UNIVERSITAT ROVIRA i VIRGILI

# AI-Based Utility-Preserving Face Image Anonymization

*Author:*
Younas KHAN




DOCTORAL THESIS
2025

Younas KHAN

# AI-Based Utility-Preserving Face Image Anonymization

DOCTORAL THESIS

*Supervisors:*
Josep DOMINGO-FERRER
David SÁNCHEZ

Doctoral Programme in Computer Science and Mathematics of Security,
Department of Computer Engineering and Mathematics



UNIVERSITAT ROVIRA i VIRGILI

June, 2025

UNIVERSITAT ROVIRA i VIRGILI

FEM CONSTAR que aquest treball, titulat “AI-based utility-preserving face image anonymization”, que presenta Younas Khan per a l’obtenció del títol de Doctor, ha estat realitzat sota la nostra direcció al Departament d’Enginyeria Informàtica i Matemàtiques d’aquesta universitat.

---

HACEMOS CONSTAR que el presente trabajo, titulado “AI-based utility-preserving face image anonymization”, que presenta Younas Khan para la obtención del título de Doctor, ha sido realizado bajo nuestra dirección en el Departamento de Ingeniería Informática y Matemáticas de esta universidad.

---

WE STATE that the present study, entitled “AI-based utility-preserving face image anonymization”, presented by Younas Khan for the award of the degree of Doctor, has been carried out under our supervision at the Department of Computer Engineering and Mathematics of this university.

---

El/s director/s de la tesi doctoral
El/los director/es de la tesis doctoral
Doctoral Thesis Supervisor/s

DOMINGO FERRER JOSEP - 33890313C

Digitally signed by DOMINGO FERRER JOSEP - 33890313C
Date: 2025.05.21 16:19:52 +02'00'

David Sánchez Ruenes - DNI 47763566H (TCAT)

Digitally signed by David Sánchez Ruenes - DNI 47763566H (TCAT)
DN: c=ES, o=Universitat Rovira i Virgili, 2.5.4.97=VATES-Q9350003A, ou=Empleat públic de nivell mig, sn=Sánchez Ruenes - DNI 47763566H, givenName=David, serialNumber=IDCES-47763566H, cn=David Sánchez Ruenes - DNI 47763566H (TCAT)
Date: 2025.05.21 16:55:56 +02'00'

David Sánchez Ruenes

Josep Domingo-Ferrer

# *Abstract*

The unprecedented sharing of facial images on social media platforms has created significant privacy vulnerabilities, with modern face recognition (FR) systems achieving accuracy rates exceeding 99.8% and enabling unauthorized tracking across platforms. This thesis explores how AI techniques can be repurposed to provide robust privacy protection against machine learning-based threats.

We address facial privacy challenges in two distinct contexts: social media sharing and data anonymization. For social media sharing, we introduce PRISMA (Perceptually Robust Iterative Similarity Momentum Attack), a novel protection method that generates visually imperceptible perturbations using a two-stage approach with momentum-based optimization. PRISMA produces protected images with a unique property: FR models trained on them converge successfully during training but completely fail to recognize clean, unseen images of the same individuals.

For data anonymization, we develop two complementary approaches. On the one hand, EDI-Anon (Explainability-Driven Incremental Image Anonymization) is designed as a general anonymization framework that leverages explainability techniques to identify identity-revealing pixels and applies incremental perturbation guided by continuous re-evaluation of re-identification risk. Our framework significantly improves the privacy/utility trade-off compared to conventional methods by targeting only the most privacy-sensitive pixels while preserving analytical utility. Thanks to its general design, it supports a wide variety of image perturbation techniques.

On the other hand, EDIALS (Explainability-Driven Image Anonymization in Latent Space) builds on a similar idea, but operates in the latent feature space using adversarial autoencoders. EDIALS employs specialized architectures and implements novel feature distortion mechanisms that selectively modify identity-revealing latent features. Our comprehensive evaluation demonstrates that EDIALS achieves near-random-guess level anonymization while retaining remarkably high utility for demographic analysis tasks, substantially outperforming conventional and state-of-the-art approaches. By manipulating the structured representation of facial attributes in the latent space rather than directly modifying pixels, EDIALS provides an optimal balance between privacy protection and analytical utility preservation.

# *Resum*

La compartició sense precedents d'imatges facials a les plataformes de xarxes socials ha creat vulnerabilitats importants en la privadesa, amb sistemes moderns de reconeixement facial (RF) que superen taxes d'encert del 99,8% i permeten el seguiment no autoritzat entre plataformes. Aquesta tesi explora com les tècniques d'intel·ligència artificial poden ser reorientades per oferir una protecció robusta contra amenaces basades en aprenentatge automàtic.

Abordem els reptes de la privadesa facial en dos contextos diferents: la compartició a les xarxes socials i l'anonimització de dades. Per a la compartició a les xarxes socials, introduïm PRISMA (Perceptually Robust Iterative Similarity Momentum Attack), un nou mètode de protecció que genera pertorbacions visualment imperceptibles mitjançant un enfocament de dues fases amb optimització basada en moments. PRISMA produeix imatges protegides amb una propietat única: els models de RF entrenats amb elles convergeixen correctament durant l'entrenament, però fracassen completament a l'hora de reconèixer imatges netes i inèdites de les mateixes persones.

Per a l'anonimització de dades, desenvolupem dos enfocaments complementaris. D'una banda, EDI-Anon (Explainability-Driven Incremental Image Anonymization) fa servir tècniques d'explicabilitat per identificar píxels reveladors de la identitat i aplica pertorbacions incrementals guiades per una reavaluació contínua del risc de reidentificació. El nostre marc millora significativament el compromís entre privadesa i utilitat respecte dels mètodes convencionals, centrant-se només en els píxels més sensibles a la privadesa i preservant la utilitat analítica.

D'altra banda, EDIALS (Explainability-Driven Image Anonymization in Latent Space) amplia aquest concepte operant a l'espai de característiques latents mitjançant autocodificadors adversarials. EDIALS fa servir arquitectures especialitzades i desplega nous mecanismes de distorsió de característiques que modifiquen selectivament les característiques latents reveladores de la identitat. La nostra avaluació exhaustiva demostra que EDIALS aconsegueix una anonimització efectiva propera a la reidentificació aleatòria, mentre manté una utilitat molt elevada, que supera àmpliament els enfocaments convencionals de l'estat de la tècnica. Manipulant la representació estructurada d'atributs facials en l'espai latent en comptes de modificar píxels directament, EDIALS forneix un equilibri òptim entre la protecció de la privadesa i la preservació de la utilitat analítica.

# *Resumen*

La difusión sin precedentes de imágenes faciales en plataformas de redes sociales ha generado importantes vulnerabilidades de privacidad, ya que los modernos sistemas de reconocimiento facial (RF) alcanzan tasas de precisión superiores al 99.8% y permiten el rastreo no autorizado entre plataformas. Esta tesis explora como las técnicas de inteligencia artificial pueden reorientarse para proporcionar una protección robusta frente a amenazas basadas en aprendizaje automático.

Abordamos los desafíos de privacidad facial en dos contextos distintos: el intercambio de imágenes en redes sociales y la anonimización de datos. Para el intercambio en redes sociales, presentamos PRISMA (Perceptually Robust Iterative Similarity Momentum Attack), un método novedoso de protección que genera perturbaciones visualmente imperceptibles mediante un enfoque de dos etapas con optimización basada en momentos. PRISMA produce imágenes protegidas con una propiedad única: los modelos de RF entrenados con ellas convergen correctamente durante el entrenamiento, pero fallan completamente para reconocer imágenes limpias y no vistas de los mismos individuos.

