3.3.4. An ideal anonymization method should effectively prevent a successful MIA, turning a potentially effective attack (100% success rate on the test set) into a failure (0% success rate on the target test set). These metrics provide insight into how well the anonymization process protects against potential privacy breaches by adversaries attempting to infer membership in the dataset.

## 4.4.2 Baseline and Ablation Study

To evaluate the effectiveness of the proposed generative image anonymization framework, we use the image reconstruction network with the proposed co-training scheme as a baseline for comparison.

Additionally, we conducted an ablation study to assess the impact of several factors on the anonymization process, including the effects of  $\mathcal{L}_{S_{ut}}$ ,  $\mathcal{L}_{S_{id}}$ , the margin  $m$  in  $\mathcal{L}_{S_{id}}$ , and the loss weight  $\lambda$ . To evaluate the effect of  $\mathcal{L}_{S_{ut}}$  and  $\mathcal{L}_{S_{id}}$ , we performed experiments



using only  $\mathcal{L}_{S_{ut}}$ , only  $\mathcal{L}_{S_{id}}$ , and both together. For the margin  $m$  in  $\mathcal{L}_{S_{id}}$ , we varied the values by setting  $m = 0.0$ ,  $m = -0.5$ , and  $m = -1.0$  to observe its influence on identity removal. Lastly, we examined the effect of the loss weight  $\lambda$  by adjusting its value in  $\lambda\mathcal{L}_{S_{ut}} + (1 - \lambda)\mathcal{L}_{S_{id}}$ , ranging from 0.1 to 0.9 to find the optimal balance between utility preservation and identity removal.

### 4.4.3 Implementation Details

#### Reconstruction Network.

The reconstruction network is a combination a custom-designed encoder and the StyleGAN2 [Karras et al., 2020] network. It processes both input and output images at a resolution of  $512 \times 512$  pixels. The weights for the pixel-wise loss  $\lambda_{\text{pixel}}$ , perceptual loss  $\lambda_{\text{perceptual}}$ , and generator loss  $\lambda_G$  are all set to 1. All other experimental settings follow those outlined in the StyleGAN2 paper [Karras et al., 2020] for training the reconstruction network.

#### Identity and Utility Network.

We utilize iResNet50 [Duta et al., 2021] for multi-class identity classification and DenseNet121 [Iandola et al., 2014] for multi-label pathology classification. The experimental settings from Chapter 3 are maintained for training both networks.

#### Latent Code Optimization.

For latent code optimization, both the input and the output anonymized images have a resolution of  $512 \times 512$ , with a batch size of 1. The latent code  $W$  is a 512-dimensional vector. The margin  $m$  in  $\mathcal{L}_{S_{id}}(X_R, X_A)$  is set by default to -0.7. We set  $\lambda_{id}$  and  $\lambda_{ut}$  to 0.5. The Adam optimizer [Kingma et al., 2014] is used with a learning rate of  $1 \times 10^{-3}$  and a weight decay of  $1 \times 10^{-2}$ , and training is conducted for a maximum of 1000 epochs or until the value of the identity removal loss,  $\mathcal{L}_{S_{id}}(X_R, X_A)$ , falls below  $1 \times 10^{-8}$ .

## 4.5 Results

### 4.5.1 Reconstruction Results

We begin by evaluating the effectiveness of the latent code mapping with the image reconstruction network, which involves finding the latent code of real images in the  $W$  latent space.

**Quantitative Results** Table 4.1 presents a quantitative comparison of the reconstruction performance. The "Co-training" approach refers to our proposed method where both

**Tab. 4.1.:** The results of generative reconstruction network pre-training. The "Co-training" refers to the proposed co-training scheme, where both the encoder and StyleGAN2 generator are optimized together, while the " $\mathcal{E}$ -training" represents the traditional encoder-based StyleGAN inversion, where the encoder is trained with a fixed, pre-trained StyleGAN generator.

Method	PSNR $\uparrow$	SSIM $\uparrow$	MS_SSIM $\uparrow$	IW_SSIM $\uparrow$	LPIPS $\downarrow$
Co-training	24.319	0.968	0.999	0.602	0.340
$\mathcal{E}$ -training	19.477	0.936	0.995	0.363	0.446

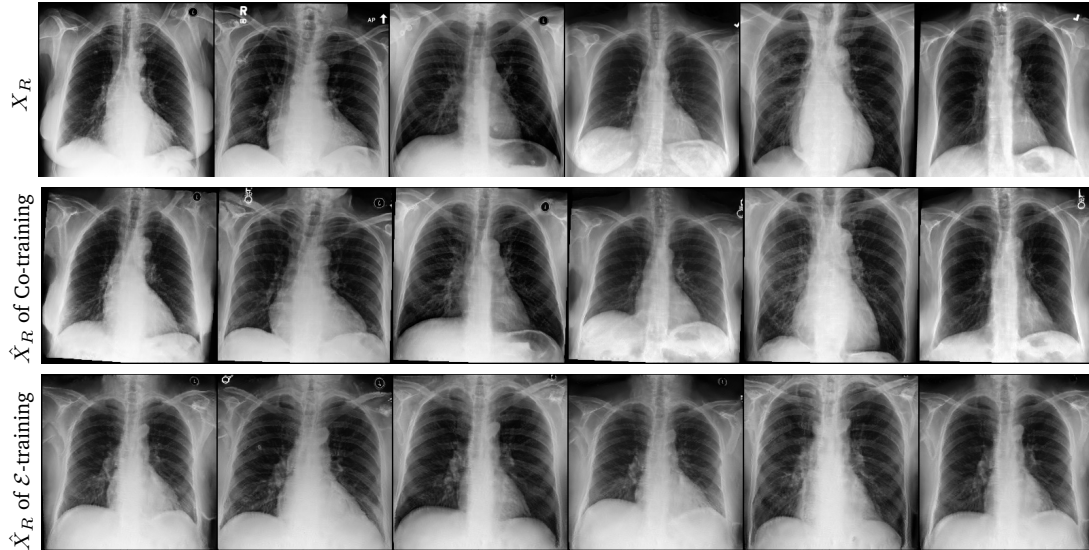
the encoder and StyleGAN2 generator are optimized together, while the " $\mathcal{E}$ -training" approach represents the traditional encoder-based StyleGAN inversion, where the encoder is trained with a fixed, pre-trained StyleGAN generator. The results demonstrate that the co-training scheme yields superior reconstruction accuracy, as evidenced by the higher PSNR and IW\_SSIM scores. In contrast, the  $\mathcal{E}$ -training approach shows a noticeable decline in performance, particularly in these key metrics, underscoring the advantages of joint optimization in improving reconstruction quality.

Compared to the reconstruction results presented in Section 3.4.1, while the SSIM and MS\_SSIM metrics show no discernible differences, there is a noticeable decline in the PSNR and LPIPS metrics. This outcome illustrates the reconstruction-editability trade-off. When the latent code  $Z \in \mathbb{R}^{256 \times 16 \times 16}$  in Section 3.4.1 is highly entangled, it can achieve high-fidelity reconstruction images but loses the capability for fine-grained editability. As a result, optimizing the latent code  $Z$  leads to global changes in the image. Conversely, the latent code  $W \in \mathbb{R}^{512}$  of StyleGAN2, being highly disentangled, sacrifices some fidelity in reconstruction but offers superior editability. The disentangled nature of the  $W$  space allows for targeted modifications, enabling the selective anonymization of specific features without affecting the entire image. This balance is crucial for effective anonymization, where maintaining utility while protecting privacy is paramount.

**Qualitative Results** As shown in Fig. 4.4, the image reconstruction network using the proposed co-training scheme effectively reproduces input images with finer details and higher fidelity. In contrast, the reconstruction results from the  $\mathcal{E}$ -training method exhibit noticeable discrepancies when compared to the original images. This visual comparison highlights the improvements in reconstruction quality achieved through the co-training approach.

## 4.5.2 Anonymization Results

In this section, we quantify the utility preservation and identity elimination achieved by the anonymization methods.



**Fig. 4.4.:** Reconstruction results of the latent code mapping. The first row displays the real images  $X_R$ . The second row displays the reconstructed images  $\hat{X}_R$  using the proposed co-training scheme, where both the encoder and StyleGAN2 generator are jointly optimized. The third row shows the reconstructed images  $\hat{X}_R$  generated by the traditional encoder-based StyleGAN inversion method, where the encoder is trained separately with a fixed, pre-trained StyleGAN generator.

**Utility Preservation.** To assess utility preservation, we use a pre-trained utility classifier, trained and tested on the real dataset, as our baseline model (refer to row 1 in Table 4.3). The results of this utility network pre-training on four predicted labels are shown in Table 4.2. These results highlight the utility network’s predictive capabilities across specific pathologies, showing that the classifier performs well in identifying Pleural Effusion, outperforming its predictions for Lung Opacity and Atelectasis.

**Tab. 4.2.:** Results of the utility network pre-training on four predicted labels ‘Acc’: accuracy, ‘AP’: average precision, ‘AUROC’: area under the ROC curve, ‘F1’: F1 score.

Label	Acc $\uparrow$	AP $\uparrow$	AUROC $\uparrow$	F1 $\uparrow$
Lung Opacity	0.6734	0.5093	0.6664	0.2717
Atelectasis	0.7325	0.5111	0.7561	0.3398
Pleural Effusion	0.7770	0.7303	0.8391	0.6975
Support Devices	0.7764	0.7746	0.8363	0.7362

For the  $X_A \rightarrow X_R$  Utility Evaluation (rows 3-9 in Table 4.3), the performance of the classifiers trained on various anonymized datasets shows minimal difference compared to the baseline. This indicates the effectiveness of the proposed anonymization approach in preserving utility, as the anonymized datasets maintain a similar level of data utility compared to their real counterparts.

In the  $X_R \rightarrow X_A$  Utility Evaluation (rows 11-17 in Table 4.3), the results show that the pre-trained utility classifier performs similarly when tested on both the real dataset and

various anonymized datasets. These findings further validate the effectiveness of the proposed anonymization approach in preserving utility.

Compared with the Utility Evaluation results for the reconstructed images produced by the AE-GAN model (rows 2,10 in Table 4.3) and the baseline results (row 1 in Table 4.3), it is evident that the AE-GAN model may lose some utility information during training, leading to a degradation in the Utility Evaluation results for the reconstructed images. This suggests that incorporating the utility-preserving loss during the training of the AE-GAN model could help mitigate the loss of utility information.

**Tab. 4.3.:** Results of utility network for pre-training and utility preservation. 'Acc': accuracy, 'AP': average precision, 'AUROC': area under the ROC curve, 'F1': F1 score.

	Methods	Acc $\uparrow$	AP $\uparrow$	AUROC $\uparrow$	F1 $\uparrow$
Pre-training	/	0.740	0.631	0.774	0.511
$X_A \rightarrow X_R$ Utility Evaluation	Co-training	0.718	0.586	0.737	0.451
	$\mathcal{L}_{S_{ut}}$	0.726	0.600	0.747	0.493
	$\mathcal{L}_{S_{id},m=0.0}$	0.719	0.583	0.734	0.449
	$\mathcal{L}_{S_{id},m=-0.5}$	0.714	0.579	0.731	0.434
	$\mathcal{L}_{S_{id},m=-0.7}$	0.696	0.559	0.713	0.424
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=0.0}$	0.731	0.608	0.752	0.517
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.5}$	0.727	0.593	0.749	0.515
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.7}$	0.713	0.578	0.737	0.531
	$X_R \rightarrow X_A$ Utility Evaluation	Co-training	0.705	0.560	0.719
$\mathcal{L}_{S_{ut}}$		0.732	0.625	0.767	0.497
$\mathcal{L}_{S_{id},m=0.0}$		0.699	0.559	0.716	0.452
$\mathcal{L}_{S_{id},m=-0.5}$		0.685	0.531	0.682	0.438
$\mathcal{L}_{S_{id},m=-0.7}$		0.665	0.490	0.647	0.397
$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=0.0}$		0.725	0.614	0.759	0.478
$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.5}$		0.735	0.624	0.767	0.498
$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.7}$		0.737	0.623	0.764	0.498

**Identity Elimination.** As shown in Table 4.4, the anonymized datasets exhibit a low risk of singling out in the identity semantic space  $S_{id}$  (rows 1-7). This indicates that the anonymization methods are effective at removing identifiable information. However, the anonymized datasets show a higher risk of singling out in the latent space (row 8-14). This is expected, as the latent space encompasses both identity and utility information.

For the Linkability Risk, we use the pre-trained identity classifier trained and tested on the real dataset as the baseline model (row 1 in Table 4.5).

As indicated in Table 4.5, the anonymized datasets exhibit a high Inner Risk and a low Outer Risk. This is expected: Inner Risk reflects the linkability relationships within the anonymized images themselves, while Outer Risk reflects the similarity between anonymized images and their real counterparts. A high Inner Risk coupled with a low Outer Risk suggests that while the anonymized images retain relative linkability

**Tab. 4.4.:** Results based on privacy metrics. 'DCR': distance to closest real image', 'NNDR': nearest neighbor distance ratio', 'LC': local cloaking', 'HR': hidden rate. ' $N_R$ ': median number of real images between a real image and its anonymization, ' $N_A$ ' median number of anonymized images between a real image and its anonymization, 'Dist': the distance between each real image and its anonymization,  $S_{id}$ : the identity semantic feature space of the identity classifier,  $W$ : the latent space of StyleGAN2.

Space	Methods	DCR↑	NNDR↑	HR↑	LC↑		Dist↑
					$N_R$	$N_A$	
$S_{id}$	Co-training	17.765	0.984	0.936	41	58	20.945
	$\mathcal{L}_{S_{ut}}$	17.661	0.986	0.936	41	56	20.827
	$\mathcal{L}_{S_{id},m=0.0}$	17.853	0.985	1.000	147	809	22.978
	$\mathcal{L}_{S_{id},m=-0.5}$	17.664	0.985	1.000	1416	1635	27.136
	$\mathcal{L}_{S_{id},m=-0.7}$	17.186	0.984	1.000	1546	1639	28.039
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=0.0}$	17.803	0.986	1.000	130	764	22.879
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.5}$	17.634	0.985	1.000	1399	1635	27.111
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.7}$	17.120	0.985	1.000	1529	1639	27.952
$W$	Co-training	0.000	0.000	0.061	0	0	0.000
	$\mathcal{L}_{S_{ut}}$	0.679	0.144	0.122	0	0	0.679
	$\mathcal{L}_{S_{id},m=0.0}$	0.601	0.129	0.183	0	0	0.601
	$\mathcal{L}_{S_{id},m=-0.5}$	2.476	0.476	0.183	0	0	2.476
	$\mathcal{L}_{S_{id},m=-0.7}$	4.008	0.673	0.670	0	0	4.008
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=0.0}$	0.711	0.151	0.183	0	0	0.711
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.5}$	2.654	0.504	0.183	0	0	2.654
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.7}$	4.298	0.698	1.463	0	0	4.298

relationships (e.g., images from the same patient remain associated with each other), the linkability relationship between anonymized images and their real counterparts is effectively removed (e.g., anonymized images do not resemble their real counterparts in terms of identity).

