UTS UNIVERSITY OF TECHNOLOGY SYDNEY

A Study on Image Privacy Protection in Response to Artificial Intelligence Technology

by Hanyu Xue

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Dr. Bo Liu

University of Technology Sydney
Faculty of Engineering and Information Technology

July 2023

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Hanyu Xue, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: 20 July 2023

A Study on

Image Privacy Protection in Response to Artificial Intelligence Technology

A thesis submitted in partial fulfilment of the requirements
for the degree of

Doctor of Philosophy
in
Analytics

by

Hanyu Xue

to

School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
NSW - 2007, Australia

June 2023

Abstract

This thesis addresses the pressing issue of image privacy protection in the era of data sharing and artificial intelligence (AI) technology. The motivation behind this research stems from the need to develop effective methods to counter the privacy risks posed by AI. The thesis aims to develop a framework for image privacy protection and evaluate the effectiveness of the proposed methods.

The research begins by highlighting the existing challenges in image privacy protection, including inadequate protection against AI adversaries, a lack of comprehensive privacy definition and quantification, limited applicability of existing methods, and the need to strike a better balance between privacy and utility. These challenges underscore the need for innovative approaches to safeguard personal information from both human and AI adversaries.

The thesis makes several contributions in different areas of image privacy protection. Firstly, it addresses privacy concerns in social media platforms and proposes an image privacy protection framework using adversarial perturbations. The framework defines privacy information, identifies private objects, and applies imperceptible adversarial perturbations to hide private information while preserving image utility.

Secondly, the thesis focuses on provable image privacy protection and introduces a differentially private image (DP-Image) framework. This framework redefines differential privacy in the context of image data and perturbs image feature vectors in the latent space to ensure privacy protection. The proposed mechanism maintains the utility of images while protecting sensitive information against both human and AI adversaries.

Thirdly, the research explores user-centric privacy protection, empowering users to have control over their privacy while benefiting from computer vision applications. Novel privacy protection mechanisms tailored to individual preferences and customizable privacy levels are proposed, demonstrating the feasibility and benefits of user-centric approaches in safeguarding privacy.

Lastly, the thesis addresses the challenge of high-quality image de-identification by developing techniques that remove or obfuscate identifying information while preserving image quality. The utilization of the decoder and attribute optimization techniques enhances the effectiveness and usability of anonymized datasets.

In conclusion, this thesis makes significant contributions to the field of image privacy protection by addressing privacy concerns in social media platforms, provable privacy protection, user-centric privacy protection, and high-quality image de-identification. The proposed frameworks, mechanisms, and techniques advance the state-of-the-art in image

privacy protection and provide valuable insights into preserving privacy in the era of data sharing and AI technology.

Dedication

To my wife, your unwavering love, patience, and understanding has been my anchor during this challenging period. Your support and belief in me have propelled me forward, even when faced with unprecedented obstacles.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my principal supervisor Dr. Bo Liu and my co-supervisor Dr. Tianqing Zhu, for their guidance, support, and invaluable expertise throughout my PhD journey. Their mentorship has been instrumental in shaping my research and academic growth.

I extend my appreciation to the faculty and staff of the University of Technology of Sydney, whose dedication to education and research has created an inspiring academic environment. Their commitment to excellence has provided me countless opportunities for learning and collaboration.

I would like to acknowledge the support and encouragement I received from my colleagues and fellow researchers. Their collaboration, stimulating discussions and friendship have made this journey truly rewarding.

My heartfelt thanks go to my family and friends for their unwavering support, understanding, and encouragement throughout my PhD studies. Their love, patience, and belief in my abilities have been a constant source of motivation.

Lastly, I would like to express my gratitude to the funding agencies and scholarships that have supported my research, enabling me to pursue my academic aspirations.

This thesis would not have been possible without the contributions and support of all those mentioned above, and for that, I am forever grateful.

List of Publications

Publications of the author during the PhD study are listed as the following:

1. **Xue, H.**, Liu, B., Din, M., Song, L., and Zhu, T. (2020, June). Hiding private information in images from AI. In ICC 2020-2020 IEEE International Conference on Communications (ICC) (pp. 1-6). IEEE.

2. **Xue, H.**, Liu, B., Ding, M., Zhu, T., Ye, D., Song, L., and Zhou, W. (2021). Dp-image: differential privacy for image data in feature space. arXiv preprint arXiv:2103.07073.

3. **Xue, H.**, Liu, B., Yuan, X., Ding, M., and Zhu, T. (2023). Face image de-identification by feature space adversarial perturbation. Concurrency and Computation: Practice and Experience, 35(5), e7554.

4. **Xue, H.**, Yuan, X., Liu, B., Ding, M., and Zhu, T. (2023). Diffussion-based face dataset anonymization. Ready to submit.

Table of Contents

List of Publications **ix**

List of Figures **xv**

List of Tables **xix**

1 Introduction **1**

1.1 Motivation and Objective of the Thesis 2
1.2 Existing challenges 3
1.3 Contributions of the thesis 4
1.3.1 Privacy concerns in social media platforms 4
1.3.2 Provable image privacy protection 5
1.3.3 User-centric privacy protection 5
1.3.4 High-quality image de-identification 6
1.3.5 Summary 6
1.4 Overview of the thesis 7

2 Preliminary and Related Work **9**

2.1 Privacy Definition Based on General Data Protection Regulation (i.e. GDPR) 10
2.2 Preliminary 11
2.2.1 Adversarial examples 11
2.2.2 Generative Adversarial Networks (GANs) 14
2.2.3 Style-Based Generative Adversarial Networks (StyleGAN) 17
2.2.4 StyleGAN-based autoencoder 18
2.2.5 Diffusion model 20
2.3 Related Work 23
2.3.1 Traditional protection methods 23
2.3.2 Adversarial examples 24

2.3.3 GAN-based inpainting 26
2.3.4 Differential privacy 27
2.3.5 Diffusion model 28

3 Hiding Private Information in Images From AI 31
3.1 Preface 32
3.2 Introduction 32
3.2.1 Motivation 32
3.2.2 Contributions 33
3.2.3 Overview of the work 34
3.3 System Model and Problem Formulation 35
3.3.1 System Model 35
3.3.2 Problem Formulation 38
3.4 Adversarial Perturbation based Image Privacy Protection Algorithm . . . 39
3.4.1 AP-based Image Privacy Protection 39
3.4.2 Evaluation Metrics 40
3.5 Experiment and Discussions 42
3.5.1 Experiment Settings 42
3.5.2 The Experiment Results 42
3.6 Conclusion and future works 45

4 DP-Image (DP-Image: Differential Privacy for Image Data in Feature Space) 47
4.1 Preface 48
4.2 Introduction 48
4.2.1 Motivation 48
4.2.2 Contributions 51
4.2.3 Overview of the work 52
4.3 Preliminaries 52
4.3.1 Differential Privacy 52
4.3.2 Rényi Differential Privacy 54
4.3.3 Privacy Amplification by Iteration 54
4.3.4 Deep Learning in Image Applications 55
4.4 Our proposed DP-Image Framework 57
4.4.1 Adversary Model 57
4.4.2 DP-Image Framework: Protecting Image Privacy in Feature Space 58

4.4.3 DP-Image Definition . . . 59
4.4.4 DP-Image Mechanism . . . 59
4.5 Experiments . . . 61
4.5.1 Experiment Setup . . . 61
4.5.2 Performance of the DP-Image Mechanism . . . 66
4.6 Discussions . . . 71
4.6.1 Privacy Analysis . . . 71
4.6.2 Disentangle Identity-related Features in Image . . . 71
4.6.3 Image Privacy vs. Database Privacy . . . 73
4.7 Related Work . . . 73
4.8 Conclusion and Future Work . . . 74

5 Face Image De-identification by Feature Space Adversarial Perturbation (FSAP) 77
5.1 Preface . . . 78
5.2 Introduction . . . 78
5.2.1 Motivation . . . 78
5.2.2 Contributions . . . 80
5.2.3 Overview of the work . . . 80
5.3 Related Work . . . 82
5.4 Feature Space Adversarial Perturbation Based Face Image De-identification Framework . . . 84
5.4.1 Problem Formulation . . . 84
5.4.2 FSAP Framework . . . 85
5.4.3 Adversarial Perturbation Generation Process . . . 87
5.5 Experiments . . . 89
5.5.1 Experiment Setup . . . 89
5.5.2 Performance Evaluation . . . 91
5.5.3 Discussions . . . 94
5.6 Conclusion . . . 96
5.7 Future work . . . 97

6 Diffusion De-ID: Diffussion-based Face Dataset Anonymization 99
6.1 Preface . . . 100
6.2 Introduction . . . 100
6.2.1 Motivation . . . 100

6.2.2 Contributions . . . 102
6.2.3 Overview of the work . . . 103
6.3 Related Work . . . 104
6.4 Preliminaries-Conditional DDIM . . . 105
6.5 Methodology . . . 107
6.5.1 Modules . . . 108
6.5.2 Fake dataset generation . . . 108
6.5.3 Pairing . . . 109
6.5.4 Anonymization . . . 109
6.6 Experiments . . . 111
6.6.1 Experiment settings . . . 111
6.6.2 Comparison to state-of-the-art (SOTA) . . . 114
6.6.3 Ablation study . . . 120
6.7 Conclusions . . . 122

7 Conclusion and Future Work **123**
7.1 Conclusion . . . 123
7.2 Future work . . . 124

Bibliography **127**

List of Figures

Figure **Page**

1.1 Research Approaches Path Map. 7

2.1 An adversarial example. [1] 12
2.2 The structure of a conditional adversarial network [2]. 16
2.3 The generator structure of a Style-based generative adversarial network [3]. 18
2.4 The structure of a StyleGAN-based autoencoder [4]. 19
2.5 The graphical model of DDPMs [5]. 20
2.6 The graphical model of DDIM [6]. 21

3.1 Image privacy protection framework. 35
3.2 The Faster R-CNN framework. 36
3.3 Diagram of AP-based image privacy protection algorithm. 39
3.4 Illustration of AP-based image privacy protect algorithm. 43
3.5 The detection results after privacy protection: (a) Image without Protection; (b) Blur; (c) Mosaic; (d) AP-based. 43
3.6 The hide/keep ratio; (a) Fixed iteration number $N = 1$; (b) Fixed $\epsilon = 0.4$; (c) Fixed $\epsilon = 0.2$ 45

4.1 A brief comparison of different image privacy protection methods. Our approach not only protects the privacy of an image by generating a photo-realistic alternative, but it also provides a controllable way for privacy preservation. . 51
4.2 Architecture of the DP-Image Framework. The Iterative DP-Image (IDPI) theorem presented in Section 4.4.4 will demonstrate the concepts of the required privacy level, denoted as ϵ, and sensitivity, represented by Δf. . . . 57
4.3 The framework of generating feature space vector using pSp encoder and reconstructing image with StyleGAN generator. 63
4.4 Distribution of the values in latent space feature vector Z. 63

4.5 Generated faces with proposed DP-Image method. 64

4.6 The DP-image visual results with reference (ϵ, δ)-DP. The x axis and y axis are the image iteration number and reference ϵ in logarithm. From left to right, they are original image and corresponding generated images with different T and same $\sigma = 0.2$. The sensitivity $\Delta f = 3840$. $\delta = 10^{-8}$ by the natural settings on deep learning. 65

4.7 The image visual results for IDPI with different iteration numbers. Above, we give the possible visual results of face images under the $\sigma = 0.2$ setting. From left to right, the first column is original face images, and all other face images are generated with the different iteration numbers shown above. 65

4.8 The image visual results for IDPI images with different σ. Here we give the possible visual results under $T = 10^2$ setting. From left to right, the first column is original images, and all other face images are generated with the different σ shown above. 66

4.9 The 20 image examples for the heat map in Fig. 4.10. 66

4.10 The confidence values of pair-wise image comparison for the 20 example images. Microsoft Face API holds high accuracy on face recognition. Any two faces with different identities will be given a lower score, e.g. confidence ≤ 50. 67

4.11 We evaluate the DP-image performance with Microsoft Face API (a) and the ArcFace (b). (a) The x axis and y axis are iteration number T and Microsoft face API confidence score respectively. (b) The x axis and y axis are iteration number T and ArcFace score respectively. We set $\sigma = 0.2, \sigma = 0.4, \sigma = 0.6$ to compare the confidence score with different settings. 68

4.12 We evaluate the DP-image performance with Microsoft Face API (a) and the ArcFace (b). (a) The x axis and y axis are iteration number T and FPPSR on Microsoft face API respectively. (b) The x axis and y axis are iteration number T and FPPSR on ArcFace respectively. We set $\sigma = 0.2, \sigma = 0.4, \sigma = 0.6$ to compare the FPPSR with a different setting. 69

4.13 Generated faces with noises added to identity-related features of Z. We give the possible visual results of adding noise on image identity features. From left to right, they are original images, the faces protected by IDPI$(X_0, \sigma = 0.2, T = 10)$, IDPI$(X_0, \sigma = 0.2, T = 20)$, IDPI$(X_0, \sigma = 0.2, T = 30)$, IDPI$(X_0, \sigma = 0.2, T = 40)$, IDPI$(X_0, \sigma = 0.2, T = 50)$ 72

5.1 Face deidentification results. From left to right, (a) Original image; (b) Blur noise; (c) Pixelation noise; (d) Fawkes [7]; (e) DeepPrivacy [8]; (f) CIAGAN [9]; (g) feature space adversarial perturbation (**Ours**). 81
5.2 Feature Space Adversarial Perturbation (FSAP) based privacy protection framework. 85
5.3 The visual results of the ablation study. The first row is the original image. The second row and third row are the de-identity images generated by ID loss and Bi-loss framework, respectively. 91
5.4 Parameters Analysis. The values of λ_I (or λ_P) range from 0.01 to 0.1 with a step size of 0.01. The maximum iteration number $N = 30$. The threshold $\tau = 0.8$. 94
5.5 The confidence of commercial API. The y-axis indicates the confidence score of the MS API and the FACE++ API. The blue violin plot signifies the outcomes derived from the Microsoft API, whereas the orange violin plot corresponds to the results obtained from the Face++ API. The dashed line denotes the threshold applied to each API, wherein facial images falling below this threshold are construed as emanating from disparate identities. Notably, our methodology exhibits efficacy for both APIs. 95
5.6 Failure examples. Output 1 refers to the generated images without protection. Output 2 shows the de-identity image with our method. 96

6.1 Visual quality compared to DeepPrivacy [10], CIAGAN [11], FALCO [12]. **Ours**, along with the FALCO approach, exhibits the capability of concealing sensitive information within the background. Notably, **Ours** also demonstrates the ability to retain a richer set of attributes and intricate details. . . 102
6.2 Our proposed protection framework. Our protection framework optimizing the trainable anonymized latent vector $\mathbf{z}_a^i \in \mathbb{R}^{512}$ using two loss functions, $\mathcal{L}_{id}$ and $\mathcal{L}_{att}$. This optimization aims to obfuscate the identity of the synthetic image $\mathbf{x}_A^i$ while maintaining the facial attributes. 107
6.3 Comparative analysis of visual effects: Our method vs. FALCO [12], DeepPrivacy [10], and CIAGAN [11], in the context of original images undetectable by the dlib [13] frontal face detector. 113
6.4 Comparative analysis of visual effects: Our method vs. FALCO [12], DeepPrivacy [10], and CIAGAN [11]. 119
6.5 Visual quality of ablation study. **Ours** compared to FALCO [12], DeepPrivacy [10], CIAGAN [11]. 121

List of Tables

Table **Page**

3.1 $L2$ and ALD_p score compared with classical methods 44
3.2 The SSIM score compared with classical methods 44

4.1 Utility comparisons of the proposed privacy protection methods with traditional privacy protection methods. 70

5.1 Ablation Study. We use the same ID distance threshold, $\tau = 0.8$, for all settings in this table. The second column is the protection rate under Face Recognition Library. The third column is the detection rate by using *dlib* [13]. The hyperparameters are set to $\lambda_I = 0.02, \lambda_P = 0.008$, and the maximum iteration number $N = 100$. 92
5.2 Privacy evaluation. The values in this table are the successful protection rates (SPRs). The generation mode of Fawkes [7] is set to *high*, which is the highest privacy level authors recommended. The threshold of Face Recognition is $\delta = 0.6$ and the threshold of FaceNet is $\delta = 1.1$ according to [14]. 93
5.3 Utility evaluation. The face detection network used in this table is *dlib*. The Landmarks Distances are generated under the Face Recognition Library. . . 93

6.1 Anonymized image dataset evaluation results. Face Detection Accuracy (Face Dete), Identity Anonymity (ID Anon), Image Quality, and Attributes Classification. 114
6.2 Facial attribute preservation results 115
6.3 Ablation study evaluation results. Face Detection Accuracy (Face Dete), Identity Anonymity (ID Anon), Image Quality, and Attributes Classification. . . . 122

CHAPTER 1

INTRODUCTION

The thesis aims to tackle the pressing issue of image privacy protection in the era of data sharing and AI technology. This introduction provides an overview of the motivation and objectives of the research, highlighting the need for effective methods to counter the privacy risks posed by AI. It discusses the limitations of existing approaches and outlines the two-fold objectives of the thesis: developing a framework for image privacy protection and evaluating the effectiveness of the proposed methods.

1.1 Motivation and Objective of the Thesis

In today's era of data sharing, where people frequently upload and share personal photos and videos on social platforms, the need to protect image privacy has become increasingly critical. The vast amount of private information embedded in these images, such as faces, license plates, locations, and email addresses, poses significant risks if exploited by adversaries. Moreover, the emergence of artificial intelligence (AI) technology, powered by deep learning methods, further amplifies these privacy risks. AI can automatically extract and detect sensitive information from images, leading to potential privacy breaches and misuse of personal data.

Traditional image privacy protection methods, which assume human adversaries, have relied on techniques like blurring, pixelation, and mosaic. However, with the rise of AI as an adversary, these methods prove insufficient in safeguarding privacy. To address this pressing issue, the objective of this thesis is to conduct a comprehensive study on image privacy protection in response to AI technology.

The main motivation behind this research is the urgent need to develop effective image privacy protection methods that can counter the privacy risks posed by AI. The existing approaches, such as adversarial perturbations and generative adversarial networks (GANs), have shown promise but still face limitations. The thesis aims to overcome these limitations and make significant contributions to the field of image privacy protection.

The major objectives of this thesis are two-fold: (i) Develop a framework for image privacy protection that can conceal private information from AI while remaining imperceptible to human observers. This framework will address the challenge of defining and quantifying image privacy, taking into account the multiple private objects often present in social network images. (ii) Evaluate the effectiveness and performance of the proposed image privacy protection methods using real-life image datasets. The evaluation will demonstrate the capability of the methods in safeguarding individuals' privacy against both human and AI adversaries.

By achieving these objectives, this thesis aims to advance the field of image privacy protection and provide practical solutions to address the privacy risks associated with AI technology. The research outcomes will contribute to the development of robust and efficient methods for protecting personal privacy in the context of image sharing and usage.

1.2 Existing challenges

The field of image privacy protection have the following existing challenges:

- **Inadequate protection against AI adversaries:** With the emergence of deep learning techniques, AI can automatically collect and detect private and sensitive information from images. Traditional privacy protection methods designed for human adversaries, such as blurring, pixelation, and mosaic, are not effective against AI adversaries. The existing methods lack the ability to mislead AI models without causing significant visual differences perceptible to human eyes.

- **Lack of comprehensive privacy definition and quantification:** Image privacy is a complex concept that is yet to be clearly defined and quantitatively measured. The existing research does not provide a comprehensive framework to define and measure image privacy, which hinders the development of effective protection mechanisms.

- **Limited applicability and controllability of adversarial perturbation-based methods:** Adversarial perturbation-based methods have shown potential in privacy protection, but they are often designed for specific models and applications. These methods lack controllability over the visual appearance of the perturbed images, making them ineffective against human adversaries. There is a need for more general and controllable approaches.

- **Lack of provable privacy measurements for GAN-based methods:** Generative adversarial network (GAN)-based image manipulation techniques have been proposed for privacy protection. However, these methods lack provable privacy measurements, making it challenging to assess their effectiveness in safeguarding privacy.

- **Limited effectiveness of existing differential privacy (DP) methods for images:** Existing DP methods applied to images have limitations in capturing the key features of images from a privacy perspective. The current approaches, such as pixel-level differential privacy and DP noise layer in deep neural networks, do not effectively address image privacy protection as their primary objective.

- **Balancing privacy and utility:** Privacy protection methods should aim to hide private information while preserving the utility and meaningfulness of the images.

The existing approaches need to strike a better balance between privacy preservation and maintaining the usefulness of the images for various applications, such as demographics analysis and social media sharing.

- **Insufficient image privacy protection in the context of high-quality content:** In the domain of image privacy protection, the inadequacy of existing advanced methodologies, specifically those reliant on GAN-based inpainting, has become apparent. These techniques are frequently impeded by the inherent limitations of the GAN model's generation capacity, rendering them ineffective in addressing intricate real-world scenarios.

Addressing these challenges is crucial for the development of effective image privacy protection methods that can safeguard personal information from both human and AI adversaries.

1.3 Contributions of the thesis

The detailed contributions are summarized in the following subsections.

1.3.1 Privacy concerns in social media platforms

This part addresses the issue of privacy concerns in social media platforms and proposes an image privacy protection framework using adversarial perturbations. The main contributions of Chapter 3 are:

- **Development of an image privacy protection framework:** The chapter presents a framework for protecting private information in images, specifically targeting the risks posed by AI. The framework defines the privacy information in images, identifies private objects, and applies adversarial perturbations to hide private information while preserving the utility of the images.

- **Proposing an adversarial perturbation-based image privacy protection scheme:** The chapter introduces a scheme that adds adversarial perturbations to sensitive parts of images, effectively hiding multiple private objects while minimizing the impact on non-private objects. This scheme provides an imperceptible privacy protection method for images shared on social media platforms.

1.3.2 Provable image privacy protection

This part focuses on provable image privacy protection and presents a differentially private image (DP-Image) framework. The main contributions of Chapter 4 are:

- **Definition of DP-Image framework:** The chapter redefines differential privacy concepts in the context of image data and proposes a DP-Image framework that adds differential privacy noise to image feature vectors in the latent space. This framework addresses the challenge of applying differential privacy to image data, considering the unique characteristics and privacy risks associated with images.
- **Design of DP-Image protection mechanism:** The chapter develops a provable and adjustable privacy protection mechanism within the DP-Image framework. This mechanism perturbs the meaningful information carried by an image to mislead adversaries while ensuring the reconstructed image maintains utility and context consistency.
- **Implementation and evaluation:** The chapter implements the proposed DP-Image protection mechanisms on a real-life image dataset and demonstrates their effectiveness in safeguarding individuals' privacy. The evaluation shows that DP-Image protection maintains the usefulness of images for applications such as demographics analysis while protecting sensitive information against both human and AI adversaries.

1.3.3 User-centric privacy protection

This part focuses on user-centric privacy protection in the context of computer vision technology. The main contributions of Chapter 5 are:

- **User-centric privacy protection:** The chapter is on empowering users to have control over their privacy while still benefiting from computer vision applications.
- **Novel privacy protection mechanisms:** The chapter proposes novel privacy protection mechanisms tailored to the needs and preferences of individual users. These mechanisms provide customizable privacy levels and enable users to manage their privacy in the context of visual data.
- **Implementation and validation:** The chapter implements the user-centric privacy protection mechanisms and evaluates their effectiveness in protecting privacy

in computer vision applications. The results demonstrate the feasibility and benefits of user-centric approaches in safeguarding privacy.

1.3.4 High-quality image de-identification

This part addresses the challenge of high-quality image de-identification. The main contributions of Chapter 6 are:

- **High-quality image de-identification:** The chapter focuses on developing techniques for de-identifying images while preserving their quality. By removing or obfuscating identifying information from images, the framework enables the sharing and analysis of images without compromising privacy.
- **Face anonymization with novel model:** The utilization of the DDIM decoder and attribute optimization techniques enhances the effectiveness and usability of the anonymized datasets, which enables the training of robust facial models for various practical applications.

1.3.5 Summary

In summary, these four parts collectively contribute to the field of image privacy protection by addressing different aspects of the problem, including privacy concerns in social media platforms, provable privacy protection, user-centric privacy protection, and high-quality image de-identification. The proposed frameworks, mechanisms, and techniques contribute to advancing the state-of-the-art in image privacy protection and provide valuable insights into preserving privacy in the era of data sharing and deep learning.

1.4 Overview of the thesis

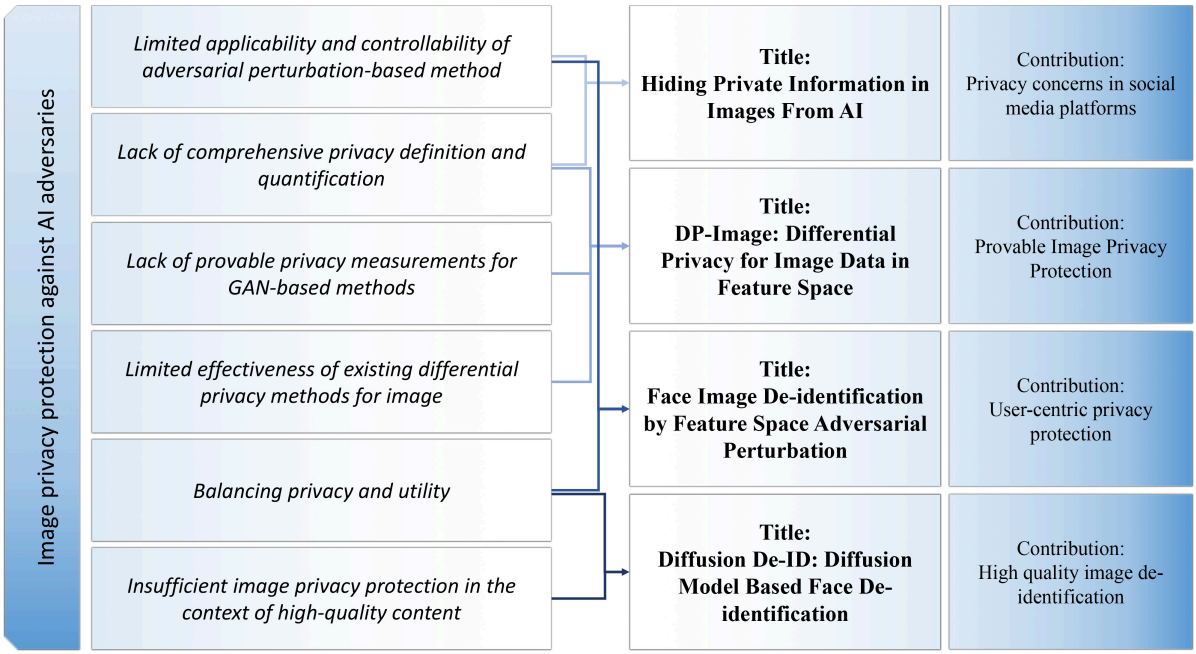


Figure 1.1: Research Approaches Path Map.

The thesis is devoted to addressing the challenges discussed in previous sections, progressing from Chapter 3 to Chapter 6 to continuously refine methods for protecting privacy in the field of image privacy.

In Chapter 3, the gradient mechanism in pattern recognition networks introduces adversarial noise, creating the first line of defence to hide privacy information within images from neural networks. If the neural network breaches this initial defence, we employ the mechanism of classification networks in Chapter 4 through Chapter 6 to misclassify the privacy information, providing attackers with the wrong data. These three chapters embody this approach, refining the effectiveness of this protective method to varying degrees.

Chapter 4 specifically explores the effectiveness of using differential privacy as a protective measure for image privacy. The goal is to find a credible protection method for image privacy, enhancing the trustworthiness of the safeguard.

Chapter 5 and Chapter 6 share a similar objective of balancing privacy and usability. However, the two chapters use different networks, with the method in Chapter 5 proving more effective when dealing with low-resolution images and the approach in Chapter 6 being more effective for high-resolution images.

In summary, these chapters contribute to addressing existing challenges in image privacy. The specific details can be found in Fig. 1.1. This thesis is organized as follows:

- Chapter 1 introduces the motivation and contribution of the thesis.
- Chapter 2 reviews the related works for image privacy protection and the background knowledge for the related AI technics.
- Chapter 3-Chapter 6 give four approaches in targeting the thesis objectives. Fig. 1.1 shows the detail of the approaches path map.
- Chapter 7 summarize the whole thesis.

CHAPTER 2

PRELIMINARY AND RELATED WORK

In this chapter, we will present the privacy definition provided by GDPR [15] and discuss various existing methods for protecting image privacy. These methods include traditional methods, adversarial examples, GAN-based inpainting, differential privacy, and diffusion models.

2.1 Privacy Definition Based on General Data Protection Regulation (i.e. GDPR)

Recently, our human society has witnessed a rapid increase in the use of image data, which poses high risks of privacy leakage, as images usually contain a large number of sensitive data that might visually reveal personal information [16]. For example, heavy concerns on the privacy risks were raised on uploading people's face pictures to a popular APP called FaceApp [17], which can edit a person's face by changing his/her gender, age, ethnicity, etc. In fact, the instinct to protect privacy in image data is not new in our human society.

In this light, privacy issues have become hot topics in government debates and legislation, as many countries have launched privacy acts and laws. For example, the European General Data Protection Regulation (GDPR) [15] took effect on 25 May 2018. It emphasizes the protection of "personal data", interpreted as "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person". According to this definition, images contain a variety of personal identifiers, such as people's faces, text, and license plates. Therefore, effective privacy protection methods for image data are in urgent need.

2.2 Preliminary

This section presents a concise introduction to Adversarial Examples, Generative Adversarial Networks (GANs), and Diffusion Models. As these techniques will be employed for privacy protection in our subsequent research, it is pertinent to include this background knowledge within this chapter.

2.2.1 Adversarial examples

This section aims to provide an overview of a seminal technique in machine learning called adversarial examples (AEs) [1]. The section is organized as follows: Sect. 2.2.1.1 offers a succinct introduction to the fundamental concept of adversarial examples. Sect. 2.2.1.2 introduces the algorithm of one of the most acceptable methods used in image privacy protection, Fast Gradient Sign Method (FGSM). Sect. 2.2.1.3 shows the advanced protection scheme using AEs.

2.2.1.1 What is an adversarial example?

An adversarial example [1] refers to a specially crafted input data sample that has been intentionally designed to deceive a machine learning model. It is created by making slight modifications to the original input data in order to cause the model to misclassify or produce an incorrect output. These modifications are often imperceptible to human observers but can have a significant impact on the model's behaviour.

The goal of generating adversarial examples is to exploit vulnerabilities or weaknesses in the machine learning model's decision-making process. By introducing carefully calculated perturbations, an adversary can manipulate the model's predictions, potentially leading to erroneous results or compromising the model's performance.

Adversarial examples have garnered significant attention as they highlight the vulnerabilities of machine learning models and raise concerns regarding their robustness and reliability. Understanding and mitigating the impact of adversarial examples is crucial for improving the security and trustworthiness of machine learning systems in various domains, including computer vision, natural language processing, and autonomous systems. One typical example is shown in Fig. 2.1.

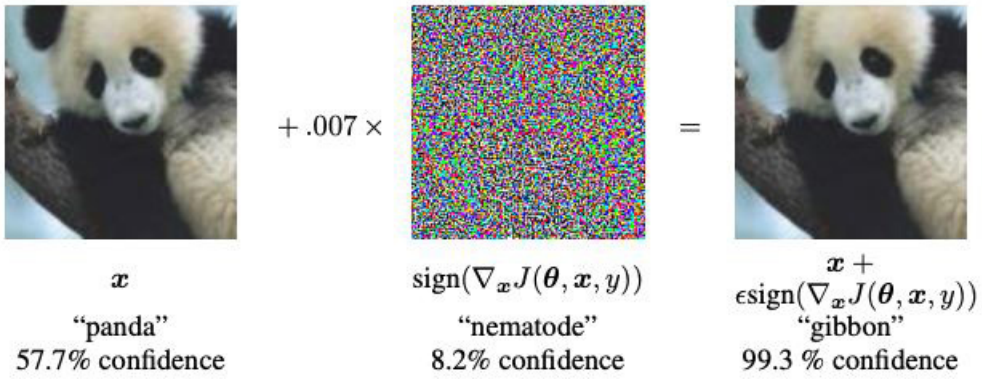


Figure 2.1: An adversarial example. [1]

2.2.1.2 Fast Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM) [1] is a popular algorithm used to generate adversarial examples in the field of machine learning. It is a simple yet effective technique for crafting adversarial perturbations.

The FGSM algorithm operates by taking advantage of the gradient information of the loss function with respect to the input data. It perturbs the input data by adding a small perturbation in the direction of the sign of the gradient. The magnitude of the perturbation is determined by a hyperparameter called the epsilon value, denoted as ϵ. The larger the value of ϵ, the more noticeable the perturbation becomes.

Mathematically, given an input sample $\mathbf{x}$, a target class y, and a loss J, the FGSM algorithm generates an adversarial example $\mathbf{x}_{adv}$ by computing the perturbation as follows:

$$\mathbf{x}_{adv} = \mathbf{x} + \epsilon * \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)), \tag{2.1}$$

where J denotes the loss function, and $\nabla_{\mathbf{x}}$ represents the gradient of the loss function with respect to the input $\mathbf{x}$. The sign function $\text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y))$ extracts the sign of the gradient to determine the direction in which the perturbation should be added.

By leveraging the Fast Gradient Sign Method (FGSM), we can strategically generate adversarial examples that exploit the decision boundaries of the model. This approach allows us to induce potential misclassifications and generate privacy-preserving outputs.

2.2.1.3 Advanced protection scheme

In this subsection, based on the section 2.2.1.2 above, we will provide a detailed explanation of the most accepted and advanced methodology employed for privacy protection using adversarial examples.

Algorithm 1: Iterative class obfuscation scheme (COS).

Data: $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$.
Result: $\mathbf{x}^{pr} \in \mathbb{R}^{h \times w \times c}$.
$\delta \mathbf{x}_0 \leftarrow 0$;
$\mathbf{x}_0^{pr} \leftarrow \mathbf{x}$;
$n \leftarrow 0$;
while $n \leq N$ **do**
  $\delta \mathbf{x}_n = -\epsilon \text{sign}(\nabla_{\mathbf{x}_n^{pr}} \mathcal{L}_{COS})$; /* Loss maximum */
  $\mathbf{x}_{n+1}^{pr} = \delta \mathbf{x}_n + \mathbf{x}_n^{pr}$;
  $n += 1$;
end
$\mathbf{x}^{pr} = \mathbf{x}_n^{pr}$.

Algorithm 2: Iterative class replacement scheme (CRS).

Data: $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$, $\bar{\mathbf{x}} \in \mathbb{R}^{h \times w \times c}$.
Result: $\mathbf{x}^{pr} \in \mathbb{R}^{h \times w \times c}$.
$\delta \mathbf{x}_0 \leftarrow 0$;
$\mathbf{x}_0^{pr} \leftarrow \mathbf{x}$;
$n \leftarrow 0$;
while $n \leq N$ **do**
  $\delta \mathbf{x}_n = \epsilon \text{sign}(\nabla_{\mathbf{x}_n^{pr}} \mathcal{L}_{CRS})$; /* Loss minimum */
  $\mathbf{x}_{n+1}^{pr} = \delta \mathbf{x}_n + \mathbf{x}_n^{pr}$;
  $n += 1$;
end
$\mathbf{x}^{pr} = \mathbf{x}_n^{pr}$.