Para la anonimización de datos, desarrollamos dos enfoques complementarios. Por un lado, EDI-Anon (Explainability-Driven Incremental Image Anonymization) aprovecha técnicas de explicabilidad para identificar píxeles que revelan la identidad y aplica perturbaciones incrementales guiadas por una reevaluación continua del riesgo de reidentificación. Nuestro marco mejora significativamente el compromiso entre privacidad y utilidad en comparación con los métodos convencionales, al dirigirse únicamente

a los píxeles más sensibles desde el punto de vista de la privacidad, preservando la utilidad analítica.

Por otro lado, EDIALS (Explainability-Driven Image Anonymization in Latent Space) extiende este concepto operando en el espacio de características latentes mediante autocodificadores adversarios. EDIALS emplea arquitecturas especializadas e implementa nuevos mecanismos de distorsión de características que modifican selectivamente las características latentes reveladoras de identidad. Nuestra evaluación exhaustiva demuestra que EDIALS logra una anonimización efectiva cercana a la reidentificación aleatoria, manteniendo una utilidad notablemente alta, que supera sustancialmente los enfoques convencionales del estado de la técnica. Manipulando la representación estructurada de atributos faciales en lugar de modificar píxeles directamente, EDIALS proporciona un equilibrio óptimo entre la protección de la privacidad y la preservación de la utilidad analítica.

# *Acknowledgements*

First and foremost, I would like to express my sincere gratitude to my supervisors, Prof. Josep Domingo-Ferrer and Prof. David Sánchez, for their invaluable guidance, support, and patience over the past three years. Their expertise, constructive feedback, and belief in my work have been crucial in shaping this dissertation and helping me overcome the numerous challenges I encountered.

A special thanks to the CRISES group, particularly to Dr. Jesús A. Manjón, who was always there to assist with machines, logistics, and anything else I needed, and to Dr. Rami Haffar for his ideas, motivation, and availability as both a mentor and a friend.

To my family—my parents, whose sacrifices and beliefs shaped my path; Abba Jan for his dedication, and Moor Jana for her unconditional love; my wife, whose unwavering support and understanding kept me grounded during the hardest times; my brothers, who always motivated me; and my friends, who were there through it all, offering encouragement and sanity.

Lastly, I would like to thank Prof. Anna Monreale for her hospitality during my time in the KDD research group in Pisa (Italy) in 2024-2025.

This work was funded by the European Commission (project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and D. Sánchez), MCIN/AEI/10.13039/501100011033. Y.K. acknowledges financial support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 945413 and from the Universitat Rovira i Virgili.

# Contents

**Abstract** **v**

**Resum** **vii**

**Resumen** **ix**

**Acknowledgements** **xi**

**1 Introduction** **1**
1.1 Objectives . . . . . 3
1.2 Thesis structure . . . . . 5

**2 Background and state of the art** **7**
2.1 Background . . . . . 7
2.1.1 Machine learning for facial recognition . . . 7
2.1.2 Adversarial machine learning . . . . . 9
Transferability in adversarial examples . . . 10
2.1.3 Explainability in machine learning . . . . . 11
2.1.4 Adversarial autoencoders . . . . . 13
2.2 State of the art in image protection . . . . . 15
2.2.1 Protection against model training on public images . . . . . 15
Fawkes . . . . . 15

Lowkey . . . . . 16
FoggySight . . . . . 17
2.2.2 Image anonymization techniques . . . . . 18
Conventional approaches . . . . . 18
Differential privacy-based methods . . . . . 20
GAN-based approaches . . . . . 21
2.3 Conclusion . . . . . 24

**3 Perceptually robust iterative similarity momentum attack (PRISMA) 25**
3.1 Contributions and plan of this chapter . . . . . 26
3.2 PRISMA . . . . . 27
3.2.1 Stage 1: Target feature computation . . . . . 27
3.2.2 Stage 2: Adversarial example generation . . 30
3.3 Empirical results . . . . . 36
3.3.1 Evaluation dataset . . . . . 37
3.3.2 Parameter configuration and FR models . . . 39
3.3.3 Experimental setup . . . . . 41
Model training configuration . . . . . 41
Baseline methods . . . . . 41
3.3.4 Privacy protection assessment . . . . . 42
Protecting all classes . . . . . 43
Protecting underage identities . . . . . 45
3.3.5 Impact of unprotected training images . . . 48
3.3.6 Robustness against reverse attacks . . . . . 49
3.3.7 Perceptual quality . . . . . 52
3.3.8 Runtime analysis . . . . . 54
3.3.9 Impact of PRISMA parameters . . . . . 55
3.4 Conclusion . . . . . 57

**4 Explainability-driven incremental image anonymization (EDI-Anon) 59**
4.1 Contributions and plan of this chapter . . . . . 60
4.2 EDI-Anon . . . . . 61
4.2.1 EDI-Anon modules . . . . . 62
FR model . . . . . 62
Explainer . . . . . 64
Anonymization function . . . . . 65
Privacy threshold . . . . . 66
4.2.2 EDI-Anon workflow and algorithm . . . . . 67
4.3 Empirical results . . . . . 69
4.3.1 Evaluation dataset . . . . . 71
4.3.2 Evaluation metrics . . . . . 71
4.3.3 Conventional methods . . . . . 73
4.3.4 EDI-Anon . . . . . 77
On EDI-Anon constants . . . . . 79
4.3.5 Comparisons . . . . . 83
Comparison with conventional methods . . 83
Comparison with DeepPrivacy2 . . . . . 85
Runtime comparison . . . . . 86
4.4 Conclusion . . . . . 88

**5 Explainability-driven image anonymization in latent space (EDIALS) 89**
5.1 Contributions and plan of this chapter . . . . . 90
5.2 EDIALS . . . . . 90
5.2.1 Image explainer . . . . . 91
5.2.2 Adapting ABELE to facial recognition . . . . 91
5.2.3 EDIALS design . . . . . 93

5.3 Empirical results . . . . . 97
5.3.1 Evaluation dataset and metrics . . . . . 98
5.3.2 Conventional anonymization methods . . . 99
5.3.3 EDIALS results . . . . . 102
5.3.4 Comparison with DeepPrivacy2 . . . . . 105
5.3.5 Runtime analysis . . . . . 107
5.3.6 Impact of the distortion factor . . . . . 107
5.3.7 Impact of image resolution . . . . . 109
5.4 Conclusion . . . . . 110

**6 Conclusions and future work** **113**
6.1 Contributions and publications . . . . . 114
6.2 Limitations and future work . . . . . 115

**Bibliography** **119**

# List of Figures

2.1 Workflow of conventional image anonymization techniques . . . . . 19

3.1 Sample images protected by Fawkes and Lowkey with their strongest privacy protection settings ('high' and LPIPS=0.05, respectively). Skin artifacts are visible in the faces of the second and third rows. . . . . 26

3.2 PRISMA Stage 1: Model Training and Target Feature Computation. After training an FR model on dataset $X$, each image $x_i$ is processed to extract logits and identify target classes using NTIC. Target images from these classes are collected, and their features are extracted and averaged to compute $F_{target}$ for each class. . . . . 29

3.3 PRISMA Stage 2: Adversarial Example Generation. For each input image $x_i$, the algorithm iteratively: (1) transforms the image, (2) extracts features, (3) computes dual losses ($L_{sim}$ and $L_{PSNR}$), (4) generates gradients, (5) updates using momentum, and (6) clips the result. The process continues until the peak signal-to-noise ratio (PSNR) falls below the threshold, producing the final adversarial image $x_{adv_i}$. 31

3.4 Visual comparison of privacy protection methods: original images (top row) and their protected versions using PRISMA, Fawkes, and Lowkey (subsequent rows) . . . . . 53