In the MIA attack, attackers achieve high accuracy (Acc: 0.968) in both the  $X_R \rightarrow X_A$  and  $X_A \rightarrow X_R$  scenarios during training, demonstrating their effectiveness on the test dataset. However, during the evaluation stage, their performance significantly drops, with an accuracy of 0. This indicates a considerable challenge in correctly predicting the membership status of anonymized images in the  $X_R \rightarrow X_A$  attack and real images in the  $X_A \rightarrow X_R$  attack.

### 4.5.3 Ablation Study

**Effect of  $\mathcal{L}_{S_{ut}}$ .** As shown in Table 4.3, the Utility Evaluation results using the utility-preserving loss alone (row 3, 11) are better than those for the reconstructed images produced by the AE-GAN model (row 2, 10). Additionally, the Utility Evaluation result using both the utility-preserving loss and the identity removal loss (row 7-9, 15-17) are better than those using the identity removal loss alone (row 4-6, 12-14). This highlights

**Tab. 4.5.:** Results of identity network for pre-training and linkability risk. 'Pre': pre-training, 'F1': F1 score, 'P':precision, 'R':recall, 'Acc': accuracy, 'TAR': true accept rate, 'FAR': false accept rate.

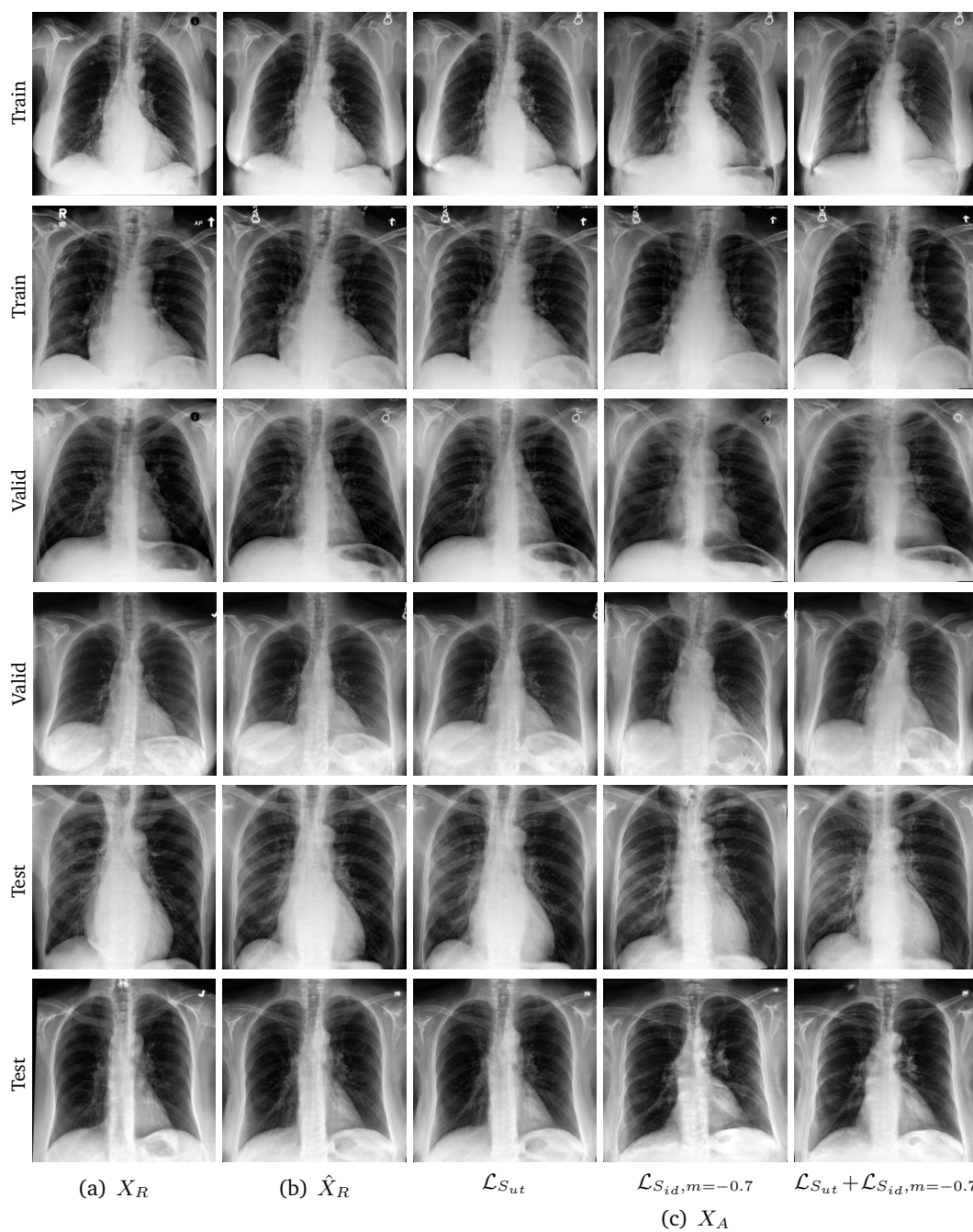
	Methods	F1↑	P↑	R↑	Acc↑	TAR↑	FAR↓
Pre-training	/	0.983	1.000	0.966	0.983	0.966	0.034
Inner risk	Co-training	0.611	0.995	0.441	0.720	0.441	0.559
	$\mathcal{L}_{S_{ut}}$	0.665	0.990	0.500	0.748	0.500	0.500
	$\mathcal{L}_{S_{id},m=0.0}$	0.617	0.964	0.454	0.718	0.454	0.546
	$\mathcal{L}_{S_{id},m=-0.5}$	0.709	1.000	0.549	0.774	0.549	0.451
	$\mathcal{L}_{S_{id},m=-0.7}$	0.855	1.000	0.746	0.873	0.746	0.254
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=0.0}$	0.609	0.984	0.441	0.717	0.441	0.559
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.5}$	0.722	0.996	0.566	0.782	0.566	0.434
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.7}$	0.887	1.000	0.798	0.899	0.798	0.202
Outer risk	Co-training	0.139	1.000	0.075	0.075	0.075	0.925
	$\mathcal{L}_{S_{ut}}$	0.136	1.000	0.073	0.073	0.073	0.927
	$\mathcal{L}_{S_{id},m=0.0}$	0.000	0.000	0.000	0.000	0.000	1.000
	$\mathcal{L}_{S_{id},m=-0.5}$	0.000	0.000	0.000	0.000	0.000	1.000
	$\mathcal{L}_{S_{id},m=-0.7}$	0.000	0.000	0.000	0.000	0.000	1.000
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=0.0}$	0.000	0.000	0.000	0.000	0.000	1.000
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.5}$	0.000	0.000	0.000	0.000	0.000	1.000
	$\mathcal{L}_{S_{ut}} + \mathcal{L}_{S_{id},m=-0.7}$	0.000	0.000	0.000	0.000	0.000	1.000

the effectiveness of the utility-preserving loss  $\mathcal{L}_{S_{ut}}$  in capturing utility-related information during the latent code optimization process.

The effectiveness of the utility-preserving loss  $\mathcal{L}_{S_{ut}}$  is further validated by the visualization results displayed in Fig. 4.5. In this figure, the anonymized images optimized using only the utility preservation loss  $\mathcal{L}_{S_{ut}}$  (column 3) exhibit greater visual similarity to their real counterparts (column 1) compared to those optimized using only the identity removal loss  $\mathcal{L}_{S_{id}}$  (column 4), as well as those optimized with both the utility preservation loss  $\mathcal{L}_{S_{ut}}$  and the identity removal loss  $\mathcal{L}_{S_{id}}$  (column 5). Additionally, the anonymized images optimized with both the  $\mathcal{L}_{S_{ut}}$  and the  $\mathcal{L}_{S_{id}}$  (column 5) appear more realistic than those optimized with the  $\mathcal{L}_{S_{id}}$  alone (column 4).

**Effect of  $\mathcal{L}_{S_{id}}$ .** As shown in Table 4.4, the anonymized images produced using the identity removal loss alone (row 3-5, 11-13) perform better than the reconstructed images produced by the AE-GAN model (row 1, 9). Additionally, the results of using both the identity removal loss and the utility-preserving loss (row 6-8, 14-16) are better than using the utility-preserving loss alone (row2, 10), particularly in the 'LC' and 'Dist' metrics. This indicates that the identity removal loss effectively increases the distance between the identity semantic features of the real and anonymized images.

Similar results are also shown in Table 4.5, where the incorporation of the identity removal loss significantly increases the Inner Risk while reducing the Outer Risk. This



**Fig. 4.5.:** Anonymization results. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The corresponding reconstructed images  $\hat{X}_R$  are displayed in the second column. The anonymized images  $X_A$ , optimized with only the utility-preserving loss  $\mathcal{L}_{S_{ut}}$ , optimized with only the identity removal loss  $\mathcal{L}_{S_{id}}$ , and those optimized with both the utility-preserving loss  $\mathcal{L}_{S_{ut}}$  and the identity removal loss  $\mathcal{L}_{S_{id}}$ , are displayed in the last three columns, respectively.

suggests that the loss  $\mathcal{L}_{S_{id}}$  can push the anonymized images belonging to the same patient closer together while severing the linkability to their real counterparts.

The effectiveness of the identity removal loss  $\mathcal{L}_{S_{id}}$  is further validated by the visualization results displayed in Fig. 4.5. In this figure, the anonymized images optimized with  $\mathcal{L}_{S_{id}}$  alone (column 4) are quite different from their real counterparts (column 1). Additionally, the anonymized images optimized with both the  $\mathcal{L}_{S_{ut}}$  and the  $\mathcal{L}_{S_{id}}$  (column 5) appear more different from their real counterparts compared to those optimized with  $\mathcal{L}_{S_{ut}}$  alone (column 3). The corresponding difference maps, calculated as  $|X_R - \hat{X}_R|$  and  $|X_R - X_A|$ , are presented in Fig. C.1, illustrating the variations between the real, reconstructed, and anonymized images.

**Effect of the Margin  $m$  in  $\mathcal{L}_{S_{id}}$ .** As shown in Table 4.4, there is a clear increase in the metric of 'LC' and 'Dist' in the  $S_{id}$  feature space as the margin  $m$  decreases. This suggests that decreasing the margin  $m$  corresponds to enlarging the distance between the identity semantic features of the real and anonymized images.

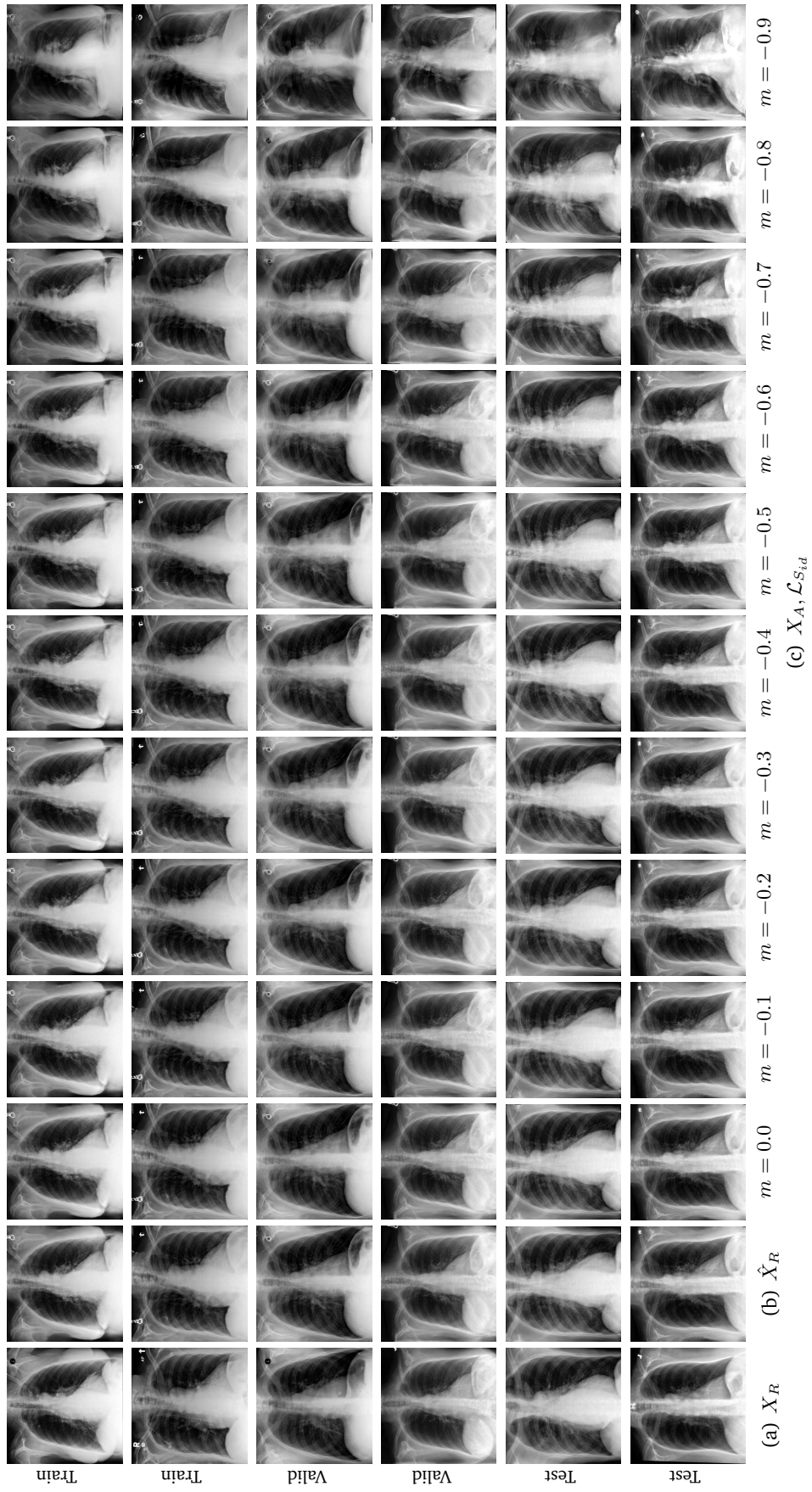
Similar results are observed in Table 4.5, where anonymized images belonging to the same patient are pushed closer together, indicated by an increasing Inner Risk and a consistently low Outer Risk as the margin  $m$  decreases.

The role of the margin  $m$  is further validated by the visualization results displayed in Fig. 4.6 and Fig. 4.7, where anonymized images gradually become more visually distinct from their real counterparts as the margin  $m$  decreases. The corresponding difference maps, calculated as  $|X_R - \hat{X}_R|$  and  $|X_R - X_A|$ , are presented in Fig. C.2 and Fig. C.3, illustrating the variations between the real, reconstructed, and anonymized images.