Two distinct approaches are commonly employed in the context of utilizing adversarial examples for privacy protection. The first approach aims to obfuscate the privacy category that necessitates protection, effectively concealing its true nature as a privacy category. Conversely, the second approach involves replacing the privacy category with an alternative label, resulting in its misclassification as the desired category. For the sake of brevity, we shall refer to the former approach as the class obfuscation scheme (COS) and the latter approach as the class replacement scheme (CRS). Detailed descriptions of the algorithms corresponding to each method can be found in Alg. 1 and Alg. 2, respectively.

The algorithmic flow for the iterative methods are presented in Alg. 1 and Alg. 2. In these algorithms, the variable $\mathbf{x} \in \mathbb{R}^{h \times w \times c}$ denotes an image from the image dataset, containing sensitive information that requires protection. $\mathbf{x}^{pr} \in \mathbb{R}^{h \times w \times c}$ represents the image with the sensitive information protected. The parameter n signifies the current

iteration number, while N represents the maximum number of iterations allowed. The magnitude of noise added in each iteration is denoted by ϵ. The sign function is utilized to determine the sign of its argument and returns an integer value accordingly. The sign function syntax is defined as sign$(\cdot)$, where $\cdot$ can be any valid numerical expression. The return value of the sign function is 1 if the $\cdot$ is greater than 0, 0 if it equals 0, and -1 if it is less than 0. The operator ∇ corresponds to the gradient operator. Additionally, the loss functions for the two algorithms are denoted as $\mathcal{L}_{COS}$ and $\mathcal{L}_{CRS}$ respectively, with their mathematical expressions as follows:

$$\mathcal{L}_{COS} = J(\theta, \mathbf{x}^{pr}, y_{\mathbf{x}}); \mathcal{L}_{CRS} = J(\theta, \mathbf{x}^{pr}, y_{\bar{\mathbf{x}}}) \tag{2.2}$$

Let θ denote the model parameters. The target class associated with $\mathbf{x}$ is represented as $y_{\mathbf{x}}$, while $y_{\bar{\mathbf{x}}}$ corresponds to the target class associated with $\bar{\mathbf{x}}$. Here, $\bar{\mathbf{x}} \in \mathbb{R}^{h \times w \times c}$ denotes the image used for replacement in the CRS scheme. The loss function used in the neural network is denoted as $J(\theta, \mathbf{x}^{pr}, y)$.

2.2.2 Generative Adversarial Networks (GANs)

This section aims to provide an overview of generative adversarial networks (GANs) and their unique designs.

Generative adversarial networks (GANs) [18] is a type of deep learning model that consists of two components: a generator and a discriminator. GANs are designed to generate realistic data samples that are similar to a given training dataset.

The generator is responsible for generating new samples by transforming random noise or input data into output samples that resemble the training data. It learns to capture the underlying patterns and distributions of the training data in order to generate realistic samples.

The discriminator, on the other hand, acts as a binary classifier that distinguishes between the generated samples from the generator and the real samples from the training dataset. It learns to differentiate between real and fake samples by optimizing its classification accuracy.

During training, the generator and discriminator are pitted against each other in a two-player minimax game. The generator aims to generate samples that can fool the discriminator into classifying them as real, while the discriminator aims to classify the real and generated samples accurately:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x})} \left[\log D(\mathbf{x}) \right] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} \left[\log(1 - D(G(\mathbf{z}))) \right]. \tag{2.3}$$

Where D and G represent the discriminator and generator, respectively. Through this adversarial process, both the generator and discriminator improve their performance iteratively.

The ultimate goal of GANs is to achieve a state where the generator can produce high-quality samples that are indistinguishable from real data while the discriminator struggles to differentiate between the generated and real samples. GANs have demonstrated impressive capabilities in generating realistic images, audio, text, and other types of data.

Besides generating realistic samples, GANs have also been used for various applications, such as data augmentation, image and video synthesis, style transfer, anomaly detection, and privacy protection. GANs offer a powerful framework for capturing complex data distributions and have the potential to revolutionize several fields that require realistic data generation and manipulation.

2.2.2.1 Conditional Generative Adversarial Nets (cGANs)

Conditional generative adversarial networks (cGANs) [2] is an extension of the original generative adversarial networks (GANs) [18] that incorporate additional conditioning information during the training and generation process. The structure of a simple conditional adversarial network is displayed in Fig. 2.2

In cGANs, both the generator and discriminator receive additional input in the form of conditional information, typically in the form of extra data or labels. This conditioning information guides the generation process, allowing the generator to generate samples that correspond to specific conditions or classes.

The generator takes random noise as input along with the conditional information and generates samples that match the given condition. For example, in image generation, the conditional information could be a class label, and the generator can produce samples representing different class objects.

The discriminator, which also receives conditional information, tries to differentiate between the real and generated samples while considering the given condition. It learns to classify the samples as real or fake, taking into account the conditioning information.

The training of cGANs involves a game between the generator and discriminator, similar to traditional GANs. The generator aims to produce samples that not only fool the discriminator but also adhere to the given condition, while the discriminator aims to

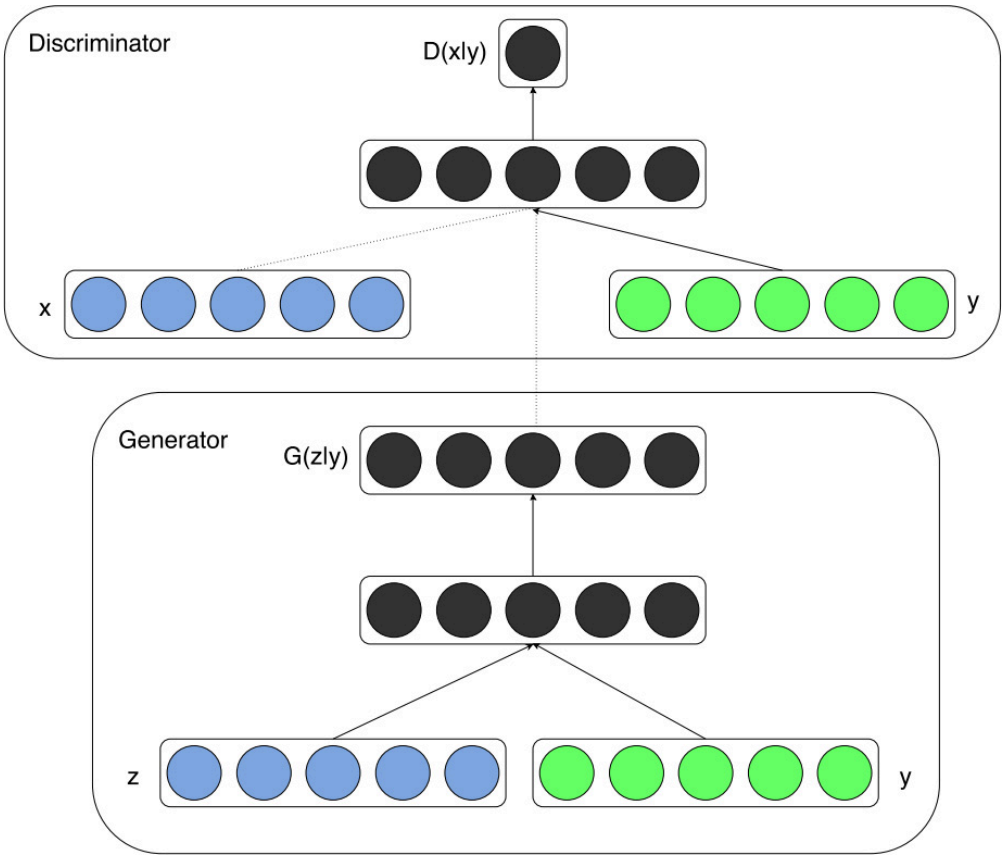


Figure 2.2: The structure of a conditional adversarial network [2].

classify the real and generated samples accurately:

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\log D(\mathbf{x}\mid\mathbf{y})\right] + \mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z})}\left[\log(1-D(G(\mathbf{z}\mid\mathbf{y})))\right]. \tag{2.4}$$

The conditional nature of cGANs allows for more controlled and targeted generation. It enables tasks such as image-to-image translation, where the generator takes an input image from one domain and generates a corresponding output image in another domain while maintaining the desired characteristics or conditions.

cGANs have been successfully applied in various domains, including image synthesis, text-to-image generation, image inpainting, and style transfer. They provide a powerful framework for conditional data generation, allowing for fine-grained control and manipulation of generated samples based on the provided conditions or constraints.

Building upon the cGAN framework, two notable methods, namely Conditional Identity Anonymization Generative Adversarial Network (CIAGAN) [11] and DeepPrivacy [10], introduced the integration of autoencoder within the feature space of images. CIAGAN demonstrated the ability to anonymize faces and bodies, generating high-quality images and videos. On the other hand, DeepPrivacy took into account factors

such as pose and background to generate images with improved realism and privacy preservation.

2.2.3 Style-Based Generative Adversarial Networks (StyleGAN)

Style-based generative adversarial networks (StyleGAN) is a type of generative adversarial network (GAN) architecture that focuses on generating high-quality and realistic images with enhanced control over the generated styles and details.

StyleGAN introduces a novel approach to disentangling the latent space representation of the generator. Unlike traditional GANs where the latent vector directly controls the generated image, StyleGAN incorporates style vectors that control different aspects of the image generation process, such as the global styles, object styles, and details.

The architecture of StyleGAN consists of two main components: the generator and the discriminator. The generator starts with a random noise vector as input and transforms it through several intermediate layers. In each intermediate layer, the generator combines the learned styles from style vectors with the input noise vector to produce feature maps that capture different levels of image details. Finally, the feature maps are transformed into the final synthesized image.

The detailed structure of the generator is shown in Fig. 2.3. Instead of directly feeding the latent code through the input layer, StyleGAN involves mapping the input to an intermediate latent space $\mathcal{W}$. This intermediate space controls the generator through adaptive instance normalization (AdaIN) at each convolution layer.

$$\text{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i}, \tag{2.5}$$

where $\mathbf{x}_i$ represents the feature maps and $\mathbf{y}$ represents the style to be transferred.

Additionally, StyleGAN incorporates Gaussian noise after each convolution and before applying nonlinearity. The transformation of the input is performed using a learned affine transform denoted as “A”, and per-channel scaling factors are applied to the noise input denoted as “B”. The mapping network f consists of 8 layers, while the synthesis network g consists of 18 layers, with two layers dedicated to each resolution ($4^2 - 1024^2$). The output of the final layer is converted to RGB using a separate 1×1 convolution.

The discriminator, as in other GANs, tries to distinguish between real and generated images. However, StyleGAN is designed to provide feedback at multiple resolutions, allowing for better control over the generated details.

One notable feature of StyleGAN is its ability to generate images with varied styles and qualities by manipulating the style vectors. By modifying the style vectors, users can

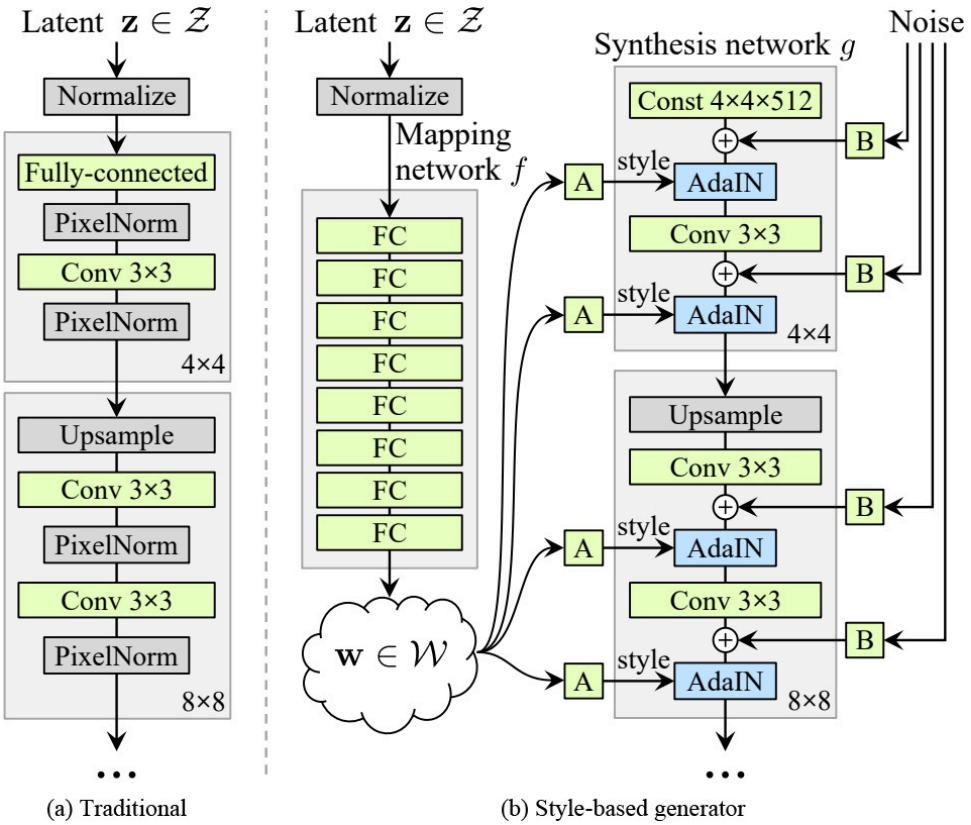


Figure 2.3: The generator structure of a Style-based generative adversarial network [3].

control attributes like facial expressions, hairstyles, or lighting conditions in generated images, giving them a high degree of control over the generated output.

StyleGAN has been widely used in various applications, including image synthesis, face generation, and image editing. Its ability to generate high-quality images with fine-grained control over styles and details has made it popular among researchers and artists alike.

2.2.4 StyleGAN-based autoencoder

Expanding on the capabilities of StyleGAN, two recent studies [4, 19] have focused on the development of an encoder specifically tailored for image manipulation within the StyleGAN framework. These studies aim to enable precise control and manipulation of images by designing an encoder that accurately captures the latent representations associated with different visual attributes and styles present in the images.

The proposed encoders in both studies employ a combination of adversarial training and perceptual loss for training. Adversarial training aligns the encoded latent codes

with the distribution of latent codes generated by the pre-trained StyleGAN. Meanwhile, the perceptual loss ensures that the encoded representations preserve the visual quality and attributes of the original images. By successfully encoding real images into the latent space of StyleGAN, the developed encoders facilitate various image manipulation tasks. Users can manipulate specific attributes, such as age, expression, or hairstyle, by modifying the corresponding latent codes. The modified latent codes are then fed into the StyleGAN generator to generate the desired manipulated images.

Since the two studies share a similar basic framework, we will provide a brief overview of the main contributions and foundational structure of one of them, known as *pixel2style2pixel* (pSp) [4].

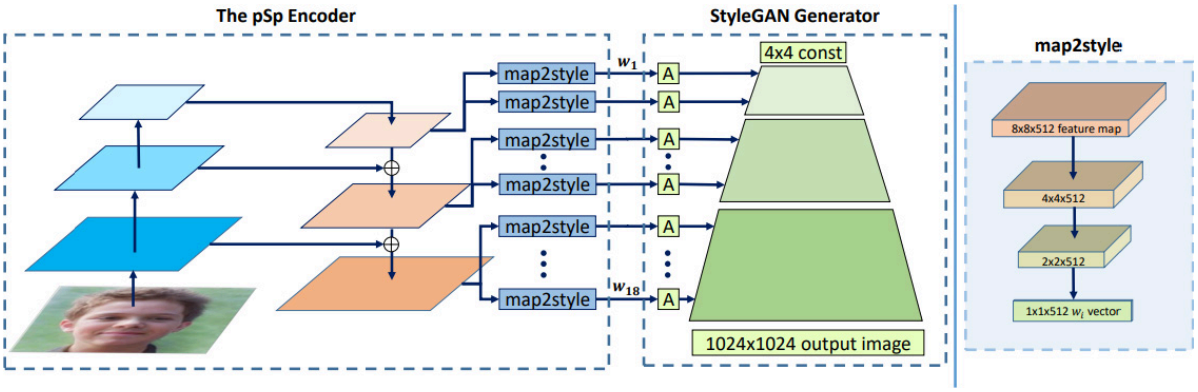


Figure 2.4: The structure of a StyleGAN-based autoencoder [4].

Fig. 2.4 illustrates the architecture of the pSp framework. The authors leverage the representational power of the pre-trained StyleGAN generator and the $\mathcal{W}+$ latent space to achieve their objectives. To accomplish this, they adopt a Feature Pyramid Network [20] as the underlying structure, generating three levels of feature maps that capture different levels of detail: coarse, medium, and fine. These feature maps are then processed by an intermediate network, referred to as *map2style* (depicted in Fig. 2.4), to extract the corresponding styles. In line with the hierarchical representation, the aligned styles are subsequently fed into the generator according to their respective scales, effectively completing the transformation from input pixels to output pixels while leveraging the intermediate style representation. This encoder facilitates the direct reconstruction of real input images, enabling manipulations in the latent space without requiring computationally intensive optimization. Therefore, GAN-based autoencoders are a crucial branch that cannot be disregarded, particularly in the realm of privacy protection. A recent paper [12] proposes a novel approach to de-identify individuals in facial image datasets while preserving the facial attributes. Rather than training custom

neural networks from scratch, they operate directly in the latent space of a pre-trained StyleGAN.

2.2.5 Diffusion model

According to the research publication "Denoising Diffusion Probabilistic Models," [5] the diffusion models are defined as: "A diffusion model or probabilistic diffusion model is a parameterized Markov chain trained using variational inference to produce samples matching the data after finite time"

In diffusion models for images, the goal is to model the conditional distribution of pixel values given the previous values in the image. This is typically achieved by formulating the evolution of the pixel values as a diffusion process, where each pixel is updated based on the neighbouring pixels and a noise term.

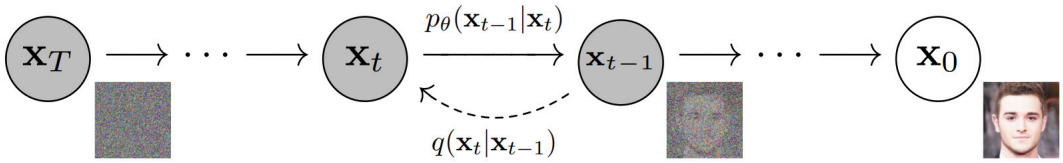


Figure 2.5: The graphical model of DDPMs [5].

Unlike other generative models such as GANs and most traditional-style VAEs that encode input data into a low-dimensional space, diffusion models maintain a latent space of the same size as the input. In the forward process, DDPMs progressively add noise to the image until it is completely degraded into pure Gaussian noise. Assuming an ideal forward process, where a real input image $\mathbf{x}_0$ undergoes T rounds of Gaussian noise addition, resulting in a purely Gaussian noise image $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Consequently, each step of noise addition can be formally expressed as the following probability function:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) \sim \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{2.6}$$

where β_t is the coefficient associated with the noise.

As a result, the cumulative noise in the image $\mathbf{x}_0$ after t processing steps can be represented as another Gaussian noise:

$$q(\mathbf{x}_t|\mathbf{x}_0) \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I}), \tag{2.7}$$

where

$$\alpha_t = \prod_{i=1}^{t}(1-\beta_i). \tag{2.8}$$

In the reverse process, i.e., learning the distribution $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$, the noise is gradually removed to generate a realistic image. To train a DDPM network, Ho et al. [21] give the distribution of $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \sim \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t)$, and propose a learnable function $\epsilon_\theta(\mathbf{x}_t, t)$ to predict the noise added in each step. Despite requiring a large number of noise injection and denoising steps to generate samples, DDPMs exhibit superior image fidelity and diversity compared to other types of generative models.

The training of diffusion models involves maximizing the likelihood of observed images. This is typically done by iteratively applying the diffusion process to a given image and learning the parameters of the diffusion model through gradient-based optimization. The training involves sampling from the diffusion process and applying an inverse transform to map samples from the observed data space to a latent space. This latent space is often modelled using an autoregressive neural network. The detailed training algorithms are shown in Alg. 3 and Alg. 4.

Algorithm 3: DDPMs training

Data: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
$t \sim \text{Uniform}(\{1, \dots, T\})$
$\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
Take gradient descent step on
$\nabla_\theta \left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t) \right\|^2$
until converged

Algorithm 4: DDPMs sampling

Data: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
  $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)) + \sigma_t \mathbf{z}$
end
return $\mathbf{x}_0$

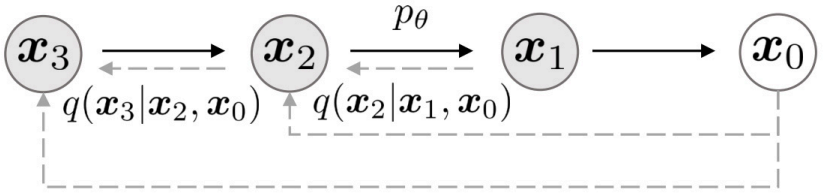


Figure 2.6: The graphical model of DDIM [6].

Diffusion models have a notable drawback: they rely on a lengthy sequence of diffusion steps to produce desired outcomes, resulting in slow generation speed. However, Denoising Diffusion Implicit Models (DDIMs) [6] address this limitation by relaxing the requirement for the diffusion process to adhere to a Markov chain, distinguishing them from DDPMs.

Let's briefly explore the principles of DDIM. The derivation in DDPM is closely related to DDIM, and it can be summarized as follows:

$$p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \rightarrow p(\mathbf{x}_t \mid \mathbf{x}_0) \rightarrow p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_0) \approx p(\mathbf{x}_{t-1} \mid \mathbf{x}_t). \tag{2.9}$$

This progression occurs step by step. However, the final result exhibits two key characteristics:

- The loss function depends solely on $p(\mathbf{x}_t \mid \mathbf{x}_0)$.
- The sampling process depends solely on $p(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$.

One significant insight from this is that DDPM, as a latent variable model, allows for various choices of inference distributions. As long as the inference distribution satisfies the marginal distribution condition (representing the characteristics of the diffusion process), these inference processes do not necessarily need to be Markov chains. Consequently, DDIM eliminates $p(\mathbf{x}_t \mid \mathbf{x}_{t-1})$ from the entire derivation process. In the DDIM paper, the inference distribution is defined as follows:

$$q_\sigma(\mathbf{x}_{1:T}|\mathbf{x}_0) = q_\sigma(\mathbf{x}_T|\mathbf{x}_0)\prod_{t=2}^{T} q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0). \tag{2.10}$$

It is necessary for $q_\sigma(\mathbf{x}_T|\mathbf{x}_0)$ to satisfy $\mathcal{N}(\sqrt{\alpha_T}\mathbf{x}_0,(1-\alpha_T)\mathbf{I})$, and for all $t \geq 2$:

$$q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1};\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\frac{\mathbf{x}_t-\sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}},\sigma_t^2\mathbf{I}). \tag{2.11}$$

In this case, the variance $\sigma_t^2 \in \mathbb{R}$. Additionally, the mean of the distribution $q_\sigma(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)$ is defined as a composite function depending on $\mathbf{x}_0$ and $\mathbf{x}_t$. Hence, it holds for all t:

$$q_\sigma\left(\mathbf{x}_t|\mathbf{x}_0\right) = \mathcal{N}\left(\mathbf{x}_t;\sqrt{\alpha_t}\mathbf{x}_0,\left(1-\alpha_t\right)\mathbf{I}\right). \tag{2.12}$$

Based on the derived inference distribution, DDIM can achieve the same optimization objective as DDPM without requiring a forward process. Consequently, during the generation phase, $\mathbf{x}_{t-1}$ can be obtained as follows:

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t-\sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t,t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta(\mathbf{x}_t,t) + \sigma_t\epsilon_t. \tag{2.13}$$

The paper further defines σ_t^2 as:

$$\sigma_t^2 = \eta\cdot\tilde{\beta}_t = \eta\cdot\sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{(1-\alpha_t/\alpha_{t-1})}. \tag{2.14}$$

If $\eta = 1$, then $\sigma_t^2 = \tilde{\beta}_t$, resulting in the generation process being the same as DDPM. Another case is when $\eta = 0$. In this situation, the generation process becomes deterministic without any random noise. The paper refers to this type of model as DDIM. Once the initial random noise $\mathbf{x}_T$ is determined, the sample generation process in DDIM becomes deterministic. Furthermore, DDIM does not explicitly define a forward process, which means that DDIM can employ fewer sampling steps, effectively expediting the generation process. This streamlined approach further accelerates the relatively computationally demanding sampling procedure inherent to DDPMs.

2.3 Related Work

In this section, we will provide an overview of the related work in image privacy protection. The discussion encompasses various approaches, including traditional methods for image privacy protection, protection methods employing adversarial perturbations, GAN-based protection methods, and the application of differential privacy. Additionally, we will explore the utilization of diffusion models for image editing, which will be employed to enhance image privacy protection.

2.3.1 Traditional protection methods

Traditionally, image privacy protection methods have focused on human adversaries and employed techniques such as blurring, pixelation, and mosaic to obfuscate sensitive information.

One prevalent technique is face blurring [22–27], which employs blurring algorithms or filters to obscure the facial features of individuals in an image. By intentionally degrading the details of the faces, it becomes challenging to identify specific individuals while still retaining the overall structure of the image.

Addressing the contemporary privacy concerns arising from the widespread deployment of surveillance cameras, a recent paper [22] introduces a novel requirement for protected images, emphasizing the need for them to retain useful information beyond privacy considerations. The authors present an efficient privacy protection scheme tailored for video surveillance systems. In response to the identified concerns, they propose a privacy-sensitive region detection algorithm capable of identifying not only human faces and bodies but also personal accessories that may potentially reveal identifiable information.

To mitigate privacy risks, the authors further introduce a silhouette-blur-guided privacy protection algorithm. This algorithm effectively obfuscates the silhouettes of sensitive objects while concurrently reducing high spatial frequencies, ensuring anonymity for the general public and preserving valuable behavioural information. The paper underscores the significance of automatic methods in de-identifying individuals present in surveillance videos, striking a balance between privacy preservation and the retention of visual information necessary for human behaviour analysis.

Another approach is pixelation [28–32], where the facial region is replaced with large pixels, concealing facial details. Recently, an intriguing article [33] has emerged, showcasing an innovative method that integrates differential privacy (DP) [34] with

pixelation to safeguard privacy. This article primarily centres on tackling privacy issues associated with the sharing of image data. It introduces a pioneering approach that combines pixelization with differential privacy to address these concerns effectively.

Nevertheless, these methods achieve privacy protection at the cost of compromising facial attributes and image quality. Furthermore, recent studies [35, 36] have indicated that the de-identification effects of blurring and pixelation, while perceptually effective to human observers, can be circumvented by deep learning algorithms, thereby diminishing their efficacy in preserving privacy.

The rapid advancement of technology has rendered conventional privacy protection methods increasingly inadequate in the face of the capabilities exhibited by neural networks. This evolutionary shift has stimulated a heightened demand for the development of efficient privacy preservation techniques, thereby presenting a formidable challenge to researchers in this domain. Consequently, the imperative to devise innovative approaches that can effectively mitigate the emerging challenges posed by neural networks underscores the profound significance of our research endeavours.

2.3.2 Adversarial examples

In recent years, the quest for privacy preservation in the context of deep neural networks (DNNs) has led to the discovery of a technique known as adversarial samples. This section explores the applications of adversarial examples (AEs) as a means to safeguard the privacy of images. Extensive research [37–46] in this domain highlights the importance of utilizing AEs in achieving image privacy protection objectives. The primary concept revolves around generating an adversarial example that represents the image to be shared, thereby preserving privacy.

These approaches propose unique frameworks and utilize AEs to counter Deep Neural Network (DNN) models. For instance, Fawkes [37] demonstrate the effectiveness of AEs in altering images' feature space to provide high protection against user recognition. Oh *et al.* [38] introduce a game theoretic framework for person obfuscation using adversarial image perturbation. Rajabi *et al.* [39] propose two novel schemes, Universal Ensemble Perturbation (UEP) and k-Randomized Transparent Image Overlays (k-RTIO), to protect against automated face recognition. Xue *et al.* [40] develop a framework that adds imperceptible perturbations to sensitive image areas to hide private information from AI while remaining visible to humans. Hu *et al.* [41] propose the Adversarial Makeup Transfer GAN (AMT-GAN) to generate natural-looking adversarial face images with strong attack ability. Liu *et al.* [44] introduce the Stealth algorithm, which generates

adversarial examples to protect privacy by rendering automatic detection systems unable to identify objects in images. Additionally, Liu *et al.* [43, 45] propose an architecture using AEs to safeguard image privacy against deep learning image classification tools. Li *et al.* [46] propose the AnonymousNet framework, leveraging deep learning techniques for facial semantic extraction, attribute selection, photo-realistic image generation, and adversarial perturbation to enhance privacy protection in the context of image obfuscation.

These methods can be broadly categorized as Noise-based Adversarial Examples (NAE). They incorporate constraints on the range of noise added to images to minimize any adverse effects on image usability. Consequently, these methods excel in effectively reducing the chances of human perception, thereby preserving essential information within the protected images. Furthermore, with an increasing number of noise iterations, these methods possess an expansive search space, enabling them to effectively counter deep neural networks, particularly in the context of white-box scenarios, where the success rate of protection exceeds 90%. However, obtaining a white-box model in real-world scenarios is often challenging, thereby imposing limitations on the practical application of these methods. When confronted with black-box neural networks, some approaches [37] resort to increasing the magnitude of noise to enhance the success rate of protection. Nevertheless, even with compromised noise concealment, the success rate of protection still falls short compared to other techniques like image inpainting.

Consequently, certain studies have discarded the limitations on the noise norm in AEs and instead leveraged the underlying principles to explore the AE in the semantic space of images. These approaches are commonly known as Unrestricted Adversarial Examples (UAE). Initially, UAE, like other AEs, was predominantly employed to launch attacks on classification neural networks [47–49], causing misclassifications. However, researchers found the potential of the UAE in the realm of image privacy protection and started utilizing them for anonymizing facial images.

A paper [50] introduces a protection method that utilizes UAE to add sunglasses to faces in images. Similarly, another paper [51] employs the same approach to add hats to faces in images. An improvement upon these methods is presented in another paper [52], which uses UAE to add random patterns to protect image privacy. These three methods focus on adding noise to specific regions in the images in a semantically reasonable manner. However, such methods may affect the transferability of protection effectiveness across different models. Moreover, these significant modifications also impact the user experience to some extent.

Subsequently, with the emergence of generative models [53, 54], more image protection methods have been proposed by combining UAE and generative models [55–58]. Among them, two papers [57, 58] introduce a novel approach where faces in images are disguised with makeup to deceive deep neural networks used for facial recognition. This protection method, which employs makeup as a noise, demonstrates the use of UAE to search for adversarial examples in semantically understandable ways, aligning with human cognition of facial privacy protection. This greatly enhances the user experience.

2.3.3 GAN-based inpainting

The concept of GAN-based inpainting was introduced as a means to generate content that effectively conceals sensitive information or the identity of an image while preserving the quality of the original image [59].

In recent years, the remarkable image generation capabilities of GANs [18] have spurred their adoption as a novel tool for image editing [60–77]. Notably, numerous methodologies [60, 66, 71, 73] have embraced autoencoder architectures, wherein the GAN's generator serves as the decoder component, enabling the production of high-fidelity images. The process involves encoding the input image to extract semantic features and decoding through the GAN to reconstruct the output image. Crucially, these approaches integrate semantic continuity constraints during GAN training, facilitating seamless manipulation of diverse image attributes within the GAN's semantic space.

Two notable papers in the field of privacy preservation for facial images are DeepPrivacy [10] and CIAgan [11]. DeepPrivacy [10] introduces a novel generator architecture based on conditional generative adversarial networks (cGAN) [2] specifically designed for face anonymization. It considers the existing background and sparse pose annotations to generate realistic anonymized faces while ensuring the removal of privacy-sensitive information from the original face. The model is trained using a progressive growing training technique that gradually increases the resolution from 8×8 to 128×128, resulting in improved image quality and reduced training time.

On the other hand, CIAgan [11] proposes a model for anonymizing images and videos by removing identification characteristics while preserving necessary features for computer vision tasks such as detection, recognition, and tracking. The model also leverages cGAN [2] to generate realistic anonymized images and videos. Notably, it introduces an identity control discriminator to ensure controlled identity changes during anonymization.

Both approaches contribute to privacy preservation by offering methods that hide individuals' identities while producing visually convincing and realistic anonymized images. The works address the challenges of privacy-sensitive information removal and the generation of new identities in the context of face anonymization. Additionally, they emphasize the importance of maintaining realism and temporal consistency in the generated outputs for effective utilization in detection and recognition systems.

2.3.4 Differential privacy

Differential privacy [78, 79] is a concept and framework for ensuring the privacy of individuals when analyzing or processing sensitive data. It provides a mathematical definition and a set of techniques to quantify and control the privacy risk associated with sharing or releasing data.

At its core, differential privacy aims to protect the privacy of individuals by introducing randomness into the data analysis process. The main idea is to add controlled noise or perturbation to the data or the results of computations to prevent identifying specific individuals or leaking sensitive information.

The concept of differential privacy guarantees that the presence or absence of any individual's data has a limited effect on the output or conclusions drawn from the data analysis. In other words, even if an adversary has access to the output or aggregated results, they cannot accurately determine whether a specific individual's data was included in the analysis.

Differential privacy provides a rigorous and quantifiable privacy guarantee by defining a privacy parameter, often denoted as ϵ (epsilon), representing the maximum allowable privacy loss. A smaller value of ϵ indicates a higher level of privacy protection. By controlling the amount of noise added to the data or computations based on ϵ, the privacy of individuals can be preserved while still providing meaningful and accurate analysis results.

Several initial studies have applied differential privacy (DP) in images. Fan proposed ϵ-differential private methods in the pixel level of the image [80], and in Singular Value Decomposition (SVD) [81], respectively. Lecuyer et al. [82] proposed the PixelDP framework, including a DP noise layer in the DNN.

A more detailed explanation of DP will be introduced in Chapter 4.

2.3.5 Diffusion model

Diffusion models have gained attention due to their ability to generate high-quality and diverse images, surpassing the limitations of traditional generative models. They have been used to generate realistic images, create artistic visual effects, and explore the latent space of image data.

In the realm of guided image generation tasks such as stroke painting, stroke-based editing, and image composition, diffusion models have been employed by Meng *et al.* [83]. These tasks involve starting with a guided image where certain properties (e.g., shapes and colours) are preserved while smoothing out deformations through the progressive addition of noise (forward process of the diffusion model). Subsequently, the resulting image is denoised (reverse process) to generate a realistic image based on the given guidance. The authors utilize a generic diffusion model to synthesize images by solving the reverse stochastic differential equation (SDE), eliminating the need for customized datasets or training modifications. Nichol *et al.* [84] train a diffusion model conditioned on text descriptions and explore the effectiveness of classifier-free and CLIP-based guidance for image inpainting. They find that the former option yields better results and additionally fine-tune the model specifically for image inpainting, enabling image modifications based on text input. Avrahami *et al.* [85] employing latent diffusion models for local image editing with text. An adaptive-to-time mask (representing the region to edit) and the image are encoded using a variational autoencoder (VAE) into the latent space, where the diffusion process occurs. Each sample is iteratively denoised while guided by the text within the region of interest. Inspired by Blended Diffusion [86], the image is combined with the masked region in the latent space, which is noised at the current time step. Finally, the sample is decoded using the VAE to generate the new image. This method demonstrates superior performance while maintaining comparable speed. In an inpainting method presented by Lugmay *et al.* [87], the authors adopt an agnostic approach to mask form. They utilize an unconditional diffusion model for this task but modify its reverse process. The image at step $t-1$ is generated by sampling the known region from the masked image, while the unknown region is obtained by applying denoising to the image generated at step t. Through this procedure, the authors observe that the unknown region possesses the correct structure but may be semantically incorrect. To address this, they repeat the proposed step multiple times and, at each iteration, replace the previous image from step t with a new sample generated from the denoised version obtained at step $t-1$. The first approach for editing specific regions of images based on natural language descriptions is introduced in [86]. Users can specify

the regions to be modified using masks. This method uses CLIP guidance to generate an image according to the input text. However, the authors observed that combining the output with the original image at the end does not yield globally coherent images. To address this, they modify the denoising process by applying the mask to the latent image after each step while incorporating the noisy version of the original image.