4.1 Workflow of EDI-Anon, illustrating the iterative process of detecting identity-revealing pixels through explainability techniques, applying targeted perturbations, and continuously re-evaluating RIR until the privacy threshold is reached. . . . . 63
4.2 Evaluation results of conventional anonymization methods for different privacy parameters. Within each group, from left to right, the bars correspond to RIR, $F_1$-Age, $F_1$-Gender, and $F_1$-Race. . . . . 75
4.3 Evolution of RIR and utilities for EDI-Anon across iterations with different anonymization functions until reaching a 1% privacy threshold . . . . . 78
4.4 Evolution of RIR and utilities for EDI-DP-Blur across iterations with different IDPS Percentages until reaching a 1% privacy threshold . . . . . 81
4.5 Evolution of RIR and utilities for EDI-DP-Blur across iterations with different $\Delta$-parameter values until reaching a 1% privacy threshold . . . . . 82
4.6 Average utility comparison between EDI-Anon and its conventional counterparts at equal RIRs . . . . . 84

5.1 EDIALS workflow: Input $x$ from dataset $X$ is encoded by $\mathcal{E}$ into latent representation $z$, then processed through *neighgen* to create neighborhood $H$. After validation by $\sigma$ and decoding by $\mathcal{D}$, a decision tree $c$ extracts rule $r$ and features $Z_r$ for targeted distortion. The green pathways illustrate how $z_{\text{distorted}}$ is incrementally and iteratively modified by $\alpha$ and decoded until $b$ misclassifies (i.e., fails to re-identify) the resulting $x_{\text{distorted}}$, which is then added to $X_{\text{anon}}$. 95

# List of Tables

3.1 Privacy protection comparison against different FR models trained on images protected with PRISMA, Lowkey, and Fawkes. Accuracy is reported on both protected training images (Protected) and clean probe images (Clean). Lower accuracy on clean images indicates better privacy protection. . . . . 44
3.2 Privacy protection comparison when protecting only minors. . . . . 47
3.3 Impact of unprotected training data on the protection afforded by PRISMA, Lowkey, and Fawkes using ConvNeXt Base. . . . . 49
3.4 Vulnerability to reverse attacks in the minors' classes scenario (53 identities). . . . . 51
3.5 Impact of varying PRISMA parameters. Results shown for ConvNeXt. . . . . 55
3.6 Impact of NTIC and target class image count on PRISMA's performance using ConvNeXt. . . . . 57

4.1 Number of individuals for the various tasks and labels in the MTF dataset . . . . . 71
4.2 Average utility comparison between DeepPrivacy2 and EDI-Anon methods for equal RIR $\approx$ 16% . . . . 86

4.3 Average runtime per image for conventional methods, their EDI-Anon counterparts and DeepPrivacy2 . . . . . 87

5.1 Performance metrics across different privacy-preserving transformations . . . . . 103
5.2 EDIALS performance (with $\alpha = 0.5$) vs. original images. . . . . 104
5.3 Performance comparison between DeepPrivacy2 and EDIALS at comparable RIR levels . . . . . 106
5.4 Impact of distortion factor $\alpha$ on efficiency and utility preservation . . . . . 108
5.5 Impact of image resolution on runtime and utility at 0.42% RIR . . . . . 109

# List of Algorithms

1 PRISMA Stage 1: Model Training and Target Feature Computation . . . . . 30
2 PRISMA Stage 2: Adversarial Example Generation . 35
3 Explainability-Driven Incremental Anonymization . 70
4 EDIALS . . . . . 98

# List of Abbreviations

| | |
|---|---|
| **ABELE** | **A**dversarial **B**lack box **E**xplainer generating **L**atent **E**xemplars |
| **AI** | **A**rtificial **I**ntelligence |
| **CNN** | **C**onvolutional **N**eural **N**etwork |
| **dB** | **d**eci**B**els |
| **DL** | **D**eep **L**earning |
| **DP-Blur** | **D**ifferential **P**rivacy-**B**lurring |
| **DP-based** | **D**ifferential **P**rivacy-**b**ased |
| **DP-Pix** | **D**ifferential **P**rivacy-**P**ixelation |
| **DSSIM** | **S**tructural **D**issimilarity **I**ndex |
| **EDI-Anon** | **E**xplainability-**D**riven **I**ncremental **I**mage **A**nonymization |
| **EDIALS** | **E**xplainability-**D**riven **I**mage **A**nonymization in **L**atent **S**pace |
| **FDF** | **F**lickr **D**iverse **F**aces |
| **FGSM** | **F**ast **G**radient **S**ign **M**ethod |
| **FN** | **F**alse **N**egatives |
| **FP** | **F**alse **P**ositives |
| **FR** | **F**acial **R**ecognition |
| **GAN** | **G**enerative **A**dversarial **N**etwork |
| **GDPR** | **G**eneral **D**ata **P**rotection **R**egulation |
| **IDPS** | **I**dentity-**D**isclosing **P**ixels **S**et |
| **LIME** | **L**ocal **I**nterpretable **M**odel-Agnostic **E**xplanations |
| **LPIPS** | **L**earned **P**erceptual **I**mage **P**atch **S**imilarity |

| | |
|---|---|
| **MI-FGSM** | **M**omentum **I**terative **F**ast **G**radient **S**ign **M**ethod |
| **MTF** | **M**ulti-**T**ask **F**ace |
| **NTIC** | **N**umber of **T**op **I**ncorrect **C**lasses |
| **PI-Net** | **P**erceptual **I**ndistinguishability-**Net** |
| **PRISMA** | **P**erceptually **R**obust **I**terative **S**imilarity **M**omentum **A**ttack |
| **PSNR** | **P**eak **S**ignal-to-**N**oise **R**atio |
| **PSNR-HVS-M** | **P**eak **S**ignal-to-**N**oise **R**atio-**H**uman **V**isual **S**ystem-**M**etric |
| **RIR** | **R**e-**I**dentification **R**isk |
| **RGB** | **R**ed **G**reen **B**lue |
| **TP** | **T**rue **P**ositives |

*I dedicate this thesis to my family,*
*who witnessed both my stumbling steps and*
*determined strides in this intellectual odyssey.*
*To my wife, Mashal, and our three boys, Yousaf,*
*Yamin, and Yusha:*
*In your eyes, I found the reflection of possibility*
*when my own vision faltered.*

# Chapter 1

# Introduction

In today's digital landscape, the unprecedented sharing of facial images on social media platforms has created significant privacy vulnerabilities. The exponential growth in facial image sharing is staggering: Instagram alone witnesses over 95 million images and videos shared daily (Gajic, 2023), with facial content receiving 38% more likes and 32% more comments than non-facial posts (Bakhshi, Shamma, and Gilbert, 2014). This phenomenon extends across all major platforms—Facebook, TikTok, LinkedIn, and countless others—where billions of users regularly share images containing facial biometric data. The ubiquitous adoption of high-resolution smartphone cameras and automated image-sharing features has normalized this behavior, making facial data one of the most prolifically shared forms of biometric information. This widespread sharing of facial data creates substantial privacy risks, as facial recognition (FR) systems can re-identify and link images to enable unauthorized tracking and user profiling across platforms. Current FR systems achieve accuracy rates exceeding 99.8% on standard benchmarks (Deng et al., 2019), thereby transforming casual image sharing into a serious privacy threat.

The scale of this phenomenon is particularly alarming for vulnerable populations. Wright and Bernal (2018) report that parents have posted an average of 1,300 images and videos of their children online by the time they reach age 13, creating extensive databases of biometric data for individuals who cannot consent to its collection. Commercial entities have already exploited this vulnerability, with services such as PimEyes searching through billions of facial images across the Internet, while the Clearview AI database has grown to over 30 billion facial images (Liu, 2023).

The privacy threats associated with facial images arise in two distinct contexts, each requiring a specialized approach to protection. On the one hand, when users share images on social platforms, these images can be collected by third parties to train FR models without consent. This collection and training enable systematic tracking and profiling of individuals across contexts and platforms, with models trained on scraped images later recognizing the same individuals in different settings, potentially enabling surveillance at scale. Current protection methods attempt to disrupt this process by perturbing the published images, but many fail to prevent FR model generalization, that is, the ability of trained models to recognize clean, unseen images of the same individuals in different contexts. Furthermore, existing protection methods often introduce noticeable distortions that compromise visual quality, which is the priority when sharing face images online.