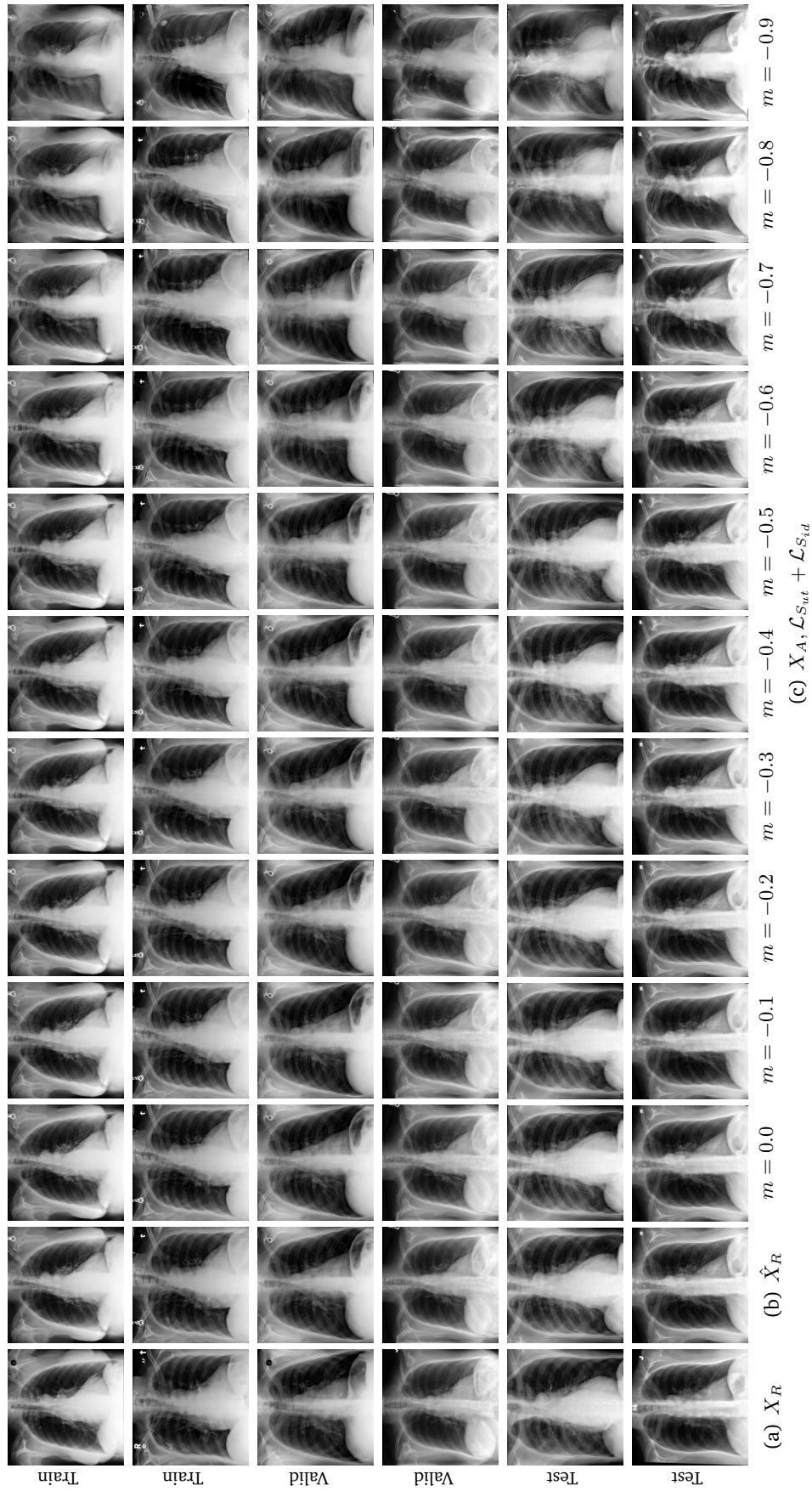
**Effect of the Loss Weight** As shown in Fig. 4.8, the weights of the utility preservation loss  $\mathcal{L}_{S_{ut}}(X_R, X_A)$  and the identity removal loss  $\mathcal{L}_{S_{id}}(X_R, X_A)$  play an important role in controlling the trade-off between identity removal and utility preservation. As the  $\lambda$  value increases from 0.1 to 0.9, the anonymized images become increasingly similar to their real counterparts due to the greater emphasis on utility preservation. Conversely, as the  $\lambda$  value decreases from 0.9 to 0.1, the anonymized images become more dissimilar from their real counterparts with increased distortion, reflecting a higher emphasis on identity removal. The corresponding difference maps, calculated as  $|X_R - \hat{X}_R|$  and  $|X_R - X_A|$ , are presented in Fig. C.4, illustrating the variations between the real, reconstructed, and anonymized images.

Interestingly, as shown in Fig. 4.7, the margin  $m$  in the identity removal loss  $\mathcal{L}_{S_{id}}(X_R, X_A)$  also influences the trade-off between identity removal and utility preservation. As the margin  $m$  decreases from 0.0 to -0.9, the anonymized images increasingly diverge from





**Fig. 4.6.:** Anonymization results. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The corresponding reconstructed images  $\hat{X}_R$  are displayed in the second column. The anonymized images  $X_A$ , optimized with only the identity removal loss  $\mathcal{L}_{S_{id}}$  but with varying margins  $m$ , are displayed in the last ten columns, respectively.



**Fig. 4.7.:** Anonymization results. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The corresponding reconstructed images  $\hat{X}_R$  are displayed in the second column. The anonymized images  $X_A$ , optimized with both the utility-preserving loss  $\mathcal{L}^{S_{ut}}$  and the identity removal loss  $\mathcal{L}^{S_{id}}$  but with varying margins  $m$ , are displayed in the last ten columns, respectively.

their real counterparts, with greater distortion, highlighting a stronger emphasis on identity removal.

#### 4.5.4 Disentanglement Property of Latent Space

As highlighted in Section 4.3, the disentanglement property of the latent space is crucial to the proposed approach, as it enables precise control over identity removal and utility preservation. Without this disentanglement, identity and utility attributes would become highly entangled, meaning that even minor modifications to one attribute could lead to significant and unintended changes in the other.

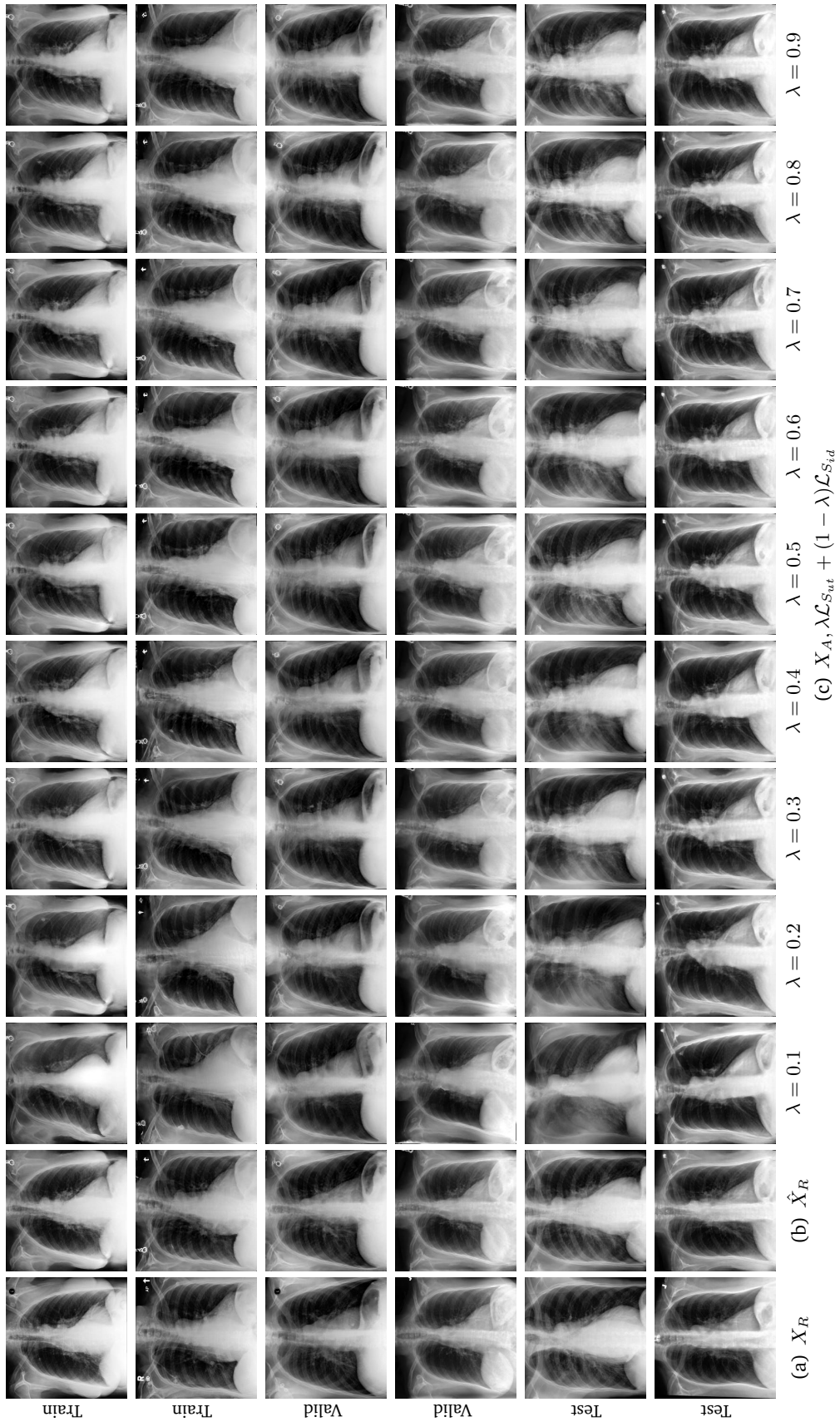
We analyze the disentanglement property of the latent codes by selectively mixing channels from two distinct sets of latent codes. Specifically, given two real images,  $X_R^A$  and  $X_R^B$ , a new image is generated by blending portions of their latent codes. The process (illustrated in Fig. 4.9,) begins by copying the first 4 channels from the latent code of real image  $X_R^B$  into the latent code of  $X_R^A$ , then continues with a stride of 4 channels, alternating between the two latent codes, until all 16 channels are covered. As demonstrated in Fig. 4.10, copying the early channels (e.g. first 8 channels) of real image  $X_R^A$  with those from real image  $X_R^B$  results in images that closely resemble real image  $X_R^B$ , indicating that these early channels carry significant visual information of  $X_R^B$ . However, as the copying process advances to the later channels (e.g. last 8 channels) of real image  $X_R^A$ , the generated images increasingly resemble real image  $X_R^A$ , reflecting the diminishing influence of the copied channels from real image  $X_R^B$ .

Unlike the latent codes  $Z$  used in Chapter 3, where mixing latent codes from two images produced overlapping and less distinct images, the latent code  $W$  exhibits strong disentanglement properties. This not only ensures that the anonymization process effectively balances utility preservation with identity elimination, as evidenced by the experiments in Section 4.3, but also allows the mixed latent code  $W$  to generate visually meaningful and realistic images.

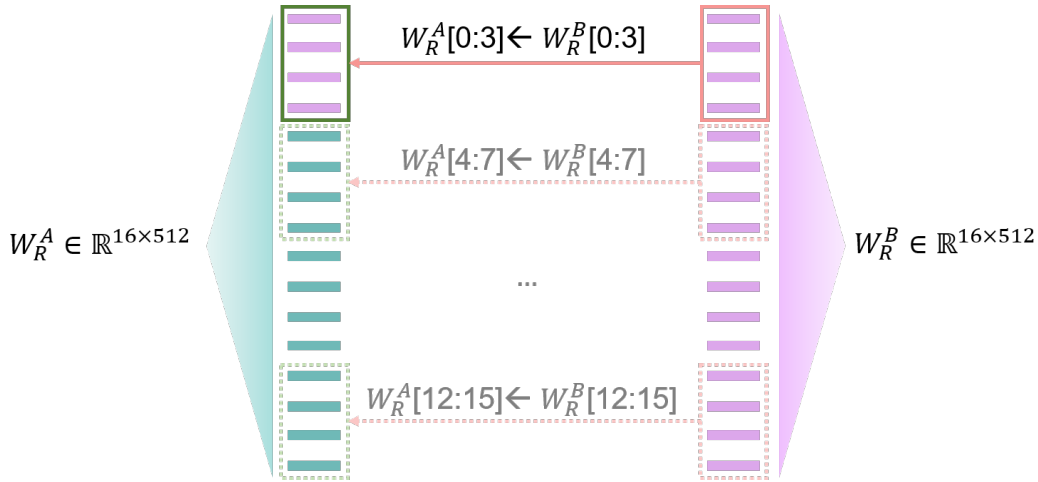
## 4.6 Discussion and Limitations

### 4.6.1 Discussion

**Disentangled Latent Space:** The disentangled nature of the  $W$  latent space offers the flexibility to effectively remove identity information while minimizing the impact on utility-related features. As discussed in [Liu et al., 2023c], there are typically three key embedding spaces in StyleGAN: the latent space  $W$ , the extended latent space  $W^+$ , and



**Fig. 4.8.:** Anonymization results. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The corresponding reconstructed images  $\hat{X}_R$  are displayed in the second column. The anonymized images  $X_A$ , optimized with both the utility-preserving loss and the identity removal loss  $\lambda \mathcal{L}_{S_{ut}} + (1 - \lambda) \mathcal{L}_{S_{id}}$  with margin  $m = -0.7$  and  $\lambda$  ranging from 0.1 to 0.9, are displayed in the last nine columns, respectively.