A recent study by Konpat *et al.* [88] introduced a novel approach called the Diffusion Autoencoders. The authors' main objective is to utilise a trainable encoder to uncover intricate semantics, complemented by the adoption of DPM as a decoder to encapsulate the residual stochastic variations. The proposed methodology demonstrates the remarkable capability to encode diverse images into a latent code, consisting of two distinctive components: the first component, characterized by its semantic significance, exhibits a linear structure, while the second component proficiently captures fortuitous intricacies, thereby facilitating an exceptionally precise reconstruction. Consequently, this proficiency endows the system with the capacity to effectively address challenging applications, such as attribute manipulation in real images, a task that has proven to be elusive for conventional GAN-based approaches.

The intersection of image privacy and diffusion models presents a promising area for exploration as a novel model. There are still many aspects that need to be explored. For example, a recent article by Xiao *et al.* [89] discussed training a diffusion model for use as a facial privacy model. However, the model is constrained by the extensive training requirements of the diffusion model, leading to various limitations.

Moreover, the novel approach presented in this study provides a valuable tool for image privacy protection. By utilizing the Diffusion Autoencoders, sensitive information within images can be encoded and subsequently reconstructed with high fidelity, allowing for improved safeguarding of private visual content.

CHAPTER 3

Hiding Private Information in Images From AI

3.1 Preface

In this chapter, we explore the prevalent privacy concerns that pervade popular social media platforms, leaving individuals susceptible to privacy breaches. Our main focus is to address the specific privacy issues that arise from the uploading of images on prominent social networking sites like Facebook or Google Earth street view. Diverging from traditional protective measures, we recognize the convolutional neural network (CNN) as the primary threat in this context. As a result, we have developed a protective framework that relies on adversarial noise as the central safeguarding mechanism. This innovative approach ensures that the minimal noise introduced to the images remains undetectable by human observers, enhancing their overall usefulness while preserving the joy of sharing pictures. Through meticulous testing conducted on standard street view images, we have successfully validated the effectiveness of our approach, which exhibits exceptional capabilities surpassing those of conventional methods.

3.2 Introduction

3.2.1 Motivation

In such an era of data sharing, many people would like to share their life photos and videos on social software with friends or strangers. For example, Every 60 seconds on Facebook, 136,000 photos are uploaded [90]. However, people may not have noticed that these images and videos contain a large amount of private information [91–93] such as the faces, vehicle license plates, locations, email addresses, etc. If such information is used by adversaries, it may have a detrimental effect on the users [92]. Meanwhile, the newly emerging deep learning techniques further increase the privacy risks for online photo sharing. Artificial intelligence (AI) aided by deep learning methods can automatically collect and detect private and sensitive information from social networks. For example, DNNs can automatically search meaningful information in images and exploit an outcome to perform targeted advertisements [94]. DNNs can even extract user's private information, such as fingerprints [95], addresses, family members [96, 97], etc. This brings more risks to personal privacy, while the traditional privacy-preserving method seems powerless when facing large-scale deep learning tools. Therefore, the development of image privacy protection methods is in urgent need, especially when considering AI as the adversary.

Privacy protection for unstructured data such as images is much more complicated

compared with that for structured data. Traditional image privacy protection research assumes humans as an adversary. “Blurring”, “pixelation” and “mosaic” are the most commonly used methods. For example, Viola et al. [98] used a sliding window detector to identify and blur the license plates in Google Street View images. Researchers start to consider the case where AI acts as an adversary very recently. The fundamental idea is to generate a small but intentional worst-case disturbance to an original image, which misleads deep neural networks (DNNs) without causing a significant difference perceptible to human eyes. The perturbed image is called an “adversarial example” [99], and the specially generated noise is named adversarial perturbations (AP). A few papers have discussed the potential of AP in privacy protection in different applications, including image classification [100] and face recognition [101]. In [102], the authors proposed a novel stealth algorithm that makes all the objects invisible to DNNs in an image. These works cannot solve our problem thoroughly for several reasons: First, there are generally multiple private objects in the images, especially for social network images. Second, the revision of the images should be as small as possible and limited to private information to preserve the utility of the images.

3.2.2 Contributions

To overcome the above-mentioned problems, we proposed a framework for image private information protection in this chapter. It consists of three major steps: i) defining the privacy information in images, ii) identifying the private objects and their positions in images, and iii) image privacy protection using adversarial noise. Specifically, for image privacy protection, we propose to add adversarial perturbations to the sensitive parts of the images so that the private information can be hidden while the rest parts of the images are still visible to AI.

In summary, the contributions of this work are as follows:

- Developing an image privacy protection framework to hide private information from AI while the applied privacy protection is imperceptible to human eyes.

- Proposing an adversarial perturbation-based image privacy protection scheme such that it can hide multiple private objects in the image while having a minor impact on the non-private objects.

3.2.3 Overview of the work

The rest of the chapter is organized as follows. Section 3.3 discusses the system model and formulates the research problem. In Section 3.4, an AP-based image privacy protection scheme is proposed. Section 3.5 shows our experimental results. Finally, the results are concluded in Section 3.6.

3.3 System Model and Problem Formulation

In this section, we present the system model used in this work and formulate the research problem.

3.3.1 System Model

As shown in Fig. 3.1, our proposed image privacy protection framework consists of three major parts: object detection, image privacy definition, and image privacy protection.

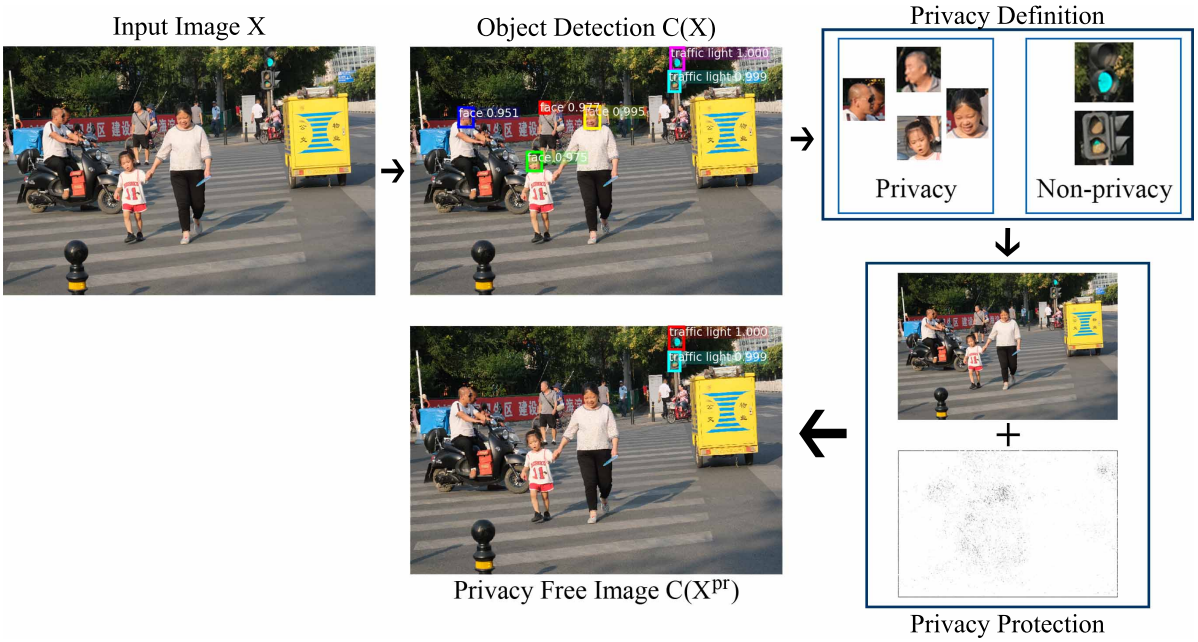


Figure 3.1: Image privacy protection framework.

3.3.1.1 Object Detection

the input image **X** will first pass the object detection module.

There are many existing frameworks for object detection, among which Faster R-CNN [103] is a widely used framework that has been cited frequently in this research area. Therefore, we adopt Faster R-CNN as our object detection module. As shown in Fig. 3.2, the Faster R-CNN detects the region containing objects by three submodules.

- Feature Extractor: a traditional convolutional neural network to perform the feature extraction.

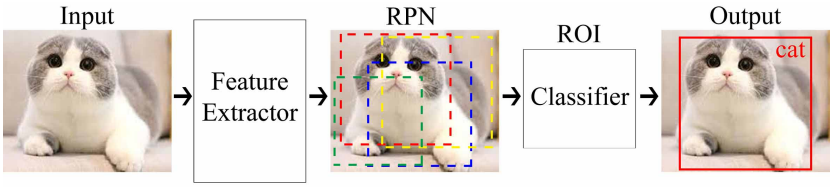


Figure 3.2: The Faster R-CNN framework.

- Region Proposal Network (RPN): RPN finds the object regions by scanning the image using different size anchors (The area RPN scans) in a slide window fashion. The outputs of RPN include a series of anchors A_a, as well as pre-classifier result P_a, i.e.:

$$A_{rpn} = (A_a | P_a) = \left(\begin{array}{cccc|c} x_1 & y_1 & w_1 & h_1 & p_1 \\ x_2 & y_2 & w_2 & h_2 & p_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_\alpha & y_\alpha & w_\alpha & h_\alpha & p_\alpha, \end{array}\right) \tag{3.1}$$

where x_i, y_i, w_i, h_i represent the up left corner x-coordinate, y-coordinate and width, and height of anchors, respectively. i is the index of the anchor ($i = 1, 2, \ldots, \alpha$). $P_a = (p_1, p_2, \ldots, p_\alpha)^T$ denotes the probabilities of anchors being positive.

- Regions of Interest (ROI) Classifier: ROI classifier output contains the location and size of each proposed region and the probability of anchors being a class (e.g. cat, dog, face):

$$A_{roi} = \left(\begin{array}{cccc|ccc} x_{a1} & y_{a1} & w_{a1} & h_{a1} & p_{11} & \ldots & p_{1m} \\ x_{a2} & y_{a2} & w_{a2} & h_{a2} & p_{21} & \ldots & p_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{an} & y_{an} & w_{an} & h_{an} & p_{n1} & \ldots & p_{nm} \end{array}\right), \tag{3.2}$$

where n is the number of anchors that ROI proposed ($n \leq \alpha$). x_{aj}, y_{aj}, w_{aj}, h_{aj} are the coordinate and size information of ROI proposed anchors. $p_{11}, \ldots, p_{nm}$ (noted as P_{ROI}) are the probability of n anchors belonging to m class respectively.

Finally, the output of the object detection module is represented as:

$$C(\mathbf{X}) = \left(\begin{array}{cccc|c} x_{a1} & y_{a1} & w_{a1} & h_{a1} & c_1 \\ x_{a2} & y_{a2} & w_{a2} & h_{a2} & c_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{an} & y_{an} & w_{an} & h_{an} & c_n \end{array}\right), \tag{3.3}$$

where

$$\forall j \in (1,n) : c_j = \begin{cases} \arg\max_i p_{ni}, & 1 \le i \le m \\ c_{bg}, & \forall p_{ni} \le threshold \end{cases}$$

It is worth noting that Faster-RCNN treats background as a class, i.e., c_{bg}. *threshold* is used to deal with the unrecognizable area that may appear. If the probability of all classes is less than *threshold*, it is recognized as the background.

3.3.1.2 Image Privacy Definition

In this module, we first define what object in the image contains individuals' private information. According to the General Data Protection Regulation (GDPR) [15], anything that can be used as a personal identifier should be treated as private information. Therefore, we propose that private objects in images should include:

- Personal identity - license plate, phone number, address, etc.
- Biometrics - face, calendar data, fingerprints, retinal scans, photos, etc.
- Electronic records - cookies, IP locations, mobile device IDs, social network activity records

According to this definition, all classes in the object detection output are divided into two subsets: $\mathbf{C}_{private}$ is the set of private classes, and $\mathbf{C}_{non-private}$ includes non-private classes.

3.3.1.3 Image Privacy Protection

a small adversarial perturbation $\delta\mathbf{X}$ targeting on private objects is applied to generate the privacy-free image $\mathbf{X}^{pr}$, so that only non-private information can be detected when

passing $\mathbf{X}^{pr}$ through an object detector, i.e.,

$$C(\mathbf{X}^{pr}) = \left(\begin{array}{cccc|c} x_{a1} & y_{a1} & w_{a1} & h_{a1} & c_1^{pr} \\ x_{a2} & y_{a2} & w_{a2} & h_{a2} & c_2^{pr} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{an} & y_{an} & w_{an} & h_{an} & c_n^{pr} \end{array}\right), \tag{3.4}$$

where $\forall c_j \in \mathbf{C}_{private} : c_j^{pr} = c_{bg}$.

3.3.2 Problem Formulation

Based on the above-described framework, our target is to fool the network by changing the class of the private objects to 'bg', while the non-private objects are recognized as their original classes. Meanwhile, the added noise δX should be small so that it is imperceptible for humans. Hence, the problem can be formulated as follows:

$$\arg\min_{\delta\mathbf{X}} \left\|\delta\mathbf{X}\right\|_2 \tag{3.5}$$

$$\text{s.t.: } \forall c_j \in \mathbf{C}_{private} : c_j^{pr} = c_{bg} \tag{3.6}$$

$$\forall c_j \in \mathbf{C}_{non-private} : c_j^{pr} = c_j \tag{3.7}$$

3.4 Adversarial Perturbation based Image Privacy Protection Algorithm

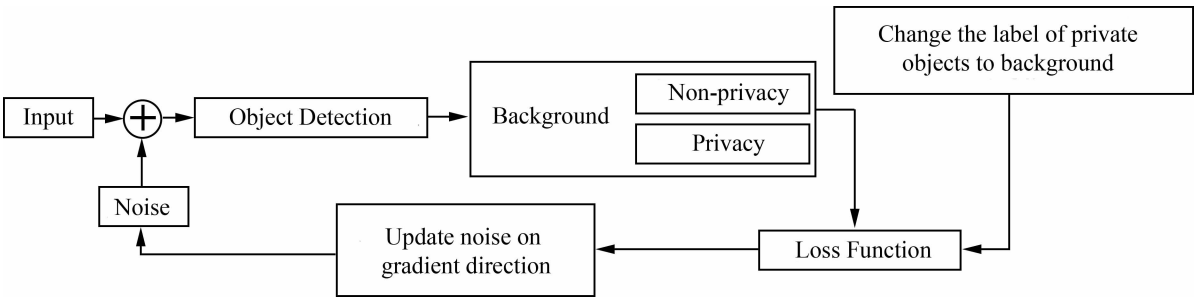


Figure 3.3: Diagram of AP-based image privacy protection algorithm.

In order to solve the image privacy protection problem, we proposed an AP-based image privacy protection algorithm in this section, along with the metrics that can be used to evaluate the performance of the algorithm.

3.4.1 AP-based Image Privacy Protection

Fig. 3.3 gives the flow chart of our algorithm. The input image is sent to the object detector along with the generated noise, and then the output objects are divided into three categories (Background, Non-privacy objects, and Privacy objects). Initially, the added noise is 0, and the object detector will find all objects in the image. Next, we replace the label of the private objects with the background and then put it into the loss function to calculate the gradient. Then the noise is updated according to the gradient. Finally, a perturbed image is generated, in which all privacy objects are treated as background by the object detector.

The key part of the algorithm is to trick the classification loss ($\mathcal{L}_{cls}$) so as to mislead the object detector recognizing the private objects to the background. We define our new loss function as shown in Eq. (3.8) to mislead the classifier so that it will reckon all private objects as background:

$$\mathcal{L}_{cls} = \frac{1}{n_a}\sum_i En(p_i, p_i^*) + \lambda \left\| \mathbf{X} - \mathbf{X}^{pr} \right\|_2, \tag{3.8}$$

where $p_i = [p_{i1}, \ldots, p_{im}]$ is the probability of the content of an anchor being recognized as each class. p_i^* is one-hot encoded ($p_i^* = [0, 0, \ldots, 1, \ldots, 0, 0]$), in which 1 appears in the position where we set the class as the correct class. p_i^* will be generated according to the

ground truth label if the object is non-private, while it will be changed to the background if the object is private. n_a is the number of anchors in the image so that the entropy will be averaged over all anchors. Next, we can use $\mathcal{L}_{cls}$ to generate the perturbation using the fast gradient sign method (FGSM) [99].

Using the targeted FGSM, the perturbation can be calculated in the direction of the gradient:

$$\delta\mathbf{X} = -\epsilon \text{sign}(\nabla_{\mathbf{X}}\mathcal{L}_{cls}) = -\epsilon \text{sign}(\frac{\mathcal{L}_{cls}}{\partial\mathbf{X}}), \tag{3.9}$$

where ϵ is the step parameter that scales the noise. Therefore, the generated image will be:

$$\mathbf{X}^{pr} = \mathbf{X} + \delta\mathbf{X} = \mathbf{X} - \epsilon \text{sign}(\frac{\mathcal{L}_{cls}}{\partial\mathbf{X}}) \tag{3.10}$$

In practice, one-step FGSM is usually not enough, so we can use an iterative version as shown in Alg. 5.

Algorithm 5: AP-based image privacy protection algorithm.

Parameters: Noise scalar ϵ.
Iteration number N.
Input: The original image $\mathbf{X}$.
Output: The released privacy-preserving image $\mathbf{X}^{pr}$.
Initialization: Overall noise $\delta\mathbf{X} = 0$, $\mathbf{X}_0^{pr} = \mathbf{X}$.
for $1 \le n \le N$ **do**
  $\delta\mathbf{X}_{n-1} = -\epsilon \text{sign}(\nabla_{\mathbf{X}}\mathcal{L}_{cls})$;
  $\delta\mathbf{X} = \delta\mathbf{X}_{n-1} + \delta\mathbf{X}$;
  $\mathbf{X}_n^{pr} = \delta\mathbf{X}_{n-1} + \mathbf{X}_{n-1}^{pr}$;
end
$\mathbf{X}^{pr} = \mathbf{X}_n^{pr}$.

3.4.2 Evaluation Metrics

In order to measure the performance of our proposed methods, we introduce the following metrics from three different aspects:

3.4.2.1 Distortion metrics

Two distortion metrics are used to measure the amount of noise added to the original image.

- L_2 computes the Euclidean distance between original and perturbed examples, i.e., $L_2 = \left\|\mathbf{X}^{pr} - \mathbf{X}\right\|_2 = \left\|\delta\mathbf{X}\right\|_2$

- Average L_p Distortion ALD_p [104]: $ALD_p = \frac{\|\mathbf{X}^{pr}-\mathbf{X}\|_p}{\|\mathbf{X}\|_p}$.
  We use ALD_∞ to measure the maximum change in all dimensions of adversarial perturbations in the simulation.

3.4.2.2 Structural Similarity (SSIM)

SSIM is a method used to measure the similarity between two digital images. Compared with the traditional image quality measurement methods, such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE), SSIM can better match the human judgment of image quality [105][106]. It can be used to quantify the extent that the perturbation is invisible to human eyes.

3.4.2.3 Private Information Hiding Ratio

$R = \bar{r}_p + \lambda\bar{r}_a$, where $\bar{r}_p$ and $\bar{r}_a$ are the average hiding ratio or keeping the ratio of privacy objects and non-privacy objects, respectively, λ is set to 0.5 for better illustration. It is used to measure whether our method can hide the proper objects in images.

3.5 Experiment and Discussions

In this section, we show our experimental results.

3.5.1 Experiment Settings

In our experiment, we use the images from the data set provided by Tribhuvanesh et al. [92]. The data set is originated from the VISPR data set. The authors selected images containing private information and pixel-annotated using 24 privacy attributes. In our experiment, we choose faces and license plates as privacy items. Hence, we filtered the images with these two annotations from the data set. And filter some non-private data from the original data set. Then, we added more street view images containing faces and license plates into the training data set for better performance.

Here we use Faster R-CNN as the object detector. Faster R-CNN requires that the input image shape is square ($1024 * 1024$ is the suggested size). The original data set contains a large number of large-size images (e.g. 7000x6000), so we make standardization on our training data set before training for better training performance. The model was trained on one GPU card, GeForce GTX 2080Ti.

3.5.2 The Experiment Results

Fig. 3.4 shows an example of our proposed algorithm. The left column shows the detection result of the original image, the next two columns are the detection result after the proposed privacy treatment and the added perturbation, respectively. As can be seen in the figure, without privacy protection, all objects in the images can be detected by a standard Faster R-CNN. After adding the adversarial noise, the detector cannot detect the privacy objects, including faces and car plates, while the non-sensitive features (e.g. traffic lights) can still be detected. The adversarial perturbation in the images are generated in the range R_p ($R_p \in (-2,2)$). In order to display the noise in the image, we normalize the R_p to R_P ($R_P \in (0,255)$). Observing the relative location of objects in the original image, the noise dots concentrate on the regions containing objects. By adding adversarial perturbation in such a small range (e.g.R_p), human eyes can hardly recognize the difference. But the object detector has been successfully fooled.

Now we compare the performance of our proposed algorithm with Blur and Mosaic. As shown in Fig. 3.5, the Blur and Mosaic's "thickness" has been carefully adjusted to just hide the sensitive information. But human eyes can easily notice those changes.

Figure 3.4: Illustration of AP-based image privacy protect algorithm.

Figure 3.5: The detection results after privacy protection: (a) Image without Protection; (b) Blur; (c) Mosaic; (d) AP-based.

Our method, while deceiving the detector from the private objects, greatly preserves the non-private information in the original image, so that naked eyes can hardly see the difference.

Next, we measure the effectiveness of our approach using the metrics mentioned in Section III.B. Table I shows the performance comparison measured by distortion metrics ($L2$ and ALD_p). The adversarial noise (AD Noise) is generated in the range R_p ($R_p \in (-2,2)$). Blur and Mosaic noise thickness has been modified to barely hide the sensitive information from Object Detector. Compared with Blur, our method is 73.5% and 79.2% lower in the $L2$ and ALD_p average scores, respectively. Also, our method is superior to Mosaic in both $L2$ and ALD_p, i.e., our algorithm is 81.4% lower in $L2$ and 85.2% lower in ALD_p.

Table 3.1: $L2$ and ALD_p score compared with classical methods

		Original	Blur	Mosaic	AD Noise
$L2$	1	0	4111	4153	765
	2	0	2008	3730	776
	3	0	3426	6168	778
	4	0	1980	2386	731
	Average	0	2881.2	4109.2	762.5
ALD_p (10^{-2})	1	0	5.55	6.27	0.97
	2	0	2.29	4.82	0.68
	3	0	4.59	6.67	0.69
	4	0	2.19	2.81	0.70
	Average	0	3.655	5.143	0.76

Table 3.2 presents the results measured by the SSIM metric. A higher score means a smaller distortion of the image. The performance of our method has increased by 192.5% compared to blur, which is an increase of 474.8% compared to Mosaic.

Table 3.2: The SSIM score compared with classical methods

	Original	Blur	Mosaic	AD Noise
1	1	0.548	0.286	0.998
2	1	0.536	0.243	0.995
3	1	0.442	0.121	0.997
4	1	0.472	0.191	0.998
5	1	0.634	0.182	0.998
6	1	0.478	0.235	0.996
Average	1	0.518	0.210	0.997

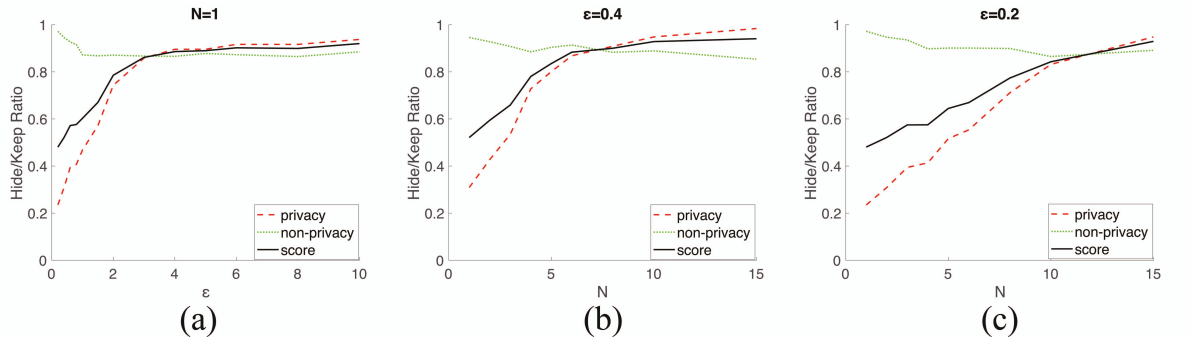


Figure 3.6: The hide/keep ratio; (a) Fixed iteration number $N = 1$; (b) Fixed $\epsilon = 0.4$; (c) Fixed $\epsilon = 0.2$

From the above results, we can see that our method gains performance improvement in balancing privacy protection and information preservation compared with classical methods. Finally, we measure the privacy protection efficiency by running the test data set with different iteration numbers (N) and noise scalar (ϵ). The adversarial noise thickness is related to the iteration numbers (N) and noise scalar (ϵ). The noise range R_p is related to $N \times \epsilon$. So, we use noise range as an index to measure our adversarial noise thickness. As can be seen in Fig. 3.6, the private information hiding rate is proportional to the noise thickness, while the non-sensitive objects keeping ratio slightly decreases with the increase of thickness. Fig. 3.6.(a) gives the change in hiding ratio with an increase of ϵ. It shows that a very small amount of noise thickness: $R_p \in (-3, 3)$ (out of 255) is enough to achieve an over 90% high hiding ratio. Fig. 3.6. (b) and Fig. 3.6. (c) show that under the same thickness (R_p), a smaller ϵ achieves a relatively higher hiding ratio, but it needs more iterations.

3.6 Conclusion and future works

The recent advancement of artificial intelligence exacerbates the privacy concern, especially for images that contain a variety of personal information. In order to solve this problem, we proposed an image privacy protection framework against AI using an AP-based privacy protection algorithm. Our results show that private objects in images can be well protected while non-private information is preserved by adding a small amount of noise. Therefore it can protect image privacy while preserving the image's utility. Moreover, the noise added can hardly be detected by the naked eye, which lens more practical value of the proposed algorithm in real-life employment.

However, the field of image privacy protection has long suffered from a comprehensive

definition for safeguarding privacy in images. This gap has restricted the quantification of privacy protection effectiveness solely to visual observation, leaving it susceptible to exploitation by AI attackers. Consequently, the importance of an approach that can be proven, quantified, and controlled becomes paramount in ensuring robust protection. The forthcoming chapter will thoroughly investigate this issue and introduce our proposed solution: DP-Image.

CHAPTER 4

DP-Image (DP-Image: Differential Privacy for Image Data in Feature Space)

4.1 Preface

In this chapter, our exploration revolves around the development of provable strategies for safeguarding image privacy. Our aim is to integrate the widely embraced techniques of differential privacy, which offer controllable and verifiable protection of sensitive data. However, the application of pixel-level differential privacy to non-structured data, such as images, often leads to excessive distortion of crucial image information. To overcome this challenge, we propose an innovative approach that introduces cyclic noise to the latent space vectors of images, effectively ensuring robust image privacy protection. Within this chapter, we unveil a protective framework that seamlessly combines differential privacy with the power of autoencoders. Through meticulous experimentation on a facial dataset, we thoroughly evaluate the framework's effectiveness in preserving facial identity. Notably, our approach outperforms alternative methods, showcasing exceptional performance in achieving privacy preservation goals.

4.2 Introduction

4.2.1 Motivation

Recently, our human society witnesses a rapid increase in the use of image data, which poses high risks of privacy leakage, as images usually contain a large number of sensitive data that might visually reveal personal information [16]. For example, heavy concerns on the privacy risks were raised on uploading people's face pictures to a popular APP called FaceApp [17], which can edit a person's face by changing his/her gender, age, ethnicity, etc. In fact, the instinct to protect privacy in image data is not new in our human society.

So the question is: *why the issue of image privacy becomes urgent now?* This is because the newly emerging deep learning techniques have exacerbated the privacy risks for image sharing and usage. In more detail, artificial intelligence (AI) aided by deep learning methods can automatically detect and collect people's private and sensitive information, thus impacting everyone's daily life.

This brings risks of personal privacy to a whole new level, while the traditional privacy-preserving methods seem powerless when facing the large-scale deep learning tools [107].

Unfortunately, privacy protection for unstructured data, such as images, is much more difficult compared with that for structured data. Structured data are usually well-

documented in an organized and systematic manner. However, in unstructured data such as images, data attributes are implicitly represented by sets of pixels covering irregular shapes and sizes. Such implicit data attribute values cannot be processed by traditional statistical methods, such as correlation, regression, etc. Hence, the underlying privacy risks in unstructured data are much less clear than those in structured one.

Due to the unstructured nature of the image data, research on image privacy protection usually takes a different approach than that for structured data. First, traditional research on image privacy protection often assumes *human adversaries*. In other words, privacy risks are usually quantified by how effectively the information in images can be picked up by human eyes and brains. As a result, "blurring", "pixelation", and "mosaic" are the most used methods to protect privacy in images, e.g. in Google Earth street view. However, we are now marching into a new age of AI, where the information contained in images can be filtered out by AI functions/models. Those AI functions, such as DNNs, seem to take a different approach to interpreting and understanding images, and they are able to re-identify the obfuscated image with high accuracy [107]. Therefore, very recently, researchers have started to consider a new scenario, where AI acts as an adversary. For example, to throw off AI adversaries when analyzing images, the authors of [99] proposed to generate a small but intentional worst-case disturbance to an original image, which can mislead DNNs in machine-learning tasks such as image classification, without causing a significant visual difference perceptible to human eyes. The perturbed image is referred to as an "adversarial example", and the specially generated noise pattern is thus named adversarial perturbations (AP). A few papers have discussed the potential of AP in privacy protection [7, 101, 102, 108–111]. Another recent trend for image privacy protection is the use of the generative adversarial network (GAN). Content generated by GAN can be used to replace or edit the original images so that the identity-related information can be removed [112–116]. Finally, there have been several initial studies that apply differential privacy (DP) in images. Fan proposed ϵ- differential private methods in the pixel level of the image [80], and in Singular Value Decomposition (SVD) [81], respectively. However, making image pixels or SVD modes indistinguishable does not make much sense in practice, and the quality of the generated image is quite low. Lecuyer et al. [82] proposed the PixelDP framework that includes a DP noise layer in the DNN. The PixelDP scheme enforces that the output prediction function is DP provided the input changes on a small number of pixels (when the input is an image). However, the purpose of PixelDP is to increase the model's robustness to adversarial examples, other than image privacy.

Overall, the aforementioned works cannot solve the image privacy protection problem thoroughly due to several reasons:

- First, image privacy is yet neither clearly defined, nor can it be quantitatively measured.
- Second, the current AP-based method is only designed for specific models and applications. It is not effective against human adversaries due to uncontrollable visual appearance.
- Third, the GAN-based image manipulation lacks provable privacy measurements.
- Finally, the existing DP methods cannot effectively capture the key features of images from the privacy perspective or were not designed for the purpose of image privacy protection.

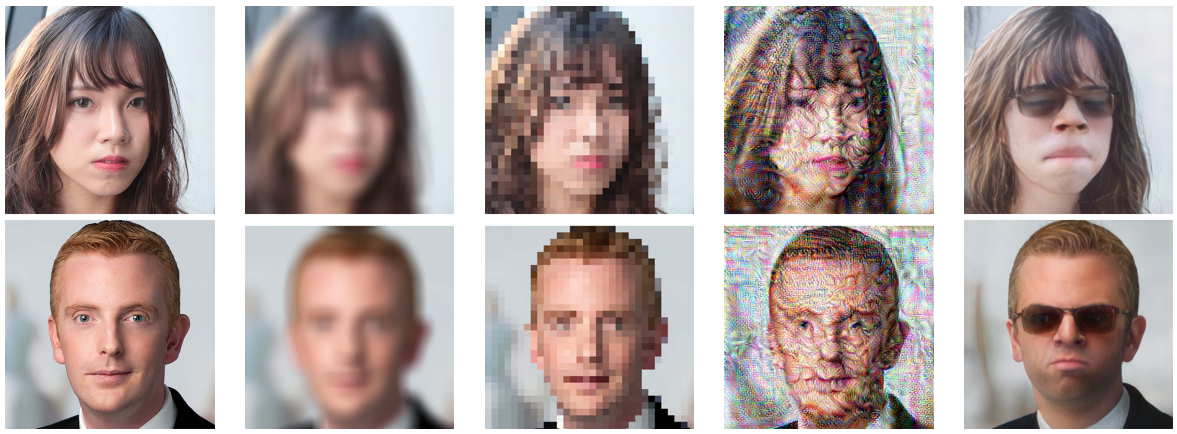
Motivated by these challenges, we propose a differentially private image (DP-Image) framework and design DP-Image mechanisms in this work. This framework redefines the traditional DP [78] and Rényi differential privacy (RDP) [117] notions, in the context of the image data. It should be noted that the application of DP on the image data is not a trivial extension, due to the vastly different application scenarios. In traditional database applications, we assume that a data curator has a database D and an attacker can either make a query on the database to obtain useful information to launch privacy attacks (i.e., the interactive setup) or get access to a synthetic version of database D (i.e., the non-interactive setup).

The important thing is that the adversary does NOT have access to the original database. Therefore, applying DP to the query answers or the synthetic data release will be able to protect the original data.

However, the most common application of images is *image publication and sharing*. It includes the case of personal image sharing on social network platforms and commercial image applications such as Google Street View. In this scenario, the original image data need to be exposed to some extent for meaningful applications. For example, it would be weird to share a synthetic photo of a cartoon character in a person's Facebook post, saying he/she took a selfie on a sunny Saturday afternoon.

Hence, in the case of the image data, the adversary gains a considerable advantage in stepping closer to the original data than the conventional scenarios, and thus, it becomes less difficult for the adversary to re-identify and collect sensitive information from published images. Therefore, the perturbation of privacy protection needs to be

added directly onto the original image, so as to mislead the adversary when he analyzes the image. To this end, we propose a DP-Image framework that adds DP noise to the image feature vector in the latent space to alter the meaningful information carried by the image, and then reconstructs the image from the perturbed vector to replace the original image. In this way, the reconstructed one will not disrupt the intended applications because it shares non-private information with the original one, which keeps the application context meaningful and consistent.



((a)) Original ((b)) Blur ((c)) Mosaic ((d)) AP(pixel) ((e)) Our

Figure 4.1: A brief comparison of different image privacy protection methods. Our approach not only protects the privacy of an image by generating a photo-realistic alternative, but it also provides a controllable way for privacy preservation.