On the other hand, in scenarios where facial images need to be anonymized before sharing them for research or analytical purposes, many conventional methods incur a systematic and significant degradation of image quality. FR models can extract identity

information even from heavily degraded images, potentially re-identifying individuals in supposedly anonymized datasets. This compromises user privacy and violates regulatory frameworks such as the General Data Protection Regulation (GDPR), which treats facial data as sensitive biometric information requiring explicit consent for processing. As a result, relevant facial features that are essential for analytical tasks might be severely distorted, thereby impairing the usefulness of anonymized images in secondary tasks. In this respect, research indicates that both humans and algorithms focus on specific facial features for re-identification, such as the eyes and forehead (Sinha et al., 2006; Royer et al., 2018), while other features that are more relevant for demographic analysis do not need to be perturbed to anonymize images. Existing methods fail to make this crucial distinction, resulting in unnecessary degradation of image utility. The challenge lies in developing anonymization techniques that effectively prevent re-identification while maintaining sufficient utility for legitimate analytical purposes such as age, gender, and race classification.

## 1.1 Objectives

This thesis aims to develop novel methods that address the aforementioned privacy challenges in facial images. Our specific objectives are:

1. **For social image sharing:** Develop a protection method that allows users to modify their images before posting them online, ensuring that unauthorized FR models trained on

these images —while appearing to learn successfully during training— will fail to generalize to clean, unseen images of the same individuals.

2. **For face image anonymization:** Create techniques that anonymize facial images by selectively targeting only identity-revealing features, while preserving other characteristics that are necessary for secondary analytical tasks such as age, gender, and race classification.

3. **For both contexts:** Develop novel methods leveraging adversarial machine learning (ML), explainability techniques, and latent space manipulation to achieve effective privacy protection while preserving utility for secondary analytical tasks. We evaluate approaches across multiple dimensions: visual imperceptibility through perceptual quality metrics, analytical utility preservation for demographic classification tasks, computational efficiency, and transferability across diverse model architectures.

To achieve these objectives, we develop three novel approaches to facial image protection. First, we introduce PRISMA (Perceptually Robust Iterative Similarity Momentum Attack), which is designed specifically for the social media sharing context. PRISMA employs a sophisticated two-stage approach that first identifies robust target features across multiple classes and then uses momentum based optimization to generate visually imperceptible perturbations. The key innovation of PRISMA is that it creates protected images that exhibit a unique behavior: FR models trained

on PRISMA-modified images converge successfully with high accuracy on the training set, yet completely fail to generalize when presented with clean, unseen images of the same individuals.

Second, we present EDI-Anon (Explainability-Driven Incremental Image Anonymization) to address the data anonymization context. Unlike conventional methods that apply uniform distortion to all pixels, EDI-Anon takes a disclosure risk-aware approach that leverages explainability techniques to identify and perturb only the pixels that contribute most to identification. This targeted approach ensures that anonymized images remain useful for analytical tasks while effectively preventing re-identification. EDI-Anon is also designed as a general image anonymization framework that supports a wide variety of image perturbation techniques.

Third, we propose EDIALS (Explainability-Driven Image Anonymization in Latent Space), which further advances the data anonymization objective by operating in the latent feature space rather than directly on pixels. This approach achieves random guess-level anonymization while retaining remarkably high utility for demographic analysis tasks.

## 1.2 Thesis structure

- **Chapter 2:** Reviews existing approaches to facial image privacy protection, including conventional methods, differential privacy-based (DP-based) techniques, and GAN-based anonymization approaches.

- **Chapter 3:** Presents PRISMA in detail, explaining its innovative two-stage process and evaluating its effectiveness at preventing model generalization while maintaining image quality.
- **Chapter 4:** Describes the EDI-Anon framework, detailing how it identifies and targets identity-revealing pixels to achieve superior privacy/utility balance in data anonymization.
- **Chapter 5:** Introduces EDIALS, explaining how it extends anonymization to the latent space for even more effective protection of analytical utility.
- **Chapter 6:** Summarizes the main contributions and outlines potential directions for future research.

# Chapter 2

# Background and state of the art

This chapter provides background and reviews the state of the art in image privacy protection. We begin by examining ML techniques for facial recognition (FR), including adversarial ML and explainability methods. We then explore privacy challenges in ML specifically related to facial data. Finally, we review existing approaches to privacy protection in facial images, categorized into conventional image perturbation techniques, differential privacy-based methods, and ML-based approaches.

## 2.1 Background

### 2.1.1 Machine learning for facial recognition

Modern FR systems predominantly employ deep learning models, particularly convolutional neural networks (CNNs), which have demonstrated superior performance compared to traditional

approaches. These models typically follow a pipeline of face detection, alignment, feature extraction, and identity classification or verification.

The breakthrough in deep learning-based FR came with DeepFace (Taigman et al., 2014), which introduced a nine-layer CNN with over 120 million parameters and achieved 97.35% accuracy on the Labeled Faces in the Wild (LFW) benchmark. Subsequent architectures like FaceNet (Schroff, Kalenichenko, and Philbin, 2015) further improved performance by employing triplet loss to learn discriminative features in a compact Euclidean space.

More recent advancements include ArcFace (Deng et al., 2019), which introduced additive angular margin loss to enhance feature discrimination. ArcFace and similar approaches have pushed accuracy on benchmarks like LFW beyond 99.8%, approaching or even surpassing human-level performance.

The advancement of FR technology has led to the proliferation of commercial FR services that operate at unprecedented scale. Companies like Clearview AI[1] have amassed databases containing over 30 billion facial images scraped from social media and other websites (Liu, 2023). Other commercial services include PimEyes[2], a search engine that allows users to find facial matches across the internet, and FaceVACS[3], a biometric technology provider offering FR capabilities to various clients.

[1] `https://www.clearview.ai`
[2] `https://pimeyes.com`
[3] `https://www.cognitec.com`

### 2.1.2 Adversarial machine learning

Adversarial ML explores the vulnerability of neural networks to carefully crafted perturbations that can cause misclassification while remaining imperceptible to humans. This concept was formalized by Szegedy et al. (2013) and Goodfellow, Shlens, and Szegedy (2014), who demonstrated that adding minimal noise to images could cause deep neural networks to produce incorrect predictions with high confidence.

In the context of image anonymization, adversarial perturbations serve as an effective tool to prevent unauthorized FR models from generalizing. When adversarial examples are introduced into a model's training data, they can corrupt the learned feature representations and cause the model to associate the perturbed images with incorrect identity features (Shan et al., 2020). This means that while the model may perform well on the adversarially modified training images, it fails catastrophically when fed with clean, unmodified images of the same individuals - effectively preventing re-identification (Cherepanova et al., 2021). This mechanism forms the foundation for several privacy protection methods that aim to render user images unusable for unauthorized facial recognition purposes (Evtimov, Sturmfels, and Kohno, 2020).

The Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy, 2014) was one of the first approaches for generating adversarial examples. It computes the gradient of the loss function with respect to the input image and then perturbs the image in the direction that maximizes the loss. Iterative methods like the Basic Iterative Method (BIM) (Kurakin, Goodfellow, and

Bengio, 2018) address some of FGSM's limitations by applying smaller step sizes over multiple iterations. The Momentum Iterative Fast Gradient Sign Method (MI-FGSM) (Dong et al., 2018) further improves this approach by incorporating momentum into the iterative process.

#### Transferability in adversarial examples

Transferability is a crucial property in adversarial ML, referring to how well adversarial examples generated for one model can successfully fool other models with different architectures or training data. In FR systems, transferability is particularly important because privacy protection techniques must remain effective regardless of the specific FR model being used by an adversary.