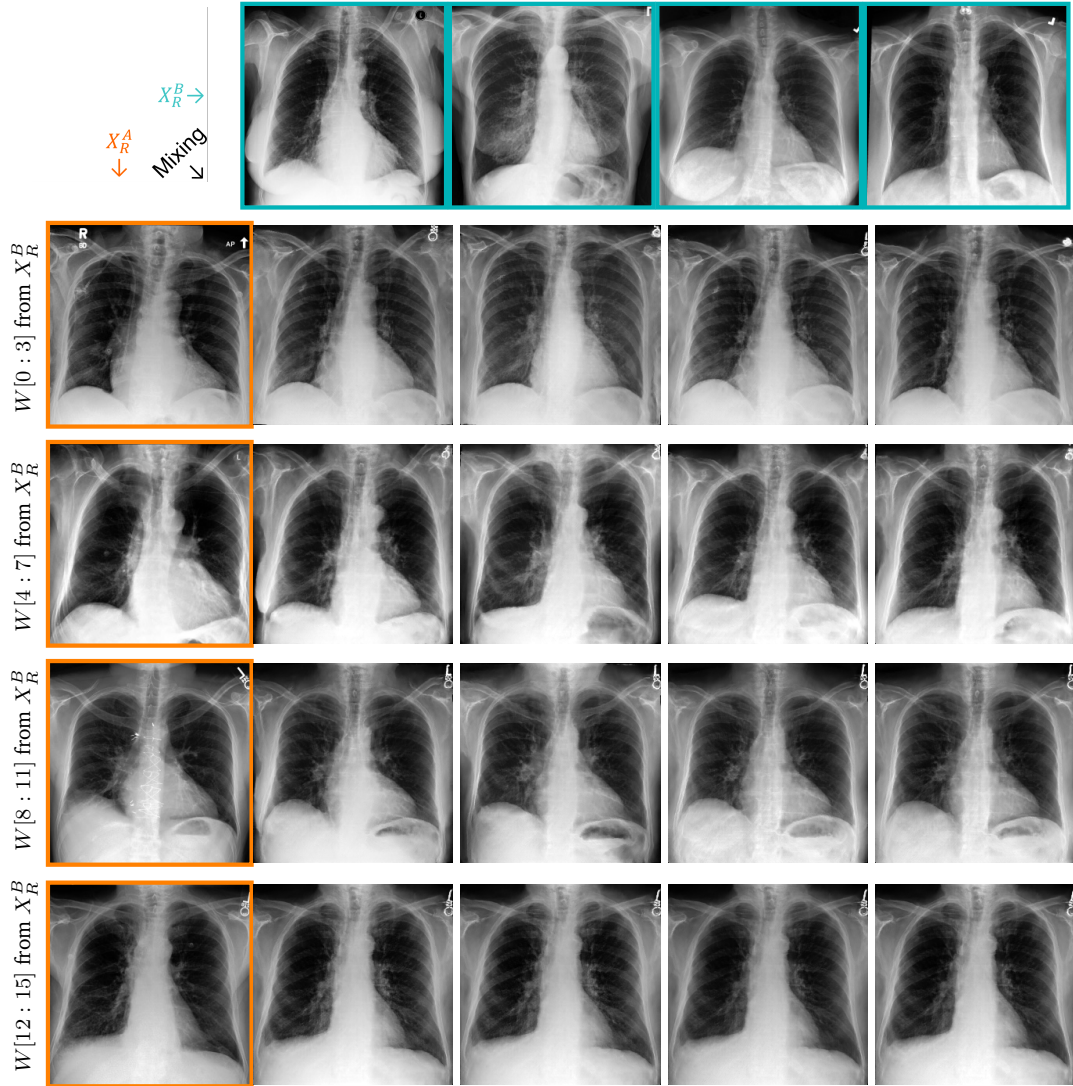
**Fig. 4.9.:** Latent code mixing. Given two real images,  $X_R^A$  and  $X_R^B$ , a new image is generated by blending their respective latent codes,  $W_R^A$  and  $W_R^B$ . The process starts by copying the first 4 channels from the latent code  $W_R^B$  into  $W_R^A$ , followed by alternating between the two latent codes with a stride of 2 channels, until all 16 channels are covered.

the feature space  $F$ , which contains spatial image representations. Moving from  $W$  to  $W^+$  to  $F$ , the editability of the latent codes decreases, while reconstruction quality improves. This highlights a clear trade-off between reconstruction fidelity and editability in GAN inversion. In our framework, we prioritize editability and reduced complexity during latent code optimization, accepting a degree of loss in reconstruction fidelity as a necessary trade-off for more effective anonymization. This choice enables more precise and controlled adjustments of identifiable features, though it comes at the cost of slight degradation in image quality.

**Enhanced Anonymization:** In this chapter, we leverage the same GMIA framework introduced in Chapter 3, but replace the latent space  $Z$  with a more disentangled latent space  $W$ . This choice is related to the reconstruction-editability trade-off, where the  $W$  latent space provides improved editability but sacrifices some reconstruction fidelity. As a result, it is unsurprising that the reconstruction quality in this chapter is less accurate compared to Chapter 3. However, the trade-off leads to improved anonymization performance, particularly in removing identity features, while still effectively preserving utility-related information.

## 4.6.2 Limitations

**Utility Preservation:** While the utility classifier used for preserving diagnostic information shows some success, it has notable limitations. First, the classifier is designed to handle only four labels, restricting its ability to retain a broader range of pathology-related information. Additionally, the dataset employed in this chapter suffers from a significant class imbalance, which we did not adequately address, resulting in suboptimal



**Fig. 4.10.:** Latent code mixing between two sets of latent codes. The images in the first column and first row represent the real images,  $X_R^A$  and  $X_R^B$ , corresponding to their respective latent codes. The remaining images are generated by selectively combining specific channels from the latent code  $W_R^B$  of  $X_R^B$  with the remaining channels from the latent code  $W_R^A$  of  $X_R^A$ .

classification performance. To improve accuracy, future work should focus on mitigating the imbalance issue through techniques like oversampling, undersampling, or using advanced loss functions that account for imbalance. Moreover, incorporating explainable AI techniques, such as Grad-CAM [Selvaraju et al., 2017], could enhance the transparency of the utility classifier's decisions, allowing for better interpretation and validation of the preserved utility.

**Dataset with Follow-up Patients:** Having access to datasets that include follow-up images for the same patients provides a key advantage for training the identity classifier. In cases where only one image per patient is available, data augmentation and multiple generations of synthetic images would be necessary to create the diversity needed for effective training. This could improve the identity classifier's robustness and the overall effectiveness of identity removal.

**Risk of Re-identification:** Despite the anonymization, there may remain a potential risk of re-identification, especially if the synthetic images are combined with other data sources that are not anonymised. Sophisticated adversaries may use advanced techniques to cross-reference anonymized images with other datasets, posing a threat to patient privacy. Ongoing research and development are required to enhance the robustness of the anonymization process.

In summary, while the proposed generative image anonymization approach offers numerous benefits in terms of privacy protection and data utility, it is not without its challenges. Addressing these limitations through continued research and collaboration with the medical community will be key to realizing the full potential of this innovative technology.

## 4.7 Conclusions

The proposed generative medical image anonymization approach presents a promising solution to the challenges of protecting patient privacy while maintaining the clinical utility of medical images. By leveraging advanced generative adversarial networks (GANs), the framework effectively removes personally identifiable information from medical images, aligning with key privacy metrics and regulatory standards. This technology enables secure data sharing and collaboration across various healthcare and research settings, fostering innovation and improving patient care.

Despite the evident benefits, the approach faces challenges, such as the need for high-quality, diverse training data and significant computational resources. Addressing these limitations through careful data curation, augmentation, and optimization of computa-

tional processes will be critical for enhancing the model's robustness and accessibility. Furthermore, expanding the utility network's capability to handle more pathology labels and adapt to highly imbalanced datasets will enhance the framework's versatility.

As the medical imaging field continues to advance, the proposed generative anonymization method offers a promising solution to the ongoing challenge of balancing patient privacy with data utility. Continued research, rigorous validation studies, and close collaboration with the medical community will be vital for refining this technology and build trust in its clinical and research applications. By addressing these key areas, generative image anonymization could become a standard practice in medical imaging, ensuring patient confidentiality while maintaining the integrity and usefulness of the data for healthcare and research. Additionally, this technique holds significant potential for generating anonymized images to enhance machine learning model training, where preserving both privacy and utility is critical.



# Conclusion

In this thesis, we first expose privacy vulnerabilities in medical data lakes by introducing a novel data exfiltration by compression (DEC) attack. To address these risks, we then propose a generative medical image anonymization framework (GMIA) designed to safeguard data privacy from the outset of the dataset. The remainder of this section summarizes each contribution and explores potential future research directions.

## 5.1 Main Contributions

### **A Novel Data Exfiltration by Compression Attack (DEC Attack)**

In Chapter 2, we propose a novel attack based on data compression, termed the DEC attack, which is agnostic to both the architecture and the task of the exported network. By presenting two distinct scenarios where the DEC attack is applicable, the research highlights the versatility of this approach, carefully balancing their inherent pros and cons. The investigation delves into various hypotheses impacting the attack's design, including the type of compression (lossy vs. lossless), the use of pretrained models, methods for integrating compression codes, the addition of noise, and the application of common mitigation strategies. Furthermore, the thesis demonstrates that advanced learned image compression methods, such as those proposed by [Mentzer et al., 2020], are particularly effective for generating high-quality medical images with small compression codes and exhibit resilience to Gaussian noise within the network. Extensive experiments on two public medical image datasets (LiTS and BraTS) empirically validate the DEC attack in real-world scenarios, marking the first known evaluation of an exfiltration attack on medical image datasets. The findings reveal that the DEC attack poses a substantial threat to medical data lakes, enabling efficient extraction of a large number of medical images without significantly compromising the network's performance in image segmentation tasks.

### **A Novel Generative Medical Image Anonymization Framework (GMIA)**

In Chapter 3, we propose a novel medical image anonymization framework, termed GMIA, which addresses the critical balance between identity removal and utility preservation in the anonymization process. The GMIA framework first extracts a precise latent representation of the original image using a learning-based image reconstruction network. It then separately trains a utility network and an identity network to accurately identify

the utility and identity features of the image. Finally, through latent code optimization, the framework effectively removes identity information using an identity removal loss, while preserving utility using a utility preservation loss. Extensive experiments on the MIMIC-CXR-JPG dataset validate the effectiveness of the GMIA framework, emphasizing the critical role of the latent code's disentanglement properties. These findings highlight the need to explore a more disentangled latent space to enhance the efficacy of latent code optimization-based anonymization method.

### **Designing an Encoder for Disentangled Latent Space**

In Chapter 4, we design a simple yet effective encoder that leverages the disentangled latent space of StyleGAN2 [Karras et al., 2020], significantly improving the efficiency of the latent code optimization process. Unlike the latent space discussed in Chapter 3, the latent space in Chapter 4 exhibits superior disentanglement properties, allowing for more distinct separation of utility and identity features. As a result, during the optimization process, the utility-preserving and identity-removing loss functions have a reduced impact on each other, leading to more precise control over the balance between maintaining utility and removing identity information. This contribution enhances the overall effectiveness of the proposed anonymization framework by improving the disentanglement of the latent space. Extensive experiments on the MIMIC-CXR-JPG dataset demonstrate that this approach not only aligns with key privacy metrics but also adheres to regulatory standards, ensuring robust and compliant medical image anonymization.

## **5.2 Future research**

In this thesis, we have presented several contributions to the field of data privacy, through both data exfiltration by compression (DEC) attack and generative medical image anonymization framework (GMIA), which nonetheless call for further inquiries.

### **Improving DEC Attack with Reduced Size of Compression Network and Latent Codes**

Although the DEC attack presented in Chapter 2 is highly effective in extracting entire datasets from a privacy data lake, there are several areas for potential improvement. First, the network's size could be further reduced to better conceal the attack, but this must be carefully balanced against the risk of diminishing the compression network's effectiveness. Reducing the network size typically impairs its ability to carry out the compressed latent codes. Consequently, it becomes essential to also reduce the size of the latent codes proportionally. However, reducing the size of the latent codes without compromising their representational quality requires enhancing the disentanglement of these codes. Improved disentanglement ensures that each channel of the latent

code operates more independently, minimizing mutual interference and preserving the critical information needed for accurate image reconstruction, even with smaller latent code sizes. Finally, it is recommended to conduct additional experiments to verify the effectiveness of mitigation strategies based on model fine-tuning. This approach has the potential to enhance privacy protection against existing data exfiltration attacks while maintaining the utility of the model for its target tasks.

### **Improving and Generalizing GMIA with Robust Key Components**

The GMIA framework has demonstrated that our generative anonymization approach is effective. However, there are several ways for potential improvement. First, the reconstruction quality needs further improvement. The use of the disentangled latent space of StyleGAN2 [Karras et al., 2020], while beneficial for creating independent and non-redundant information channels, comes at the cost of reduced reconstruction ability. This trade-off arises because disentanglement inherently reduces the redundancy that often aids in more accurate image reconstruction.

Second, both the utility and the identity networks should prioritize explainability and adaptability across a broader range of downstream tasks. To enhance the utility network's explainability, it can be generalized to tackle more explainable tasks, such as landmark detection or mask segmentation. Integrating explainability techniques will also help localize and highlight the important regions of an image that drive the decision-making process of the network. Furthermore, extending the GMIA framework to different datasets and tasks by training corresponding utility and identity networks is crucial for further validating its effectiveness and versatility across diverse applications.

Third, the identity network must be capable of handling datasets where only a single image is available per patient. To train the identity network without compromising its generalizability, various strategies should be employed, such as utilizing multiple generative models and applying data augmentation techniques to create diverse images for each patient. This approach ensures robust identity representation even in limited-data scenarios.

### **Building a Large-scale Dataset for Medical Image Anonymization Research**

While the MIMIC-CXR-JPG dataset [Johnson et al., 2019] provides a strong foundation for studying medical image anonymization, there is a need for a more specialized benchmark dataset to advance research in this domain. Such a dataset should be able to evaluate critical privacy risks, including linkability <sup>1</sup> and inference risk <sup>2</sup> highlighted in [Party, 2014].

---

<sup>1</sup>The ability to link at least two records concerning the same data subject or a group of data subjects in **two different databases**

<sup>2</sup>Inference risk refers to the possibility to deduce, with significant probability, the value of an attribute from the values of a set of other attributes.

To address linkability risk, the dataset should include an auxiliary dataset that shares non-sensitive attributes with the target private dataset. This auxiliary set would enable researchers to assess how easily anonymized data can be linked back to individuals based on these shared characteristics.

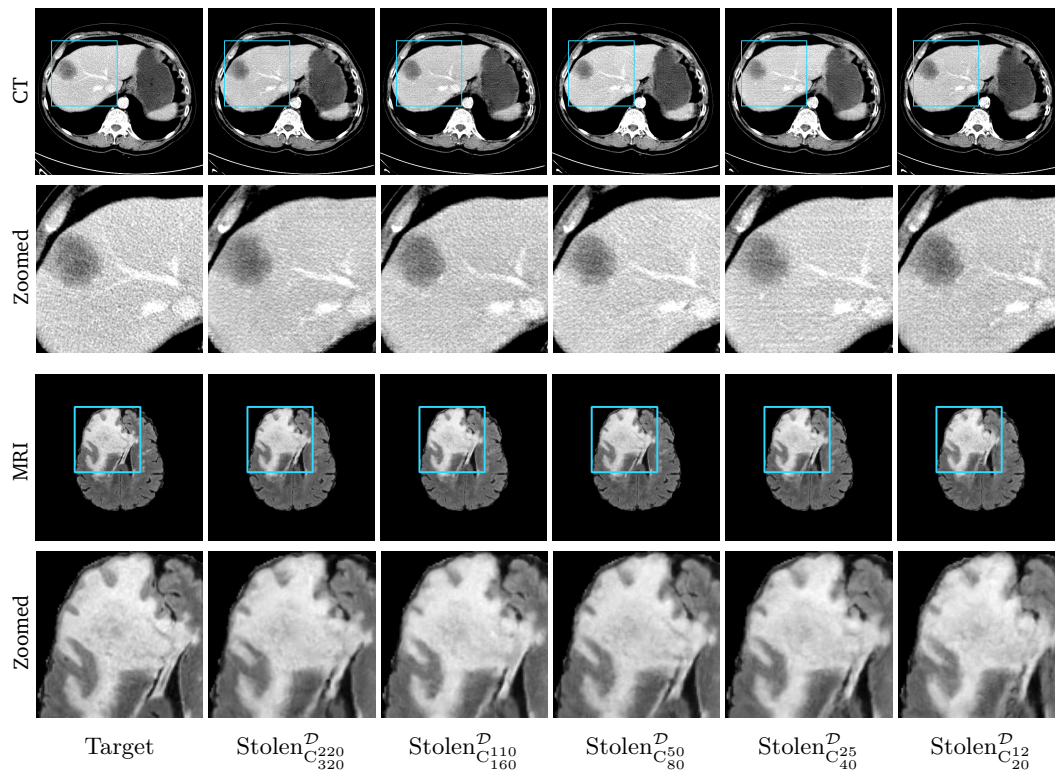
For evaluating inference risk, the dataset should contain both sensitive attributes, which serve as inference targets, and non-sensitive attributes, which could be used for inferential analysis. This setup would allow for a more rigorous assessment of how effectively the anonymization process prevents private information from being deduced based on publicly available data.