4.2.2 Contributions

To elaborate on this in a more concrete manner, we show a comparison of our method with several traditional methods in Fig. 4.1. As can be seen from the figure, traditional methods, such as blur and mosaic, considerably destroy the utility of face images. If such images were used in, e.g., demographics analysis, the application would be most likely to be unfruitful. Adding adversarial perturbation on the pixel domain cannot guarantee privacy protection when facing humans as an adversary. Looking at the DP images generated by our method,

while the identity of the faces has been changed, they are still useful for a possible demographics analysis or a straightforward Facebook post. Besides, all sensitive information, such as gender, race, age, etc., are generated randomly through the DP-Image process, which greatly protects an individual's privacy, against both human and AI adversaries.

The major contributions of this work are summarized as follows:

- We propose a DP-Image definition based on the notions of DP and RDP.
- We propose a DP-Image protection mechanism with provable and adjustable privacy levels.
- We design a DP-Image framework that enables us to apply the DP-Image protection mechanism at the image feature level, utilizing advanced deep learning and GAN techniques.
- We implement the proposed DP-Image protection mechanisms on a real-life image dataset and show its effectiveness in safeguarding people's privacy.

4.2.3 Overview of the work

The structure of the remainder of the chapter is as follows. In Section 4.3, we discuss the preliminary knowledge for our work, including DP, RDP, Privacy Amplification and deep learning in image applications. In Section 4.4, we first define the adversary model. Then, we proceed to formulate the notion of DP-Image and present our DP-Image protection mechanisms. In Section 4.5, we conduct experiments to validate our proposed scheme using a case study of face image privacy. We analyze the privacy guarantee of our proposed framework and schemes in Section 4.6. Section 4.7 discusses the related work, and Section 4.8 concludes our work.

4.3 Preliminaries

4.3.1 Differential Privacy

Informally, DP [78, 79] is a definition of privacy that guarantees that the likelihood of seeing an output on a given original dataset is close to the likelihood of seeing the same output on another dataset that differs from the original one in any single row. Here, an output could be a synthetic dataset, a statistical summary table, or a simple answer to a query, etc. Since this single row is arbitrary, this definition aims to make an adversary unable to be certain about whether a particular individual is in the original dataset or not. In other words, differential privacy provides any individual in the dataset with plausible deniability - the ability to say, "I am not in that original dataset" - and hence an individual can claim that he/she has nothing to do with the output. This is supported

by mathematical proofs showing an employed method does provide the DP property. Generally speaking, the basic idea of a DP mechanism is to introduce randomness into the original dataset so that any individual's information cannot be inferred by an adversary looking at the released output.

A key property of DP is that we can add outputs of several DP algorithms, and the result is still DP. However, this property incurs a loss in the privacy budget. The maximum impact is the sum of the privacy budgets for the involved outputs. By appropriately setting the privacy budget of constituent algorithms, we can ensure that the overall privacy budget of the data release algorithm is within the bound of a target privacy budget.

Another property of DP is that if the dataset is divided into disjoint subsets, such that the addition or removal of a row affects at most one subset, then we can assign each dataset a privacy budget of ϵ, and still maintain an overall privacy budget of ϵ. These two properties are referred to as the sequential and parallel composition of DP.

A final useful property of DP is that the output of a DP algorithm can be post-processed without affecting privacy, provided it is done independently of the original dataset. For example, averaging, rounding, or any change to the numbers will not impact the privacy of the data. This means that an analyst can do any amount of data post-processing on a released DP dataset and cannot reduce its privacy guarantee.

The DP treatment usually involves adding noise to give uncertainty to the impact of any single individual. We need to add noise on a scale relative to how much any individual could make a difference. For example, if we add noise to counts of individuals, then any single individual only influences that count by a maximum value of 1. So, the noise is on a scale relative to that 1.

Following the above brief discussion on DP, here we show the formal definition of DP.

Definition 4.3.1. (ϵ-Differential Privacy) [78, 79]: A randomized mechanism $\mathcal{M}$ gives ϵ-differential privacy if for any neighboring datasets $\mathcal{R}$ and $\mathcal{R}'$ differing on one element, and all sets of output S:

$$Pr[\mathcal{M}(\mathcal{R}) \in S] \leq \exp(\epsilon) \cdot Pr[\mathcal{M}(\mathcal{R}') \in S]. \tag{4.1}$$

(ϵ, δ)-DP is a relaxation of ϵ-DP that allows a δ additive term in its definition.

Definition 4.3.2. ((ϵ, δ)-Differential Privacy) [78, 79]: A randomized mechanism $\mathcal{M}$ gives ϵ-differential privacy if for any neighboring datasets $\mathcal{R}$ and $\mathcal{R}'$ differing on one

element, and all sets of output S:

$$Pr[\mathcal{M}(\mathcal{R}) \in S] \leq \exp(\epsilon) \cdot Pr[\mathcal{M}(\mathcal{R}') \in S] + \delta. \tag{4.2}$$

4.3.2 Rényi Differential Privacy

Rényi differential privacy (RDP) [117] proposed by Ilya Mironov gives a convenient way of tracking cumulative privacy loss when applying differentially private mechanisms. RDP measures the Rényi divergence [118] between the output distribution on two neighbouring datasets.

Definition 4.3.3. (Rényi Divergence) [117]: P and Q are probability distribution over $\mathcal{R}$, the Rényi divergence for $\alpha > 1$ is defined as:

$$D_\alpha(P \parallel Q) \triangleq \frac{1}{\alpha - 1} \log E_{x \sim Q} \left(\frac{P(x)}{Q(x)} \right)^\alpha. \tag{4.3}$$

All the logarithm is natural, $P(x)$ is the density of P at x.

Definition 4.3.4. ((α, ϵ)-RDP) [117]: A randomized mechanism $\mathcal{M}$ gives (α, ϵ)-RDP if for any neighboring datasets $\mathcal{R}$ and $\mathcal{R}'$ holds that:

$$D_\alpha(\mathcal{M}(\mathcal{R}) \parallel \mathcal{M}(\mathcal{R}')) \leq \epsilon. \tag{4.4}$$

RDP can easily transfer to (ϵ, δ)-differential private.

Lemma 4.3.1. (RDP to (ϵ, δ)-DP) [117]: If a mechanism $\mathcal{M}$ satisfies (α, ϵ)-RDP, it also satisfies $(\epsilon + \frac{\log 1/\delta}{\alpha - 1}, \delta)$-DP for any $0 < \delta < 1$. Besides, $\mathcal{M}$ also satisfies pure ϵ-DP, for $\alpha = \infty$.

4.3.3 Privacy Amplification by Iteration

Privacy amplification by iteration has been proven to have its superiority for differentially private optimization algorithms, such as DP-SGD (Differentially-Private Stochastic Gradient Descent) [119]. The paper reveals the iterative privacy budget under natural settings for DP mechanisms.

Lemma 4.3.2. [119]: Assuming $X, X' \in \mathcal{R}$, let $\mathcal{M}(X)$ be obtained from X by iterating T times:

$$X_{t+1} \doteq f_{t+1}(X_t) + m_{t+1} \tag{4.5}$$

where $\{f\}_{t=1}^{T}$ are contractive maps and $\{m\}_{t=1}^{T} \sim \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$. Let $\mathcal{M}(X')$ be obtained from X' under the same process. Then $\mathcal{M}$ satisfies:

$$D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X')) \leq \frac{\alpha \left\| X - X' \right\|^2}{2T\sigma^2}. \tag{4.6}$$

4.3.4 Deep Learning in Image Applications

Deep learning has profoundly changed the landscape of computer vision and image processing. The most advanced algorithms in this field are based on DNNs, usually with a convolutional structure. Convolutional Neural Network (CNN) [120] consists of neurons with learnable weights and biases. Each neuron receives some input, performs a dot product, and optionally goes through its nonlinear activation function. The entire network still implements a distinguishable score function: from the original image pixels on one end to the class score on the other end.

There are several mostly used CNN architectures. AlexNet [121] was submitted to the ImageNet ILSVRC Challenge in 2012, and its performance far exceeded the second place [120]. It is this work that popularized CNN in computer vision. Szegedy et al. Google proposed GoogLeNet [122] in 2014. It introduced an Inception module, which significantly reduced the number of parameters in the network (4M, compared to 60 million for AlexNet). GoogLeNet also has multiple subsequent versions, such as Inception-v3 [123], Inception-v4 [124]. Karen Simonyan and Andrew Zisserman proposed VGGnet [125], which shows that network depth is a key factor for superior performance. ResNet [126] was developed by Kaiming He et al, which has a special skip connection and extensive use of batch normalization functions. ResNets are currently widely used in practice.

The image applications include two important categories: extracting information from images, constructing images with semantic meanings. Deep learning has improved the performances from both ends.

For information extraction, deep learning has been used for image classification [121, 124, 125], object detection [127], recognition [128], tracking [129], and semantic segmentation [130], etc. And it outperforms traditional methods in all these tasks. Outputs of DNNs in these applications contain rich information such as the type and position of objects, identity and actions of people, thus making DNNs and privacy issues highly relevant.

On the other hand, deep learning has also been used to generate synthetic images. For example, GAN [131] invented by Ian Goodfellow and his colleagues in 2014, is able

to learn to generate new data with the same statistics as the training set. Following this initial work, some milestone works in GAN were developed, such as cGAN [132], StyleGAN [133, 134], etc. While GAN could simply generate random new outputs that are similar to the training data, Variational Autoencoders (VAEs) [135] can help to explore variations on data you already have. Moreover, there have also been researched to combine VAEs and GANs together, which can generate compelling results for complex datasets such as images [136]. This VAEs architecture can be used to process an input image, by converting it into a smaller, dense representation, which the decoder network can use to convert it back to the original image.

4.4 Our proposed DP-Image Framework

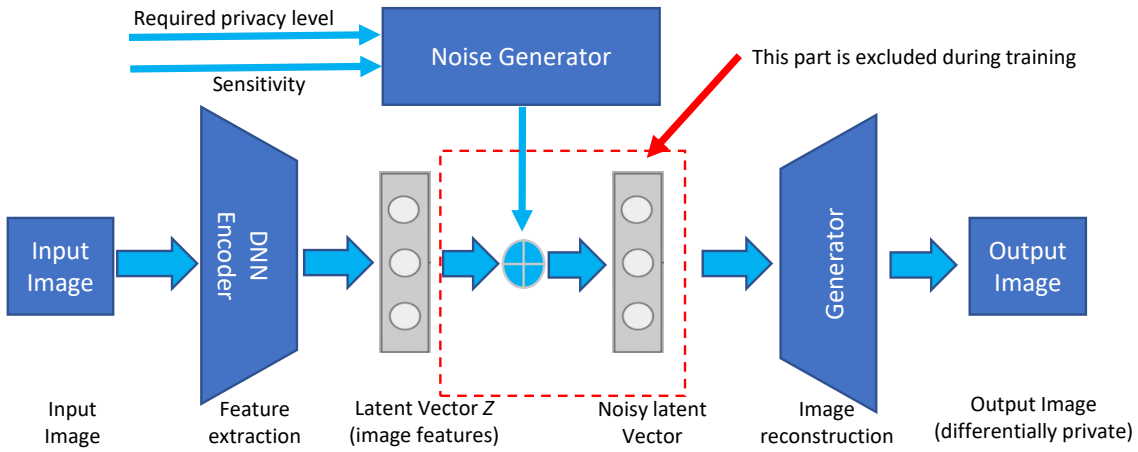


Figure 4.2: Architecture of the DP-Image Framework. The Iterative DP-Image (IDPI) theorem presented in Section 4.4.4 will demonstrate the concepts of the required privacy level, denoted as ϵ, and sensitivity, represented by Δf.

In this section, we present the adversary model considered in this thesis and then introduce our proposed DP-Image framework.

4.4.1 Adversary Model

As discussed in Section 4.2, we focus on the mainstream image application: *image publication and sharing scenario*. This is a non-interactive setup, which includes the case of personal image sharing on social network platforms and commercial image applications such as Google Street View.

In this scenario, the adversary can either manually search or use a web crawler and AI-based tools to automatically identify and collect sensitive information from published images. Assisted with the advanced deep learning tools, the adversary can obtain different types of information from images, for example:

- Class or identity of an image using classification or face recognition tools.
- Object contained in the image via object detection.
- Specific features in the image like numbers or text by natural language processing (NLP) or other feature extractors.

The adversary has two significant differences compared with the traditional database adversary: 1) he has access to the image; 2) he extracts information from the image using an advanced query method based on deep learning tools.

On the other hand, the users would like to publish or share their photos without personal information leakage. Moreover, the users would like information coverage satisfies two key criteria: 1) the images should look realistic and natural after modification, so "Blur" or "Mosaic" is not preferable; 2) the amount of noise should be controllable, so the users can decide the trade-off between privacy and utility.

4.4.2 DP-Image Framework: Protecting Image Privacy in Feature Space

As discussed in Section 4.2, in the GDPR, privacy information is defined as "personal data that are related to an identified or identifiable natural person." Here, the emphasis on privacy is laid on personal identities. On the other hand, the long history of image processing techniques has proved that the transform domain of an image can provide more useful information than the original pixel domain. Also, the adversary can obtain more detailed information by extracting a feature vector $f(X)$ from the image X. This vector $f(X)$ represents certain features in images. In this case, a privacy protection scheme that can blur the feature vector $f(X)$ will be more effective than a scheme that works in the pixel domain.

Based on the above argument, we design an image privacy protection framework in the feature space. The three main modules are feature extraction, noise generator, and image reconstruction, as shown in Fig. 4.2. And our motivations for designing these modules in Fig. 4.2 are explained as follows.

1. Feature extraction is crucial to feature-level privacy protection. The challenge stems from the difficulty in directly extracting the features from pixel values in the original image form, let alone identifying privacy-related features. In order to solve this problem, we propose to use an advanced DNN as the encoder for feature extraction. The output of this encoder will be a feature extraction network that can map an image X to a vector Z in the latent (feature) space.

2. The noise generator is used to inject noise into the feature vector. And the amount of noise can be controlled by parameters.

3. The image reconstruction module will transform the perturbed feature vector back to the image domain. This generator can be trained in the same way as GAN.

4.4.3 DP-Image Definition

In our *image publication and sharing scenario*, the attacker aims to find personal information from the images they have access to. And image privacy protection aims to ensure that the adversary cannot learn too much about each individual,Äôs private information even though he/she can see the image data. This is in stark contrast to the conventional privacy protection methods that might provide an additional layer of protection by not granting the adversary access to the original format of the raw data.

In this sense, starting with the DP and RDP notions the definition of image differential privacy can be written as:

Definition 4.4.1. (α,ϵ)-RDP-Image): A randomized mechanism $\mathcal{M}$ gives (α,ϵ)-RDP-Image if and only if for any two image data X, X', $\mathcal{M}$ satisfies:

$$D_\alpha(\mathcal{M}(X) \parallel \mathcal{M}(X')) \leq \epsilon. \tag{4.7}$$

The above definition is very similar to that of the traditional RDP, except that we are enforcing DP for given different input samples, instead of requiring, there is only a single item (pixel) difference in two images. Also, we expect the output images of $\mathcal{M}(X)$ and $\mathcal{M}(X')$ to be indistinguishable from a privacy perspective other than in the pixel domain.

Moreover, as we plan to add noise in the feature space, a generalized sensitivity in RDP of the feature vector $f(X)$ is defined as follows:

Definition 4.4.2. (Sensitivity in Image Feature Space): For any two images in the dataset, each consisting of n pixels: $X = (x_1, \ldots, x_n)$, $X' = (x'_1, \ldots, x'_n)$, if f is the function that maps the image to its feature space, the sensitivity Δf is defined as the maximum differences in $f(X)$ produced by two different images:

$$\Delta f \doteq \sup_{X,X'} ||f(X) - f(X')||_2. \tag{4.8}$$

This indicates the largest difference between the feature vectors of the two images.

4.4.4 DP-Image Mechanism

Following Definition 4.4.1, to satisfy the DP-Image requirement, we implement the DP-Image by adding DP noise on the feature space in an iterative manner.

Algorithm 6: DP-Image Algorithm by Iteration.

Parameters: Noise coefficient σ, Iteration number T.
Input: The original image X.
Output: The released privacy-preserving image X_T.
for $0 \leq t < T$ **do**
  $z_t = f(X_t)$
  $z_{t+1} = f_c(z_t) + \mathcal{N}(0, \sigma^2 \mathbb{I}_d)$
  $X_{t+1} = g(z_{t+1})$
end
$X_T = X_{t+1}$.

Where $f(\cdot)$ is a function that maps image X to its feature space vector $z \in \mathbb{R}^m$. $f_c(\cdot)$ is the clipping function. $g(\cdot)$ is a function that maps the feature space vector back to the image. $\mathcal{N}(0, \sigma^2 \mathbb{I}_d)$ are i.i.d random variables drawn from the Gaussian distribution.

It is worth noting that $f_c(\cdot)$ is a contractive map, which is also said to be 1-Lipschitz.

Definition 4.4.3. (Contraction): For a Banach space $(\mathcal{B}, \|\cdot\|)$, a function $f : \mathcal{B} \to \mathcal{B}$ is said to be contractive (1-Lipschitz) if for all $x, y \in \mathcal{B}$:

$$\left\| f(x) - f(y) \right\| \leq \left\| x - y \right\|. \tag{4.9}$$

Based on the iterative DP-Image mechanism above and privacy amplification by iteration (see Lemma 4.3.2), we conclude the privacy budget of iterative DP-Image (IDPI).

Theorem 4.4.1. (**Iterative DP-Image (IDPI)**): Assuming a start image $X_0 \in \mathcal{R}$, and its corresponding output image X_T. $f(\cdot)$ is a map function that maps the image into its feature space, $f_c(\cdot)$ is 1-Lipschitz. Then, for every $\sigma > 0, \alpha > 1$, IDPI(X_0, σ, T) satisfies $(\alpha, \frac{\alpha \Delta f^2}{2T\sigma^2})$-RDP. By definition, it also satisfies $(\epsilon + \frac{\log(1/\delta)}{\alpha - 1}, \delta)$-DP for any $0 < \delta < 1$, where $\epsilon = \frac{\alpha \Delta f^2}{2T\sigma^2}$, and satisfies pure $(\frac{\alpha \Delta f^2}{2T\sigma^2})$-DP, for $\alpha = \infty$.

Proof. Here, we give the simplest version of the proof. First, according to the equation in Lemma (4.3.2). $\mathcal{M}(X)$ satisfies $(\alpha, \frac{\alpha \Delta f^2}{2T\sigma^2})$-RDP:

$$\begin{aligned} D_\alpha(\mathcal{M}(f(X)) \parallel \mathcal{M}(f(X'))) &\leq \frac{\alpha \left\| f(X) - f(X') \right\|^2}{2T\sigma^2} \\ &\leq \frac{\alpha \sup ||f(X) - f(X')||_2^2}{2T\sigma^2} \\ &= \frac{\alpha \Delta f^2}{2T\sigma^2} \quad \text{(Definition 4.4.2)} \end{aligned}$$

As claimed.

Then the rest of the proof follows from the post-processing property of DP. Hence, we can conclude that if the feature vector is treated with IDPI, then the reconstructed image X_T generated by mechanism $\mathcal{M}$ satisfies the (α, ϵ)-RDP as defined in Definition 4.4.1.

■

4.5 Experiments

In this section, we conduct experiments to validate the effectiveness of the proposed DP-Image scheme. As facial recognition is the most widely used privacy sensitive application at present, we use face images that contain personal identity information as the dataset in our experiments.

4.5.1 Experiment Setup

4.5.1.1 Dataset

In our experiment, we used the Flickr-Faces-HQ (FFHQ) dataset [137], which is a high-quality image dataset of human faces. The dataset consists of $70K$ high-quality PNG images with a resolution of 1024×1024, and contains considerable variation in terms of age, ethnicity, and image background.

4.5.1.2 Evaluation metrics

We use two groups of metrics to evaluate the performance: 1) privacy metrics to measure the privacy protection performance: including the **Face Privacy Protection Success Rate (FPPSR)** and **Identity Similarity Score (ISS)**; 2) utility metrics that can validate the utility of the perturbed images: including l_2**-distance**, **Average** l_p **Distortion (ALD)**, **Structural Similarity (SSIM)**, and **Frechet Inception Distance (FID)**.

1. Face privacy protection success rate (FPPSR): the average ratio of successful face de-identification. It is obtained via using face recognition systems, e.g., ArcFace [138] and Microsoft Azure Face Recognition API [139] to check whether the perturbed image is recognized as a different person from the original image. The mechanisms of these two systems are quite similar. ArcFace can be treated as a white-box setting as it is open-source, while Microsoft API represents a black-box setting.

2. Identity Similarity Score (ISS): this is also obtained via using a face recognition system. However, rather than using the binary outcome of "Yes" or "No", the soft value that shows the similarity between the perturbed image and the original image is used to measure to what degree the privacy has been preserved.

3. Distortion metrics: two distortion metrics are used to measure the amount of noise added to the original image.

   - l_2-distance computes the Euclidean distance between original and perturbed examples, i.e., $l_2 = \|\mathbf{Y} - \mathbf{X}\|_2$

   - ALD_p [104]: $ALD_p = \frac{\|\mathbf{Y}-\mathbf{X}\|_p}{\|\mathbf{X}\|_p}$.
     We use ALD_∞ to measure the maximum change in all dimensions of the perturbed images.

4. SSIM: SSIM is a method used to measure the similarity between two digital images. Compared with the traditional image quality measurement methods, such as peak signal-to-noise ratio (PSNR) and mean squared error (MSE), SSIM can better match the human judgment of image quality [105] [106]. It can be used to quantify the extent that the perturbation is invisible to human eyes. A high score means that two images are structurally similar.

5. FID [140]: FID captures the similarity between two images based on the deep features calculated using the Google Inception v3 model. A lower score indicates that the two images are more similar.

4.5.1.3 Generation of the latent space vector Z

For an image, its generated latent space vector Z needs to satisfy two important requirements: 1) Z is a good representation of the image features; 2) the original image can be recovered by feeding Z to the generator network.

To achieve this target, we use the pSp framework proposed in [141]. It is an encoder that directly generates a series of style (feature) vectors which can be fed into a pre-trained StyleGAN generator.

As shown in Fig. 4.3, the dimension of Z generated by the pSp framework is $(512, 18)$, consisting of 18 style vectors, each with a length of 512 elements. These styles are extracted from the corresponding feature map generated from a ResNet backbone, where

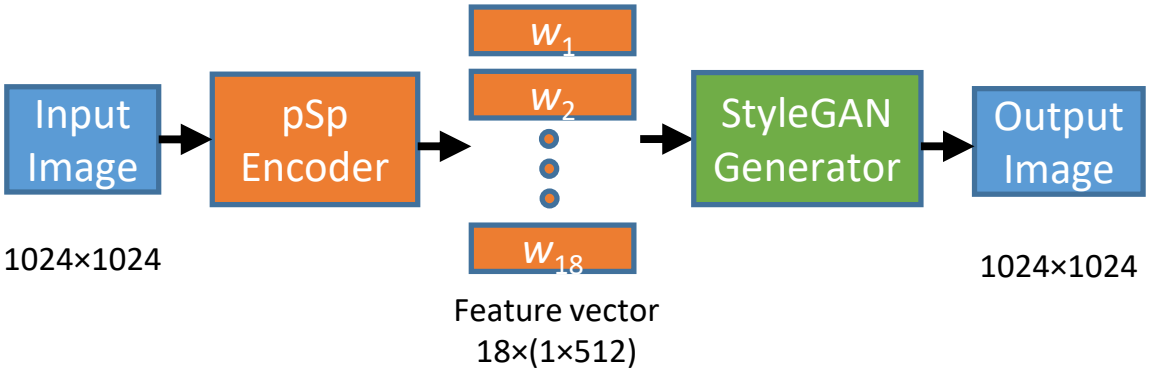


Figure 4.3: The framework of generating feature space vector using pSp encoder and reconstructing image with StyleGAN generator.

style vectors $(1-3)$ are generated from the small feature map, style vectors $(4-7)$ from the medium feature map, and style vectors $(8-18)$ from the largest feature map.

The distribution of the generated latent space vector values is exhibited in Fig. 4.4. It can be observed that the elements of the Z are in the range of $[-20,20]$, with most values near 0. In practice, the dimension is decided by the feature extraction network.

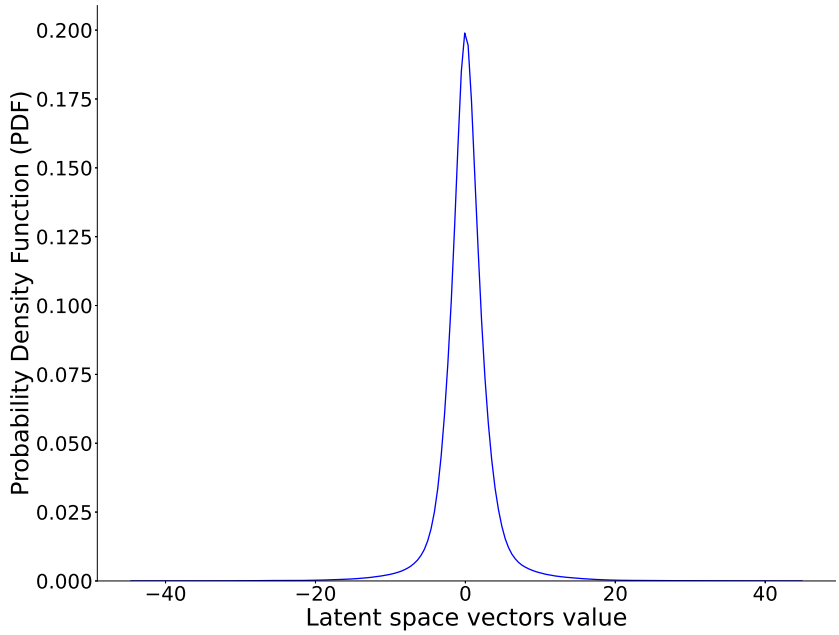


Figure 4.4: Distribution of the values in latent space feature vector Z.

Generally speaking, the quality of the recovered image improves if we have a higher

dimension of Z. However, it comes at the expense of more involved feature extraction of images.

Figure 4.5: Generated faces with proposed DP-Image method.

4.5.1.4 Image reconstruction using StyleGAN2

For the image reconstruction, we use the state-of-the-art StyleGAN2 [134] generator, which is trained on the FFHQ dataset. The parameters of the generator are fixed after the training process, and the output image is 1024×1024. It is worth noting that although

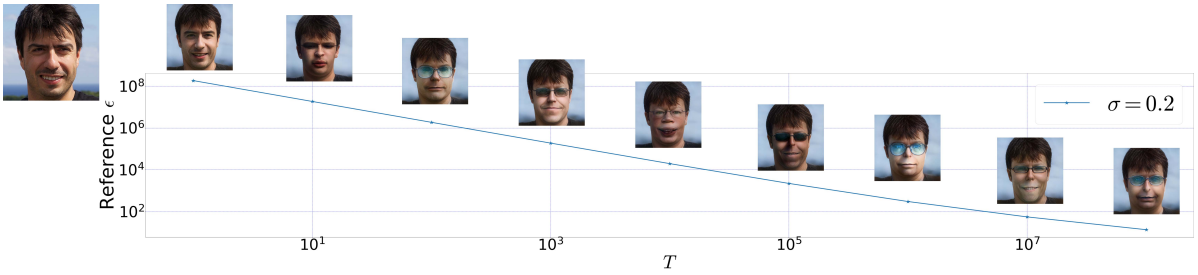


Figure 4.6: The DP-image visual results with reference (ϵ, δ)-DP. The x axis and y axis are the image iteration number and reference ϵ in logarithm. From left to right, they are original image and corresponding generated images with different T and same $\sigma = 0.2$. The sensitivity $\Delta f = 3840$. $\delta = 10^{-8}$ by the natural settings on deep learning.

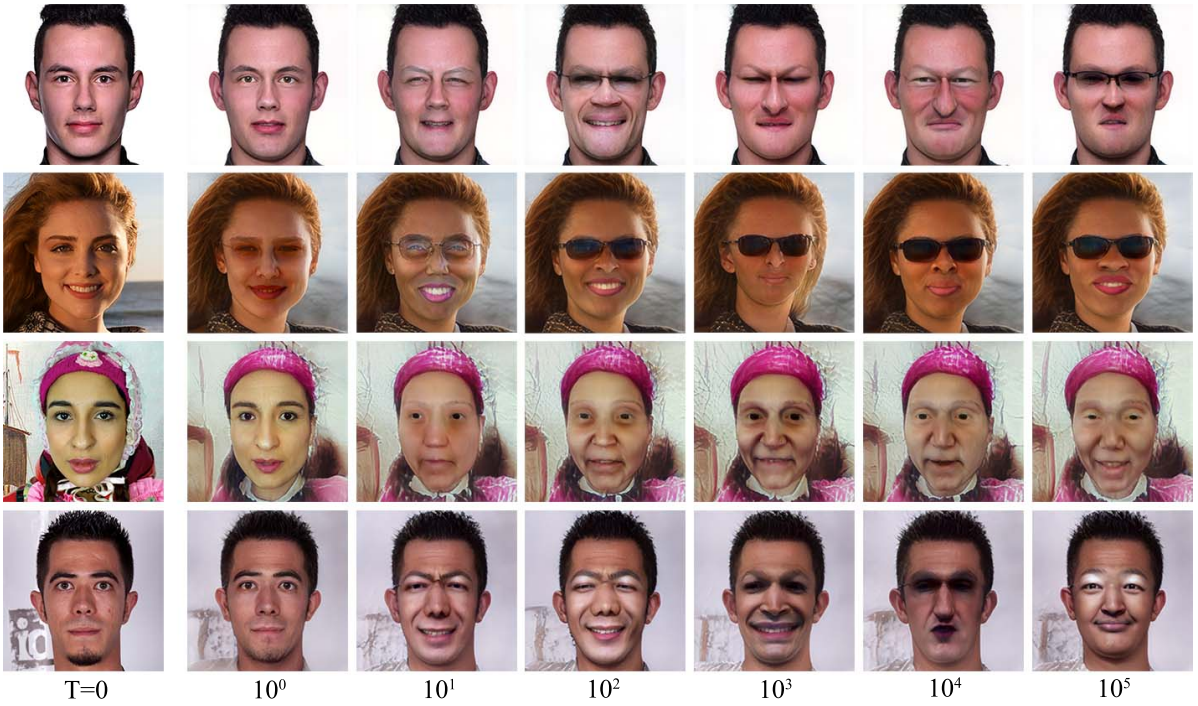


Figure 4.7: The image visual results for IDPI with different iteration numbers. Above, we give the possible visual results of face images under the $\sigma = 0.2$ setting. From left to right, the first column is original face images, and all other face images are generated with the different iteration numbers shown above.

StyleGAN2 can accept a simplified $(512, 1)$ vector as input, it will lose many fine details in high-resolution images [141]. Therefore, we use the $(512, 18)$ vector generated by the pSp framework in our experiments.

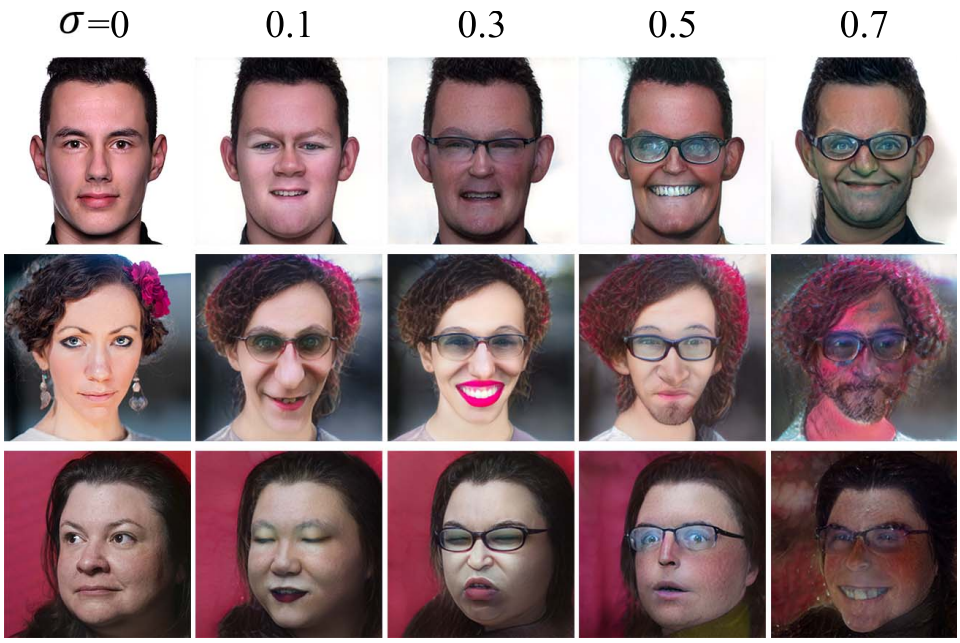


Figure 4.8: The image visual results for IDPI images with different σ. Here we give the possible visual results under $T = 10^2$ setting. From left to right, the first column is original images, and all other face images are generated with the different σ shown above.

Figure 4.9: The 20 image examples for the heat map in Fig. 4.10.

4.5.2 Performance of the DP-Image Mechanism

4.5.2.1 Sensitivity

The sensitivity is defined as the maximum element-wise difference between the feature vectors produced by two different images, as shown in Eq. (4.8).

In this scenario, as the encoder framework is non-convex and does not have a closed-form expression, Through observing the distribution of the feature space values (see Fig. 4.4), we get an empirical sensitivity by clipping the feature vector into the range $[-20, 20]$, which keeps more than 99% of the image features intact. In this case, according

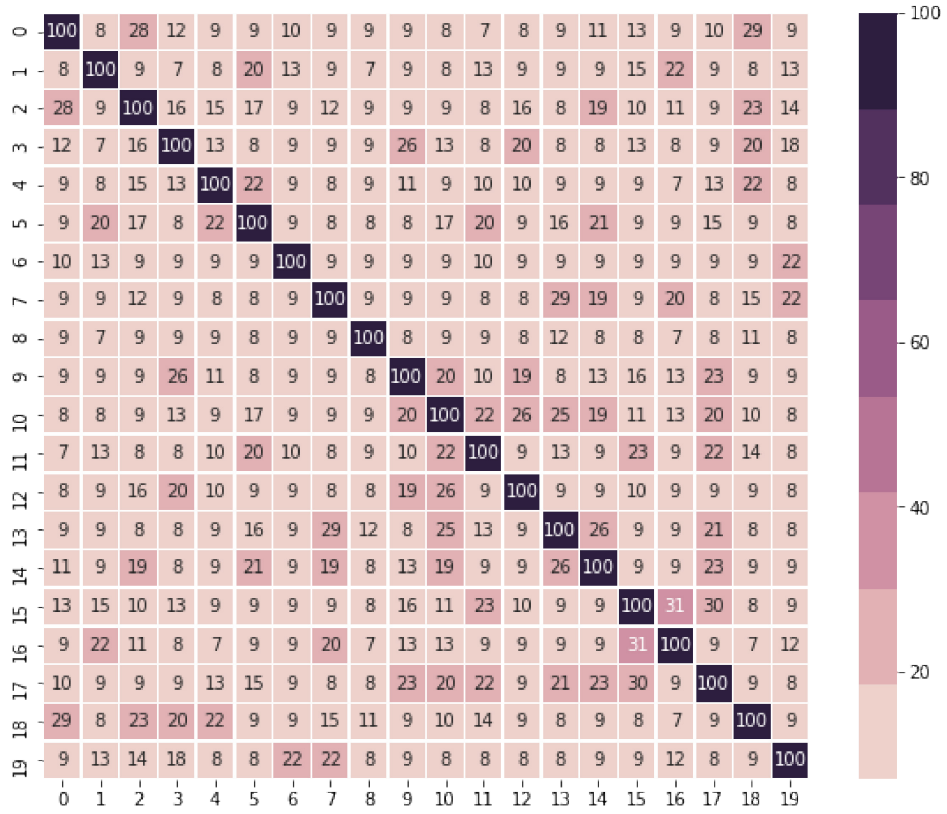


Figure 4.10: The confidence values of pair-wise image comparison for the 20 example images. Microsoft Face API holds high accuracy on face recognition. Any two faces with different identities will be given a lower score, e.g. confidence ≤ 50.

to Definition 4.4.2, the sensitivity Δf can be calculated:

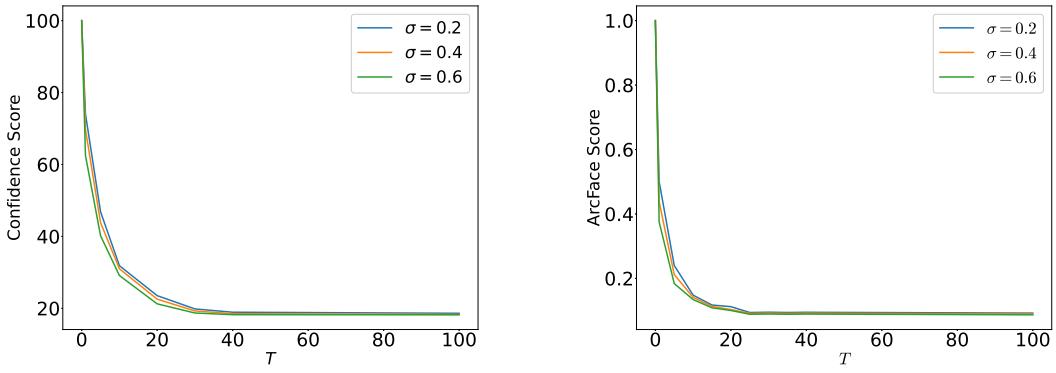
$$\Delta f \doteq \sup_{X,X'} ||f(X) - f(X')||_2 \quad \text{(Definition 4.4.2)} \tag{4.10}$$

$$= ||20\mathbb{I}_d, -20\mathbb{I}_d||_2 \tag{4.11}$$

$$= 3840. \tag{4.12}$$

Where $\mathbb{I}_d$ is an all-ones vector with the same dimension of image feature vector (shape (18,512) in this scheme).