Several factors influence transferability in FR systems:

- **Model architecture similarity**: Adversarial examples transfer more effectively between models with similar architectures (Liu et al., 2017).
- **Training data overlap**: Models trained on similar datasets show higher transferability of adversarial examples (Tramèr et al., 2017).
- **Feature space alignment**: The degree to which different models learn similar feature representations affects transferability across architectures (Dong et al., 2018).

To enhance transferability, researchers have developed ensemble-based approaches that optimize perturbations across multiple models simultaneously (Liu et al., 2017), as well as momentum-based

methods that stabilize gradient directions during optimization (Dong et al., 2018).

### 2.1.3 Explainability in machine learning

Explainability methods help reveal which parts of an input image contribute most significantly to a model's prediction.

Gradient-based methods are fundamental explainability techniques for image classifiers. These approaches compute the gradient of the output prediction with respect to the input image pixels, identifying which pixels most strongly influence the model's decision. The simplest form, vanilla gradient visualization (Simonyan, Vedaldi, and Zisserman, 2014), computes the partial derivative of the output class score with respect to each input pixel. However, vanilla gradients often produce noisy explanations and can suffer from gradient saturation in deep networks (Adebayo et al., 2018).

Several refinements have been proposed to address limitations in basic gradient methods:

- **SmoothGrad** (Smilkov et al., 2017) reduces noise by averaging gradients computed for multiple slightly perturbed versions of the input image. This averaging approach significantly improves the visual clarity of explanations compared to vanilla gradients, though it inherits many of the base method's structural limitations (Adebayo et al., 2018).

- **Integrated Gradients** (Sundararajan, Taly, and Yan, 2017) addresses gradient saturation issues by integrating gradients along a path from a baseline image to the input. Additionally, it satisfies desirable axiomatic properties like sensitivity and implementation invariance, providing more theoretically grounded explanations. However, the method's effectiveness is sensitive to the choice of baseline, and the resulting attributions can be unstable in high-dimensional inputs (Ghorbani, Abid, and Zou, 2019).

- **Grad-CAM** (Selvaraju et al., 2017) produces class-discriminative localizations using gradients that flow into the final convolutional layer. Unlike pixel-level gradient methods, Grad-CAM provides coarser but more interpretable region-based explanations, making it easier to understand which areas of an image are most important. This higher-level representation improves interpretability, although it trades off spatial precision and may blur fine-grained attributions (Omeiza et al., 2019).

- **Local Interpretable Model-agnostic Explanations (LIME)** (Ribeiro, Singh, and Guestrin, 2016) treats the target model as a complete black box. It works by generating perturbed versions of the input by turning "on" or "off" different image regions (superpixels), observing how these perturbations affect the model's output, and training an interpretable surrogate model on this dataset. The key benefit of LIME is its model-agnostic nature, allowing it to work with any classifier without requiring access to gradients or model internals. However, its explanations are sensitive to the choice of

segmentation and perturbation strategy, leading to potential instability across runs (Laugel et al., 2018).

- **Adversarial example-based explanation** (Haffar et al., 2021) offers a further refinement by generating adversarial examples to explain image misclassification. Their approach uses gradient-based optimization to identify the minimal changes needed to correct misclassifications, thereby highlighting the features most responsible for the incorrect prediction. Unlike previous perturbation-based methods, this technique frames explanation as an optimization problem and reveals actionable insights for model improvement.

- **Adversarial Black box Explainer generating Latent Exemplars (ABELE)** (Guidotti et al., 2020) represents a significant advancement by operating in the latent space rather than directly on pixels. ABELE uses an adversarial autoencoder (AAE) to map images into a lower-dimensional latent space, where it then generates exemplars and counter-exemplars to explain model decisions. This latent space operation provides several key advantages over pixel-based methods: it can capture more complex feature interactions, offers more semantically meaningful explanations, and enables controlled manipulation of high-level image attributes rather than individual pixels.

### 2.1.4 Adversarial autoencoders

AAEs (Makhzani et al., 2015) combine the autoencoder architecture with adversarial training inspired by Generative Adversarial

Networks (GANs). An autoencoder consists of two primary components:

- An **encoder** that compresses the input data into a lower-dimensional representation (the latent space)
- A **decoder** that attempts to reconstruct the original input from this compressed representation

The network is trained to minimize the reconstruction error, typically using a loss function such as mean squared error between the input and its reconstruction (Mienye and Swart, 2025).

AAEs extend traditional autoencoders by adding a discriminator network that attempts to distinguish between samples from a prior distribution (often a normal distribution) and encoded data points. This adversarial component forces the encoder to generate latent representations that follow the prior distribution, making the latent space more structured and organized.

The latent space of a well-trained AAE exhibits several important properties:

- **Feature disentanglement**: In an ideal latent space, different dimensions would correspond to independent semantic features (Higgins et al., 2017). While perfect disentanglement remains challenging, modern autoencoders achieve partial disentanglement where certain dimensions correspond more strongly to specific attributes.
- **Semantic organization**: Research has shown that directions in the latent space often correspond to semantically meaningful transformations (Radford, Metz, and Chintala, 2016).

For example, moving along certain vectors might change attributes like age, expression, or head pose, while other directions might modify other features of the image.

## 2.2 State of the art in image protection

### 2.2.1 Protection against model training on public images

Several approaches have been developed to protect users' (facial) images from being effectively used in unauthorized FR model training.

#### Fawkes

Shan et al. (2020) introduced Fawkes, which uses "cloaking" to protect users' images before they are uploaded to social media. Fawkes perturbs an image such that its feature representation shifts toward a selected target identity while maintaining visual similarity under a Structural Dissimilarity Index (DSSIM)-based perceptual budget.

The method proceeds in two steps. First, it selects a target identity $T$ whose centroid is maximally distant from any of the user's images, that is,

$$T = \arg \max_{k=1..K} \min_{x \in X_U} Dist(\Phi(x), C_k),$$

where $\Phi$ denotes the feature extractor, $C_k$ is the centroid of candidate class $k$, and $X_U$ is the set of user images.

Second, for each user image $x$ and a randomly sampled target image $x_T$, it solves

$$\min_{\delta} Dist(\Phi(x_T), \Phi(x \oplus \delta(x, x_T))) + \lambda \cdot \max(|\delta(x, x_T)| - \rho, 0),$$

where $\oplus$ denotes the cloaked image, $\lambda$ controls the penalty weight, and $\rho$ is the perceptual budget measured using DSSIM.

This optimization pushes the cloaked image's features toward the target's features. When unauthorized FR models train on cloaked images, they learn distorted representations, causing clean images to be misrecognized due to feature space mismatch.

**Lowkey**

Cherepanova et al. (2021) proposed Lowkey, which incorporates the entire FR pipeline into the optimization process and employs an ensemble of multiple models to improve protection against unauthorized FR systems. Unlike Fawkes, Lowkey incorporates face detection, alignment, and resizing operations directly into the adversarial optimization.

For each input image $x$, Lowkey generates a perturbed version $x'$ by solving

$$\max_{x'} \frac{1}{2n} \sum_{i=1}^{n} \left( \frac{\|f_i(A(x)) - f_i(A(x'))\|_2^2 + \|f_i(A(x)) - f_i(A(G(x')))\|_2^2}{\|f_i(A(x))\|_2} \right) - \alpha \cdot \text{LPIPS}(x, x'),$$

where $f_i$ is the $i$-th ensemble model, $A$ represents the preprocessing pipeline (detection, extraction, $112 \times 112$ resizing, alignment), $G$ applies Gaussian smoothing, and $\alpha$ weights the Learned

Perceptual Image Patch Similarity (LPIPS)-based perceptual loss.