Overall, developing such a dataset would create a more comprehensive platform for evaluating and improving medical image anonymization techniques, addressing both linkability and inference vulnerabilities.

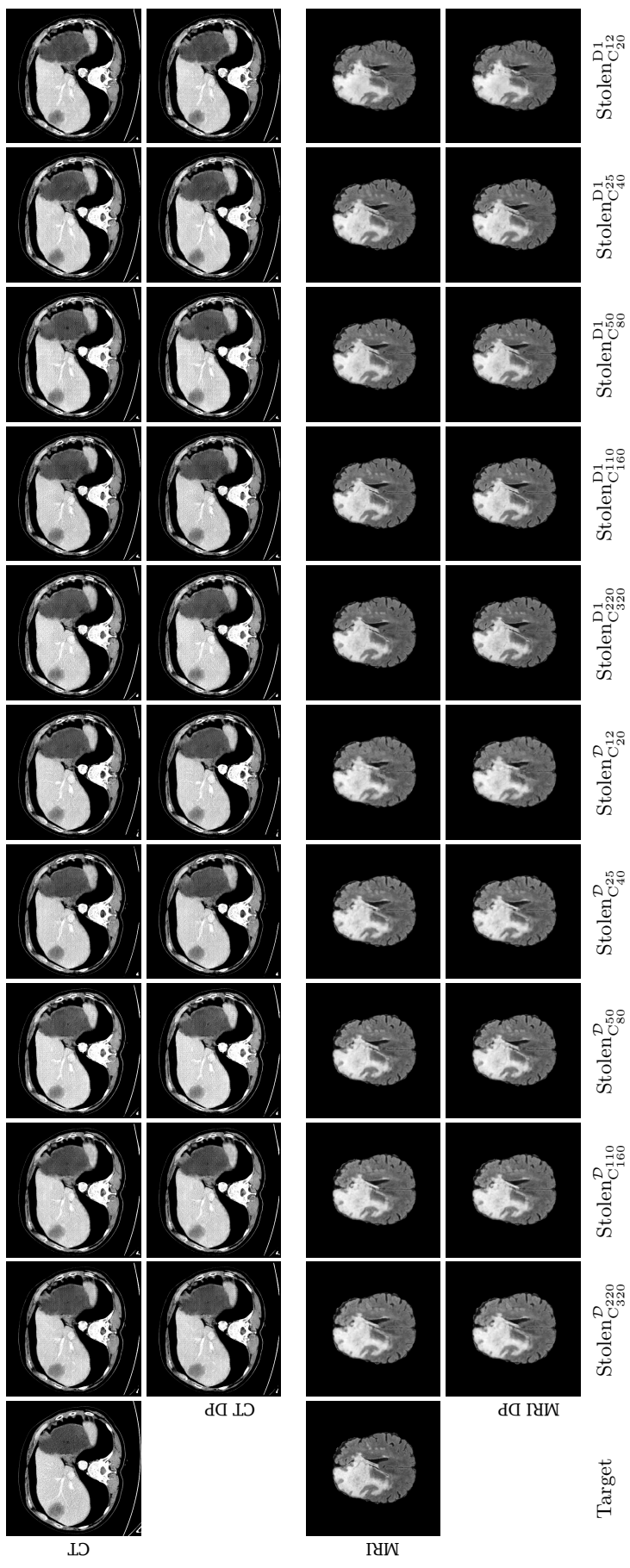
## Appendix of Chapter2

As shown in Fig. A.1, the reconstructed images generated using steganography code exporting strategy exhibit no significant differences compared to their original counterparts. By allocating the least significant 16 bits (half of the 32-bit float32) of each network parameter to store the compression codes, the steganography technique modifies the model's parameter values, potentially impacting accuracy. However, the results in Fig. A.1 confirm that the model's reconstruction ability remains largely unaffected by this approach. These findings validate the decision to use no more than the least significant 16 bits for compression code allocation without compromising the network's performance.

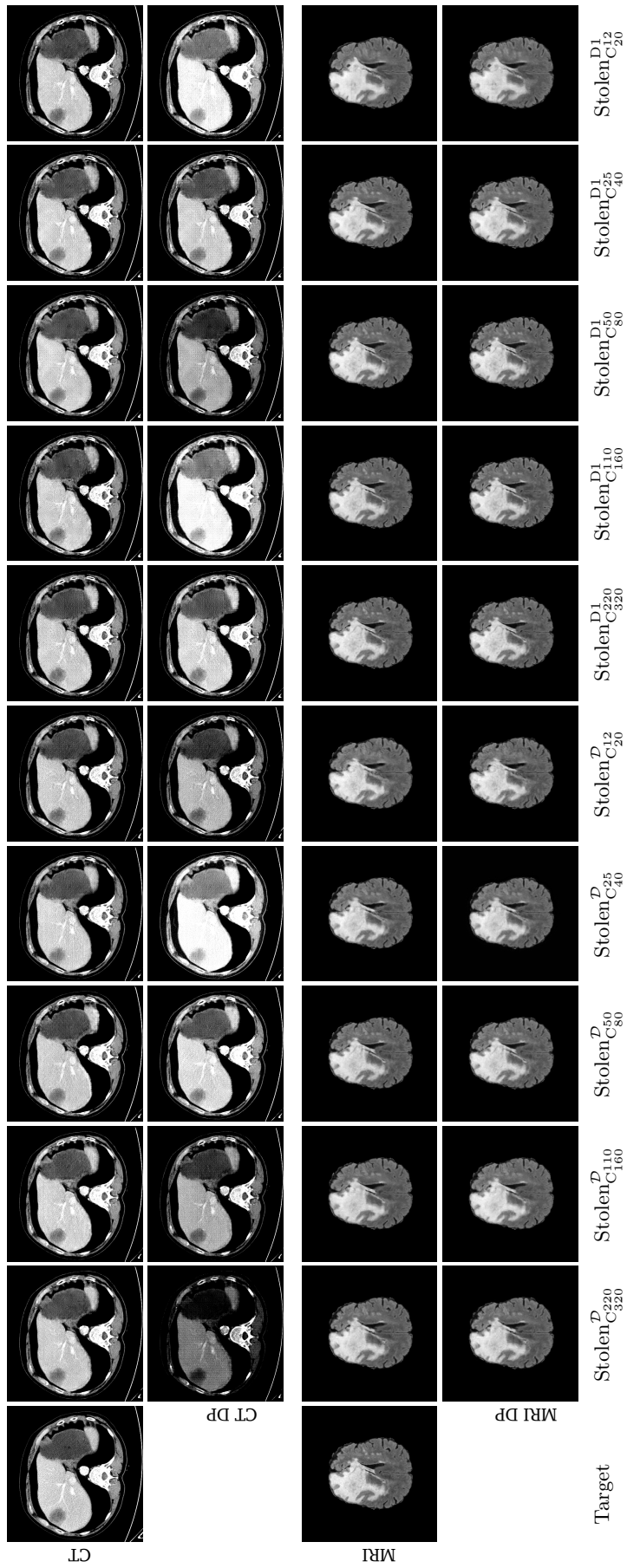
Figures A.2 and A.3 show the image reconstructions before and after applying Differential Privacy (DP) in the EP and IT scenarios, respectively. These figures complement Fig. 2.13 by providing a more detailed visual comparison of the exfiltrated images before and after DP is applied. In the EP scenario (Fig. A.2), where noise affects only the decoded latent codes, which follow a Gaussian distribution with a standard deviation, the reconstructed images after DP application show no discernible artifacts. This demonstrates the attack's resilience to DP. In the IT scenario (Fig. A.3), where noise perturbs both the compression codes and the decoder, the stolen images still retain acceptable quality, though slight variations in image intensity are observable in the CT images generated by decoder D1. In contrast, MR images show no noticeable differences, indicating that the attack decoder is more sensitive to noise in CT images than in MR images.



**Fig. A.1.:** Lossy image reconstructions on CT (row 1, 2) and MRI (row 3, 4) images after applying steganography, where the row 2, 4 provide a zoomed-in view of the bounding box region of the row 1, 3. The leftmost column represents the target images, while the subsequent five columns show the stolen images reconstructed by the decoder D through the utilization of steganography for code exporting.



**Fig. A.2.:** Lossy image reconstructions before and after applying Differential Privacy in EP Scenario on CT and MRI images.

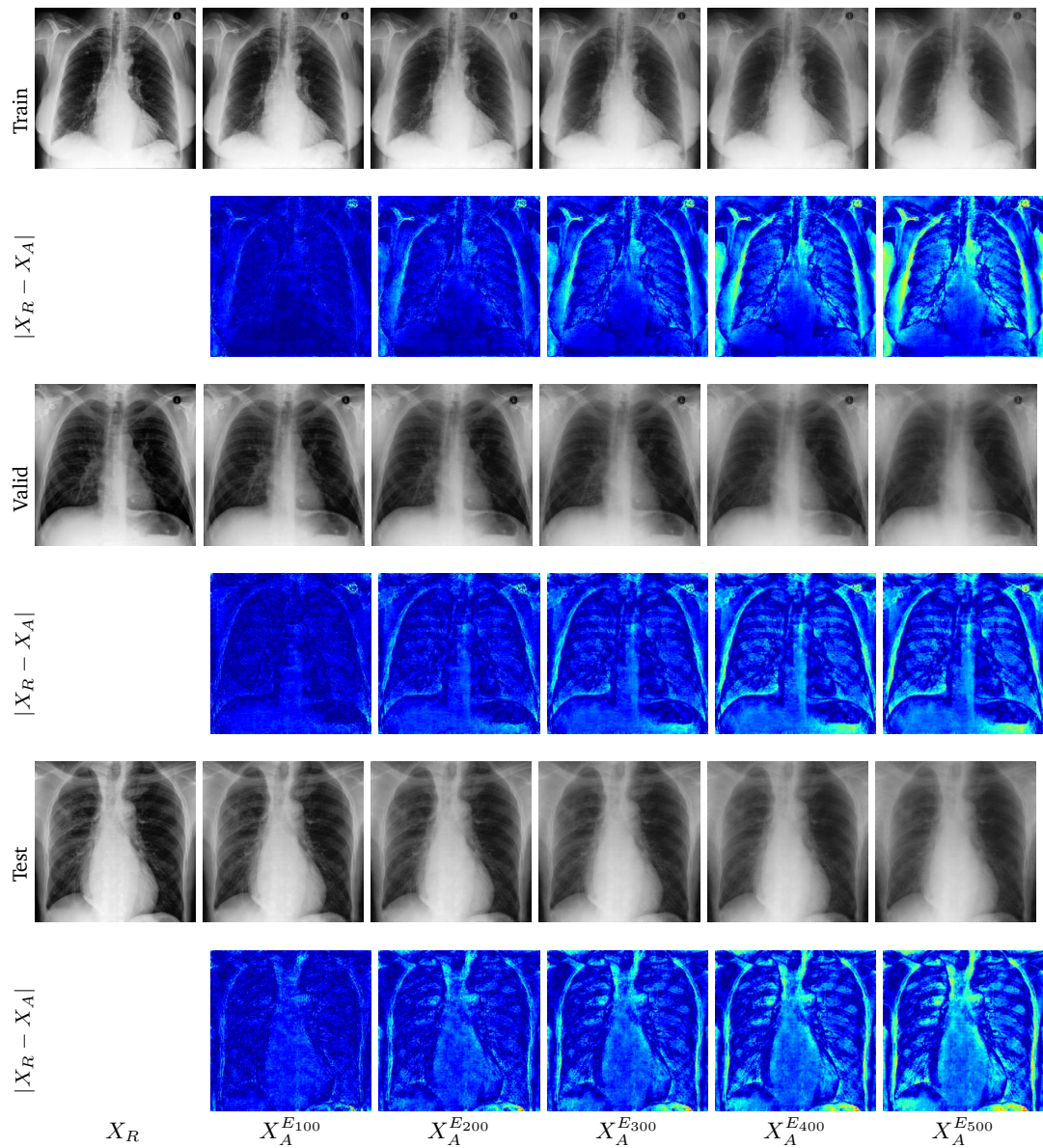


**Fig. A.3.:** Lossy image reconstructions before and after applying Differential Privacy in IT Scenario on CT and MRI images.



## Appendix of Chapter 3

Fig. B.1 presents the real images  $X_R$ , anonymized images  $X_A$ , and the difference maps  $|X_R - X_A|$  across different optimization epochs. As the optimization progresses, a noticeable difference emerges in the outlining regions of the lungs in the anonymized images—areas that contain the principal biometric information in chest X-rays. This suggests that the proposed anonymization framework effectively identifies and targets the lung outlines as key identity-related features within the chest X-ray images.