Δf scales the noise added to the image features. By IDPI definition, with a large sensitivity, the image will need more iterations to achieve a smaller ϵ. However, from the utility perspective, the image could be more real-looking if we adopt a smaller clipping bound.



((a)) Microsoft Azure Face Recognition API confidence score

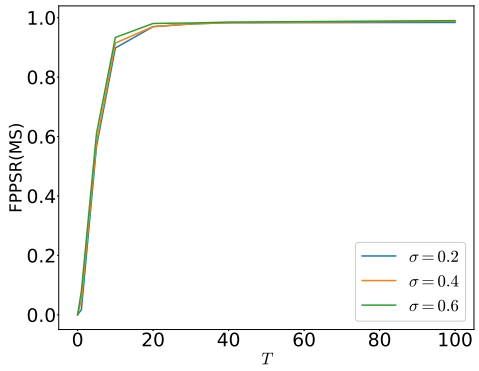
((b)) ArcFace score

Figure 4.11: We evaluate the DP-image performance with Microsoft Face API (a) and the ArcFace (b). (a) The x axis and y axis are iteration number T and Microsoft face API confidence score respectively. (b) The x axis and y axis are iteration number T and ArcFace score respectively. We set $\sigma = 0.2, \sigma = 0.4, \sigma = 0.6$ to compare the confidence score with different settings.

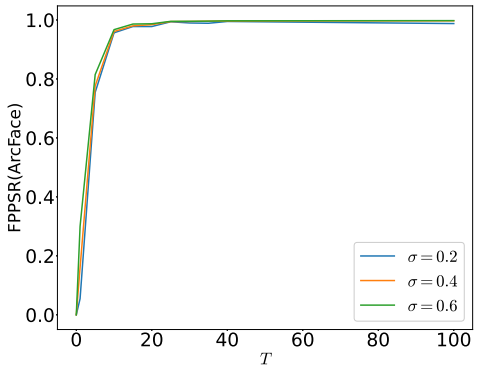
4.5.2.2 Privacy protection performance

To test the performance of our proposed algorithm, we randomly choose a few images from the dataset and apply our DP-Image scheme on the faces of the image. As shown in fig. 4.5, the original images are given on the top row, all the other four rows are the possible released privacy-preserved face images. All faces are generated under a setting of IDPI ($X_0, \sigma = 0.2, T = 10^4$). In fact, for each column, there is only one image generated from the original image above, i.e. the face identities of the second row are generated from the first face, the third row are generated from the second face, and so on. From the DP definition, all of the four proposed images have the probability of being the 'real' generated image. In other words, the adversary cannot distinguish the queried face by the naked eye, even though he is able to access the whole dataset.

Fig. 4.6 gives the reference ϵ under different IDPI iteration numbers. Next, we show a series of experiment results to display the visual influence of different IDPI settings. Fig. 4.7 shows the images with different iteration numbers, i.e. IDPI($X_0, \sigma = 0.2, T = [10^0, 10^1, 10^2, 10^3, 10^4, 10^5]$). In this case, we fixed the σ value to observe how the image changes as the iteration number T increases. With a larger iteration number, e.g. 10^5, the faces do not get much distortion. Moreover, the IDPI provides good and stable utility during the iteration.



((a)) FPPSR of Microsoft Azure Face Recognition API



((b)) FPPSR of ArcFace

Figure 4.12: We evaluate the DP-image performance with Microsoft Face API (a) and the ArcFace (b). (a) The x axis and y axis are iteration number T and FPPSR on Microsoft face API respectively. (b) The x axis and y axis are iteration number T and FPPSR on ArcFace respectively. We set $\sigma = 0.2, \sigma = 0.4, \sigma = 0.6$ to compare the FPPSR with a different setting.

Then, we fix the IDPI parameter T and adjust the privacy budget for each step σ. Fig. 4.8 shows the images with various σ values. The faces exhibit poorer utility with a higher σ. Because a larger perturbation in one step will lose too much feature information to maintain the utility of the semantic features. Thus, σ serves as a parameter to control the utility of the whole image.

To quantify the performance of our proposed scheme, we test the IDPI on the advanced DNN threat model. The face recognition models are used to measure the distance of image features. As the DP perturbation is added to the image feature space, the distance measurement on the feature space is an essential evaluation in our scheme. First, we use a commercial face recognition system Microsoft Face API as an attacker to evaluate our proposed algorithm. The Microsoft API can yield a confidence score for any pair of faces being the same person. In order to better understand the score of the API, we first generate the confidence score values of pair-wise image comparison for the 20 example images (see Fig. 4.9) in Fig. 4.10.

After that, we make experiments for the ISS metric. In Fig. 4.11(a), we generate DP-images on the FFHQ dataset under different settings and get the Microsoft API confidence scores. The average Microsoft API confidence scores decrease to lower than 50 within the first 5 iteration for all σ. Basically, Microsoft API uses a confidence score of 50 as a threshold for image identity.

Then we attack IDPI with one state-of-the-art face recognition system ArcFace. ArcFace uses DNN (e.g. Resnet) to extract image features and then measure the distance between different features by using Additive Angular Margin Loss. Nowadays, ArcFace is one of the top face recognition models in the deep learning area and being implemented in many applications. Our results (see Fig. 4.11(b))show that IDPI can reduce the average ArcFace score to lower than 0.31, which is the empirical threshold in using ArcFace as face identifier, within 5 iterations for all σ.

Besides, the average confidence scores and the ArcFace score are stable at around 18 and 0.09, respectively, in the following iterations, which indicates a way of choosing better IDPI parameters in practice.

Next, we conducted a quantitative evaluation of the dataset using the FPPSR privacy evaluation metric introduced in 4.5.1.2. We evaluate our method on both ArcFace and Microsoft Azure Face Recognition APIs. We make the same settings with the ISS metric and count the success rate by using the threshold introduced in previous experiments (i.e. Microsoft Azure Face Recognition API: 50, ArcFace: 0.31). Fig. 4.12 shows that the FPPSR was stable at over 99% for both models.

Table 4.1: Utility comparisons of the proposed privacy protection methods with traditional privacy protection methods.

	ISS(ArcFace 0.31)				ISS(Microsoft API 50)			
	l_2	ALD_∞	SSIM	FID	l_2	ALD_∞	SSIM	FID
Blur	71783	1.1502	0.5895	265.8	71533	1.1462	0.6029	220.9
Mosaic	70439	1.1386	0.5705	377.5	70439	1.1386	0.5705	377.5
Adversarial (pixel)	71657	1.1011	0.2014	391.1	71778	1.1029	0.2141	393.7
Ours	73383	1.2567	0.5050	**155.7**	73408	1.2531	0.5400	**152.6**

4.5.2.3 Utility

Table 4.1 compares the utilities of the proposed DP-Image with several traditional perturbation methods when they have approximately the same ISS. As ArcFace and Microsoft API perform a bit differently, we use both privacy metrics on these methods. We try to keep the ISS of ArcFace (i.e. 0.31) and Microsoft API (i.e. 50) at a similar level. It shows that the proposed DP-Image mechanism can well preserve the perceptual visualization of the image (lowest FID). On the other hand, the DP-Image method yields a similar performance in the sense of l_2, ALD_∞ and SSIM, as these metrics are evaluated on the pixel domain.

4.6 Discussions

4.6.1 Privacy Analysis

Our experiments in Section 4.5 show that it is feasible to achieve DP for images in feature space. And we can control the amount of noise and the appearance of reconstructed images by adjusting the IDPI parameters. Moreover, the noise added to the features can not be reverted back so as to reveal the original image, thus providing the capability to protect image privacy in a data-sharing scenario.

It is worth noting that although IDPI can theoretically achieve strong ϵ-DP (e.g. $\epsilon \leq 10$), the infinite iteration will cost vast of computing power, which naturally decreases the effectiveness of the image. In our experiment, only a relatively small amount of noise can cause the image identity to change. There are two major reasons for this phenomenon: 1) although using i.i.d random variable vector $\mathcal{N} = (N_1, ..., N_m)$ drawn from the Gaussian distribution can guarantee DP [79], the bound is quite loose, especially when the dimension m is high; 2) while the DP noise is added on the latent vector $f(X)$, we do not need to make $\mathcal{M}(X)$ completely indistinguishable from $\mathcal{M}(X')$ in practice. Our target is to make the identities of reconstructed images from X and X' close to each other, and hence a small variation on the feature vector is enough.

The ultimate purpose of image privacy protection is to prevent de-identification, i.e., removing personal identifiers in the image. To achieve this target, we need to consider two different types of adversaries: human and AI. For a human adversary, the generated image needs to satisfy two conditions: 1) the appearance of the image has changed; 2) the new image needs to be realistic and natural, so that an adversary who has not seen the original image does not know that the new image is a generated instance. In this sense, our proposed DP-Image method that manipulates images in the feature space is an appealing option. For the AI adversary, it does not look at the appearance, but the features instead. In this case, the ISS and FPPSR are the primary performance metrics. And our experiments show that the DP-Image scheme that adds random noises can generate real-looking face images.

4.6.2 Disentangle Identity-related Features in Image

One interesting discussion is that can we add noises only to the identity-related features in the feature space. To answer this question, the precondition is that we are able to extract the identity-related features from images.

In terms of the pSp framework that we use as the encoder, we found that the style vectors $(1-7)$ are more identity-related (representing the structure of the faces), while the rest of the style vectors $(8-18)$ are not highly related to identity (representing colour, pose, etc.) Therefore, we investigate the case of adding noise only to identity-related feature vectors of Z, which are layers $(1-7)$.

As shown in Fig. 4.13, adding noise only to identity-related features in Z could greatly reduce the amount of sensitivity (i.e. from 3840 to 2394 by definition). In the meantime, the effectiveness of privacy persevering keeps the same level.

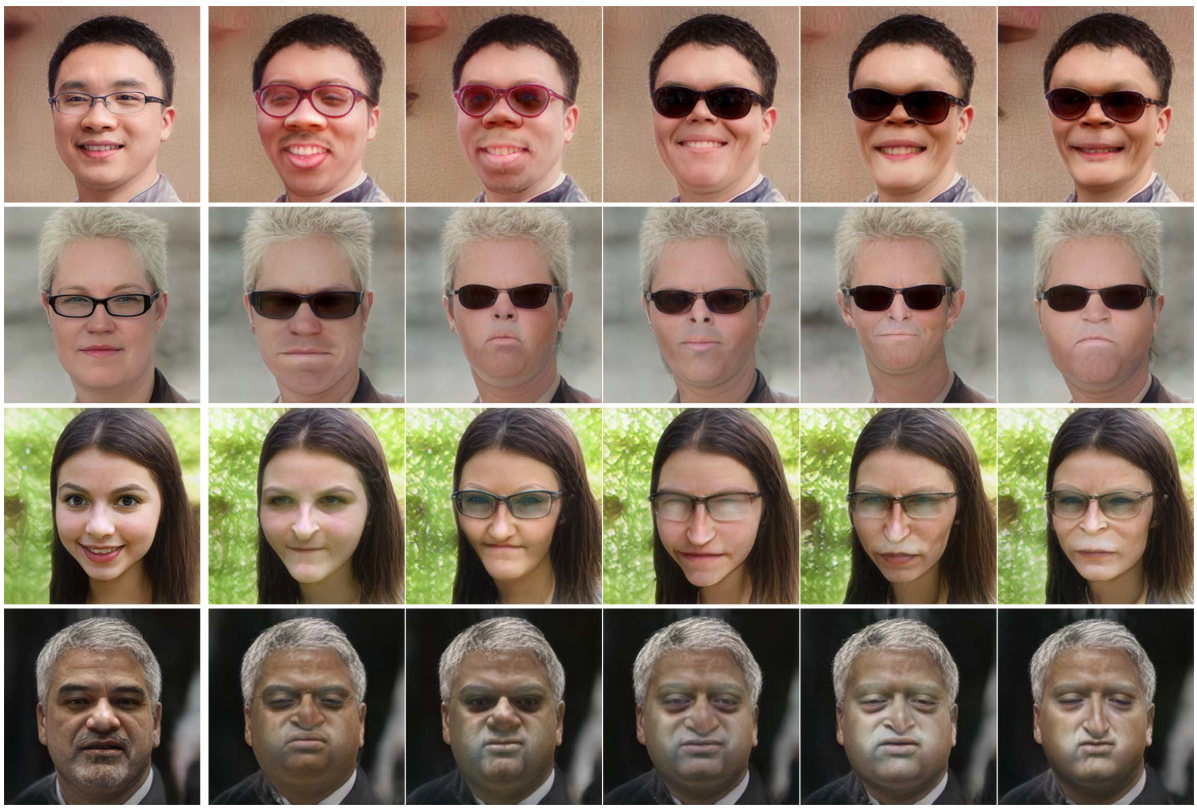


Figure 4.13: Generated faces with noises added to identity-related features of Z. We give the possible visual results of adding noise on image identity features. From left to right, they are original images, the faces protected by IDPI($X_0, \sigma = 0.2, T = 10$), IDPI($X_0, \sigma = 0.2, T = 20$), IDPI($X_0, \sigma = 0.2, T = 30$), IDPI($X_0, \sigma = 0.2, T = 40$), IDPI($X_0, \sigma = 0.2, T = 50$)

The pSp framework is not specifically designed for extracting identity-related features, therefore the discussions in this subsection are preliminary. The extraction of identity-related features is a topic of its own that attracts many recent research interests. If we can combine the latest extraction technique with state-of-the-art image autoencoders in the future, it will further enhance the performance of the proposed DP-Image mechanism.

4.6.3 Image Privacy vs. Database Privacy

From the experiments, we can see that image privacy is quite different from conventional data privacy.

1. The protection target of database privacy is the existence of a certain record in the database or not; while image privacy protection targets the identity-related information in the image.

2. It is easy to add random noise to the data record and generate a meaningful DP perturbed record. However, for an image, adding random noise might lead to distorted and unrealistic pictures.

3. The performance of DP-Image needs to be judged by both visual and numerical metrics, as we may need to consider both humans and AI as adversaries.

4.7 Related Work

As much of the related work on differential privacy and deep learning has already been discussed in Section 4.3, here we only focus on the works specifically on image privacy protection.

The image privacy issue first attracted people's attention, along with the booming of social networks developing. The proliferation of social networks generated massive photos flooding the internet that contains sensitive information. For example, Pesce et al. [142] use photo tags to attack users and get their privacy. The image privacy issue becomes more server with the widely spread of facial recognition systems, as people start to worry that organizations might use their faces for profiling or social control.

The previous mainstream method to combat the image privacy attack is using access control on sensitive content. For example, Vyas et al. [143] use annotation data to predict the privacy preferences of users and control the shared content. Moreover, Squicciarini et al. [144] proposed collaborative privacy management that can let users collaborative control their photos. Similarly, to deal with the privacy issue in facial recognition systems, the current countermeasure is simply banning [145]. The access control-based method has several limitations. It only gives "Yes" or "No" options for the use of images, while we need to use part of the information in applications such as Google Street View. And it cannot automate privacy protection based on the privacy information of the image itself, requiring human participation.

Some more recent image privacy researches focus on changing the identity-related information in images. The main techniques are obfuscation and inpainting. Simple obfuscation has been proved to be ineffective against DNN-based recognizer [107, 146]. In order to improve the robustness against deep learning aided attacks, adversarial example-based methods have been proposed to mislead the neural networks [7, 101, 102, 109–111]. Also, there are some researchers who start to use GAN to generate content to replace the sensitive information in the images [112–116]. For example, Sun et al. [112] proposed GAN-based head inpainting to remove the original identity. Finally, there have recently been a few attempts to combine the DP notion with image privacy. Fan [80] proposed an ϵ- differential private method in the pixel level of the image. However, making image pixels indistinguishable does not make much sense in practice, and the quality of the generated image is quite low. In another work from the same author [81], metric privacy is defined in the image transformation domain, but the quality of the obfuscated image is still low. Lecuyer et al. [82] proposed the PixelDP framework that includes a DP noise layer in the DNN. PixelDP scheme enforces that the output prediction function is DP provided the input changes on a small number of pixels (when the input is an image). However, the purpose of PixelDP is to increase robustness to adversarial examples, other than image privacy.

4.8 Conclusion and Future Work

In this chapter, we have proposed a DP-Image framework IDPI that can protect sensitive personal information in images. The major contributions are two-fold. First, we present the novel notion of DP-Image. Second, we propose a method to perturb the image by adding noise to its feature vector in the latent space. The effectiveness of the proposed framework is validated by extensive experiments on the face image dataset FFHQ.

Image privacy and the broader topic of unstructured data privacy is an interesting research direction with many open problems.

- A more comprehensive image feature extraction network that can be applied to different types of images and even videos.
- Investigate how the image feature vector decides the various semantic attributes of images and whether we can cleanly extract these attributes.
- Based on the above understanding, we may be able to concisely control some of the attributes which we believe are crucial from the privacy perspective.

- Investigate DP-Image mechanisms in more complicated images that contain multiple sensitive objects.

In the upcoming chapter, we will focus on comprehending the specific manifestations of various attributes within the semantic space of images, alongside precise control over the "clean" protection efficacy linked to distinct features. Specifically, we will classify the latent attributes inherent in images into privacy and non-privacy, thereby facilitating precise protection of the privacy aspects for users. Undeniably, this augmentation significantly enhances both the images' usability and the overall user experience.

CHAPTER 5

Face Image De-identification by Feature Space Adversarial Perturbation (FSAP)

5.1 Preface

In this chapter, our exploration centres around a privacy protection framework that prioritizes usability. We address the limitations of existing strategies, which often overlook the impact on image usability by indiscriminately distorting information. To mitigate this issue, we propose a protective strategy that differentiates and preserves various types of information within images. Our framework optimizes the protection of privacy-sensitive information in the semantic space of images while striving to retain non-private information to the fullest extent possible. To evaluate the effectiveness of our framework, we leverage a facial dataset with a specific focus on individual identity as privacy-sensitive information. Through rigorous experimentation on this dataset, we demonstrate the efficacy of our framework and highlight its exceptional performance compared to other state-of-the-art methods.

5.2 Introduction

5.2.1 Motivation

The rapid development of Computer Vision (CV) technology has recently made the automatic processing of large-scale visual data prevalent. However, those visual data contain a large amount of personal information, leading to inadvertent disclosure of an individual's privacy. While we enjoy the benefits of advanced CV technology, including camera surveillance and video conferencing, we are reluctant to surrender our privacy and in-dignify ourselves as manipulable data records. In addition, the information that people share on the Internet is facing more powerful malicious attackers than ever before. The traditional privacy-preserving methods are less effective against deep learning-based attacking models. Therefore, new privacy preservation methods are urgently needed.

Various defence techniques and mechanisms have been proposed to enhance privacy by de-identifying face images [147]. Traditional obfuscation-based methods usually obfuscate sensitive information by blurring, pixelating, or masking an image, which are illustrated in Fig. 5.1. However, the image processed by these traditional methods can be easily and accurately detected by deep learning models. Therefore, new defence techniques and mechanisms have been designed to protect image privacy from the deep learning models, including face identity transformer, differential privacy (DP), GAN-based inpainting, and adversarial examples (AEs), etc. [132, 134, 148–153]. The face identity transformer was proposed in [148], which can perform automatic photo-

realistic password-based anonymization and deanonymization of human faces. DP-based methods can provide provable privacy guarantees but produce lower-quality images by including DP noises to the original image or in the transformed domain of the image [149]. GAN-based inpainting was proposed to generate content to cover the sensitive information or identity of the image without degrading the quality of the original image [150]. The conditional GAN (CGAN)-based method was designed by adding labels to the generator and the discriminator for better network training [132]. Conditional Identity Anonymization Generative Adversarial Network (CIAGAN) [9] and DeepPrivacy [8], both based on CGAN, proposed to add AE noises in the feature space of the images. CIAGAN can de-identify faces and bodies while generating high-quality images and videos. DeepPrivacy can generate an image with considerations of both pose and background. Nevertheless, both CIAGAN [9] and DeepPrivacy [8] add the latent noise in a vague direction, without considering the specific features of the image.

Considering deep learning-based privacy attacks, adversarial examples (AE)-based protection methods have great potential. The earliest research on AE was proposed by Szegedy *et al.* [154]. It was shown that a small perturbation could have a considerable and negative impact on the accuracy of deep neural networks. In a recent paper [99], the author designed an adversarial example attack against deep neural networks to mislead the classifier. It was revealed that the deep neural network is different from humans in the task of image classification, and AE is an efficient method to generate noise on images that affects the deep neural networks. Many subsequent studies have focused on adversarial noise in different settings, such as adversarial noise for the convolutional neural network (CNN), deep neural network (DNN), recurrent neural network (RNN), and more robust adversarial noise [155]. There is an arms race-like relationship between attack and defence technologies in such circumstances. One of the major issues of AE is its transferability, i.e., its effectiveness on alternative black-box or unknown models. To improve AE noise's transferability, some papers [156, 157] have transferred the calculation of noise direction from the output layer to the intermediate layer of the model. This can avoid the differences between models, thereby increasing transferability. Pidhorskyi *et al.* [158] studied the potential of adding adversarial perturbations on the feature level of images.

Most of the above existing work treated AE as a threat to system security. Only very recently, researchers started to use adversarial examples (also called adversarial perturbations) as a method to protect image privacy [7, 100, 101, 109–111]. However, these methods either focus on the utility [7, 100, 101, 109, 110] or focus on the privacy

protection [111], which is hard for users to choose a good trade-off.

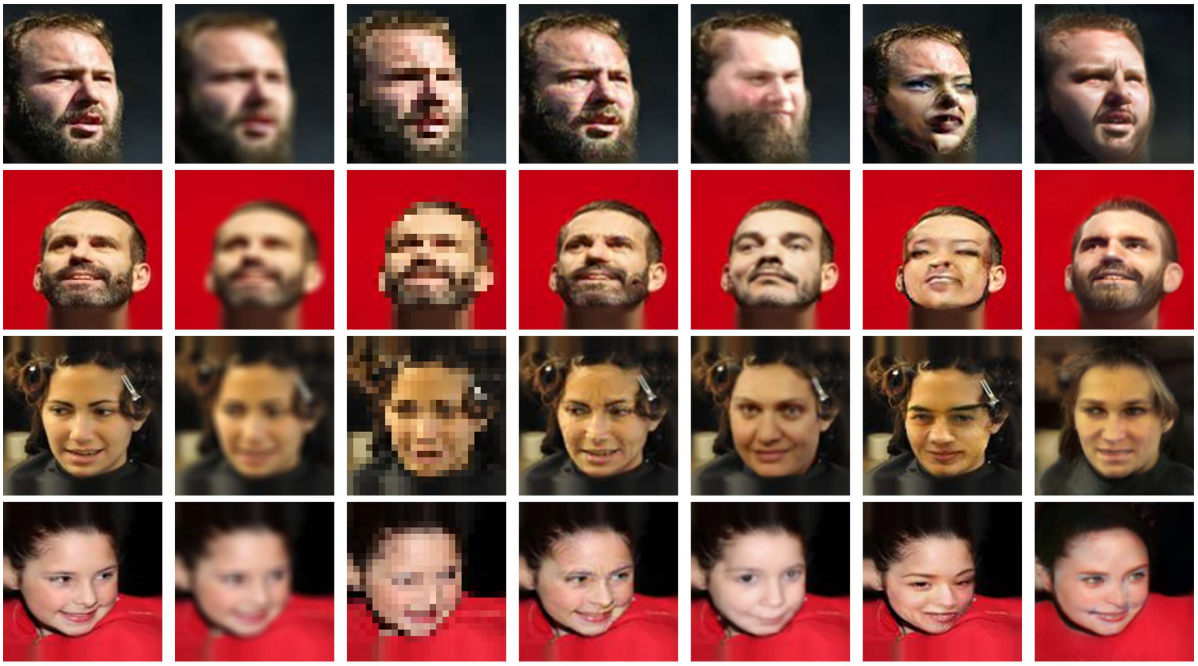
5.2.2 Contributions

To overcome the problems mentioned above, we propose a novel face image de-identification framework, where latent noise is generated based on the gradient directions concerning the identity and the attributes of an individual face image, which can accurately de-identify the image. Moreover, we use a pre-trained model as a decoder that can map the perturbed feature vector back to an image (i.e., the generated AE). The main contributions of our work are summarized as follows:

- We propose a novel face image de-identification framework based on feature space adversarial perturbations referred to as the FSAP framework for short. This framework can preserve face identity information against automated recognition by DNNs while keeping a high utility of the image.
- We propose a feature space adversarial perturbation generation algorithm. By alternative updating the perturbation according to the specially designed ID loss and attribute loss items, we can successfully direct the noise to identity-related information while ensuring the other attributes remain similar. In addition, feature space manipulation can provide good transferability of the generated perturbation. Furthermore, users can select a trade-off between privacy and utility according to their own needs by adjusting the parameters.
- We implement the proposed image protection framework and methods on a real-life image dataset and show its effectiveness in safeguarding people's privacy.

5.2.3 Overview of the work

The structure of the remaining sections in this chapter is as follows. In Section 5.3, we provide an overview of the related work in the field, covering image privacy protection methods that utilize obfuscation, GAN-based inpainting, differential privacy (DP), and adversarial examples (AEs). Moving on to Section 5.4, we formally define the problem and subsequently present our proposed FSAP (Facial dataset anonymization with Semantic-driven Adversarial Perturbation) protection mechanisms. In Section 5.5, we conduct a series of experiments to assess the effectiveness of our proposed scheme. This includes an ablation study as well as a comparison against other state-of-the-art methods. In



(a) Original (b) Blur (c) Pixelation (d) Fawkes (e) DeepPrivacy (f) CIAGAN (g) Ours

Figure 5.1: Face deidentification results. From left to right, (a) Original image; (b) Blur noise; (c) Pixelation noise; (d) Fawkes [7]; (e) DeepPrivacy [8]; (f) CIAGAN [9]; (g) feature space adversarial perturbation (**Ours**).

the concluding part of Section 5.5, we discuss the limitations of GANs in addressing the problem at hand and outline potential avenues for future work.

5.3 Related Work

Recent image privacy researches [159–161] have focused on altering identity-related information in images via different methods, including obfuscation, GAN-based inpainting, differential privacy (DP) and adversarial examples (AEs).

The main techniques under investigation are obfuscation and inpainting. Simple obfuscation has been shown to be ineffective against DNN-based recognizers [107, 146]. Therefore, some researchers have started to use GAN and AE to generate content to replace the sensitive information in the images [112–116, 162]. For example, Sun *et al.* [112] proposed GAN-based head inpainting to remove the original identities. Hukkelås *et al.* [8] proposed a CGAN-based architecture to anonymize faces without destroying the data distribution of the original image. Deb [163] proposed a framework to generate face masks based on GAN and AE techniques. Valeriia [164] proposed a method to optimize the AE method in privacy protection. Zhang [165] proposed a face protection framework against DNN by filtering the AE methods.

Furthermore, there have been some recent attempts to combine the DP with image privacy [166]. Fan [80] proposed an ϵ-differential private method at the pixel level of the image. However, making image pixels indistinguishable does not make much sense in practice, and the generated images are of low quality. In another work from the same author [81], metric privacy is defined in the image transformation domain, but the obfuscated images are still of low quality. Lecuyer *et al.* [82] proposed the PixelDP framework, which includes a DP noise layer in the DNN. The PixelDP scheme forces the output prediction function to be DP, provided that the input changes on a small number of pixels (when the input is an image). However, the purpose of PixelDP is to increase robustness to adversarial examples rather than image privacy. Chen *et al.* [149] proposed a variant of DP by considering a perceptual similarity of the facial images, named perceptual indistinguishability (PI)-Net, which can achieve image obfuscation while ensuring PI.

To achieve a good trade-off between privacy and accessibility for face de-identification, reversible privacy protection has been studied in the literature [167–169]. Pan *et al.* [167] proposed a Multi-factor Modifier (MfM) based on conditional encoder and decoder framework, which achieves multi-factor facial de/re-identification. Based on a deep generative model, a personalized and reversible de-identification method was designed in [168] to control the direction and degree of identity change by introducing a user-specific password and an adjustable parameter. You *et al.* [169] proposed a reversible privacy

protection framework with an encoder and decoder using U-Net architecture to generate high-quality protected images without visible facial features. The original images are encoded with embedded face information before uploading onto the cloud.

Video-related de-identification has also been investigated in [170–172]. Unlike the face image de-identity, it requires to be modified seamlessly without causing any visual distortions or artifacts. In [170], a multi-task extension of GAN was formulated to eliminate privacy-sensitive information of a video and detect privacy-preserving actions. In [171], a feed-forward encoder-decoder network architecture was proposed conditioned on the high-level representation of a facial image. By coupling the latent space of the auto-encoder with a trained classifier network, a rich latent space with embedded identity and attribute information can be achieved.

Compared to the existing method for face image de-identification, to the best of our knowledge, ours is the first method that generates feature space AE noise in an optimization style. And compared to the state-of-the-art techniques, ours achieves compelling results in privacy, utility, and transferability.

5.4 Feature Space Adversarial Perturbation Based Face Image De-identification Framework

In this section, we elaborate on the design of the proposed Feature Space Adversarial Perturbation (FSAP) based image de-identification framework. Under this framework, we further propose privacy protection methods against CNN face recognition.

5.4.1 Problem Formulation

Let $x \in \mathbb{R}^{h \times w \times c}$ denote a face image with c channels, each having a size of height h and width w. A CNN encoder $f_E(x)$ can generate a latent vector W of the face image x and a decoder $f_D(W)$ can reconstruct the face feature W to the output face image $\hat{x} \in \mathbb{R}^{h \times w \times c}$. If both the encoder and decoder are ideal, we should have $\hat{x} = x$, i.e., $f_D(f_E(x)) = x$.

Our ultimate goal is to find a noise ΔW in the latent space that can perturb the identity of the input face image such that the output face image $\hat{x}$ has the following characteristics:

- **De-identification.** The perturbed image $\hat{x}$ is likely to be recognized as a different face from the input image x by an arbitrary CNN face recognizer $f_I(\cdot)$, i.e., $\mathcal{L}_I(x, \hat{x}; f_I) > \tau$, where $\mathcal{L}_I(\cdot)$ indicates the identity loss. τ denotes the threshold of the two face images being recognized as different identities.
- **Maintaining the Utility.** While targeting de-identifying the face identity, the perturbed image $\hat{x}$ should suffer from the minimum attribute loss to the input image x, i.e., $\min \mathcal{L}_P(x, \hat{x})$, where $\mathcal{L}_P(\cdot)$ indicates the attribute loss. To the naked eye, the perturbed image $\hat{x}$ should have similar properties to the input image x, so that it would be difficult for humans to distinguish which image is real or an impostor.

To summarize this process into an optimization problem, we have

$$\begin{aligned} &\min \ \mathcal{L}_P(x, \hat{x}) \\ &\text{s.t. } \mathcal{L}_I(x, \hat{x}; f_I) > \tau. \end{aligned} \tag{5.1}$$

The optimization problem (5.1) tries to maximize the dissimilarity of the face image identities, while minimizing the similarity of the face image attributes. To achieve this goal, we design novel architecture, referred to as the Feature Space Adversarial Perturbation Based Face Image De-identification Framework (FSAP Framework for short).

5.4.2 FSAP Framework

The framework for generating feature-level adversarial examples is shown in Fig. 5.2, which consists of three stages: (1) Stage 1 - encodes the input image x into the latent vector W by using a pre-trained CNN, i.e., f_E. (2) Stage 2 - updates the image latent vector W' by adding adversarial perturbation iteratively. (3) Stage 3 - the output image $\hat{x}$ is reconstructed from W' by a pretrained decoder, i.e., f_D. The output image $\hat{x}$ is an image that does not contain the identity information of the original image.

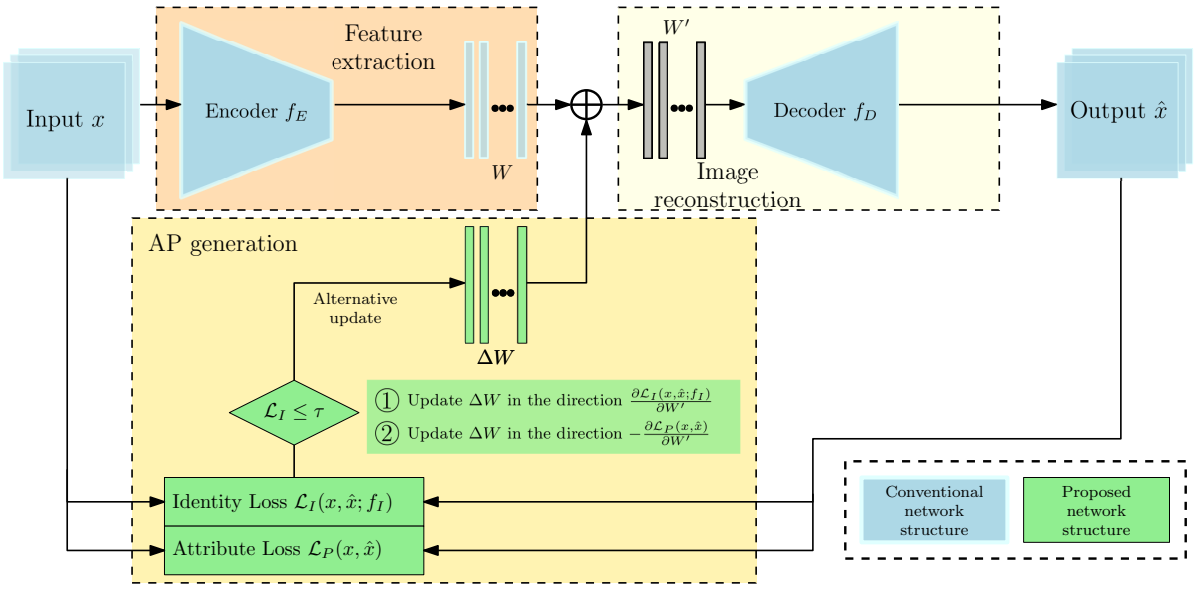


Figure 5.2: Feature Space Adversarial Perturbation (FSAP) based privacy protection framework.