The optimization includes both non-smoothed and smoothed terms: the first term measures feature distance for the directly processed perturbed image, while the second term measures distance after Gaussian smoothing. Both terms are normalized by the original feature norm and averaged across the ensemble.

Key innovations of Lowkey are: (1) pipeline integration of differentiable face detection and alignment, (2) ensemble attack against four extractors to enhance attack transferability, (3) smoothing optimization to create perturbations robust to denoising preprocessing, and (4) application of LPIPS loss for better perceptual quality.

The ensemble approach and Gaussian smoothing significantly improve attack transferability by optimizing against multiple diverse architectures with smoothing robustness. Lowkey demonstrated substantial improvements over Fawkes, achieving better protection against commercial FR systems, where Fawkes showed limited effectiveness.

**FoggySight**

Evtimov, Sturmfels, and Kohno (2020) developed FoggySight, which generates privacy-preserving images by optimizing perturbations that maximize the distance between an image's feature representation and its identity's feature subspace in the FR model. It employs three modeling approaches:

- Affine hulls that create an expanded feature space by combining existing image features;

- Class centers that use the average of all feature vectors of an identity;
- Convex hulls that define the minimal region encompassing all feature vectors of an identity.

The protected images can be shared online, and other users can contribute decoy images targeting the same feature space regions. However, FoggySight requires access to the adversary's FR model for computing feature distances, which makes it impractical against unknown model architectures.

### 2.2.2 Image anonymization techniques

Image anonymization methods aim to produce non-reidentifiable images that could be released or shared for secondary use.

#### Conventional approaches

Conventional image anonymization techniques typically follow a one-shot workflow, where an input image undergoes uniform perturbation according to a predefined parameter to produce an anonymized output in a single step, as illustrated in Figure 2.1.

The earliest and most widely used methods for preventing face-based re-identification are pixelation and blurring (Zhao and Stasko, 1998).

**Pixelation** reduces the amount of information by segmenting the image into $n \times n$ blocks and setting all the pixels within each block to the same intensity, resulting from averaging neighboring pixels. As demonstrated by Newton, Sweeney, and Malin (2005),

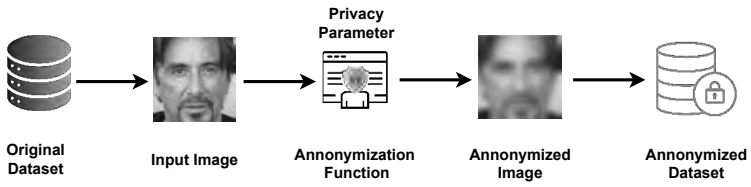



FIGURE 2.1: Workflow of conventional image anonymization techniques

standard pixelation with commonly used parameters offers limited protection against modern FR systems, which can often recover identity information from pixelated faces.

**Gaussian blurring** is used to reduce the sharpness and clarity of an image by introducing Gaussian noise (Neustaedter, Greenberg, and Boyle, 2006). This noise is applied by using a Gaussian kernel of size $k$, which averages the pixel values with those of their neighbors. Since more weight is given to the pixels closer to the center of the kernel, the resulting images have a smoother appearance.

**Motion blur** is a variant of Gaussian blurring that applies directional distortion to simulate movement during image capture (Schuler et al., 2015). Unlike standard Gaussian blur, which applies uniform blurring in all directions, motion blur creates a directional smearing effect that can more effectively disrupt certain facial features while potentially preserving others.

**Facial feature masking** focuses on occluding specific facial regions considered most important for recognition, such as the eyes or mouth. These techniques superimpose black circles or boxes on the images to emulate sunglasses or face masks (Fraser, Craig,

and Parker, 1990; Haig, 1986). Research by Sadr, Jarudi, and Sinha (2003) suggests that other features, such as eyebrows, also play a significant role in FR, indicating that effective masking would need to target multiple facial regions.

### Differential privacy-based methods

$\epsilon$-Differential privacy (DP) (Dwork et al., 2006) offers a formal privacy guarantee that ensures that the presence or absence of any individual's data in a dataset cannot be reliably determined from the output of an algorithm up to a factor exponential to $\epsilon$. In the context of image anonymization, DP has been enforced by adding calibrated noise to image pixels, also following the one-shot workflow illustrated in Figure 2.1.

Fan (2018) proposed DP-Pixelation (DP-Pix), which applies Laplacian noise to pixelated images. The process involves:

1. Pixelating the input image to reduce sensitivity
2. Adding Laplacian noise calibrated based on privacy parameter $\epsilon$ and a measure of sensitivity

DP-Pix is formalized as:

$$\text{DP-Pix}_{(i,j)} = P_{(i,j)} + Lap\left(\frac{GS}{\epsilon}\right),$$

where $\text{DP-Pix}_{(i,j)}$ is the pixel value in the DP-pixed image, $P_{(i,j)}$ is the pixel value in the pixelated image, $GS$ is the global sensitivity, and $\epsilon$ is the privacy parameter.

Fan (2019) extended this method by adding a Gaussian blur post-processing step, resulting in DP-Blurring (DP-Blur). This method aims to produce smoother images by:

1. Applying DP-Pix;
2. Adding a Gaussian blur with kernel size $k$.

DP-Blur is formalized as

$$\text{DP-Blur}_{(i,j)} = \text{DP-Pix}_{(i,j)} + B_{(i,j)}.$$

Although DP approaches should offer theoretical privacy guarantees, their application to facial images faces significant challenges. A meaningful implementation of the DP principles requires noise calibrated to the maximum possible change to any individual's image. This means that most (if not all) pixels might be replaced with maximum/minimum RGB values in a face image, thereby rendering DP-protected images unusable. To retain image utility, practical implementations of DP limit noise to a small subset of $m$ pixels, but this compromises the theoretical guarantees that DP aims to provide.

**GAN-based approaches**

Recent advances in ML have led to Generative Adversarial Network (GAN)-based image anonymization methods that generate privacy-preserving synthetic images (Goodfellow et al., 2020). These approaches represent a paradigm shift from perturbation-based methods to complete face replacement using generative models.

**StyleGAN-2** (Khorzooghi and Nilizadeh, 2023) employs latent vectors as replacements for original images to generate new ones. To enhance image quality, multiple latent vectors are provided to the model, and the anonymized image with the best quality is selected.

**Perceptual Indistinguishability-Net (PI-Net)** (Chen et al., 2021) employs an encoder-decoder architecture to produce perceptually indistinguishable obfuscated images. Similar to StyleGAN-2, PI-Net operates in the latent space; however, instead of replacing original images with latent vectors, it extracts the vector from the original image and uses it to generate a single anonymized version. This ensures better preservation of facial structure and expression compared to StyleGAN-2's random sampling approach.

More sophisticated methods leverage conditional GANs (cGANs) for image anonymization:

- **CIAGAN** (Maximov, Elezi, and Leal-Taixé, 2020) preserves background information by incorporating it into the GAN generator while using an autoencoder to extract a latent vector that captures pose and facial features. It introduces a one-hot identity vector into the generator's bottleneck to ensure that the generated face corresponds to an identity not present in the dataset.

- **DeepPrivacy** (Hukkelås, Mester, and Lindseth, 2019) generates replacement faces that preserve factors like pose, facial expression, and lighting without identity constraints. The method uses keypoint-guided generation to ensure facial structure consistency.

- **DeepPrivacy2** (Hukkelås and Lindseth, 2023) enhances the original DeepPrivacy approach by ensuring that the generated face does not appear in the dataset and extending anonymization to the entire body. The system employs three specialized generators and uses style-based U-Net architecture following StyleGAN2 principles, formulated as an image inpainting task:

  $$\tilde{I} = G(I \odot M, M, z) \odot M + (1 - M) \odot I$$

  where $I$ is the original image, $M$ indicates anonymization regions, $G$ is the generator, and $\odot$ denotes element-wise multiplication.