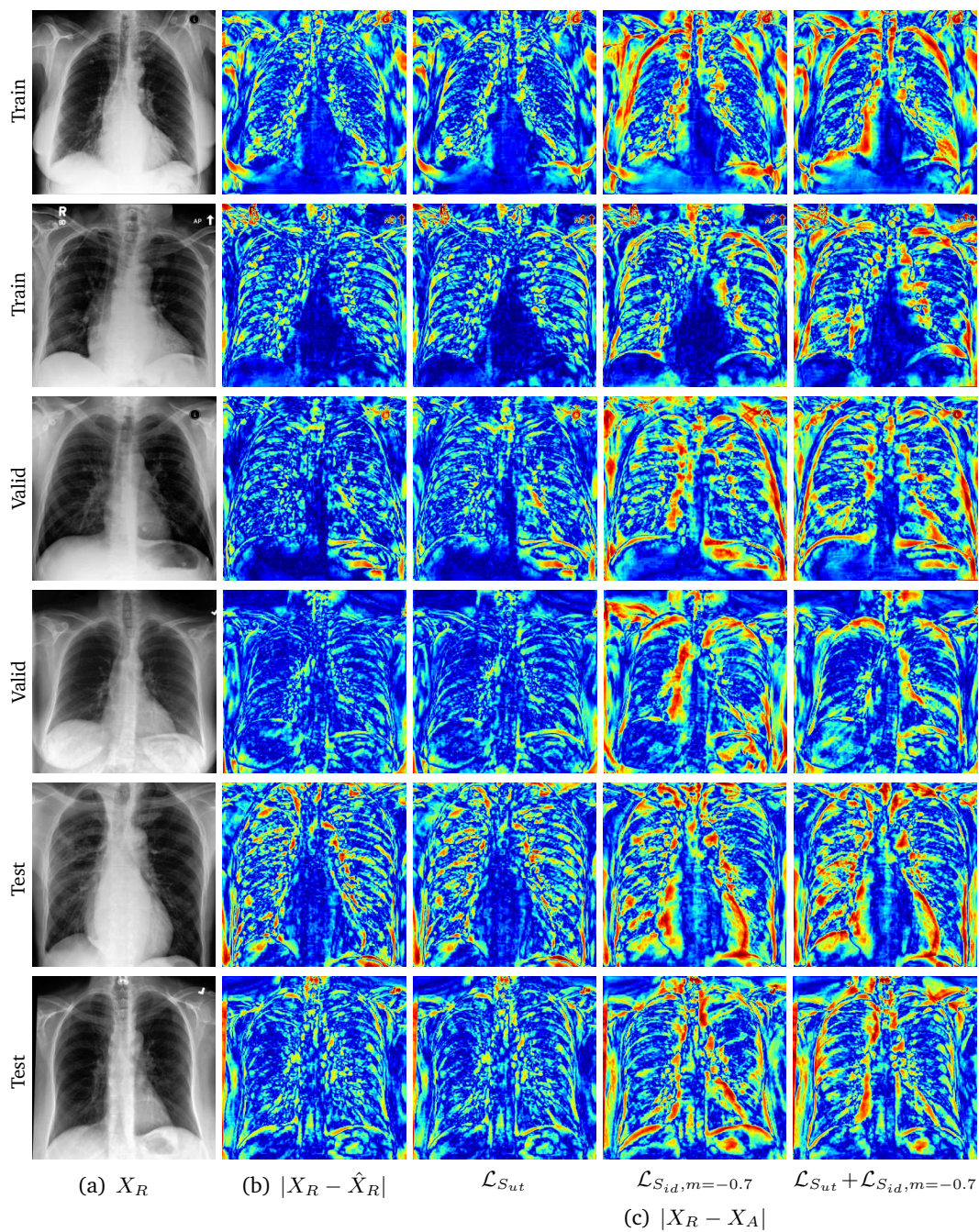
**Fig. B.1.:** Anonymization results. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The anonymized images  $X_A$  at different optimization epochs are shown in the last five columns. For instance,  $X_A^{E100}$  denotes the anonymized images  $X_A$  at 100th optimization epoch. Difference maps  $|X_R - X_A|$  denotes that the anonymized images primarily focus on outlining the lungs, which contain the principal biometric information in chest X-rays.

## Appendix of Chapter 4

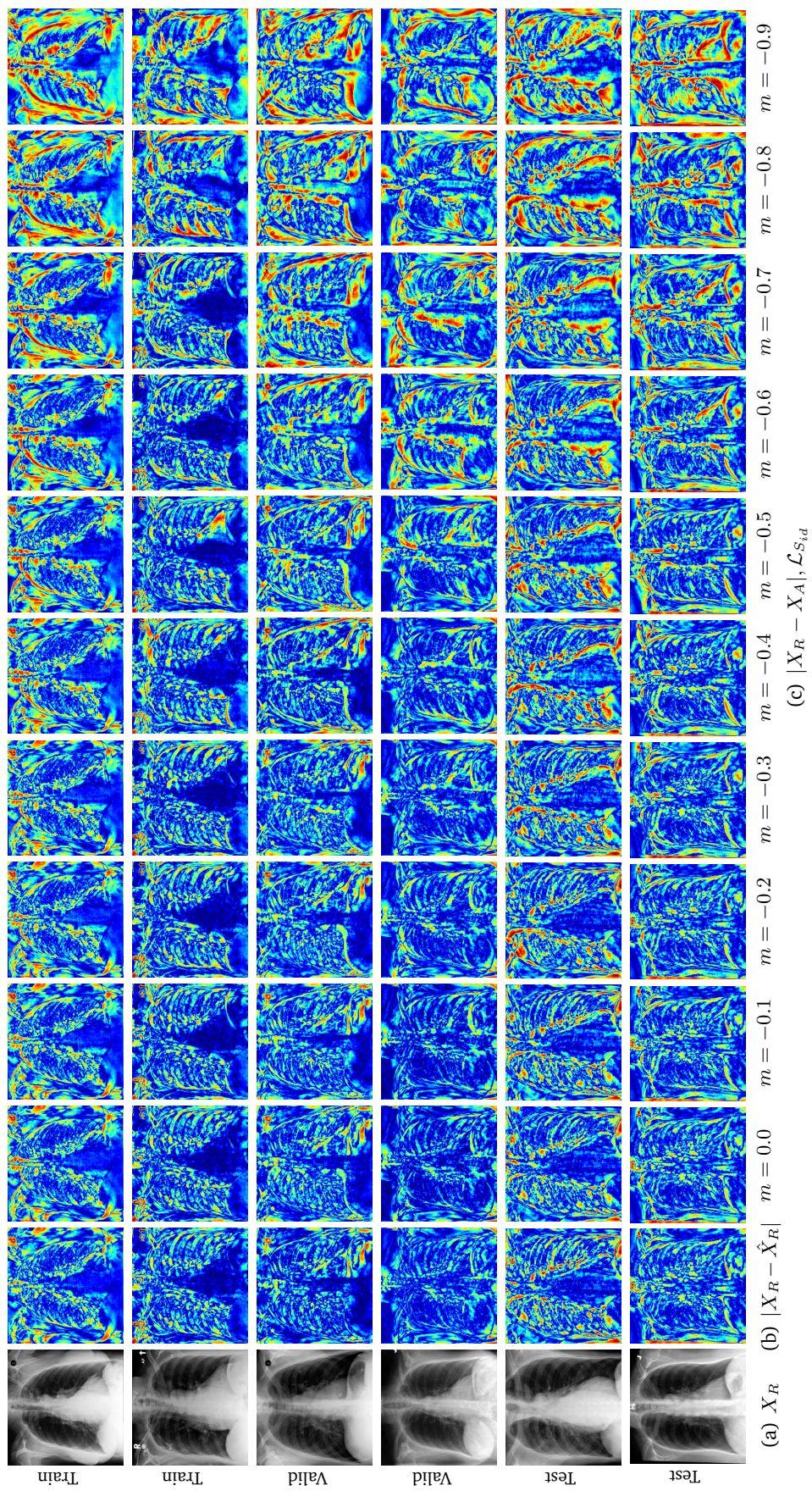
Fig. C.1 complements Fig. 4.5 by visualizing the difference maps between the real image  $X_R$  and the reconstructed image  $\hat{X}_R$ , calculated as  $|X_R - \hat{X}_R|$ , and between the real image  $X_R$  and the anonymized image  $X_A$ , calculated as  $|X_R - X_A|$ . In Fig. C.1, there are no noticeable intensity changes in the difference maps for  $|X_R - \hat{X}_R|$  and  $|X_R - X_A|$  when only the utility-preserving loss function is optimized. However, when the identity-removing loss function is solely optimized, the difference map for  $|X_R - X_A|$  shows a significant brightening around the lung outlines, which contain the primary biometric information in chest X-rays. When both the utility-preserving and identity-removing loss functions are applied, the difference map reveals a marked difference along the lung outlines while maintaining minimal changes within the lung region, highlighting an effective trade-off between identity removal and utility preservation.

Fig. C.2 and C.3 complement Fig. 4.6 and 4.7, respectively, by further visualizing the difference maps between the real image  $X_R$  and the reconstructed image  $\hat{X}_R$ , as well as between the real image  $X_R$  and the anonymized image  $X_A$ . As shown in Fig. C.2, as the margin  $m$  decreases, the difference maps exhibit more pronounced intensity variations along the lung outlines. This reinforces the role of the margin  $m$  in enhancing the distinction between the anonymized image and its original counterpart from an identity perspective. The same observation is evident in Fig. C.3, where the intensity variations along the lung outlines become more noticeable as the margin  $m$  decreases. Comparing Fig. C.2 and C.3, we observe that the difference maps maintain intensity variations around the lung region while reducing them within the lung area, further illustrating the trade-off between identity removal and utility preservation.

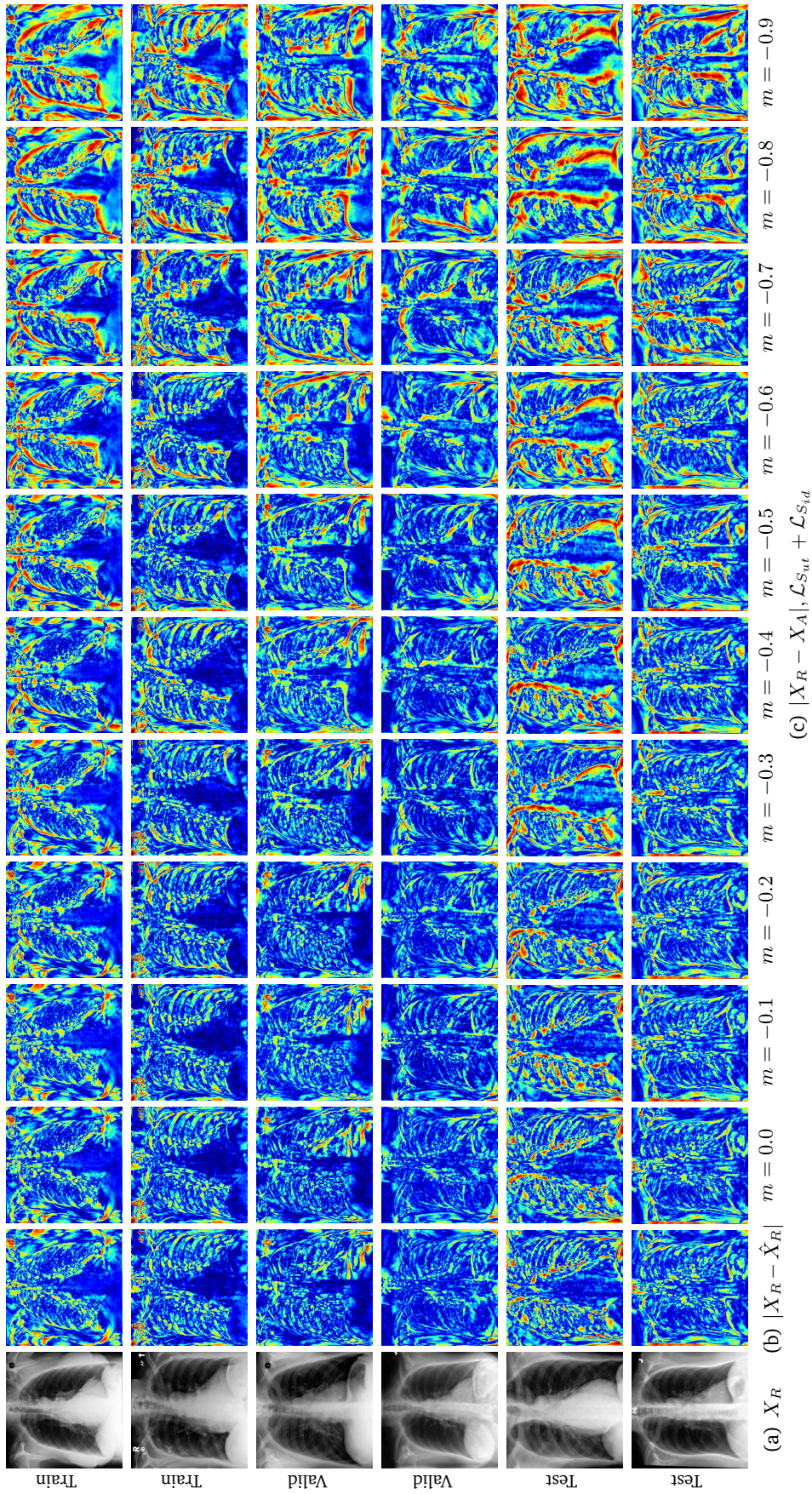
Finally, Fig. C.4 complements Fig. 4.8 by visualizing the difference maps between the real images and their corresponding reconstructions and anonymizations. As shown in Fig. C.4, as the weight of the utility-preserving loss  $\mathcal{L}_{S_{ut}}(X_R, X_A)$  increases, the intensity variations around the lung region become less prominent. Conversely, as the weight of the identity-removing loss  $\mathcal{L}_{S_{id}}(X_R, X_A)$  increases, the intensity variations around the lung region become more pronounced. This indicates that the loss weight  $\lambda$  plays a crucial role in controlling the balance between identity removal and utility preservation.