5.4.2.1 Feature extraction.

In order to extract the face image features, we adopt the pixel2style2pixel (pSp) [141] encoding framework, which can be used to solve various image-to-image translation tasks and is compatible with StyleGAN2 architecture. We use the intermediate layer between the encoder and the decoder as our latent code, denoted as W. Here, W contains 18 style vectors, with each vector of length 512. The encoder extracts the feature maps of the input image in three levels (low, medium, high) through a typical CNN (e.g., ResNet). These feature maps then were mapped to the latent vector. The process can be formulated

as:

$$W = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_6 \\ z_7 \\ \vdots \\ z_{18} \end{bmatrix} = \begin{bmatrix} s_1(k_{high}(x)) \\ s_2(k_{high}(x)) \\ s_3(k_{medium}(x), k_{high}(x)) \\ \vdots \\ s_6(k_{medium}(x), k_{high}(x)) \\ s_7(k_{low}(x), k_{medium}(x), k_{high}(x)) \\ \vdots \\ s_{18}(k_{low}(x), k_{medium}(x), k_{high}(x)) \end{bmatrix} = f_E(x), \tag{5.2}$$

where $s_n(k), n \in [1,2,\ldots,18]$ is a fully convolutional network to map the feature maps k into the style vector z_n. The feature maps k have three different dimensions, i.e., $\dim(k_{high}) < \dim(k_{medium}) < \dim(k_{low})$, and are built with a nested structure. $z_n \in \mathbb{R}^{512}, n \in [1,2,\ldots,18]$ are the style vectors corresponding to the 18 layers of the latent vector W. The input image x is an RGB face image, and f_E is the encoder.

5.4.2.2 Adversarial perturbation generation.

The latent vector W can be used to ideally reconstruct an output image $\hat{x}$ that is close to the original image x. We now start to train an AP to change the face identity, i.e., train an AP (i.e., ΔW) that can generate a face ID loss above the threshold, while minimizing the face attribute loss.

The two loss items are calculated as described below:

- *ID Loss.* The ID loss $\mathcal{L}_I(x, \hat{x}; f_I)$ is to measure the identity similarity of the two faces. This loss function maps the input image x and the output image $\hat{x}$ into the face feature space, also known as face embedding. We adopt the most widely used method cosine similarity to compute the face embedding loss (i.e., ID loss), which is defined as:

$$\mathcal{L}_I(x, \hat{x}; f_I) = 1 - \frac{f_I(x) \cdot f_I(\hat{x})}{\left\| f_I(x) \right\|_2 \cdot \left\| f_I(\hat{x}) \right\|_2}. \tag{5.3}$$

- *Attribute Loss.* The attribute loss $\mathcal{L}_P(x, \hat{x})$ measures the attribute similarity of the two faces. The loss function is a combination of three typical losses, including MSE ($\mathcal{L}_m$), LPIPS [173] ($\mathcal{L}_l$), and SSIM ($\mathcal{L}_s$), and is defined as follows:

$$\mathcal{L}_P(x, \hat{x}) := \{\lambda_1 \mathcal{L}_m(x, \hat{x}), \lambda_2 \mathcal{L}_l(x, \hat{x}), \lambda_3 \mathcal{L}_s(x, \hat{x})\}. \tag{5.4}$$

MSE ($\mathcal{L}_m$), Mean Square Error, takes the pixel loss of the input and output images, which controls the amount of noise added to the image. LPIPS [173] ($\mathcal{L}_l$), Learned Perceptual Image Patch Similarity, takes the perceptual loss from the perceptual latent distance of the input and output images and measures the perceptual similarity of the images, which controls the visual quality of the image. SSIM ($\mathcal{L}_s$), Structural Similarity Index Measure, controls the structural similarity of two images. The combination of these three losses can guarantee the utility of the image from different levels.

The details of the AP generation algorithm will be described in Subsection 5.4.3.

5.4.2.3 Image reconstruction.

We adopt the StyleGAN2 [134] synthesis network as the generator. Unlike a traditional decoder, which uses the latent vector as the bottom layer of the network, StyleGAN2 generates the images from a constant vector. The latent vectors were fed to 18 layers of the network to affect the identity of the face. In order to de-identify the face image, the perturbed latent vector $W' = W + \Delta W$ is thus fed to each layer of the synthesis network as well. The process can be formulated as follows:

$$\hat{x} = f_D(W'), \tag{5.5}$$

where f_D is the reconstruction network that decodes the modified latent vector W' back to the RGB face image.

5.4.3 Adversarial Perturbation Generation Process

Traditional adversarial perturbations can be grouped into two major categories: target and non-target. Target attacks require that the model classifier misclassifies the AEs to a specific class for malicious purposes. The non-target attacks only require AEs to be misclassified with any wrong label to avoid detection. In the context of privacy protection, non-target AP just pushes the image away from the current identity. Therefore, it is a better option than target AP.

In addition, we want to minimize the AP so that the utility of the image can be kept as much as possible. While this is often a non-convex optimization problem. Some approximate methods have been developed. The fast gradient sign method (FGSM) proposed by Goodfellow [99] is a widely adopted method of generating AE/AP. The AP generated by FGSM and its variants is superior to other traditional methods when facing the white-box model. However, traditional FGSM has less transferability when

facing the black-box model. Some studies [156, 157] have found that adding noise to the feature space can improve the transferability of AP. The argument was that the existing recognition networks generally map the pictures to latent space vectors through the CNN to recognize images. Therefore, the noise added to the latent space vector will have better transferability. To sum up, our proposed method adds noise to the latent vector of the input image x in a non-target manner.

Also, our method differs from the conventional GAN-Based methods that add the latent noise in a vague direction. The conventional GAN-Based methods use a large dataset to train the network and one network to process all the data. Even with a large dataset and time-consuming training, the generalisation ability of the network is inversely proportional to the accuracy. Whereas our method can accurately add latent noise on the identity-related information of an individual image. The proposed latent noise is generated based on the gradient direction with regard to the two losses described before, i.e. $\mathcal{L}_I$ and $\mathcal{L}_P$.

The perturbation ΔW takes the update from the following two losses alternately:

$$\Delta W_1 = \lambda_I \text{sign}\left(\frac{\partial \mathcal{L}_I(x,\hat{x};f_I)}{\partial W'}\right); \tag{5.6}$$

$$\Delta W_2 = -\left(\lambda_1 \text{sign}\left(\frac{\partial \mathcal{L}_m(x,\hat{x})}{\partial W'}\right) + \lambda_2 \text{sign}\left(\frac{\partial \mathcal{L}_l(x,\hat{x})}{\partial W'}\right) + \lambda_3 \text{sign}\left(\frac{\partial \mathcal{L}_s(x,\hat{x})}{\partial W'}\right)\right), \tag{5.7}$$

where ΔW_1 and ΔW_2 are the latent noises regarding the identity and attributes of the input image, respectively. The first loss function, $\mathcal{L}_I(x,\hat{x};f_I)$, measures the face identity dissimilarity between the input image x and the output image $\hat{x}$ with an arbitrary CNN face recognizer f_I. The second loss function, $\mathcal{L}_P(x,\hat{x})$, computes the face attribute loss between the input image x and the output image $\hat{x}$. When updating, the gradient is accumulated on the potential identity free face latent vector $W' = W + \Delta W$. It is worth noting that to simplify the experiments, we use the same attribute update rate λ_P to replace λ_1, λ_2, and λ_3. ΔW_2 can be rewritten to:

$$\Delta W_2 = -\lambda_P\left(\text{sign}\left(\frac{\partial \mathcal{L}_m(x,\hat{x})}{\partial W'}\right) + \text{sign}\left(\frac{\partial \mathcal{L}_l(x,\hat{x})}{\partial W'}\right) + \text{sign}\left(\frac{\partial \mathcal{L}_s(x,\hat{x})}{\partial W'}\right)\right), \tag{5.8}$$

The AP generation algorithm is shown in Algorithm 7.

Algorithm 7: AP generation algorithm based on the FSAP framework.

Parameters: ID update rate: λ_I; Attribute update rate: λ_P; Maximum iteration number: N; Adjustable iteration number: k; ID distance threshold: τ.
Input: The original image x.
Output: The released privacy-preserving image $\hat{x}$.
Stage 1:
Obtain the latent vector $W = f_E(x)$.
Stage 2:
Initialization: latent noise $\Delta W = 0$; Perturbed latent vector $W'_0 = W$.
for $1 \leq n \leq N$ **do**
  $\hat{x}_n = f_D(W'_n)$;
  $W'_n = W'_n + \Delta W_1$;
  for $1 \leq i \leq k$ **do**
    $\hat{x}_i = f_D(W'_n)$;
    $W'_n = W'_n + \Delta W_2$;
    $i = i + 1$;
  end
  if $\mathcal{L}_I(x, \hat{x}_n; f_I) > \tau$ **then**
    Break
  end
  $n = n + 1$;
  $\hat{x} = \hat{x}_n$.
end

5.5 Experiments

In this section, we carry out extensive experiments to verify the effectiveness of the proposed method. We also compare our method with the state-of-the-art face de-identification methods, i.e., GAN-based: CIAGAN [9], DeepPrivacy [8], and AP-based: Fawkes [7].

5.5.1 Experiment Setup

5.5.1.1 Dataset.

In this experiment, human faces are selected as the object of image privacy protection because they contain a large amount of identifiable information and have been the main concern in the field of image privacy protection. The face images for our experiments come from two well-known public face image datasets, i.e., FFHQ [137] and CelebA [174].

- The FFHQ dataset contains 70,000 high-quality PNG images with a resolution of 1024×1024 and considerable variation in terms of age, ethnicity, and image

background.

- The CelebA dataset contains 202,599 face images covering large pose variations and background clutter.

5.5.1.2 Experimental settings.

In this work, we adopt a pSp encoder [141] pre-trained on the FFHQ dataset for feature extraction. The ID Discriminator f_I used is pre-trained on the state-of-the-art face recognition network Resnet [175] with Arcface Loss [138] on the real-life dataset. The synthesis network of StyleGAN2 [134] is pre-trained on the FFHQ dataset.

Parameter settings: $N = 100$, $k = 6$, $\tau = 0.8$, $\lambda_I = 0.02$, $\lambda_P = 0.008$. τ is the threshold of the ID distance.

5.5.1.3 Evaluation Metrics.

The following methods will be used to measure the proposed algorithm:

- *De-identification:*

  Successful protection rate (**SPR**): We define successful protection as:

$$I(x,\hat{x}) > \delta, \tag{5.9}$$

  where $I(x,\hat{x})$ is the identity distance calculated on the identifier I. δ is the threshold that recognizes the face as a different identity. SPR is then formulated as:

$$SPR = \frac{1}{m}\sum_{i=1}^{m} g_P(x_i), \tag{5.10}$$

  where

$$g_P(x) = \begin{cases} 1, & \text{if } I(x,\hat{x}) > \delta, \\ 0, & \text{otherwise,} \end{cases} \tag{5.11}$$

  with m being the number of tests.

- *Utility:*

  1) Successful detection rate (**SDR**): We defined SDR as the de-identification rate of face images that can be detected. It evaluates the utilities of computer vision tasks and is formulated as:

$$SDR = \frac{1}{m}\sum_{i=1}^{m} g_D(x_i). \tag{5.12}$$

If the face can be detected, then $g_D = 1$. Otherwise, $g_D = 0$.

2) Landmarks distance. *chin* / *nose* / *eyes* / *mouth* indicates the mean distance of the key points corresponding to each facial area. It evaluates the utility of facial analysis.

- *Distortion metrics:*

  Mean square error (**MSE**) is used to measure the distortion between two images at the pixel level.

5.5.2 Performance Evaluation

In this section, we display the results of our proposed method from three aspects. (i) Ablation Study; (ii) Results compared to other methods; (iii) Analysis of the Parameters.

5.5.2.1 Ablation Study.

Figure 5.3: The visual results of the ablation study. The first row is the original image. The second row and third row are the de-identity images generated by ID loss and Bi-loss framework, respectively.

We conduct an ablation study on the framework to confirm the effectiveness of the proposed ID loss and attribute loss as introduced in Section 5.4.2. In particular, we consider the following cases: the framework is equipped with the ID loss module only.

It is worth noting that in order to protect the face identity with the ID loss module only, the perturbation ΔW adds on the gradient ascent direction of $\mathcal{L}_I$, i.e., $\Delta W = \lambda_I \text{sign}\,(\frac{\partial \mathcal{L}_I(x,\hat{x};f_I)}{\partial W'})$, with $\text{sign}\,(\cdot)$ being a Sign function. In this case, the perturbation is added to the ID information without attribute constraints. The Bi-loss mode images

is generated based on Algorithm 7. Both the ID and attribute constraints are used to generate the perturbation.

Fig. 5.3 gives the visual results of the proposed ablation study. The quantity result of the privacy evaluation method (SPR) and the utility evaluation methods (SDR and MSE) are reported in Table 5.1.

Table 5.1: Ablation Study. We use the same ID distance threshold, $\tau = 0.8$, for all settings in this table. The second column is the protection rate under Face Recognition Library. The third column is the detection rate by using *dlib* [13]. The hyperparameters are set to $\lambda_I = 0.02, \lambda_P = 0.008$, and the maximum iteration number $N = 100$.

	SPR (↑)	SDR (↑)	MSE (↓)
ID loss solely	0.901	0.979	0.426
Bi-loss (ours)	**0.905**	**1**	**0.112**

We use the framework of Face Recognition Library for the SPR, and dlib for the SDR. The ablation result shows that compared with the ID loss-only framework, the Bi-loss framework achieves a 0.4% increase in privacy performance and a 2.1% increase in face detection. Furthermore, the distortion of the Bi-loss framework has been reduced by 73.7%. In ablation experiments, the actual number of unsuccessfully protected samples is almost the same. However, since the Face Recognition Library first uses a face detection module to find faces in an image before figuring out the face ID for that image, the images protected by the ID-loss only sometimes make the image invisible to the face detection network, which lowers the protection rate.

5.5.2.2 Results compared to other de-identification methods.

This section compares our method with the state-of-the-art face de-identification methods.

Table 5.2 shows the privacy protection evaluation results with the widely used face recognition networks. We use both the Face Recognition Library and the FaceNet [14] network trained on VGGFace2 to evaluate the SPR. The results prove that our method can better de-identify the face image under the most widely used face recognition methods. Our method is better than CIAGAN, DeepPrivacy, and Fawkes by 4.9%, 2.8%, and 26.3%, respectively, in terms of the SPR under Face Recognition Library. Under FaceNet (VGGFace2) network, Ours (thick) improves the SPR by 1.7%, 14.4%, and 39.6% compared to CIAGAN, DeepPrivacy, and Fawkes, respectively.

Table 5.3 summarizes the utility performance of our method compared with the state-of-the-art methods. We evaluate our method with the SDR, MSE and the average

Table 5.2: Privacy evaluation. The values in this table are the successful protection rates (SPRs). The generation mode of Fawkes [7] is set to *high*, which is the highest privacy level authors recommended. The threshold of Face Recognition is $\delta = 0.6$ and the threshold of FaceNet is $\delta = 1.1$ according to [14].

	Face Recognition (↑)	FaceNet (VGGFace2) (↑)
CIAGAN [9]	0.918	0.943
DeepPrivacy [8]	0.939	0.816
Fawkes [7]	0.704	0.564
Ours	**0.967**	**0.960**

distance of the face feature landmarks on the pixel level. The result shows that our approaches achieve the highest SDR. In other words, our de-identified faces lead to better performance for face detection tasks. Moreover, compared with GAN-Based methods, i.e., CIAGAN [9], DeepPrivacy [8], our methods strike a compelling score on minimizing the distance of each facial feature. Our method lowers the average face feature distance of that on CIAGAN [9] and DeepPrivacy [8] by 77% and 71%, respectively. Both our method and Fawkes [7] have the lower average face feature distance, but Fawkes achieves the score at the expense of privacy-preserving effectiveness (lowest privacy score). In addition, our method achieves the lowest pixel-level distortions. Compared with CIAGAN, DeepPrivacy and Fawkes, our distortion is decreased by 14.5%, 67.4% and 73.5% respectively.

Taking into account the above all, our method has better performance in protecting the face ID while minimizing the impact on image utility. The images being protected by our method can be used in multi-image tasks.

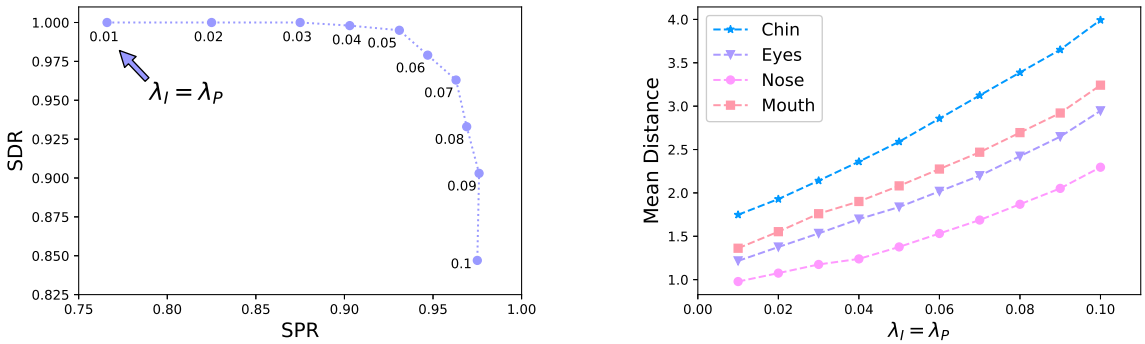
Table 5.3: Utility evaluation. The face detection network used in this table is *dlib*. The Landmarks Distances are generated under the Face Recognition Library.

	SDR (↑)	Landmarks Distance				MSE (↓)
		chin (↓)	nose (↓)	eyes (↓)	mouth (↓)	
original	1	0	0	0	0	0
CIAGAN [9]	0.9939	2.635	2.130	2.422	2.622	0.131
DeepPrivacy [8]	0.9989	2.070	1.631	1.384	2.712	0.344
Fawkes [7]	0.9990	0.720	**0.3921**	**0.389**	0.492	0.422
Ours	**1**	**0.704**	0.6664	0.484	**0.375**	**0.112**

5.5.2.3 Analysis on the Hyperparameters.

To evaluate the influence of the hyperparameters, we evaluate the perturbation performance with different controllable noise coefficients, i.e., λ_I and λ_P. Fig. 5.4(a) shows a trade-off between the SPR and the SDR under different values of λ_I and λ_P. We see that the SDR first decreases slowly when $\lambda_I = \lambda_P < 0.05$, and then decreases rapidly when $\lambda_I = \lambda_P \geq 0.05$. In contrast, the SPR increases rapidly when $\lambda_I = \lambda_P < 0.05$, and then flattens out at $\lambda_I = \lambda_P = 0.08$. In other words, with the growth of λ_I and λ_P, the success rate of protection (privacy) increases while the success rate of detection (utility) decreases.

Fig. 5.4(b) gives the mean feature distance under different hyperparameter settings. The result shows that the distance of the feature will increase when the hyperparameter increase. From the first point 0.01 to the last point 0.1, the average distance increases 135%.



((a)) The trade-off between privacy and utility under different parameters.

((b)) The analysis between the parameters and face features.

Figure 5.4: Parameters Analysis. The values of λ_I (or λ_P) range from 0.01 to 0.1 with a step size of 0.01. The maximum iteration number $N = 30$. The threshold $\tau = 0.8$.

5.5.3 Discussions

5.5.3.1 Privacy protection against the commercial network.

In this section, we test our protected images on the two most widely used commercial networks, Microsoft Face API [176] and Face Plus Plus API [177]. These two APIs are built on large face recognition networks, which use the advanced deep neural network and are trained on a large dataset. They provide several applications, including face detection and analysis, identity verification and finding similar faces. In this experiment,

we use the identity verification service to evaluate our method. In identity verification, the APIs will take the original images and the protected images as the input and output a score of confidence that indicates the probability that two faces belong to the same person. The higher the confidence, the higher chance they belong to the same person. If the input is the images without protection and the original image, the confidence score equals 1.0.

The experiment results are shown in Fig. 5.5. Microsoft Face API and Face Plus Plus use different confidence thresholds. While Microsoft Face API takes 0.5, Face Plus Plus takes the value of around 0.69. The sample with a score below the threshold is recognized as a different person in the API. The results show that our method lowers the score of 82.9% samples under the threshold of 0.5 against Microsoft Face API, which makes the network almost ineffective. And the score of 55.8% samples on Face Plus Plus API is under the threshold of 0.69, which makes the network works in a random guess. The experiment result proves that our method is transferable in different networks. Because these commercial APIs networks do not open their source code and we did this experiment in a black box mode.

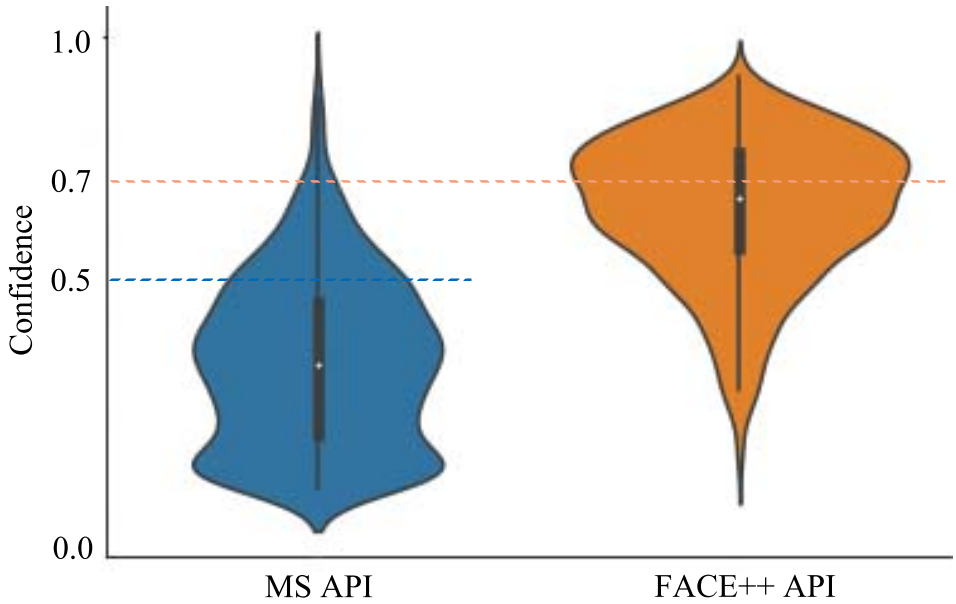


Figure 5.5: The confidence of commercial API. The y-axis indicates the confidence score of the MS API and the FACE++ API. The blue violin plot signifies the outcomes derived from the Microsoft API, whereas the orange violin plot corresponds to the results obtained from the Face++ API. The dashed line denotes the threshold applied to each API, wherein facial images falling below this threshold are construed as emanating from disparate identities. Notably, our methodology exhibits efficacy for both APIs.

5.5.3.2 Limitations.

Although our proposed framework for face identity protection can achieve compelling results in both privacy and utility, it has some room for further improvement. First, the face latent code generated by the auto-encoder architecture comes with a cost - the de-identity image quality is limited to the image-to-image translation ability of the chosen auto-encoder. Thus, de-identifying the face image with complex attributes, e.g. background, might be challenging if such examples were not synthesized well in the auto-encoder. Fig. 5.6 presents a few examples of such images.

As can be seen in Fig. 5.6, the complex attributes, e.g. background and hands, which are non-identity-related attributes, can not display correctly due to the reconstruction failure.

Figure 5.6: Failure examples. Output 1 refers to the generated images without protection. Output 2 shows the de-identity image with our method.

5.6 Conclusion

In this Chapter, we propose a novel face image de-identification framework. This framework de-identifies the face by adding feature space adversarial perturbation (FSAP). Moreover, we conduct intensive experiments to prove the effectiveness of the framework. With the latent vector W trained on the elaborate loss, the perturbed faces are equipped to reduce the risks of identity leakage under CNN face recognition technics while balancing the utility for computer vision tasks. The merits of the proposed framework are two-folded. First, compared with the GAN-based face de-identification network, instead of adding perturbations in a generalised direction, FSAP adds noise on the gradient, which ensures accuracy. Second, compared with the AP-based face de-identification

network, feature-space adversarial perturbations have better transferability among different neural network models.

5.7 Future work

In light of the limitations highlighted in Section 5.5.3.2 regarding the utilization of GAN as an image generator, our forthcoming research endeavours will be dedicated to resolving this concern in order to elevate the standard of image privacy protection. Our primary focus will centre on contemporary and pioneering generative models. Fusing these novel large-scale models with image privacy protection will constitute a pivotal facet of our future investigations. Subsequently, in the ensuing chapter, we will showcase the technical merits introduced by the new model in our research.

CHAPTER 6

DIFFUSION DE-ID: DIFFUSSION-BASED FACE DATASET ANONYMIZATION

6.1 Preface

In this chapter, our focus lies in the domain of high-quality privacy protection within image datasets. Specifically, we tackle the challenge of anonymizing facial datasets while ensuring the anonymized dataset remains conducive to effective CNN network training in downstream tasks, all the while preserving privacy. Existing approaches to address this issue predominantly rely on GANs, which may restrict image quality when handling intricate facial data. Consequently, we propose a framework that leverages diffusion models as the foundation, complemented by advanced semantic optimization techniques, to achieve high-quality anonymization of image datasets. Through rigorous validation experiments conducted on facial datasets, our approach outperforms existing state-of-the-art methods, showcasing its superior performance.

6.2 Introduction

6.2.1 Motivation

Anonymizing face image datasets plays a crucial role in preserving privacy and protecting individuals' personal information. These datasets often contain sensitive and personally identifiable data that must be safeguarded. Anonymization involves the removal of direct and indirect identifiers, ensuring that the data cannot be linked back to specific individuals. This process mitigates the risk of privacy breaches and unauthorized use of personal information. One significant benefit of anonymization is its impact on obtaining informed consent. By removing identifiable features from face image datasets, the risk of using these images without explicit consent is minimized. This practice aligns with ethical guidelines and legal requirements, promoting responsible data handling and respecting individuals' autonomy over their personal information. Another important consideration is the risk of re-identification when face image datasets are shared or linked with other data sources. Anonymizing the datasets by eliminating personally identifiable information significantly reduces the possibility of re-identifying individuals. This step ensures that individuals' privacy is maintained even when datasets are used for legitimate purposes.

Anonymization also facilitates ethical data sharing and collaboration among researchers, organizations, and institutions. Concerns regarding privacy violations and unauthorized use of personal information are mitigated by removing personal identifiers from face image datasets. Researchers can work with anonymized datasets, focusing

on data analysis and deriving insights while upholding privacy rights and promoting responsible data practices. Furthermore, compliance with privacy regulations and laws is crucial to anonymization. For instance, the General Data Protection Regulation (GDPR) [15] in the European Union mandates the protection of personal data, including face images. Anonymizing face image datasets ensures compliance with these legal obligations, reducing the risk of penalties or legal repercussions for organizations and researchers.

Several methods are commonly used for anonymizing face images. One prevalent technique is face blurring [22], which employs blurring algorithms or filters to obscure the facial features of individuals in an image. By intentionally degrading the details of the faces, it becomes challenging to identify specific individuals while still retaining the overall structure of the image. Another approach is pixelation [28], where the facial region is replaced with large pixels, concealing facial details. Face detection and masking techniques [178, 179] leverage face detection algorithms to identify the locations of faces within an image. Once the faces are detected, a solid color or a patterned mask is overlaid on the facial region, effectively concealing the facial features and maintaining individuals' anonymity. Facial landmark removal [180] is a method that involves removing or distorting facial landmarks, i.e., specific points on the face, such as eyes, nose, and mouth. By altering these landmarks, the overall facial structure is modified, making it challenging to recognize individuals while still preserving some contextual information. Nevertheless, these methods achieve privacy protection at the cost of compromising facial attributes and image quality. Furthermore, recent studies [35, 36] have indicated that the de-identification effects of blurring and pixelation, while perceptually effective to human observers, can be circumvented by deep learning algorithms, thereby diminishing their efficacy in preserving privacy.

Another prominent category of approaches to tackle this issue is primarily based on Generative Adversarial Networks (GANs) [131]. GANs offer a unique approach to face image anonymization. The techniques based on GAN create synthetic faces that preserve anonymity while maintaining the visual quality of faces. In particular, GAN-based face manipulation [181] involves controlling the attributes, including identities, of the original face in an image, which completely transforms the appearance of the individual while retaining the overall structure of the image, providing a novel approach to anonymity. The current landscape of GAN-based methodologies can be categorized into three primary classes: image inpainting [10], latent optimization [182], and attribute manipulation [183]. These approaches exhibit varying degrees of effectiveness in addressing

the challenges of face anonymization while introducing new and distinctive challenges. Generating anonymous face images with exceptional quality and usability is a significant challenge to overcome. One notable methodology, GAN-based Autoencoder [19], which has been accepted as a module in face anonymization [12], leverages a StyleGAN2-based encoder to extract latent codes from real-world face images. This encoder is trained to map images into the latent space of a pre-trained StyleGAN2 generator, which has demonstrated promising outcomes in the domain of image manipulation. However, recent studies [4, 182] have highlighted the limitations of GAN-Based Autoencoders in manipulating facial images. These limitations stem from the inherent reconstructive capabilities of GANs, which fail to capture numerous intricate details.

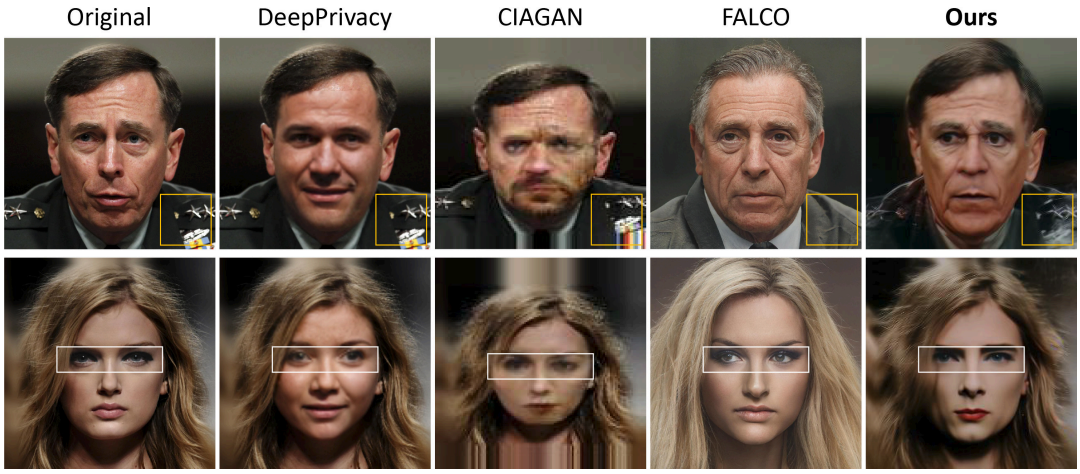


Figure 6.1: Visual quality compared to DeepPrivacy [10], CIAGAN [11], FALCO [12]. **Ours**, along with the FALCO approach, exhibits the capability of concealing sensitive information within the background. Notably, **Ours** also demonstrates the ability to retain a richer set of attributes and intricate details.

6.2.2 Contributions

Motivated by the aforementioned advancements, this chapter presents an innovative contribution: the pioneering application of a diffusion model [5] as the fundamental framework for anonymizing face datasets. By embracing the diffusion model, we aim to address the limitations encountered in previous GAN-based approaches. This, in turn, facilitates the practical implementation of face dataset anonymization and effectively safeguards the privacy of facial data. This choice stems not only from the innovative nature of the diffusion model but also from its potential to overcome the aforementioned constraints associated with existing methods. In particular, our proposed approach

leverages the DDIM decoder as the primary generator for synthesizing synthetic images. To preserve the various facial attributes present in the original dataset, we employ attribute optimization techniques on these generated images. By ensuring the faithful preservation of these comprehensive facial attributes, our approach enables the training of diverse facial models suitable for a wide range of applications. The main contributions of this chapter are summarized as follows:

- This chapter proposes a novel face dataset anonymization framework based on a Diffusion model, which surpasses other state-of-the-art methods by offering enhanced identity anonymization performance without compromising the preservation of crucial facial attributes.

- We propose a novel loss function that combines identity loss and attribute loss. This novel approach, in collaboration with a Denoising Diffusion Implicit Model (DDIM) decoder and attribute optimization techniques, empowers the generation of an anonymized dataset that exhibits enhanced effectiveness and usability.

- Through empirical evaluation and comparative analysis on datasets such as CelebA-HQ and FFHQ, we demonstrate the superior efficacy of our approach in preserving facial attributes compared to other state-of-the-art methods. Moreover, our approach stands out by enabling the training of robust facial models that maintain the integrity of essential attributes crucial for a wide range of practical applications.

6.2.3 Overview of the work

The structure of the remaining sections in this chapter is as follows. Section 6.3 provides an overview of the related work in the field. In Section 6.4, we present the necessary preliminary knowledge for our work, specifically focusing on Conditional Denoising Diffusion Implicit Models (DDIMs). Moving on to Section 6.5, we introduce our proposed framework, outlining its key components and presenting our protection mechanisms. In Section 6.6, we conduct a series of experiments to validate the effectiveness of our proposed scheme. This includes a comprehensive comparison study against other state-of-the-art methods, assessing quantitative and qualitative aspects of face image privacy. Additionally, we discuss the optimization methods employed in our approach through an ablation study.

6.3 Related Work

In recent image privacy researches [159–161], considerable attention has been paid to modifying identity-related information in images using various techniques, such as obfuscation, GAN-based inpainting [184], differential privacy (DP) [166, 185], and adversarial examples (AEs) [37, 40]. It has been demonstrated that basic obfuscation techniques are ineffective when it comes to countering deep neural network (DNN)-based recognition systems [107, 146].

GAN-based face inpainting The concept of GAN-based inpainting was introduced to generate content that effectively conceals sensitive information or the identity of an image while preserving the quality of the original image [59]. To enhance the training process, a conditional GAN (CGAN)-based approach was developed by incorporating labels into both the generator and discriminator networks [186]. Building upon the CGAN framework, two notable methods, namely Conditional Identity Anonymization Generative Adversarial Network (CIAGAN) [11] and DeepPrivacy [10], introduced the integration of autoencoder within the feature space of images. CIAGAN demonstrated the ability to anonymize faces and bodies, generating high-quality images and videos. While DeepPrivacy took into account factors such as pose and background to generate images with improved realism and privacy preservation.