  The system eliminates keypoint dependency for improved detection recall, employs style-based architecture enabling controllable anonymization, and integrates ensemble detection (dual-shot face detector (Li et al., 2019), continuous surface embeddings (Neverova et al., 2020), mask region-based CNN (He et al., 2017)) for robust face detection.

Despite their advances over conventional techniques, GAN-based methods often produce obviously synthetic faces that may compromise the social value of sharing personal images and may inadvertently alter characteristics important for demographic analysis. Additionally, most approaches focus exclusively on privacy, with inadequate attention to preserving utility for analytical tasks.

## 2.3 Conclusion

Despite their advances over conventional techniques, current privacy protection methods face significant limitations. Conventional anonymization approaches, such as pixelation, blurring, and noise addition, apply uniform distortion across all pixels, even though only specific facial features are critical for identification (Sinha et al., 2006; Sadr, Jarudi, and Sinha, 2003), thereby unnecessarily degrading utility. DP-based methods struggle to provide meaningful privacy guarantees for images without severely compromising utility. Adversarial perturbation methods like Fawkes and Lowkey, while innovative, often introduce noticeable artifacts. Moreover, many methods inadequately consider the perceptual quality of protected images, which is crucial for both social acceptance and practical usability.

## Chapter 3

# Perceptually robust iterative similarity momentum attack (PRISMA)

In today's digital landscape, images shared on social media platforms often contain faces that can be exploited by unauthorized FR systems. As discussed in Chapter 1, this creates serious privacy vulnerabilities, especially when these images are collected to train FR models without the subjects' consent. Existing protection approaches significantly degrade image quality, making them unsuitable for social sharing, or fail to prevent model generalization, leaving users vulnerable to recognition in different contexts. This chapter presents our first contribution that specifically addresses the challenge of protecting facial images for social media sharing while maintaining their visual quality.

## 3.1 Contributions and plan of this chapter

We present PRISMA (Perceptually Robust Iterative Similarity Momentum Attack), a novel method that fundamentally reimagines adversarial image protection. As detailed in Chapter 2, existing approaches like Fawkes and Lowkey demonstrate an unacceptable trade-off between privacy and quality (see Figure 3.1 for visual evidence of their artifacts). PRISMA addresses these limitations through momentum-based optimization with perceptual constraints to generate visually imperceptible perturbations that nonetheless render images unusable for FR training.




FIGURE 3.1: Sample images protected by Fawkes and Lowkey with their strongest privacy protection settings ('high' and LPIPS=0.05, respectively). Skin artifacts are visible in the faces of the second and third rows.

Our empirical results demonstrate that PRISMA significantly outperforms existing approaches in privacy protection while maintaining similar perceptual quality scores. Our experiments also uncover a critical vulnerability in existing approaches, *i.e.*, Fawkes,

and Lowkey: they can be bypassed through reverse attacks that achieve recognition rates up to 69.56%, nearly matching their protected accuracies. Finally, to support the evaluation of privacy protection mechanisms across diverse age groups, we contribute a carefully curated dataset of 416 facial images from 53 underage public figures.

The rest of this chapter is organized as follows. Section 3.2 presents our imperceptible adversarial image generation method. Section 3.3 reports the experimental results. Finally, Section 3.4 summarizes the conclusions.

## 3.2 PRISMA

PRISMA generates adversarially modified versions of images with visually imperceptible perturbations that make them unusable for training FR models. It employs a two-stage process: computation of target features followed by generation of perturbations. This separation enables precise control over optimization: it ensures that robust target features are computed across multiple images of each class before generating optimal perturbations. In this way, stable solutions are attained by preventing simultaneous shifts in target features during perturbation optimization, and consistent privacy protection is achieved across entire classes rather than in images in isolation.

### 3.2.1 Stage 1: Target feature computation

Stage 1, depicted in Figure 3.2 and formalized in Algorithm 1, computes target features that guide the perturbation generation.

The process begins with the original set of images $X$, each one labeled with its corresponding identity (class). We first train a *Face Recognition Model* on $X$, which serves the dual purpose of extracting feature vectors (through $FeatureExtractor$) and classifying images (through $Classifier$) (line 2). For each image $x_i$ within a given class, the model generates classification probabilities through its logits (lines 6–7). To create effective adversarial examples, we identify potential target classes by selecting the most probable incorrect predictions (excluding the true class). The *Number of Top Incorrect Classes (NTIC)* parameter determines how many of these incorrect classes will be selected as *target classes* (line 8) for robust feature computation. See Section 3.3.9 for empirical validation of the optimal NTIC configuration.

For each class, the algorithm processes individual images $x_i$ to identify its target classes through logits and NTIC. For each $x_i$, all images belonging to its identified target classes are collected as *target images* (line 9). The Feature Extractor extracts feature vectors from these target images, which are averaged to compute the same target features ($F_{target}$, line 11) for the original class. Unlike prior work like Fawkes (Shan et al., 2020), which selects a different target image for each input image based on maximum feature distance, our averaging approach ensures consistency across the entire class while capturing the central tendency of each target class's feature distribution. This consistency is crucial: when all images of a class are pushed toward the same averaged target features, unauthorized models learn one consistent but distorted (artificial) mapping.

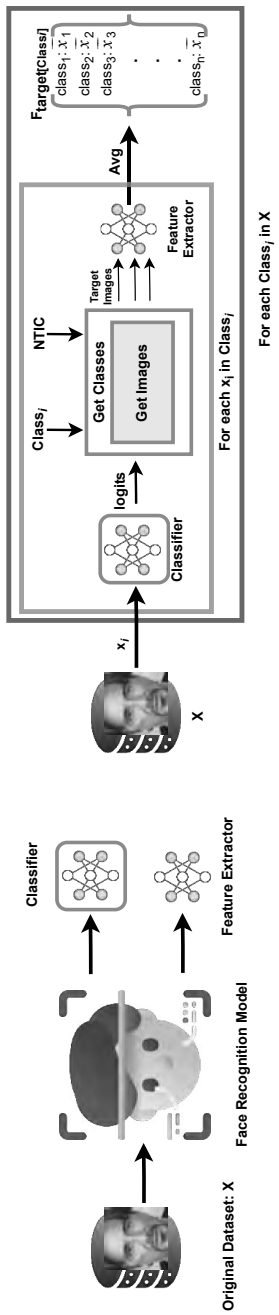



FIGURE 3.2: PRISMA Stage 1: Model Training and Target Feature Computation. After training an FR model on dataset $X$, each image $x_i$ is processed to extract logits and identify target classes using NTIC. Target images from these classes are collected, and their features are extracted and averaged to compute $F_{target}$ for each class.

**Algorithm 1** PRISMA Stage 1: Model Training and Target Feature Computation

**Inputs:** Original Dataset: $X$, Number of Top Incorrect Classes: $NTIC$
Feature Extractor, Classifier $\leftarrow$ Train.Model$(X)$
$F_{target} \leftarrow \{\}$
**for** $class \in$ Unique Classes$(X)$ **do**
    Target Images $\leftarrow$ []
    **for** $x_i \in$ Get Images$(X, class)$ **do**
        logits $\leftarrow Classifier(x_i)$
        Target Classes $\leftarrow$ Get Classes$(logits, NTIC, class)$
        Target Images.extend(Get Images(X, Target Classes))
    **end for**
    $F_{target}[class] \leftarrow$ Average(Feature Extractor(Target Images))
**end for**
**return** $F_{target}$, Feature Extractor

Then, when these unauthorized models later encounter clean, unseen images of the same identity, the significant difference between their natural features and the learned distorted features will prevent successful recognition.

### 3.2.2 Stage 2: Adversarial example generation

Stage 2 focuses on generating the adversarial examples, as depicted in Figure 3.3 and formalized in Algorithm 2. By using the target features $F_{target}$ computed in Stage 3.2.1, the algorithm processes each image in every class to create their adversarial versions.