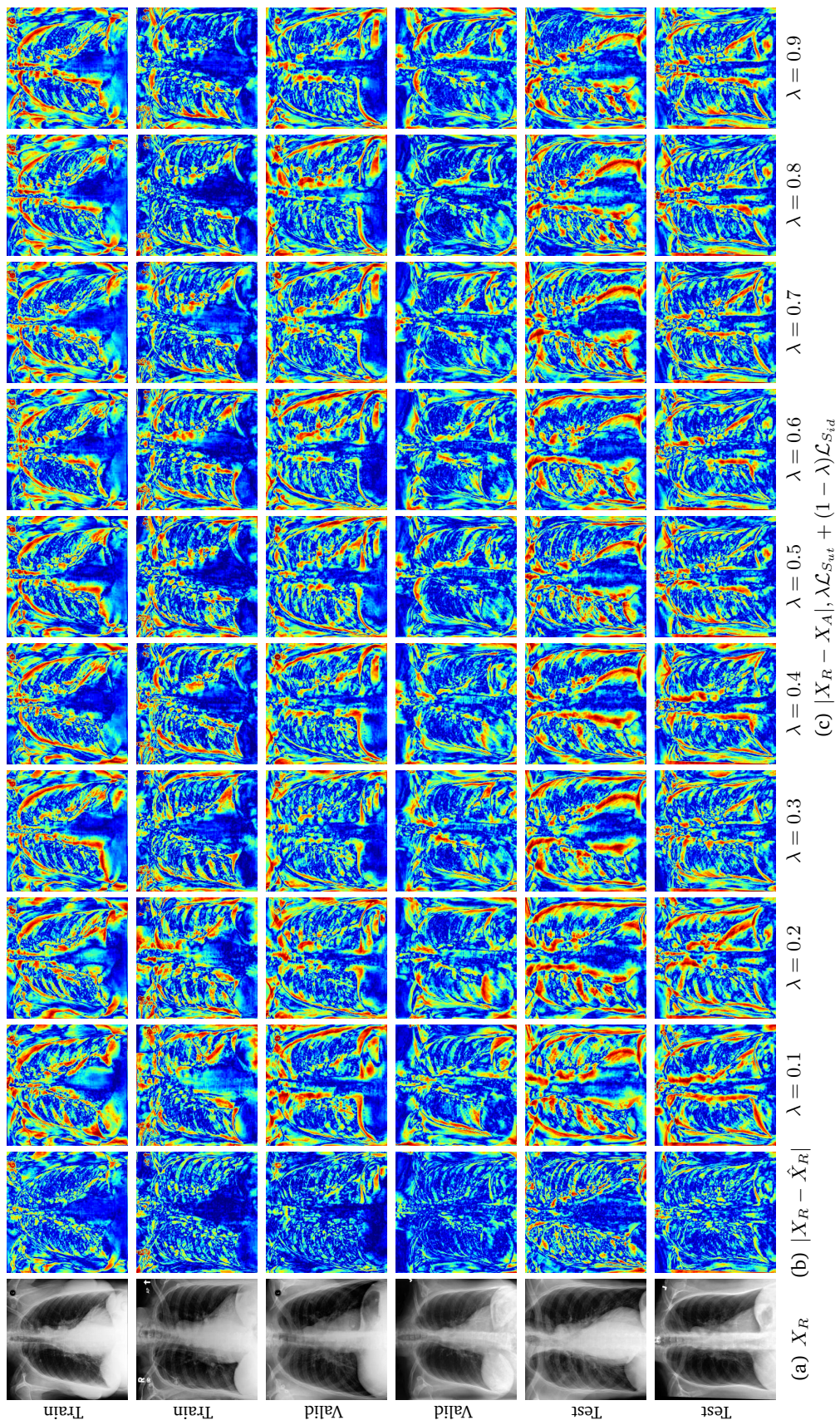
**Fig. C.1.:** Visualizations of difference maps. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The difference maps computed as  $|X_R - \hat{X}_R|$  are displayed in the second column. The difference maps computed as  $|X_R - X_A|$  are displayed in the last three columns, where the anonymized images  $X_A$  are optimized with: only the utility-preserving loss  $\mathcal{L}_{S_{ut}}$ , only the identity removal loss  $\mathcal{L}_{S_{id}}$ , and both the utility-preserving loss  $\mathcal{L}_{S_{ut}}$  and the identity removal loss  $\mathcal{L}_{S_{id}}$ , respectively.



**Fig. C.2.:** Visualizations of difference maps. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The difference maps computed as  $|X_R - \hat{X}_R|$  are displayed in the second column. The difference maps computed as  $|X_R - X_A|$  are displayed in the last ten columns, where the anonymized images  $X_A$  are optimized with only the identity removal loss  $\mathcal{L}_{S_{id}}$  using varying margins  $m$ .



**Fig. C.3.:** Anonymization results. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The difference maps computed as  $|X_R - \hat{X}_R|$  are displayed in the second column. The difference maps computed as  $|X_R - X_A|$  are displayed in the last ten columns, where the anonymized images  $X_A$  are optimized with both the utility-preserving loss  $\mathcal{L}_{S_{ut}}$  and the identity removal loss  $\mathcal{L}_{S_{id}}$  using varying margins  $m$ .



**Fig. C.4.:** Anonymization results. Real images  $X_R$  randomly selected from the training, validation, and test sets are displayed in the first column. The difference maps computed as  $|X_R - \hat{X}_R|$  are displayed in the second column. The difference maps computed as  $|X_R - X_A|$  are displayed in the last nine columns, where the anonymized images  $X_A$  are optimized with both the utility-preserving loss and the identity removal loss  $\lambda \mathcal{L}_{S_{ut}} + (1 - \lambda) \mathcal{L}_{S_{id}}$  with margin  $m = -0.7$  and  $\lambda$  ranging from 0.1 to 0.9.





# Bibliography

- [Abdal et al., 2019] Rameen Abdal, Yipeng Qin, and Peter Wonka. “Image2stylegan: How to embed images into the stylegan latent space?” In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 4432–4441 (cit. on p. 73).
- [Abdal et al., 2020] Rameen Abdal, Yipeng Qin, and Peter Wonka. “Image2stylegan++: How to edit the embedded images?” In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8296–8305 (cit. on p. 73).
- [Agustsson et al., 2019] Eirikur Agustsson et al. “Generative adversarial networks for extreme learned image compression”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2019, pp. 221–231 (cit. on p. 24).
- [Amit et al., 2023] Guy Amit, Mosh Levy, and Yisroel Mirsky. “Transpose Attack: Stealing Datasets with Bidirectional Training”. In: *arXiv preprint arXiv:2311.07389* (2023) (cit. on pp. 2, 3, 17, 24, 41, 42).
- [Anwar et al., 2017] Sajid Anwar, Kyuyeon Hwang, and Wonyong Sung. “Structured pruning of deep convolutional neural networks”. In: *ACM Journal on Emerging Technologies in Computing Systems (JETC)* 13.3 (2017), pp. 1–18 (cit. on pp. 19, 26).
- [Ateniese et al., 2015] Giuseppe Ateniese et al. “Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers”. In: *International Journal of Security and Networks* 10.3 (2015), pp. 137–150 (cit. on p. 15).
- [Baheti et al., 2021] Bhakti Baheti et al. “The brain tumor sequence registration challenge: establishing correspondence between pre-operative and follow-up MRI scans of diffuse glioma patients”. In: *arXiv preprint arXiv:2112.06979* (2021) (cit. on p. 29).
- [Baid et al., 2021] Ujjwal Baid et al. “The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification”. In: *arXiv preprint arXiv:2107.02314* (2021) (cit. on p. 28).
- [Ballé et al., 2018] Johannes Ballé et al. “Variational Image Compression with a Scale Hyperprior”. In: *6th Int. Conf. on Learning Representations (ICLR)*. 2018 (cit. on pp. 19, 25, 28).

- [Barattin et al., 2023] Simone Barattin et al. “Attribute-preserving Face Dataset Anonymization via Latent Code Optimization”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 8001–8010 (cit. on pp. 7, 48, 50, 55).
- [Bilic et al., 2023] Patrick Bilic et al. “The liver tumor segmentation benchmark (lits)”. In: *Medical Image Analysis* 84 (2023), p. 102680 (cit. on p. 28).
- [Bortsova et al., 2021] Gerda Bortsova et al. “Adversarial attack vulnerability of medical image analysis systems: Unexplored factors”. In: *Medical Image Analysis* 73 (2021), p. 102141 (cit. on pp. 3, 4).
- [Boutet et al., 2023] Antoine Boutet, Carole Frindel, and Mohamed Maouche. “Towards an evolution in the characterization of the risk of re-identification of medical images”. In: *2023 IEEE International Conference on Big Data (BigData)*. IEEE. 2023, pp. 5454–5459 (cit. on p. 52).
- [Carlini et al., 2023] Nicolas Carlini et al. “Extracting training data from diffusion models”. In: *32nd USENIX Security Symposium (USENIX Security 23)*. 2023, pp. 5253–5270 (cit. on p. 17).
- [Cheddad et al., 2010] Abbas Cheddad et al. “Digital image steganography: Survey and analysis of current methods”. In: *Signal processing* 90.3 (2010), pp. 727–752 (cit. on p. 20).
- [Chen et al., 2021] Si Chen et al. “Knowledge-enriched distributional model inversion attacks”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 16178–16187 (cit. on p. 15).
- [Chen et al., 2022] Chen Chen et al. “Practical attribute reconstruction attack against federated learning”. In: *IEEE Transactions on Big Data* (2022) (cit. on pp. 2, 3, 15, 16).
- [Chen et al., 2023] Minghui Chen et al. “Data anonymization evaluation against re-identification attacks in edge storage”. In: *Wireless Networks* (2023), pp. 1–15 (cit. on pp. 2, 3, 15, 16).
- [Cheng et al., 2017] Yu Cheng et al. “A survey of model compression and acceleration for deep neural networks”. In: *arXiv preprint arXiv:1710.09282* (2017) (cit. on pp. 19, 26).
- [Chevrier et al., 2019] Raphaël Chevrier et al. “Use and understanding of anonymization and de-identification in the biomedical literature: scoping review”. In: *Journal of medical Internet research* 21.5 (2019), e13484 (cit. on p. 2).
- [Dadoun et al., 2021] Hind Dadoun et al. “Combining Bayesian and deep learning methods for the delineation of the fan in ultrasound images”. In: *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2021, pp. 743–747 (cit. on p. 47).

- [Deng et al., 2019] Jiankang Deng et al. “Arcface: Additive angular margin loss for deep face recognition”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4690–4699 (cit. on pp. 52, 53, 60).
- [Ding et al., 2021] Wenjie Ding et al. “Beyond Universal Person Re-Identification Attack”. In: *IEEE Transactions on Information Forensics and Security* 16 (2021), pp. 3442–3455 (cit. on pp. 2, 15).
- [Du et al., 2019] Ling Du et al. “An efficient privacy protection scheme for data security in video surveillance”. In: *Journal of visual communication and image representation* 59 (2019), pp. 347–362 (cit. on p. 47).
- [Duta et al., 2021] Ionut Cosmin Duta et al. “Improved residual networks for image and video recognition”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 9415–9422 (cit. on pp. 60, 81).
- [Dwork, 2006] Cynthia Dwork. “Differential privacy”. In: *International colloquium on automata, languages, and programming*. Springer. 2006, pp. 1–12 (cit. on pp. 5, 6, 21, 27).
- [El Emam et al., 2011] Khaled El Emam et al. “A systematic review of re-identification attacks on health data”. In: *PloS one* 6.12 (2011), e28071 (cit. on pp. 2, 15).
- [Fredrikson et al., 2015] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. “Model inversion attacks that exploit confidence information and basic countermeasures”. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015, pp. 1322–1333 (cit. on p. 15).
- [Ganju et al., 2018] Karan Ganju et al. “Property inference attacks on fully connected neural networks using permutation invariant representations”. In: *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 2018, pp. 619–633 (cit. on p. 15).
- [Gao et al., 2020] Rui Gao et al. “Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 3665–3680 (cit. on p. 24).
- [Gentner et al., 2023] Tobias Gentner et al. “Data lakes in healthcare: applications and benefits from the perspective of data sources and players”. In: *Procedia Computer Science* 225 (2023), pp. 1302–1311 (cit. on p. 1).
- [Gerstner et al., 2013] Timothy Gerstner et al. “Pixelated image abstraction with integrated user constraints”. In: *Computers & Graphics* 37.5 (2013), pp. 333–347 (cit. on p. 47).
- [Goodfellow et al., 2014] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014) (cit. on pp. 60, 71).

- [Goodfellow et al., 2020] Ian Goodfellow et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144 (cit. on pp. 51, 77).
- [Gou et al., 2021] Jianping Gou et al. “Knowledge distillation: A survey”. In: *International Journal of Computer Vision* 129 (2021), pp. 1789–1819 (cit. on pp. 19, 26).
- [Guillaudeux et al., 2023] Morgan Guillaudeux et al. “Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis”. In: *NPJ Digital Medicine* 6.1 (2023), p. 37 (cit. on p. 61).
- [Guo et al., 2020] Chuan Guo, Ruihan Wu, and Kilian Q Weinberger. “On hiding neural networks inside neural networks”. In: *arXiv preprint arXiv:2002.10078* (2020) (cit. on pp. 23, 24).
- [Haim et al., 2022] Niv Haim et al. “Reconstructing Training Data From Trained Neural Networks”. In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 22911–22924 (cit. on p. 17).
- [Hatamizadeh et al., 2022] Ali Hatamizadeh et al. “Gradvit: Gradient inversion of vision transformers”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10021–10030 (cit. on p. 15).
- [Hatamizadeh et al., 2023] Ali Hatamizadeh et al. “Do gradient inversion attacks make federated learning unsafe?” In: *IEEE Transactions on Medical Imaging* (2023) (cit. on p. 15).
- [Hu et al., 202] Yueyu Hu et al. “Learning end-to-end lossy image compression: A benchmark”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.8 (0202), pp. 4194–4211 (cit. on p. 41).
- [Hu et al., 2022] Hongsheng Hu et al. “Membership inference attacks on machine learning: A survey”. In: *ACM Computing Surveys (CSUR)* 54.11s (2022), pp. 1–37 (cit. on p. 15).
- [Huang et al., 2021] Yangsibo Huang et al. “Evaluating gradient inversion attacks and defenses in federated learning”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 7232–7241 (cit. on p. 15).
- [Iandola et al., 2014] Forrest Iandola et al. “Densenet: Implementing efficient convnet descriptor pyramids”. In: *arXiv preprint arXiv:1404.1869* (2014) (cit. on pp. 60, 79, 81).
- [Ingemar et al., 2008] J Ingemar et al. *Digital Watermarking and Steganography Second Edition*. 2008 (cit. on pp. 20, 26).
- [Isola et al., 2017] Phillip Isola et al. “Image-to-image translation with conditional adversarial networks”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1125–1134 (cit. on p. 72).

- [Jeon et al., 2022] Minkyu Jeon et al. “k-SALSA: k-anonymous synthetic averaging of retinal images via local style alignment”. In: *European Conference on Computer Vision*. Springer, 2022, pp. 661–678 (cit. on pp. 7, 47, 50, 64).
- [Johnson et al., 2019] Alistair Johnson et al. “MIMIC-CXR-JPG-chest radiographs with structured labels”. In: *PhysioNet* (2019) (cit. on pp. 57, 99).
- [Kaissis et al., 2020] Georgios A Kaissis et al. “Secure, privacy-preserving and federated machine learning in medical imaging”. In: *Nature Machine Intelligence* 2.6 (2020), pp. 305–311 (cit. on pp. 2, 4–6, 15).
- [Karras et al., 2017] Tero Karras et al. “Progressive growing of gans for improved quality, stability, and variation”. In: *arXiv preprint arXiv:1710.10196* (2017) (cit. on p. 71).
- [Karras et al., 2019] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative adversarial networks”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 4401–4410 (cit. on p. 72).
- [Karras et al., 2020] Tero Karras et al. “Analyzing and improving the image quality of stylegan”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 8110–8119 (cit. on pp. 10, 70, 72, 74, 75, 77, 78, 81, 98, 99).
- [Karras et al., 2021] Tero Karras et al. “Alias-free generative adversarial networks”. In: *Advances in neural information processing systems* 34 (2021), pp. 852–863 (cit. on p. 72).
- [Katsumata et al., 2024] Kai Katsumata et al. “Revisiting Latent Space of GAN Inversion for Robust Real Image Editing”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2024, pp. 5313–5322 (cit. on p. 68).
- [Katzir et al., 2022] Oren Katzir et al. “Multi-level latent space structuring for generative control”. In: *arXiv preprint arXiv:2202.05910* (2022) (cit. on pp. 70, 74).
- [Kaviani et al., 2022] Sara Kaviani, Ki Jin Han, and Insoo Sohn. “Adversarial attacks and defenses on AI in medical imaging informatics: A survey”. In: *Expert Systems with Applications* 198 (2022), p. 116815 (cit. on p. 15).
- [Kingma et al., 2013] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013) (cit. on p. 72).
- [Kingma et al., 2014] Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on pp. 30, 56, 60, 61, 81).
- [Kocabaş et al., 2014] Övünç Kocabaş and Tolga Soyata. “Medical data analytics in the cloud using homomorphic encryption”. In: *Handbook of Research on Cloud Infrastructures for Big Data Analytics*. IGI Global, 2014, pp. 471–488 (cit. on p. 4).