Diffusion models Denoising diffusion probabilistic models (DDPMs) [5] are generative models that associate image generation with the sequential denoising process of isotropic Gaussian noise. Unlike other generative models such as GANs and most traditional-style VAEs that encode input data into a low-dimensional space, diffusion models maintain a latent space of the same size as the input. In the forward process, DDPMs progressively add noise to the image until it is completely degraded into pure Gaussian noise. Assuming an ideal forward process, where a real input image $\mathbf{x}_0$ undergoes T rounds of Gaussian noise addition, resulting in a purely Gaussian noise image $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Consequently, each step of noise addition can be formally expressed as the following probability function: $q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) \sim \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$, where β_t is the coefficient associated with the noise. As a result, the cumulative noise in the image $\mathbf{x}_0$ after t processing steps can be represented as another Gaussian noise $q(\mathbf{x}_t \mid \mathbf{x}_0) \sim \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I})$, where $\alpha_t = \prod_{i=1}^{t}(1-\beta_i)$. In the reverse process, i.e., learning the distribution $p(\mathbf{x}_{t-1} \mid \mathbf{x}_t)$, the noise is gradually removed to generate a realistic image. To train a DDPM network, Ho et al. [21] give the distribution of $p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) \sim \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \sigma_t)$, and propose a learnable function $\epsilon_\theta(\mathbf{x}_t, t)$ to predict

the noise added in each step. Despite requiring a large number of noise injection and denoising steps to generate samples, DDPMs exhibit superior image fidelity and diversity compared to other types of generative models.

A notable limitation of diffusion models lies in their reliance on an extended sequence of diffusion steps to achieve desirable outcomes, resulting in sluggish generation speed. In contrast, Denoising Diffusion Implicit Models (DDIMs) [6], in comparison to DDPMs, overcome this constraint by relaxing the requirement for the diffusion process to conform to a Markov chain. As a result, DDIMs can employ fewer sampling steps, effectively expediting the generation process. Furthermore, DDIMs possess a distinctive characteristic in that the generation of samples from random noise is deterministic, obviating the need for intermittent random noise injections. This streamlined approach further accelerates the relatively computationally demanding sampling procedure inherent to DDPMs. By leveraging a deterministic forward-backward process, DDIMs demonstrate a near-perfect reconstruction capability. In this study, we employ conditional DDIMs as the generator to anonymize human facial images.

6.4 Preliminaries-Conditional DDIM

Coditional semantic encoder We use a conditional DDIM with an additional latent code, denoted as $\bar{\mathbf{z}}$. In contrast to certain other conditional DPMs [187–189] that employ spatial 2-D latent maps, the latent code we employ is a non-spatial code with a dimension of 512. The primary objective of the Conditional Semantic Encoder is to encode all the semantic information in an image into a high-level semantic space. Consequently, the latent code $\bar{\mathbf{z}} \in \mathbb{R}^{512}$ encompasses global semantics that is not specific to any spatial regions within the image, facilitating smooth optimization.

Diffusion-based decoder The primary objective of the Diffusion-based Decoder is to model the distribution of the target dataset by training a network in the reverse process, denoted as $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$. A successfully trained network can generate a realistic image with a Gaussian noise map. There are various methods to model this distribution, and one of them is through the DDIM introduced by Song et al. [190]:

$$\begin{aligned} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) &\sim \mathcal{N}(\sqrt{\alpha_{t-1}}\mathbf{x}_0 + \sqrt{1-\alpha_{t-1}}\frac{\mathbf{x}_t - \sqrt{\alpha_t}\mathbf{x}_0}{\sqrt{1-\alpha_t}}, 0) \\ &\approx q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0) \end{aligned} \tag{6.1}$$

In this work, we employ a conditional DDIM [88] for generating images, where the input is represented by a latent variable, $\mathbf{z} = (\bar{\mathbf{z}}, \mathbf{x}_T)$. This variable consists of two com-

ponents: the high-level semantic latent code, $\bar{\mathbf{z}}$, generated by the conditional semantic encoder, and the low-level semantic latent maps, $\mathbf{x}_T$, generated through the forward process of conditional DDIM. During the image generation (reverse) process, the probability function is then written as $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \bar{\mathbf{z}})$.

Drawing from Equation 6.1, we define the generative process using the following equations:

$$p_\theta(\mathbf{x}_{0:T}|\bar{\mathbf{z}}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \bar{\mathbf{z}}) \tag{6.2}$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \bar{\mathbf{z}}) = \begin{cases} \mathcal{N}(\mathbf{f}_\theta^{(1)}(\mathbf{x}_1, \bar{\mathbf{z}}), \mathbf{0}) & \text{if } t = 1 \\ q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{f}_\theta^{(t)}(\mathbf{x}_t, \bar{\mathbf{z}})) & \text{otherwise} \end{cases} \tag{6.3}$$

where

$$\mathbf{f}_\theta^{(t)}(\mathbf{x}_t, \bar{\mathbf{z}}) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t, \bar{\mathbf{z}})) \tag{6.4}$$

where $\epsilon_\theta(\mathbf{x}_t, t, \bar{\mathbf{z}})$ is a pre-trained UNet model in [88]. Like the DDIM proposed by Song *et al.* [6], conditional DDIM follows a similar deterministic generative process:

$$\begin{aligned} \mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}(&\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_t, t, \bar{\mathbf{z}})}{\sqrt{\alpha_t}}) \\ &+ \sqrt{1-\alpha_{t-1}}\epsilon_\theta(\mathbf{x}_t, t, \bar{\mathbf{z}}) \end{aligned} \tag{6.5}$$

Diffusion-based encoder The diffusion-based encoder is the deterministic generative process of DDIM, which is used to encode an input image $\mathbf{x}_0$ into a lower semantic subcode $\mathbf{x}_T$. The generative process can be represented by the following equation:

$$\mathbf{x}_t = \sqrt{\alpha_t}\mathbf{f}_\theta(\mathbf{x}_{t-1}, t-1, \bar{\mathbf{z}}) + \sqrt{1-\alpha_t}\epsilon_\theta(\mathbf{x}_{t-1}, t-1, \bar{\mathbf{z}}) \tag{6.6}$$

Due to the restricted semantic information it possesses, $\mathbf{x}_T$ remains unchanged and does not participate in the optimization process for anonymizing image identities.

6.5 Methodology

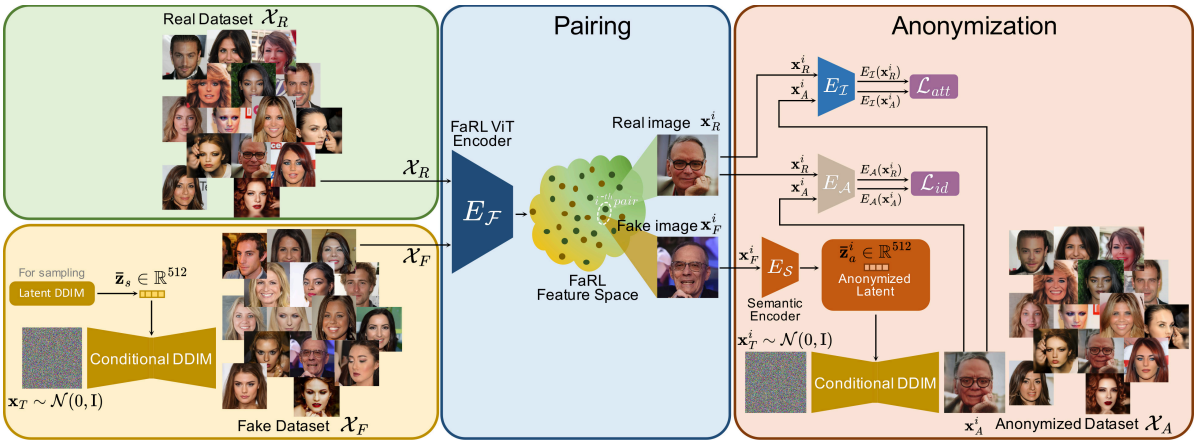


Figure 6.2: Our proposed protection framework. Our protection framework optimizing the trainable anonymized latent vector $\mathbf{z}_a^i \in \mathbb{R}^{512}$ using two loss functions, $\mathcal{L}_{id}$ and $\mathcal{L}_{att}$. This optimization aims to obfuscate the identity of the synthetic image $\mathbf{x}_A^i$ while maintaining the facial attributes.

This work proposes a novel method for anonymizing faces in a given real-face dataset based on the diffusion model to gain better image utility. The proposed method optimizes the latent codes of images in the dataset within the latent space of a pre-trained conditional diffusion model [88]. The method involves creating a fake dataset $\mathcal{X}_F$ by randomly generating a large set of fake images such that $|\mathcal{X}_F| > |\mathcal{X}_R|$, where $\mathcal{X}_R$ is the real dataset. To obtain meaningful initial values for the latent codes that will be optimized to create the anonymized version of the real dataset $\mathcal{X}_A$, the real images from $\mathcal{X}_R$ are paired with the fake ones from $\mathcal{X}_F$ in the feature space of the ViT-based FaRL [191] image attribute encoder. The method then optimizes the successfully paired latent codes using two loss functions: (1) the identity loss, denoted by$\mathcal{L}_{id}$, ensures that the fake images remain a certain distance away from the real ones in terms of identity, and (2) the attribute preservation loss, denoted by $\mathcal{L}_{att}$, pulls the fake images closer to the real ones in the feature space of the FaRL [191] image encoder. In this way, the anonymized images inherit the attribute information of the real ones while possessing a different identity.

The rest of the section is organized as follows: Sec. 6.5.1 briefly introduces the modules used in our framework, Sec. 6.5.2 shows the details of the fake dataset generation, Sec. 6.5.3 discusses the pairing, and Sec. 6.5.4 presents the anonymization process and the losses.

6.5.1 Modules

6.5.1.1 Conditional DDIM

The conditional DDIM [88] image decoder takes a latent variable $\mathbf{z} = (\bar{\mathbf{z}}, \mathbf{x}_T)$ as input, where $\bar{\mathbf{z}}$ represents the high-level "semantic" subcode and $\mathbf{x}_T$ is the low-level "stochastic" subcode inferred by reversing the generative process of DDIM. Here, we keep the low-level stochastic subcode unchanged and optimize the image on the high-level semantic subcode.

6.5.1.2 ArcFace

To measure the similarity between the identities of two face images, we employ ArcFace [138], a method that maps images to a 512-dimensional feature space related to identity. Using the identity features, we optimize the semantic latent codes $\bar{\mathbf{z}}$ of the conditional DDIM [88] to generate images that minimize the cosine similarity of the image identity features between real and fake images.

6.5.1.3 FaRL

FaRL [191] is a facial representation network trained in a contrastive manner on 20 million face image-text pairs for representing images in a meaningful and rich semantic feature space. The ViT image encoder of the FaRL framework is used to represent images in a 512-dimensional feature space and to find meaningful initial values for the latent codes that will be optimized to anonymize the real dataset. This approach provides a solid foundation for the anonymization of real datasets by capturing the underlying features of the input data and representing images in a high-dimensional feature space.

6.5.2 Fake dataset generation

Given a real face dataset $\mathcal{X}_R$, to generate a fake face dataset $\mathcal{X}_F$ such that $\mathcal{X}_F > \mathcal{X}_R$, we utilize the decoder of a conditional DDIM [88] as the generator. To sample $\bar{\mathbf{z}}$ from the latent distribution and generate a synthetic dataset using diffusion autoencoders, we employ a pre-trained latent DDIM to approximate the latent distribution of $\bar{\mathbf{z}}$. The latent distribution modeled by the latent DDIM is initially normalized to have a mean of zero and a variance of the unit. To perform unconditional sampling from a diffusion autoencoder, the process involves sampling from the latent DDIM and then unnormalizing it as $\bar{\mathbf{z}}$. Next, $\mathbf{x}_T$ is drawn from a normal distribution with a mean of zero and a covariance

matrix of $\mathbf{I}$, $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. Finally, the fake image synthesis process involves using the decoder to decode $\mathbf{z} = (\bar{\mathbf{z}}, \mathbf{x}_T)$.

6.5.3 Pairing

To pair real images in dataset $\mathcal{X}_R$ with fake images in dataset $\mathcal{X}_F$, we employ the pre-trained FaRL [191] ViT-based encoder $E_{\mathcal{F}}$. All images from both datasets are represented in a 512-dimensional feature space using the class token representation. This approach yields a robust feature representation of both datasets, which we use to train a kNN classifier. Using this classifier, we can identify the fake image in $\mathcal{X}_F$ that is closest to each real image in $\mathcal{X}_R$ based on the Euclidean distance. After generating and pairing the images, the real dataset $\mathcal{X}_R$ and the fake dataset $\mathcal{X}_F$ are paired, creating a set of pairs consisting of images from the real dataset $\mathcal{X}_R$ and images from the fake dataset $\mathcal{X}_F$.

6.5.4 Anonymization

To create an anonymized version $\mathcal{X}_A$ of the real dataset $\mathcal{X}_R$, we utilize the pairs of real-fake images obtained from the pairing section. These pairs consist of real and fake images that are semantically similar based on the FaRL image representation features. Specifically, for each pair, the real image $\mathbf{x}_R^i$ and its corresponding anonymized image $\mathbf{x}_A^i$, generated using the anonymized latent code $\mathbf{z} = (\bar{\mathbf{z}}_A^i, \mathbf{x}_T^i)$, are used to calculate the proposed losses.

The identity loss $\mathcal{L}_{id}(\mathbf{x}_A^i, \mathbf{x}_R^i)$ ensures that the identity of $\mathbf{x}_A^i$ maintains a desired distance from the identity of $\mathbf{x}_R^i$. The attribute preservation loss $\mathcal{L}_{att}(\mathbf{x}_A^i, \mathbf{x}_R^i)$ enforces the preservation of facial attributes from the original image in the anonymized image.

For each pair consisting of a real image $\mathbf{x}_R^i$ and its anonymized version $\mathbf{x}_A^i$, i ranges from 1 to the total number of images in $\mathcal{X}_R$. The anonymization is achieved by optimizing the following losses:

6.5.4.1 Identity loss

The identity loss is defined as

$$\begin{aligned} \mathcal{L}_{id} &= \left| \frac{E_{\mathcal{A}}(\mathbf{x}_A^i) \cdot E_{\mathcal{A}}(\mathbf{x}_R^i)}{\left\| E_{\mathcal{A}}(\mathbf{x}_A^i) \right\|_2 \cdot \left\| E_{\mathcal{A}}(\mathbf{x}_R^i) \right\|_2} \right| \\ &= \left| \cos(E_{\mathcal{A}}(\mathbf{x}_A^i), E_{\mathcal{A}}(\mathbf{x}_R^i)) \right|, \end{aligned} \tag{6.7}$$

where $E_{\mathcal{A}}$ is a pre-trained ArcFace [138] identity encoder used to extract identity features from facial images. $\cos(\cdot,\cdot)$ denotes the cosine similarity, which is a measure of similarity that calculates the cosine of the angle between two vectors. The loss yields a value ranging from 0 to 1. A value of 1 indicates that the vectors are identical, while a value of 0 signifies no similarity.

6.5.4.2 Attribute loss

The attribute loss is defined as

$$\mathcal{L}_{att}(\mathbf{x}_A^i, \mathbf{x}_R^i) = \left\| E_{\mathcal{I}}(\mathbf{x}_A^i) - E_{\mathcal{I}}(\mathbf{x}_R^i) \right\|_1, \tag{6.8}$$

where $E_{\mathcal{I}}$ refers to the patch-level features, specifically the 14×14 768-dimensional features, obtained from the ViT-based image encoder introduced in the FaRL [191] paper. These 14×14 768-dimensional features are subsequently flattened into $14 \cdot 14 \cdot 768$-dimensional vectors for the purpose of loss calculation.

6.6 Experiments

In this section, we aim to investigate and evaluate four crucial dimensions of our framework: face detection accuracy, identity anonymity, image quality, and facial attribute preservation. Through a series of experiments, we will systematically analyze and assess the performance of our framework in these specific areas. This section is structured as follows: Sect. 6.6.1 provides an overview of the experimental preparation conducted. Sect. 6.6.2 presents a comparative analysis of our method and other state-of-the-art (SOTA) approaches using evaluation metrics. Lastly, Sec. 6.6.3 delves into an ablation study to further investigate and analyze the individual components of our proposed method.

6.6.1 Experiment settings

6.6.1.1 Datasets

Our anonymization process is conducted on the following two datasets: (i) **CelebA-HQ** [192], which comprises $30,000$ high-resolution (1024×1024) facial images of celebrities sourced from the CelebA dataset. These images exhibit diverse demographic attributes, including age, gender, and race. Additionally, each image is annotated with 40 attribute labels. (ii) **FFHQ** [137], is a curated collection of high-quality images depicting human faces. This dataset comprises $70,000$ PNG images with a resolution of 1024×1024 pixels. Notably, FFHQ exhibits significant diversity in terms of age, ethnicity, and background settings, providing a comprehensive range of facial characteristics. Moreover, the dataset offers ample coverage of various accessories, including eyeglasses, sunglasses, hats, and more.

6.6.1.2 State-of-the-art (SOTA)

We conduct a comparative evaluation of our anonymization framework against three SOTA methods, namely FALCO [12], CIAGAN [11], and DeepPrivacy [10]. In particular, we focus on the FALCO method, which employs latent code optimization techniques for dataset anonymization, thereby sharing similar objectives with our task.

6.6.1.3 Evaluation metrics

We briefly introduce the evaluation metrics we used:

(i) **Face detection accuracy (Face Dete)** refers to the degree to which a face detection system or algorithm can accurately detect and locate faces within an image. The detection of faces plays a pivotal role in training machine learning models, serving as a fundamental component for various downstream applications, including face recognition, emotion detection, facial expression analysis, and video surveillance. Consequently, it is imperative for facial datasets to offer a strong foundational basis for face detection tasks while simultaneously preserving their diversity and robustness. In this research endeavor, we employed SOTA face detection models, namely MTCNN [193] and dlib [13], to assess the efficacy of different datasets in terms of face detection performance. A flawless anonymized dataset would exhibit an identical level of accuracy in face detection as the original dataset, thereby implying that the anonymized dataset has the potential to train face detection models that are comparable to those trained on the original dataset.

(ii) **Identity anonymity (ID Anon)** refers to the degree to which the personal identity of individuals is concealed or protected within a given system or framework. It involves safeguarding sensitive information, such as personal identifiers, to prevent the association of specific individuals with their respective data. The purpose of ensuring identity anonymity is to maintain privacy and confidentiality, particularly in scenarios where data sharing or analysis is conducted while protecting the identities of the individuals involved. This work uses FaceNet [194] to measure Identity anonymity. An ideal anonymized dataset would exhibit 100% anonymity in face identification, indicating that each face could be effectively de-identified.

(iii) **Image quality** refers to the level of visual fidelity, clarity, and overall excellence of an image. High image quality implies that the image accurately represents the original scene or subject, with minimal distortion, artifacts, or loss of details. High-quality images are visually appealing, enhance the overall user experience, and enable accurate analysis, interpretation, and communication of visual information. In this research, we use the Fréchet Inception Distance (FID) [195] to measure the image quality. The FID metric is widely used in the field of generative adversarial networks (GANs) to assess the diversity and quality of generated images by analyzing their statistical properties.

By examining the feature representations of the anonymized images, the FID metric effectively captures their visual quality and diversity. This allows for a comprehensive evaluation of the similarity between the anonymized and real images based on a comparison of their respective feature representations.

The FID metric utilizes the Inceptionv3 DNN [123], a pre-trained convolutional

neural network (CNN), to extract feature representations from two sets of images: the real image dataset and each anonymized image dataset. These feature representations are obtained from an intermediate layer of the network, typically positioned before the final classification layer.

After extracting feature representations, we compute the mean and covariance values separately for both sets of images. The mean represents the average feature vector, while the covariance matrix captures the interrelationships among different features. Then, we calculate the Fréchet distance, which measures the dissimilarity between two multivariate Gaussian distributions. In this case, it quantifies the disparity between the mean and covariance of the feature representations obtained from the real and anonymized image datasets. An improved anonymized dataset would demonstrate a lower FID (Fréchet Inception Distance) score, indicating that the faces in the image set possess higher image quality.

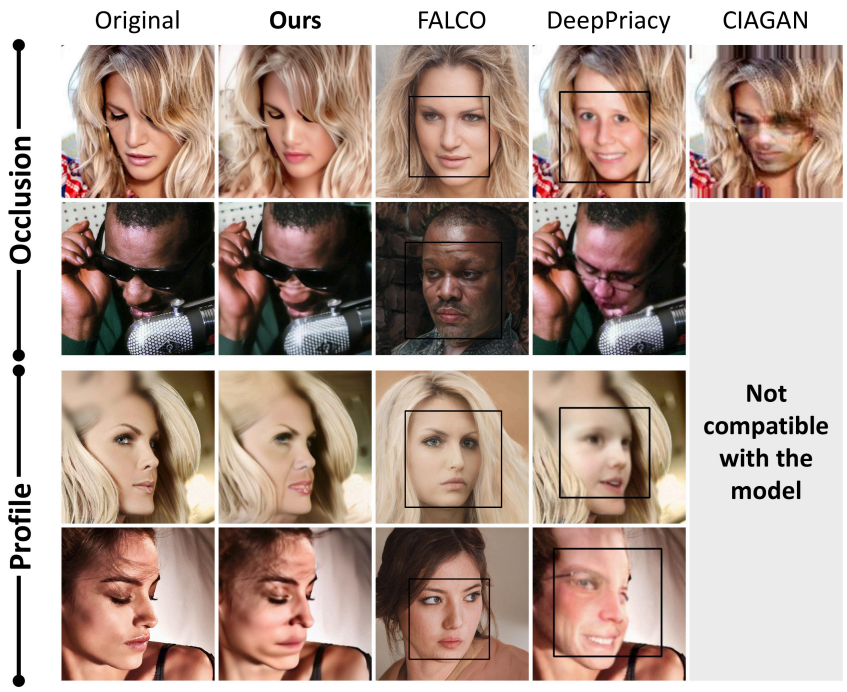


Figure 6.3: Comparative analysis of visual effects: Our method vs. FALCO [12], DeepPrivacy [10], and CIAGAN [11], in the context of original images undetectable by the dlib [13] frontal face detector.

(iv) **Facial attribute preservation** refers to the ability of a system or algorithm to retain and accurately represent the distinctive characteristics and attributes of a face after undergoing processes such as anonymization. It ensures that important facial features, such as gender, age, and facial expressions, are preserved and recognizable

even after applying privacy protection techniques. By ensuring effective facial attribute preservation, systems can balance privacy concerns with the need for accurate face characterization. It allows for the responsible use of facial data while safeguarding individual privacy and maintaining the utility of facial analysis applications. Evaluation of facial attribute preservation can be performed through qualitative and quantitative analysis. Qualitative assessment involves visually inspecting the anonymized faces to ensure that essential attributes are still identifiable. In this research, a quantitative metric, attribute classification accuracy, is used to measure the system's ability to preserve facial attributes accurately. Specifically, to ensure fair comparisons, our study adheres to the experimental configuration described in [12], wherein we employ a pre-trained MobileNetV2 model for attribute classification. This approach allows us to maintain consistency and comparability across different experiments.

	Face Dete ↑		ID Anon ↑		Image quality ↓	Attributes classification ↑	
	dlib(%)	MTCNN(%)	CASIA(%)	VGG(%)	FID	Inner face(%)	Outer face(%)
Original	98.55	99.91	0.00	0.00	0.00	85.64	85.21
CIAGAN [11]	−3.14	−0.35	**97.95**	**99.61**	31.11	80.22	80.72
DeepPrivacy [10]	+0.47	−0.13	97.29	98.12	28.32	80.29	79.55
FALCO [12]	+1.45	+0.09	96.71	97.27	28.19	80.16	79.5
Ours	**−0.29**	**−0.08**	97.55	98.21	**23.42**	**82.73**	**82.5**

Table 6.1: Anonymized image dataset evaluation results. Face Detection Accuracy (Face Dete), Identity Anonymity (ID Anon), Image Quality, and Attributes Classification.

6.6.2 Comparison to state-of-the-art (SOTA)

In this section, we present a comparative analysis of our proposed method against three SOTA techniques, namely CIAGAN [11], DeepPrivacy [10], and FALCO [12]. By benchmarking our proposed method against the most advanced and widely recognized approaches in the field, we aim to comprehensively evaluate its effectiveness and performance. This comparison enables us to assess the advancements made by our method and highlight its unique contributions in pushing the boundaries of the current SOTA.

6.6.2.1 Quantitative evaluation

In this section, a comprehensive quantitative analysis is conducted to compare the effectiveness of our proposed method with other approaches. The specific numerical comparison results are presented in Tabs. 6.1 and 6.2.

	Original	**Ours**	**FALCO** [12]	**Deep** [10]	**CIAGAN** [11]
Inner face region					
5 o Clock Shadow	85.45	81.21	**82.05**	79.56	77.60
Arched Eyebrows	86.43	**84.13**	80.01	81.58	82.03
Bags Under Eyes	85.69	**82.34**	79.54	79.15	80.55
Big Lips	85.15	**82.59**	79.70	80.54	80.62
Big Nose	85.14	**82.12**	79.60	78.72	78.69
Bushy Eyebrows	84.87	80.98	81.18	78.72	**81.85**
Eyeglasses	83.74	**81.71**	79.77	79.98	80.10
Goatee	83.46	79.38	79.32	76.08	**80.22**
Heavy Makeup	87.51	**85.33**	80.91	83.03	80.32
High Cheekbones	86.13	83.22	78.75	80.16	**84.03**
Male	85.64	**82.37**	82.19	80.17	82.02
Mouth Slightly Open	86.21	**83.03**	79.21	80.72	79.50
Mustache	83.13	**79.35**	78.47	75.88	77.13
Narrow Eyes	83.81	**80.82**	77.89	78.41	77.08
No Beard	87.22	**85.00**	82.01	83.05	80.82
Pale Skin	86.07	**84.49**	82.72	83.27	77.93
Pointy Nose	85.81	**83.26**	80.36	81.40	79.93
Rosy Cheeks	86.59	**83.28**	77.07	79.88	81.75
Smiling	86.03	**82.76**	78.55	79.76	80.26
Wearing Lipstick	87.36	**85.29**	81.27	83.12	81.80
Young	86.96	**84.55**	82.72	82.84	80.35
Outer face region					
Bald	83.03	**80.38**	77.20	75.57	79.06
Bangs	86.63	**84.02**	80.01	81.47	82.59
Black Hair	85.12	**82.63**	82.37	79.60	80.76
Blond Hair	88.35	**85.93**	81.47	84.99	82.35
Brown Hair	86.21	**84.03**	81.35	82.19	80.92
Chubby	82.61	**79.91**	76.60	75.18	79.32
Double Chin	83.32	**79.81**	75.25	74.88	79.29
Gray Hair	85.55	**83.55**	78.45	78.39	78.64
Oval Face	85.54	**83.12**	80.37	81.04	79.98
Receding Hairline	84.32	82.08	78.94	78.36	**82.14**
Sideburns	84.56	79.81	80.27	76.77	**81.42**
Straight Hair	85.94	**83.14**	81.87	81.01	80.50
Wavy Hair	86.4	**83.98**	80.27	82.88	81.39
Wearing Earrings	85.83	**84.20**	78.82	80.60	82.11
Wearing Hat	85.9	**82.72**	81.08	81.59	79.85
Wearing Necklace	85.17	**82.86**	78.24	80.47	79.32
Wearing Necktie	84.03	80.40	78.89	77.42	**82.58**

Table 6.2: Facial attribute preservation results

Face detection accuracy (Face Dete) In this paragraph, we present a detailed explanation of the face detection accuracy values. The specific numerical comparison results for face detection accuracy, including our proposed method and other approaches, are provided in the first two columns of Tab. 6.1. The first row of Tab. 6.1 represents the percentage of accurately detected faces in the original dataset, as detected by the face detectors dlib [13] and MTCNN [193]. The last row illustrates the percentage of faces detected in our anonymized dataset by the same face detectors. The intermediate rows present the detection ratios achieved by employing SOTA anonymization methods.

It is worth noting that the original dataset exhibits a few instances where certain faces were not detected by the employed face detectors, resulting in a detection ratio of 98.55% for dlib [13] and 99.91% for MTCNN [193]. This can be attributed to the diverse range of facial poses in the original dataset, encompassing extreme angles and substantial occlusions. However, these challenging examples contribute to enhancing the robustness of face detectors when training them with such diverse data. We visualize this phenomenon in Fig. 6.3 and provide a further elaborate in the Sec. 6.6.2.2.

Therefore, in the first two columns of Tab. 6.1, a detection rate closer to that of the original dataset indicates that the employed anonymization method better preserves the diversity of data. Remarkably, our proposed method outperforms other approaches, yielding detection rates of 98.26% (0.29% lower than the original dataset) and 99.83% (0.08% lower than the original dataset), respectively.

Identity anonymity (ID Anon) To evaluate the efficacy of identity anonymity, we conducted facial identity verification using the FaceNet [194] network pre-trained on two datasets, namely VGG [196] and CASIA [197]. The corresponding success rates for concealing identity information are summarized in Tab. 6.1.

Our approach achieved the second-highest success rate among the SOTA techniques. Notably, the CIAGAN method, which serves as a face inpainting technique, achieved the highest score. However, this accomplishment was accompanied by compromised image quality and a trade-off in preserving facial attributes.

Image quality The image quality results (FID [195]) of our approach in comparison to others are documented in the fifth column of Tab. 6.1.

A lower FID distance, also known as the FID score, indicates higher quality and diversity of the anonymized images. This implies that the distribution of the anonymized images closely aligns with the distribution of real images.

The image quality results are presented in Tab. 6.1. Our method demonstrated the lowest FID score of 23.42, showcasing a significant improvement of approximately 17% compared to the second lowest FALCO [12] method, which obtained a score of 28.19.

Facial attribute preservation To evaluate the metric of 'facial attribute preservation' in the anonymized datasets, we conducted training on a CNN network using the anonymized datasets as the training set. Specifically, the pre-trained MobileNetV2 model is utilized in this study. The original dataset served as the ground truth for training and evaluation purposes and is split into training and testing sets. All models trained using different datasets were evaluated on the testing set of the original dataset. To ensure fairness, consistent strategies are applied across the different datasets, including data size and training parameter settings.

The facial attribute categories are derived from the attribute labels in the CelebA dataset [192]. To compare the effectiveness of facial attribute preservation with face inpainting methods, such as DeepPrivacy [10] and CIAGAN [11], we categorize the 38 attribute labels (excluding 'Attractive' and 'Blurry') into two groups: inner face labels and outer face labels, based on the corresponding facial regions represented by each attribute category. The specific categorizations are presented in Tab. 6.2.

The quantitative results of facial attribute preservation are presented in Tabs. 6.1 and 6.2, where the values indicate the percentage of correct classifications, representing the ratio of correctly predicted labels to the total number of predictions. The original dataset is considered as the ground truth in the tables, serving as the baseline for comparing the anonymized datasets. The higher classification accuracy indicates better performance of the models trained on a particular dataset. This implies that the anonymized dataset possesses a facial attribute distribution that is closer to the ground truth, thereby enabling more effective support for various machine-learning tasks.

In Tab. 6.1, the last two columns illustrate the classification accuracies for inner face and outer face attributes, respectively. Under this setting, the original dataset achieves accuracies of 85.64% and 85.21% for inner face and outer face attributes, respectively. Our dataset achieves accuracies of 82.73% and 82.5% under the same settings. Compared with the second-best results, our dataset demonstrates an improvement of 2.44% in accuracy for inner face attributes (compared to DeepPrivacy [10]) and an increase of 1.78% for outer face attributes (compared to CIAGAN [11]).

Tab. 6.2 presents the classification accuracies for different datasets across all labels. A total of 38 label categories are utilized for evaluation. Our approach achieves the highest

prediction accuracy for 31 label categories, while CIAGAN [11] achieves the highest accuracy for 6 categories, and FALCO [12] attains the highest score for the '5 o'Clock Shadow' label. Through comparison, it is evident that our dataset exhibits superior facial attribute preservation performance.

6.6.2.2 Qualitative evaluation

In this section, we conduct a comprehensive qualitative evaluation to compare the efficacy of our proposed method with other approaches.

In the context of qualitative evaluation, visual quality serves as a crucial criterion for comparing our approach with other SOTA methods and demonstrating the observed effects. The primary visual performances utilized for presentation are displayed in Fig. 6.3 and Fig. 6.4. Notably, Fig. 6.3 focuses specifically on showcasing unique instances extracted from the original dataset, and we delve into its analysis in the subsequent discussion.

Fig. 6.3 illustrates the results of face detection using dlib's frontal face detector (dlib [13]). The face detection boxes, delineated by black boxes, indicate successful detections, while images without the black boxes signify instances where the face detection failed. It was noted that certain face images in the real dataset were not detected by the dlib frontal face detector [13]. Consequently, we conducted a comprehensive investigation of this particular subset of data, identifying two broad categories: faces turned at an extreme angle and faces occluded by obstacles. It is pertinent to mention that although these images constitute only a small fraction of the face dataset employed in our experiments, they are frequently encountered in real-world scenarios. Therefore, it becomes imperative for sophisticated machine learning models to effectively handle these unique samples during the training phase by assigning appropriate weights, thereby optimizing the overall performance.

Upon meticulous examination of these samples, an intriguing observation emerged: our anonymized dataset exhibits detections that closely resemble those of the original dataset. In contrast, when processing these images, both the FALCO [12] and the DeepPrivacy models [10] tend to treat them as normal face images, erroneously introducing facial regions that were not present in the original images. Moreover, the data preprocessing of CIAGAN [11], which relies on facial key points information, proved to be inefficient in handling these images. Thus, we demonstrate that our anonymized dataset outperforms alternative approaches in preserving the intricate nature of facial poses.

Next, we present additional visual comparison images, as depicted in Fig. 6.4. Sub-

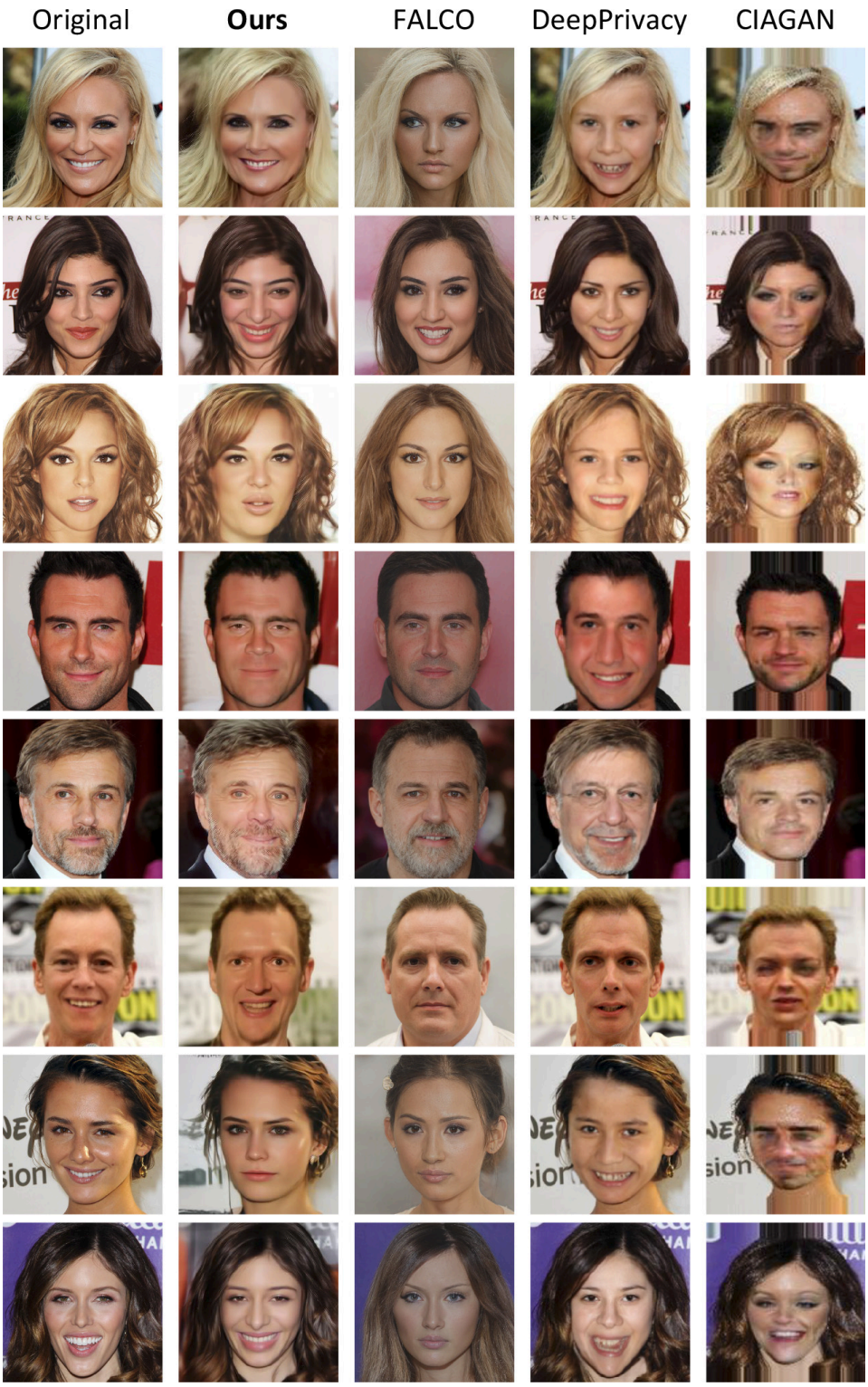


Figure 6.4: Comparative analysis of visual effects: Our method vs. FALCO [12], DeepPrivacy [10], and CIAGAN [11].

sequently, we explain the comparison results from three distinct perspectives. Firstly, regarding image quality, our generated and FLACO images demonstrate a superior level of realism compared to the images produced by others. This enhancement can be attributed to the remarkable capabilities of the generation model in retaining intricate details, such as the accurate handling of teeth processing. Secondly, concerning the diversity of generated faces, our approach showcases a higher retention of attributes from the original dataset, encompassing expressions, skin color, age, and other pertinent features. This advantage can be attributed to the inherent strengths of the Diffusion model in effectively handling complex attributes during the generation process. Lastly, we prioritize privacy protection by effectively safeguarding potentially sensitive information in the image backgrounds. This distinguishes our method from DeepPrivacy [10] and CIAGAN [11], which may overlook such privacy concerns. Concurrently, we aim to maintain the dataset's diversity to the utmost extent possible, setting our approach apart from FALCO [12], which may inadvertently compromise diversity during the generation process.

6.6.3 Ablation study

This section presents an ablation study conducted to investigate optimization methods and identify the most effective approach for generating anonymized datasets.

In the domain of image privacy protection utilizing optimization techniques, two primary categories of optimization methods have emerged: the class replacement scheme (CRS) and the class indistinguishable scheme (CIS). These methods differ in their fundamental approach to handling the classes to be protected. The CRS adopts a new label as the optimization target, aiming to induce the misclassification of the protected class into the new class by the DNN. Conversely, the CIS seeks to maximize the class loss function to achieve the incorrect classification.

For our specific task of safeguarding facial identity information, we explore the impact of using different optimization methods on the dataset. To facilitate discussions, we refer to these two approaches as the identity replacement scheme (IRS) and the identity indistinguishable scheme (IIS). More specifically, the main difference between the IRS and the IIS lies in their respective approaches to handling identity loss. In the case of the IRS, we select a target face from the fake dataset and aim to minimize the identity distance between the anonymized face and the specified target face. The identity loss of

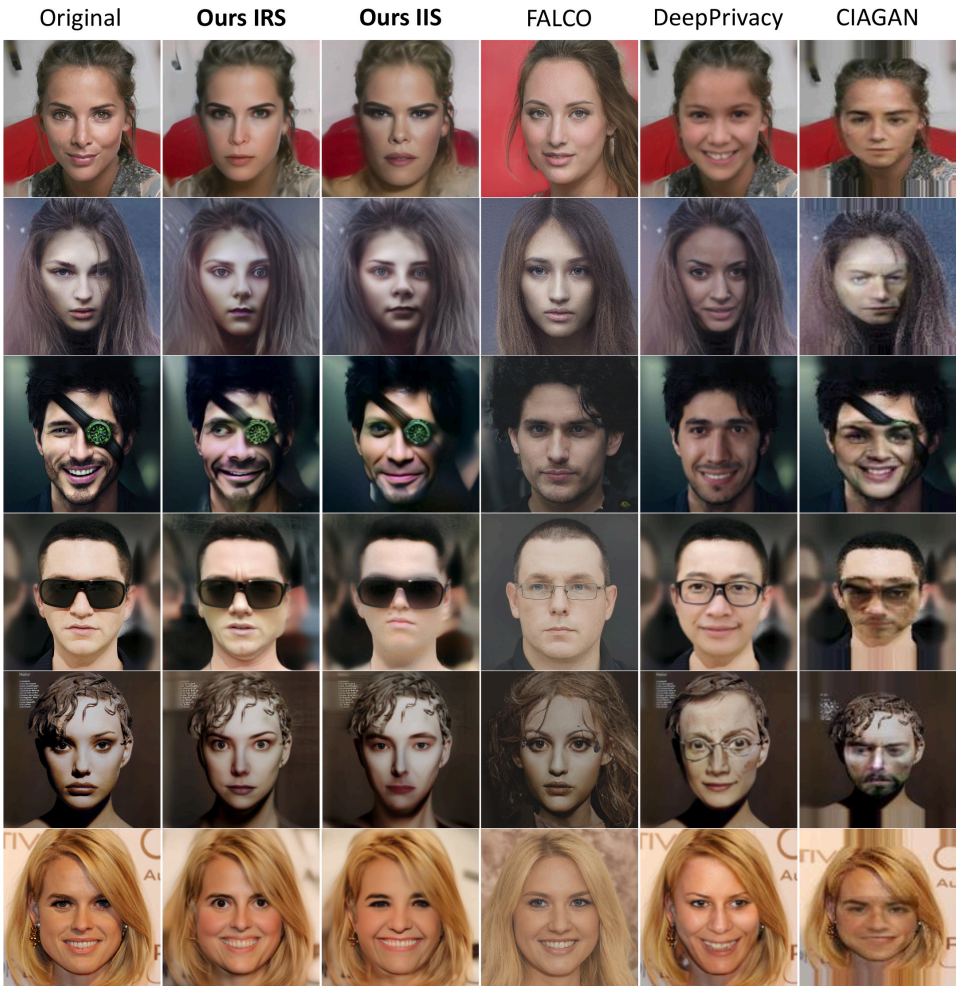


Figure 6.5: Visual quality of ablation study. **Ours** compared to FALCO [12], DeepPrivacy [10], CIAGAN [11].

the IRS is defined as follows:

$$\begin{aligned} \mathcal{L}_{IRS} &= 1 - \left| \frac{E_{\mathcal{A}}(\mathbf{x}_A^i) \cdot E_{\mathcal{A}}(\mathbf{x}_F^i)}{\left\| E_{\mathcal{A}}(\mathbf{x}_A^i) \right\|_2 \cdot \left\| E_{\mathcal{A}}(\mathbf{x}_F^i) \right\|_2} \right| \\ &= 1 - \left| \cos(E_{\mathcal{A}}(\mathbf{x}_A^i), E_{\mathcal{A}}(\mathbf{x}_F^i)) \right|, \end{aligned} \tag{6.9}$$

where $E_{\mathcal{A}}$ is a pre-trained ArcFace [138] identity encoder. $\mathbf{x}_A^i \in \mathcal{X}_A$ and $\mathbf{x}_F^i \in \mathcal{X}_F$ are the images from the anonymized dataset and the images from the fake dataset respectively. $\cos(\cdot,\cdot)$ denotes the cosine similarity. The loss yields a value ranging from 0 to 1, where 0 indicates that the vectors are identical, while 1 signifies no similarity.

In the IIS, our objective is to minimize the cosine similarity between the images from the real and anonymized datasets. This is achieved by optimizing the identity loss, as

depicted in (6.7).

	Face Dete ↑	ID Anon ↑		Image quality ↓	Attributes ↑
	dlib(%)	CASIA(%)	VGG(%)	FID	Average(%)
Ours IRS	+**0.13**	95.33	96.10	**21.17**	**83.12**
Ours IIS	−0.29	**97.55**	**98.21**	23.42	82.63

Table 6.3: Ablation study evaluation results. Face Detection Accuracy (Face Dete), Identity Anonymity (ID Anon), Image Quality, and Attributes Classification.

Based on the results presented in Tab. 6.3, each of the two optimization methods, namely the IRS and the IIS, possesses its own advantages and limitations. Specifically, IRS demonstrates certain improvements in image quality, whereas IIS outperforms IRS in preserving facial identity.

The relatively lower privacy protection rate observed in the IRS can be attributed to the possibility of selected facial features in the images being similar to those in the original dataset. We will delve deeper into this aspect in a subsequent section for further analysis. Moreover, Fig. 6.5 shows visual comparisons with SOTA methods, offering a visual performance achieved by the different approaches.

6.7 Conclusions

In this chapter, we propose a novel framework for image anonymization that specifically targets the optimization of latent space vectors within pre-trained diffusion models. Our approach leverages an identity loss function and an attribute preserving loss function to directly operate in the latent space of the pre-trained diffusion model, thereby eliminating the necessity for training intricate networks. Through extensive experimentation, we demonstrate the efficacy of our method in successfully anonymizing the identity of images while better preserving facial attributes compared to existing state-of-the-art techniques. Consequently, our approach contributes to advancing the field of image privacy by improving both de-identification and facial attribute preservation.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

In this thesis, we have tackled the pressing issue of image privacy protection in the era of data sharing and AI technology. The rapid growth of data sharing and the advancements in AI have raised significant privacy concerns, as personal information embedded in images can be easily extracted and misused by adversaries. Traditional privacy protection methods designed for human adversaries are insufficient in safeguarding privacy against AI adversaries. Therefore, our research focused on developing effective methods to counter the privacy risks posed by AI.

In Chapter 1, We introduce the motivation of this thesis and identify several challenges in the field of image privacy protection, including inadequate protection against AI adversaries, lack of comprehensive privacy definition and quantification, limited applicability and controllability of existing methods, lack of provable privacy measurements, limited effectiveness of existing differential privacy methods for images, and the need to balance privacy and utility.

Chapter 2 provides a comprehensive overview of the related work in the field, including image privacy definition, traditional protection methods, using adversarial examples in image privacy, StyleGAN-based autoencoders for image manipulation, the application of differential privacy in image analysis, and the use of diffusion models for image generation and manipulation tasks. The chapter highlights the capabilities of these approaches

and their contributions to privacy protection and high-quality image synthesis. It sets the stage for the novel approach presented in the thesis, the Diffusion Autoencoders, which combines trainable encoders and DPM decoders for precise image reconstruction and challenging image manipulation tasks.

To address these challenges, we made several contributions to this thesis. Firstly, Chapter 3 proposed an image privacy protection framework that utilizes adversarial perturbations to conceal private information from AI while remaining imperceptible to human observers. This framework provides an effective method to protect privacy in images shared on social media platforms.

In Chapter 4, we introduced a differentially private image (DP-Image) framework that redefines differential privacy concepts in the context of image data. This framework adds differential privacy noise to image feature vectors, ensuring provable privacy protection while maintaining utility and context consistency. We implemented and evaluated the proposed DP-Image protection mechanisms, demonstrating their effectiveness in safeguarding individuals' privacy against human and AI adversaries.

In Chapter 5, we focused on user-centric privacy protection and developed mechanisms that empower users to have control over their privacy in computer vision applications. These mechanisms offer customizable privacy levels and enable users to manage their privacy preferences effectively.

Lastly, Chapter 6 addressed the challenge of high-quality image de-identification by developing techniques that remove or obfuscate identifying information from images while preserving their quality. This enables the sharing and analysis of images without compromising privacy.

In conclusion, this thesis has made important contributions to the field of image privacy protection. We have addressed the challenges posed by AI adversaries, defined and quantified image privacy, developed frameworks and mechanisms for privacy protection, and evaluated their effectiveness using real-life image datasets.

7.2 Future work

In the future, we plan to incorporate more scenarios into the field of image privacy and enhance the level of protection provided by existing methods. Now, despite employing potent tools such as diffusion models as image generators to overcome the limitations of GAN models in dealing with intricate scenes, it has been observed that diffusion models impose substantial computational demands. Consequently, our objective is to enhance

the practicality of the protection framework by streamlining algorithmic procedures or investigating alternative models, such as stable diffusion.

Furthermore, it is essential to address privacy protection concerns in other multimedia formats, specifically videos. Video conferences have become increasingly prevalent with the widespread adoption of remote work after the pandemic. This surge in video usage has raised significant concerns regarding protecting sensitive information embedded within the video content, surpassing the scope of image-based privacy concerns. Consequently, safeguarding video privacy is a pivotal and urgent research area.

Finally, we direct our attention to a more expansive realm of privacy protection. Various human biological traits can be quantified and represented as images, including fingerprints, irises, and even sound, which can be converted into frequency and amplitude images. By employing the DNN networks, we can apply image privacy protection techniques to these biological features, thereby presenting a broader outlook for the field of privacy protection.

BIBLIOGRAPHY

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[2] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[3] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019.

[4] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," 2021.

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, (Red Hook, NY, USA), Curran Associates Inc., 2020.

[6] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[7] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: protecting privacy against unauthorized deep learning models," in *29th USENIX Security Symposium (USENIX Security 20)*, pp. 1589–1604, 2020.

[8] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," *International symposium on visual computing*, pp. 565–578, 2019.

[9] M. Maximov, I. Elezi, and L. Leal-Taixe, "Ciagan: Conditional identity anonymization generative adversarial networks," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2020.

[10] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," 2019.

[11] M. Maximov, I. Elezi, and L. Leal-Taixe, "CIAGAN: Conditional identity anonymization generative adversarial networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, jun 2020.

[12] S. Barattin, C. Tzelepis, I. Patras, and N. Sebe, "Attribute-preserving face dataset anonymization via latent code optimization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023.

[13] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

[14] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2015.

[15] A. Sears, F. Dechesne, I. Georgieva, T. Tani, and S. van der Hof, *EU Personal Data Protection in Policy and Practice*.
Information Technology and Law Series, T.M.C. Asser Press, 2019.

[16] M. X. Heiligenstein, "Facebook data breaches: Full timeline through 2023," *Firewall Times*, 2023.

[17] F. T. Limited, "Faceapp," *FaceApp*, 2023.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[19] O. Tov, Y. Alaluf, Y. Nitzan, O. Patashnik, and D. Cohen-Or, "Designing an encoder for stylegan image manipulation," *arXiv preprint arXiv:2102.02766*, 2021.

[20] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, 2017.

[21] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato,

R. Hadsell, M. Balcan, and H. Lin, eds.), vol. 33, pp. 6840–6851, Curran Associates, Inc., 2020.

[22] L. Du, W. Zhang, H. Fu, W. Ren, and X. Zhang, "An efficient privacy protection scheme for data security in video surveillance," *Journal of Visual Communication and Image Representation*, vol. 59, pp. 347–362, 2019.

[23] P. Agrawal and P. Narayanan, "Person de-identification in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 3, pp. 299–310, 2011.

[24] C. Thorpe, F. Li, Z. Li, Z. Yu, D. Saunders, and J. Yu, "A coprime blur scheme for data security in video surveillance," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 12, pp. 3066–3072, 2013.

[25] S. Tansuriyavong and S.-i. Hanaki, "Privacy protection by concealing persons in circumstantial video image," in *Proceedings of the 2001 workshop on Perceptive user interfaces*, pp. 1–4, 2001.

[26] D. Chen, Y. Chang, R. Yan, and J. Yang, "Tools for protecting the privacy of specific individuals in video," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1–9, 2007.

[27] E. M. Newton, L. Sweeney, and B. Malin, "Preserving privacy by de-identifying face images," *IEEE transactions on Knowledge and Data Engineering*, vol. 17, no. 2, pp. 232–243, 2005.

[28] T. Gerstner, D. DeCarlo, M. Alexa, A. Finkelstein, Y. Gingold, and A. Nealen, "Pixelated image abstraction with integrated user constraints," *Computers & Graphics*, vol. 37, no. 5, pp. 333–347, 2013.

[29] M. Gervautz and W. Purgathofer, "A simple method for color quantization: octree quantization," *Graphics gems*, pp. 287–293, 1990.

[30] P. Heckbert, "Color image quantization for frame buffer display," *SIGGRAPH Comput. Graph.*, vol. 16, p. 297,Äì307, jul 1982.

[31] M. Orchard and C. Bouman, "Color quantization of images," *IEEE Transactions on Signal Processing*, vol. 39, no. 12, pp. 2677–2690, 1991.

[32] X. Wu, "Color quantization by dynamic programming and principal analysis," *ACM Transactions on Graphics (TOG)*, vol. 11, no. 4, pp. 348–372, 1992.

[33] L. Fan, "Image pixelization with differential privacy," in *Data and Applications Security and Privacy XXXII: 32nd Annual IFIP WG 11.3 Conference, DBSec 2018, Bergamo, Italy, July 16–18, 2018, Proceedings 32*, pp. 148–162, Springer, 2018.

[34] C. Dwork, "Differential privacy," in *Automata, Languages and Programming: 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II 33*, pp. 1–12, Springer, 2006.

[35] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," 2016.

[36] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Faceless person recognition; privacy implications in social media," 2016.

[37] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng, and B. Y. Zhao, "Fawkes: Protecting personal privacy against unauthorized deep learning models," in *Proc. of USENIX Security*, 2020.

[38] S. J. Oh, M. Fritz, and B. Schiele, "Adversarial image perturbation for privacy protection a game theory perspective," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1491–1500, IEEE, 2017.

[39] A. Rajabi, R. B. Bobba, M. Rosulek, C. Wright, and W.-c. Feng, "On the (im) practicality of adversarial perturbation for image privacy," *Proceedings on Privacy Enhancing Technologies*, 2021.

[40] H. Xue, B. Liu, M. Din, L. Song, and T. Zhu, "Hiding private information in images from ai," in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2020.

[41] S. Hu, X. Liu, Y. Zhang, M. Li, L. Y. Zhang, H. Jin, and L. Wu, "Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15014–15023, June 2022.

[42] B. Liu, M. Ding, T. Zhu, Y. Xiang, and W. Zhou, "Using adversarial noises to protect privacy in deep learning era," in *2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2018.

[43] B. Liu, M. Ding, T. Zhu, Y. Xiang, and W. Zhou, "Adversaries or allies? privacy and deep learning in big data era," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 19, p. e5102, 2019.

[44] Y. Liu, W. Zhang, and N. Yu, "Protecting privacy in shared photos via adversarial examples based stealth," *Security and Communication Networks*, vol. 2017, 2017.

[45] B. Liu, J. Xiong, Y. Wu, M. Ding, and C. M. Wu, "Protecting multimedia privacy from both humansand ai," in *in Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2019.

[46] T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[47] A. Bhattad, M. J. Chong, K. Liang, B. Li, and D. A. Forsyth, "Unrestricted adversarial examples via semantic manipulation," 2020.

[48] Y. Song, R. Shu, N. Kushman, and S. Ermon, "Constructing unrestricted adversarial examples with generative models," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), vol. 31, Curran Associates, Inc., 2018.

[49] C. Xiao, J.-Y. Zhu, B. Li, W. He, M. Liu, and D. Song, "Spatially transformed adversarial examples," 2018.

[50] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Transactions on Privacy and Security*, vol. 22, pp. 1–30, jun 2019.

[51] S. Komkov and A. Petiushko, "AdvHat: Real-world adversarial attack on ArcFace face ID system," in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, jan 2021.

[52] Z. Xiao, X. Gao, C. Fu, Y. Dong, W. Gao, X. Zhang, J. Zhou, and J. Zhu, "Improving transferability of adversarial patches on face recognition with generative models," 2021.

[53] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976, 2017.

[54] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen, "Improved techniques for training gans," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[55] O. Poursaeed, T. Jiang, H. Yang, S. Belongie, and S. Lim, "Robustness and generalization via generative adversarial training," 2021.

[56] Z.-A. Zhu, Y.-Z. Lu, and C.-K. Chiang, "Generating adversarial examples by makeup attacks on face recognition," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2516–2520, 2019.

[57] N. Guetta, A. Shabtai, I. Singh, S. Momiyama, and Y. Elovici, "Dodging attack using carefully crafted natural makeup," 2021.

[58] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Z.-H. Zhou, ed.), pp. 1252–1258, International Joint Conferences on Artificial Intelligence Organization, 8 2021.
Main Track.

[59] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," 2017.

[60] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," 2016.

[61] S. Iizuka, E. Simo-Serra, and H. Ishikawa, "Globally and locally consistent image completion," *ACM Trans. Graph.*, vol. 36, jul 2017.

[62] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," 2016.

[63] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," 2019.

[64] W. Xiong, J. Yu, Z. Lin, J. Yang, X. Lu, C. Barnes, and J. Luo, "Foreground-aware image inpainting," 2019.

[65] Y. Ren, X. Yu, R. Zhang, T. H. Li, S. Liu, and G. Li, "Structureflow: Image inpainting via structure-aware appearance flow," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[66] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," 2018.

[67] Y. Song, C. Yang, Z. Lin, X. Liu, Q. Huang, H. Li, and C. C. J. Kuo, "Contextual-based image inpainting: Infer, match, and translate," 2018.

[68] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," 2018.

[69] Z. Yan, X. Li, M. Li, W. Zuo, and S. Shan, "Shift-net: Image inpainting via deep feature rearrangement," 2018.

[70] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang, "Free-form image inpainting with gated convolution," 2019.

[71] H. Liu, B. Jiang, Y. Xiao, and C. Yang, "Coherent semantic attention for image inpainting," 2019.

[72] Y. Wang, X. Tao, X. Qi, X. Shen, and J. Jia, "Image inpainting via generative multi-column convolutional neural networks," 2018.

[73] H. Liu, B. Jiang, Y. Song, W. Huang, and C. Yang, "Rethinking image inpainting via a mutual encoder-decoder with feature equalizations," 2020.

[74] Z. Yi, Q. Tang, S. Azizi, D. Jang, and Z. Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," 2020.

[75] H. Dong, X. Liang, Y. Zhang, X. Zhang, X. Shen, Z. Xie, B. Wu, and J. Yin, "Fashion editing with adversarial parsing learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8120–8128, 2020.

[76] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin, "Towards multi-pose guided virtual try-on network," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9026–9035, 2019.

[77] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7543–7552, 2018.

[78] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*, pp. 1–19, Springer, 2008.

[79] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Proceedings of the theory of cryptography conference (TCC'06)*, pp. 265–284, Springer, 2006.

[80] L. Fan, "Image pixelization with differential privacy," in *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 148–162, Springer, 2018.

[81] L. Fan, "Practical image obfuscation with provable privacy," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 784–789, IEEE, 2019.

[82] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672, IEEE, 2019.

[83] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

[84] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," 2022.

[85] O. Avrahami, O. Fried, and D. Lischinski, "Blended latent diffusion," *arXiv preprint arXiv:2206.02779*, 2022.

[86] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

[87] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

[88] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[89] Anonymous, "Diff-privacy: Diffusion-based face privacy protection," in *Submitted to The Twelfth International Conference on Learning Representations*, 2023. under review.

[90] Zephoria, "Top 20 Facebook Statistics - Updated June 2019."

[91] T. Orekondy, B. Schiele, and M. Fritz, "Towards a Visual Privacy Advisor: Understanding and Predicting Privacy Risks in Images," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[92] T. Orekondy, M. Fritz, and B. Schiele, "Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[93] S. J. Oh, M. Fritz, and B. Schiele, "Adversarial Image Perturbation for Privacy Protection – A Game Theory Perspective," in *International Conference on Computer Vision (ICCV)*, 2017.

[94] W. Meng, X. Xing, A. Sheth, U. Weinsberg, and W. Lee, "Your online interests: Pwned! a pollution attack against targeted advertising," in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 129–140, ACM, 2014.

[95] Phys.org, "Japan researchers warn of fingerprint theft from 'peace' sign," jan 2017.

[96] E. Tseng, "Computer-vision-assisted location accuracy augmentation," Apr. 5 2016. US Patent 9,305,024.

[97] W.-H. Li, F.-T. Hong, and W.-S. Zheng, "Learning to learn relation for important people detection in still images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[98] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[99] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[100] B. Liu, M. Ding, T. Zhu, Y. Xiang, and W. Zhou, "Adversaries or allies? privacy and deep learning in big data era," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 19, p. e5102, 2019.
e5102 cpe.5102.

[101] B. Liu, J. Xiong, Y. Wu, M. Ding, and C. M. Wu, "Protecting multimedia privacy from both humansand ai," in *in Proc. IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, 2019.

[102] Y. Liu, W. Zhang, and N. Yu, "Protecting privacy in shared photos via adversarial examples based stealth," *Security and Communication Networks*, vol. 2017, 2017.

[103] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015.

[104] X. Ling, S. Ji, J. Zou, J. Wang, C. Wu, B. Li, and T. Wang, "Deepsec: A uniform platform for security analysis of deep learning model," in *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 673–690, 2019.

[105] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on image processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[106] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[107] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," *arXiv preprint arXiv:1609.00408*, 2016.

[108] B. Liu, M. Ding, T. Zhu, Y. Xiang, and W. Zhou, "Adversaries or allies? privacy and deep learning in big data era," *Concurrency and Computation: Practice and Experience*, vol. 31, no. 19, p. e5102, 2019.

[109] S. J. Oh, M. Fritz, and B. Schiele, "Adversarial image perturbation for privacy protection a game theory perspective," in *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 1491–1500, IEEE, 2017.

[110] T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[111] A. Rajabi, R. B. Bobba, M. Rosulek, C. V. Wright, and W.-c. Feng, "On the (im) practicality of adversarial perturbation for image privacy," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 1, pp. 85–106, 2021.

[112] Q. Sun, L. Ma, S. Joon Oh, L. Van Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5050–5059, 2018.

[113] Y. Wu, F. Yang, Y. Xu, and H. Ling, "Privacy-protective-gan for privacy preserving face de-identification," *Journal of Computer Science and Technology*, vol. 34, no. 1, pp. 47–60, 2019.

[114] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele, "A hybrid model for identity obfuscation by face replacement," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 553–569, 2018.

[115] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Faceshifter: Towards high fidelity and occlusion aware face swapping," *arXiv preprint arXiv:1912.13457*, 2019.

[116] H.-P. Wang, T. Orekondy, and M. Fritz, "Infoscrub: Towards attribute privacy by targeted obfuscation," *arXiv preprint arXiv:2005.10329*, 2020.

[117] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, IEEE, 2017.

[118] A. Rényi *et al.*, "On measures of entropy and information," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, The Regents of the University of California, 1961.

[119] V. Feldman, I. Mironov, K. Talwar, and A. Thakurta, "Privacy amplification by iteration," *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, Oct 2018.

[120] S.-S. Kim, "Convolutional neural networks for visual recognition," *machine learning*, 2016.

[121] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[122] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[123] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016.

[124] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning.," in *AAAI*, vol. 4, p. 12, 2017.

[125] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[126] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[127] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[128] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3828–3836, 2015.

[129] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, pp. 3119–3127, 2015.

[130] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[131] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[132] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[133] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

[134] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

[135] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[136] L. Mescheder, S. Nowozin, and A. Geiger, "Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2391–2400, JMLR. org, 2017.

[137] NVIDIA, "Flickr-faces-hq (ffhq) dataset," 2019.

[138] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[139] Microsoft, "Azure facial recognition api."

[140] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in neural information processing systems*, pp. 6626–6637, 2017.

[141] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: a stylegan encoder for image-to-image translation," *arXiv preprint arXiv:2008.00951*, 2020.

[142] J. P. Pesce, D. L. Casas, G. Rauber, and V. Almeida, "Privacy attacks in social media using photo tagging networks: a case study with facebook," in *Proceedings of*

the 1st Workshop on Privacy and Security in Online Social Media, pp. 1–8, 2012.

[143] N. Vyas, A. C. Squicciarini, C.-C. Chang, and D. Yao, "Towards automatic privacy management in web 2.0 with semantic analysis on annotations," in *2009 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pp. 1–10, IEEE, 2009.

[144] A. C. Squicciarini, H. Xu, and X. Zhang, "Cope: Enabling collaborative privacy management in online social networks," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 3, pp. 521–534, 2011.

[145] T. N. Y. Times, "San Francisco Bans Facial Recognition Technology," 2019.

[146] S. J. Oh, R. Benenson, M. Fritz, and B. Schiele, "Faceless person recognition: Privacy implications in social media," *European Conference on Computer Vision*, pp. 19–35, 2016.

[147] H. Xue, B. Liu, M. Din, L. Song, and T. Zhu, "Hiding private information in images from ai," *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2020.

[148] X. Gu, W. Luo, M. S. Ryoo, and Y. J. Lee, "Password-conditioned anonymization and deanonymization with face identity transformers," *European Conference on Computer Vision*, p. 727,Äì743, 2020.

[149] J.-W. Chen, L.-J. Chen, C.-M. Yu, and C.-S. Lu, "Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6478–6487, 2021.

[150] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do, "Semantic image inpainting with deep generative models," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5485–5493, 2017.

[151] J. Cao, B. Liu, Y. Wen, Y. Zhu, R. Xie, L. Song, L. Li, and Y. Yin, "Hiding among your neighbors: Face image privacy protection with differential private k-anonymity," *2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, pp. 1–6, 2022.

[152] K. Zhang, J. Tian, H. Xiao, Y. Zhao, W. Zhao, and J. Chen, "A numerical splitting and adaptive privacy budget allocation based ldp mechanism for privacy preservation in blockchain-powered iot," *IEEE Internet of Things Journal*, pp. 1–1, 2022.

[153] M. U. Hassan, M. H. Rehmani, and J. Chen, "Anomaly detection in blockchain networks: A comprehensive survey," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2022.

[154] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[155] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing robust adversarial examples," *arXiv preprint arXiv:1707.07397*, 2017.

[156] Q. Huang, I. Katsman, H. He, Z. Gu, S. Belongie, and S.-N. Lim, "Enhancing adversarial example transferability with an intermediate level attack," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4733–4742, 2019.

[157] N. Inkawhich, W. Wen, H. H. Li, and Y. Chen, "Feature space perturbations yield more transferable adversarial examples," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7066–7074, 2019.

[158] S. Pidhorskyi, D. A. Adjeroh, and G. Doretto, "Adversarial latent autoencoders," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14104–14113, 2020.

[159] G. Zhang, B. Liu, T. Zhu, A. Zhou, and W. Zhou, "Visual privacy attacks and defenses in deep learning: a survey," *Artificial Intelligence Review*, 01 2022.

[160] Y. Zhao and J. Chen, "A survey on differential privacy for unstructured data content," *ACM Comput. Surv.*, vol. 54, sep 2022.

[161] Y. Zhao, D. Yuan, J. T. Du, and J. Chen, "Geo-ellipse-indistinguishability: Community-aware location privacy protection for directional distribution," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–11, 2022.

[162] Y. Zhao, B. Liu, T. Zhu, M. Ding, and W. Zhou, "Private-encoder: Enforcing privacy in latent space for human face images," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 3, p. e6548, 2022.

[163] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–10, 2020.

[164] V. Cherepanova, M. Goldblum, H. Foley, S. Duan, J. P. Dickerson, G. Taylor, and T. Goldstein, "Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition," *International Conference on Learning Representations*, 2021.

[165] J. Zhang, J. Sang, X. Zhao, X. Huang, Y. Sun, and Y. Hu, "Adversarial privacy-preserving filter," *Proceedings of the 28th ACM International Conference on Multimedia*, p. 1423,Äì1431, 2020.

[166] Y. Wen, B. Liu, M. Ding, R. Xie, and L. Song, "Identitydp: Differential private identification protection for face images," *Neurocomputing*, vol. 501, pp. 197–211, 2022.

[167] Y.-L. Pan, J.-C. Chen, and J.-L. Wu, "A multi-factor combinations enhanced reversible privacy protection system for facial images," *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021.

[168] J. Cao, B. Liu, Y. Wen, R. Xie, and L. Song, "Personalized and invertible face de-identification by disentangled identity information manipulation," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3334–3342, 2021.

[169] Z. You, S. Li, Z. Qian, and X. Zhang, "Reversible privacy-preserving recognition," *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021.

[170] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," *Proceedings of the european conference on computer vision (ECCV)*, pp. 620–636, 2018.

[171] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9378–9387, 2019.

[172] Y. Wen, B. Liu, J. Cao, R. Xie, L. Song, and Z. Li, "Identitymask: Deep motion flow guided reversible face video de-identification," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[173] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CVPR*, 2018.

[174] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," *Proceedings of International Conference on Computer Vision (ICCV)*, 12 2015.

[175] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[176] Microsoft, "facial recognition | microsoft azure," *Azure.microsoft.com*, 2022.

[177] Faceplusplus, "face++ - face++ cognitive services," *Faceplusplus.com*, 2022.

[178] Y. Zhang, Y. Lu, H. Nagahara, and R.-i. Taniguchi, "Anonymous camera for privacy protection," in *2014 22nd International Conference on Pattern Recognition*, pp. 4170–4175, 2014.

[179] X. Yu, K. Chinomi, T. Koshimizu, N. Nitta, Y. Ito, and N. Babaguchi, "Privacy protecting visual processing for secure video surveillance," in *2008 15th IEEE International Conference on Image Processing*, pp. 1672–1675, 2008.

[180] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan, "Generating talking face landmarks from speech," in *International Conference on Latent Variable Analysis and Signal Separation*, pp. 372–381, Springer, 2018.

[181] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[182] H. Xue, B. Liu, X. Yuan, M. Ding, and T. Zhu, "Face image de-identification by feature space adversarial perturbation," *Concurrency and Computation: Practice and Experience*, vol. 35, no. 5, p. e7554, 2023.

[183] T. Li and L. Lin, "Anonymousnet: Natural face de-identification with measurable privacy," 2019.

[184] Q. Sun, L. Ma, S. J. Oh, L. V. Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," 2018.

[185] B. Liu, M. Ding, H. Xue, T. Zhu, D. Ye, L. Song, and W. Zhou, "Dp-image: differential privacy for image data in feature space," *arXiv preprint arXiv:2103.07073*, 2021.

[186] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014.

[187] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.

[188] H. Li, Y. Yang, M. Chang, S. Chen, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Srdiff: Single image super-resolution with diffusion probabilistic models," *Neurocomputing*, vol. 479, pp. 47–59, 2022.

[189] A. Sinha, J. Song, C. Meng, and S. Ermon, "D2c: Diffusion-decoding models for few-shot conditional generation," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 12533–12548, Curran Associates, Inc., 2021.

[190] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[191] Y. Zheng, H. Yang, T. Zhang, J. Bao, D. Chen, Y. Huang, L. Yuan, D. Chen, M. Zeng, and F. Wen, "General facial representation learning in a visual-linguistic manner," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18697–18709, June 2022.

[192] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," 2015.

[193] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[194] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[195] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," 2018.

[196] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference*, 2015.

[197] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.