- [Kuang et al., 2024] Zhenzhong Kuang et al. “Facial Identity Anonymization via Intrinsic and Extrinsic Attention Distraction”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 12406–12415 (cit. on pp. 7, 48).
- [Larsen et al., 2016] Anders Boesen Lindbo Larsen et al. “Autoencoding beyond pixels using a learned similarity metric”. In: *International conference on machine learning*. PMLR. 2016, pp. 1558–1566 (cit. on p. 72).
- [Li et al., ] Huiyu Li, Nicholas Ayache, and Hervé Delingette. “Data Exfiltration by Compression Attack: Definition and Evaluation on Medical Image Data”. In: *Submitted to a journal ()* (cit. on pp. 10, 11, 14).
- [Li et al., 2006] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. “t-closeness: Privacy beyond k-anonymity and l-diversity”. In: *2007 IEEE 23rd international conference on data engineering*. IEEE. 2006, pp. 106–115 (cit. on pp. 5, 7).
- [Li et al., 2016] Dong Li et al. “Permutation anonymization”. In: *Journal of Intelligent Information Systems* 47 (2016), pp. 427–445 (cit. on pp. 5, 6).
- [Li et al., 2020a] Dong Li et al. “Privacy-preserving self-serviced medical diagnosis scheme based on secure multi-party computation”. In: *Computers & Security* 90 (2020), p. 101701 (cit. on p. 4).
- [Li et al., 2020b] Huiyu Li et al. “Deep distance map regression network with shape-aware loss for imbalanced medical image segmentation”. In: *Machine Learning in Medical Imaging: 11th International Workshop, MLMI 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Proceedings 11*. Springer. 2020, pp. 231–240 (cit. on p. 29).
- [Li et al., 2020c] Li Li et al. “A review of applications in federated learning”. In: *Computers & Industrial Engineering* 149 (2020), p. 106854 (cit. on p. 17).
- [Li et al., 2022a] Huiyu Li, Nicholas Ayache, and Hervé Delingette. “Data Stealing Attack on Medical Images: Is it Safe to Export Networks from Data Lakes?” In: *International Workshop on Distributed, Collaborative, and Federated Learning*. Springer. 2022, pp. 28–36 (cit. on pp. 10, 11, 14, 16).
- [Li et al., 2022b] Junzhuo Li et al. “Swing Distillation: A Privacy-Preserving Knowledge Distillation Framework”. In: *arXiv preprint arXiv:2212.08349* (2022) (cit. on p. 44).
- [Li et al., 2025] Huiyu Li, Nicholas Ayache, and Hervé Delingette. “Generative Medical Image Anonymization Based on Latent Code Projection and Optimization”. In: (2025) (cit. on pp. 10, 11, 70).

- [Liang et al., 2023] Haotian Liang et al. “EGIA: An External Gradient Inversion Attack in Federated Learning”. In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 4984–4995 (cit. on pp. 2, 3, 15, 16).
- [Liu et al., 2016] Weiyang Liu et al. “Large-margin softmax loss for convolutional neural networks”. In: *arXiv preprint arXiv:1612.02295* (2016) (cit. on p. 53).
- [Liu et al., 2017] Weiyang Liu et al. “Sphereface: Deep hypersphere embedding for face recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 212–220 (cit. on pp. 52, 53).
- [Liu et al., 2018] Zhuang Liu et al. “Rethinking the value of network pruning”. In: *arXiv preprint arXiv:1810.05270* (2018) (cit. on p. 19).
- [Liu et al., 2020] Tao Liu et al. “StegoNet: Turn deep neural network into a stegomalware”. In: *Proceedings of the 36th Annual Computer Security Applications Conference*. 2020, pp. 928–938 (cit. on p. 17).
- [Liu et al., 2023a] Gaoyang Liu et al. “Gradient-Leaks: Enabling Black-box Membership Inference Attacks against Machine Learning Models”. In: *IEEE Transactions on Information Forensics and Security* (2023), pp. 1–1 (cit. on pp. 2, 3, 15, 16).
- [Liu et al., 2023b] Gaoyang Liu et al. “TEAR: Exploring Temporal Evolution of Adversarial Robustness for Membership Inference Attacks Against Federated Learning”. In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 4996–5010 (cit. on p. 15).
- [Liu et al., 2023c] Hongyu Liu, Yibing Song, and Qifeng Chen. “Delving stylegan inversion for image editing: A foundation latent space viewpoint”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 10072–10082 (cit. on p. 91).
- [Liu et al., 2023d] Yimin Liu, Peng Jiang, and Liehuang Zhu. “Subject-Level Membership Inference Attack via Data Augmentation and Model Discrepancy”. In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 5848–5859 (cit. on p. 15).
- [Loshchilov et al., 2017] Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017) (cit. on p. 60).
- [Ma et al., 2022] Jun Ma et al. “Fast and low-GPU-memory abdomen CT organ segmentation: the flare challenge”. In: *Medical Image Analysis* 82 (2022), p. 102616 (cit. on p. 29).
- [Machanavajjhala et al., 2007] Ashwin Machanavajjhala et al. “l-diversity: Privacy beyond k-anonymity”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), 3–es (cit. on pp. 5, 7).

- [Meden et al., 2018] Blaž Meden et al. “k-Same-Net: k-Anonymity with generative deep neural networks for face deidentification”. In: *Entropy* 20.1 (2018), p. 60 (cit. on p. 7).
- [Melis et al., 2019] Luca Melis et al. “Exploiting unintended feature leakage in collaborative learning”. In: *2019 IEEE Symposium on Security and Privacy (SP)*. 2019, pp. 691–706 (cit. on p. 15).
- [Mentzer et al., 2020] Fabian Mentzer et al. “High-fidelity generative image compression”. In: *Advances in Neural Information Processing Systems 33* (2020), pp. 11913–11924 (cit. on pp. 24, 26, 29, 59, 97).
- [Mescheder et al., 2018] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. “Which training methods for GANs do actually converge?” In: *International conference on machine learning*. PMLR. 2018, pp. 3481–3490 (cit. on p. 77).
- [Mirza et al., 2014] Mehdi Mirza and Simon Osindero. “Conditional generative adversarial nets”. In: *arXiv preprint arXiv:1411.1784* (2014) (cit. on p. 72).
- [Mivule, 2013] Kato Mivule. “Utilizing noise addition for data privacy, an overview”. In: *arXiv preprint arXiv:1309.3958* (2013) (cit. on pp. 5, 6).
- [Neubauer et al., 2011] Thomas Neubauer and Johannes Heurix. “A methodology for the pseudonymization of medical data”. In: *International journal of medical informatics* 80.3 (2011), pp. 190–204 (cit. on p. 2).
- [Nguyen et al., 2024] Bao-Ngoc Nguyen et al. “Label-Only Model Inversion Attacks via Knowledge Transfer”. In: *Advances in Neural Information Processing Systems 36* (2024) (cit. on p. 15).
- [Orekondy et al., 2018] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. “Connecting pixels to privacy and utility: Automatic redaction of private information in images”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 8466–8475 (cit. on p. 47).
- [Ovcharenko, 2019] Illia Ovcharenko. *Lung segmentation for chest X-Ray images*. 2019 (cit. on p. 58).
- [Packhäuser et al., 2022] Kai Packhäuser et al. “Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest X-ray data”. In: *Scientific Reports* 12.1 (2022), p. 14851 (cit. on pp. 46, 52).
- [Party, 2014] Article 29 Data Protection Working Party. “Opinion 05/2014 on Anonymisation Techniques”. In: (2014) (cit. on pp. 5, 6, 8, 99).



- [Pennisi et al., 2023] Matteo Pennisi et al. “A privacy-preserving walk in the latent space of generative models for medical applications”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 422–431 (cit. on pp. 7, 47, 50, 64).
- [Pfitzner et al., 2021] Bjarne Pfitzner, Nico Steckhan, and Bert Arnrich. “Federated learning in a medical context: a systematic literature review”. In: *ACM Transactions on Internet Technology (TOIT)* 21.2 (2021), pp. 1–31 (cit. on p. 4).
- [Pham et al., 2021] Hieu H Pham et al. “Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels”. In: *Neurocomputing* 437 (2021), pp. 186–194 (cit. on p. 54).
- [Pidhorskyi et al., 2020] Stanislav Pidhorskyi, Donald A Adjeroh, and Gianfranco Doretto. “Adversarial latent autoencoders”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 14104–14113 (cit. on p. 73).
- [Radford et al., 2015] Alec Radford, Luke Metz, and Soumith Chintala. “Unsupervised representation learning with deep convolutional generative adversarial networks”. In: *arXiv preprint arXiv:1511.06434* (2015) (cit. on p. 71).
- [Reynaud et al., 2024] Hadrien Reynaud et al. “EchoNet-Synthetic: Privacy-preserving Video Generation for Safe Medical Data Sharing”. In: *arXiv preprint arXiv:2406.00808* (2024) (cit. on p. 7).
- [Richardson et al., 2021] Elad Richardson et al. “Encoding in style: a stylegan encoder for image-to-image translation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 2287–2296 (cit. on pp. 71, 73).
- [Russakovsky et al., 2015] Olga Russakovsky et al. “Imagenet large scale visual recognition challenge”. In: *International journal of computer vision* 115 (2015), pp. 211–252 (cit. on p. 60).
- [Selvaraju et al., 2017] Ramprasaath R Selvaraju et al. “Grad-cam: Visual explanations from deep networks via gradient-based localization”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 618–626 (cit. on p. 95).
- [Seo et al., 2024] Juwon Seo et al. “Generative Unlearning for Any Identity”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 9151–9161 (cit. on pp. 7, 48).
- [Shokri et al., 2017] Reza Shokri et al. “Membership inference attacks against machine learning models”. In: *2017 IEEE symposium on security and privacy (SP)*. IEEE. 2017, pp. 3–18 (cit. on p. 61).
- [Simonyan et al., 2014] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014) (cit. on pp. 51, 77).

- [Sutskever et al., 2013] Ilya Sutskever et al. “On the importance of initialization and momentum in deep learning”. In: *International conference on machine learning*. PMLR. 2013, pp. 1139–1147 (cit. on p. 60).
- [Sweeney, 2002] Latanya Sweeney. “k-anonymity: A model for protecting privacy”. In: *International journal of uncertainty, fuzziness and knowledge-based systems* 10.05 (2002), pp. 557–570 (cit. on pp. 5, 7, 47).
- [Tov et al., 2021] Omer Tov et al. “Designing an encoder for stylegan image manipulation”. In: *ACM Transactions on Graphics (TOG)* 40.4 (2021), pp. 1–14 (cit. on pp. 71, 73, 76).
- [Tsubota et al., 2023] Koki Tsubota, Hiroaki Akutsu, and Kiyoharu Aizawa. “Universal deep image compression via content-adaptive optimization with adapters”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 2529–2538 (cit. on p. 41).
- [Ullah et al., 2018] Faheem Ullah et al. “Data exfiltration: A review of external attack vectors and countermeasures”. In: *Journal of Network and Computer Applications* 101 (2018), pp. 18–54 (cit. on p. 15).
- [Wang et al., 2003] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. “Multiscale structural similarity for image quality assessment”. In: *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. IEEE. 2003, pp. 1398–1402 (cit. on pp. 29, 61, 63).
- [Wang et al., 2004] Zhou Wang et al. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612 (cit. on pp. 61, 79).
- [Wang et al., 2010] Zhou Wang and Qiang Li. “Information content weighting for perceptual image quality assessment”. In: *IEEE Transactions on image processing* 20.5 (2010), pp. 1185–1198 (cit. on p. 79).
- [Wang et al., 2022a] Xiuling Wang and Wendy Hui Wang. “Group property inference attacks against graph neural networks”. In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. 2022, pp. 2871–2884 (cit. on p. 15).
- [Wang et al., 2022b] Zhibo Wang et al. “Poisoning-assisted property inference attack against federated learning”. In: *IEEE Transactions on Dependable and Secure Computing* (2022) (cit. on p. 15).
- [Wei et al., 2022] Tianyi Wei et al. “E2Style: Improve the efficiency and effectiveness of StyleGAN inversion”. In: *IEEE Transactions on Image Processing* 31 (2022), pp. 3267–3280 (cit. on pp. 71, 74).
- [Yang et al., 2019] Jiwei Yang et al. “Quantization networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019, pp. 7308–7316 (cit. on pp. 19, 26).

- [Zari et al., 2022] Oualid Zari et al. “Membership inference attack against principal component analysis”. In: *International Conference on Privacy in Statistical Databases*. Springer. 2022, pp. 269–282 (cit. on p. 15).
- [Zhang et al., 2018] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595 (cit. on pp. 61, 63).
- [Zhang et al., 2021] Wanrong Zhang, Shruti Tople, and Olga Ohrimenko. “Leakage of dataset properties in {Multi-Party} machine learning”. In: *30th USENIX security symposium (USENIX Security 21)*. 2021, pp. 2687–2704 (cit. on p. 15).
- [Zhang et al., 2022a] Hanwei Zhang et al. “Deep Neural Network Attacks and Defense: The Case of Image Classification”. In: *Multimedia Security 1 (2022)*, pp. 41–75 (cit. on p. 15).
- [Zhang et al., 2022b] Zhikun Zhang et al. “Inference Attacks Against Graph Neural Networks”. In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022, pp. 4543–4560 (cit. on pp. 2, 3, 15, 16).
- [Zhang et al., 2023] Zeping Zhang et al. “Analysis and Utilization of Hidden Information in Model Inversion Attacks”. In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 4449–4462 (cit. on p. 15).
- [Zhao et al., 2021] Zilong Zhao et al. “Ctab-gan: Effective table data synthesizing”. In: *Asian Conference on Machine Learning*. PMLR. 2021, pp. 97–112 (cit. on p. 61).
- [Zhou et al., 2021] Xingchen Zhou et al. “Deep model poisoning attack on federated learning”. In: *Future Internet* 13.3 (2021), p. 73 (cit. on pp. 3, 4).
- [Zhu et al., 2017] Jun-Yan Zhu et al. “Unpaired image-to-image translation using cycle-consistent adversarial networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2223–2232 (cit. on p. 72).
- [Zhu et al., 2023] Tianqing Zhu et al. “Label-Only Model Inversion Attacks: Attack With the Least Information”. In: *IEEE Transactions on Information Forensics and Security* 18 (2023), pp. 991–1005 (cit. on pp. 2, 3, 15, 16, 25, 41).