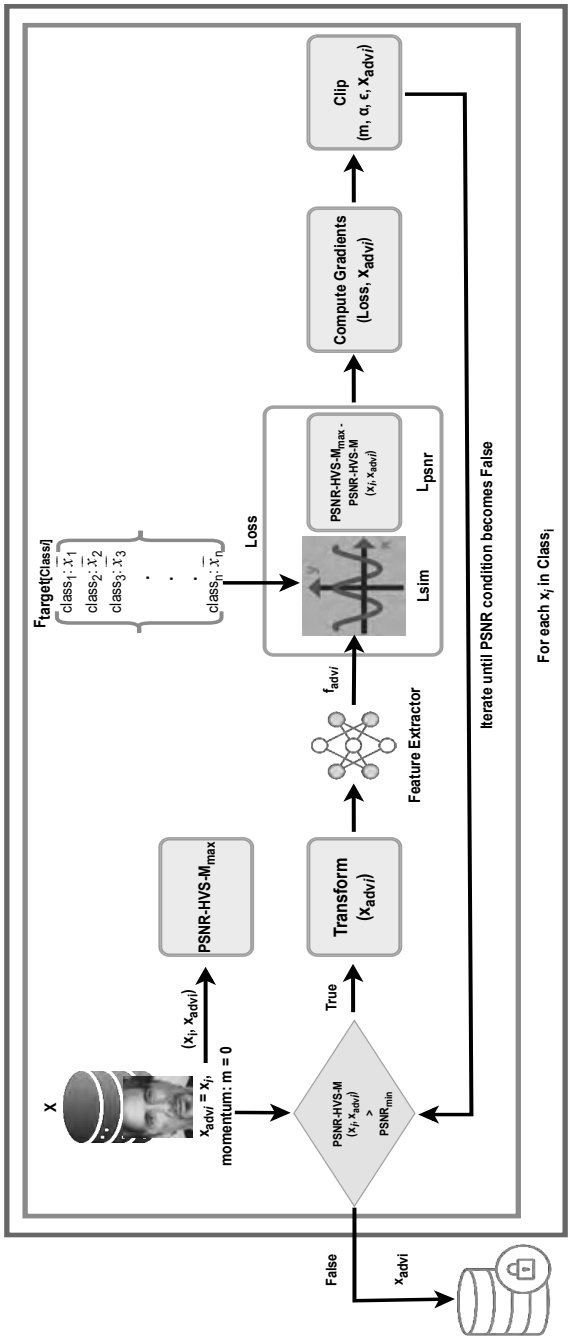



FIGURE 3.3: PRISMA Stage 2: Adversarial Example Generation. For each input image $x_i$, the algorithm iteratively: (1) transforms the image, (2) extracts features, (3) computes dual losses ($L_{sim}$ and $L_{PSNR}$), (4) generates gradients, (5) updates using momentum, and (6) clips the result. The process continues until the peak signal-to-noise ratio (PSNR) falls below the threshold, producing the final adversarial image $x_{adv_i}$.

We formulate the adversarial example generation as an optimization problem guided by a total loss function that balances two objectives: feature space manipulation ($L_{sim}$) and perceptual quality preservation ($L_{psnr}$). Through an iterative process, each step draws the adversarial example closer to the target features in the feature space while meeting the constraint of the perceptual quality threshold $\text{PSNR}_{\text{min}}$.

Although complete convergence to the target features may not be achieved due to this constraint, this approach ensures that the final adversarial example maintains a balance between privacy protection and visual quality. The optimization minimizes the cosine similarity between the features of our modified images and their corresponding target features.

For feature space manipulation, $L_{sim}$ minimizes the cosine similarity between adversarial ($f_{adv_i}$) and target features ($F_{target}$), guiding the optimization to push images away from their true class representations. In prior work, Fawkes employs Projected Gradient Descent (Kurakin, Goodfellow, and Bengio, 2018), which computes gradients with respect to image perturbations and applies SSIM-based clipping at each iteration. However, these direct gradient updates often lead to suboptimal local minima in the feature space, thereby limiting the effectiveness of the generated adversarial examples. We overcome these limitations by adapting the Momentum Iterative Fast Gradient Sign Method (MI-FGSM) (Dong et al., 2018). MI-FGSM introduces momentum into the gradient updates, which stabilizes the optimization by maintaining consistent gradient directions across iterations. Our approach uses this momentum term ($m$) to accumulate the history of loss gradients across iterations using a decay factor $\mu$ that controls how much

previous gradients influence the current update. While MI-FGSM was originally designed for classification attacks, we modified it for feature space manipulation, enabling more stable optimization and better escape from poor local minima that could compromise privacy protection.

For the second objective, *i.e.* perceptual quality preservation, we employ PSNR-HVS-M (Peak Signal-to-Noise Ratio-Human Visual System-Metric) (Ponomarenko et al., 2007; Egiazarian et al., 2006) in $L_{psnr}$, which improves upon previous approaches by incorporating both the human visual system's contrast sensitivity function and complex masking effects. This accounts for frequency-domain interactions where spatial frequencies affect visibility, and spatial masking where local features influence artifact perception in neighboring regions. This captures nuances not considered by simpler metrics like SSIM (Wang et al., 2004), or supervised approaches like LPIPS (Zhang et al., 2018) used in Lowkey. The $L_{psnr}$ loss measures the deviation from PSNR-HVS-$M_{max}$, which is calculated when the adversarial image equals the original image at initialization.

Our optimization strategy is further enhanced to ensure transferability across different FR systems. Since these systems often employ varying preprocessing approaches, we incorporate common augmentation steps during each iteration through random horizontal flips for pose variation robustness, aspect ratio-preserving resize for scale invariance, and color jitter for lighting variation resilience (Zhuchkov, 2021). We then apply Gaussian smoothing to these transformed images, with parameter $\sigma$ controlling the intensity of the blur and kernel size determining the size of the neighborhood considered for blurring. This mitigates

high-frequency artifacts in adversarial perturbations, which can potentially be filtered out by FR models during preprocessing. This approach aligns with findings from Lowkey, as the smoothing reduces noise sharpness and improves attack transferability across models. This comprehensive preprocessing pipeline, combined with our momentum-based optimization, enables PRISMA to generate adversarial perturbations that are both imperceptible and effective across different recognition systems.

In the following, we explain the workflow of Stage 2 described in Algorithm 2. For each class in the dataset (line 5), we initialize the adversarial image $x_{adv_i}$ as a copy of the original image $x_i$ and set $m$ to zero (line 7). We then compute PSNR-HVS-$M_{\text{max}}$ between $x_i$ and $x_{adv_i}$ at initialization for use in the $L_{psnr}$ calculation (line 8). The iterative process is controlled by $\text{PSNR}_{\text{min}}$, a user-provided threshold that determines the acceptable lower bound for perceptual quality. While the PSNR-HVS-M value between $x_i$ and their respective $x_{adv_i}$ remains above $\text{PSNR}_{\text{min}}$ (line 9), the following tasks are performed:

1. Apply the sequence of transformations and Gaussian smoothing to $x_{adv_i}$ (line 10), and extract features using the Feature Extractor (line 11).

2. Compute the total loss (line 14) as

$$L = L_{sim} + L_{psnr}.$$

3. Update the adversarial example. First, update the momentum (line 16) as

$$m = \mu m + g,$$

**Algorithm 2** PRISMA Stage 2: Adversarial Example Generation

1: **Constants:** $\sigma$ (blur intensity), kernel size, $\alpha$ (step size), $\epsilon$ (max perturbation), $\mu$ (decay factor)
2: **Inputs:** Original Dataset: $X$, Feature Extractor, Target Features: $F_{target}$, Minimum PSNR-HVS-M: $\text{PSNR}_{\text{min}}$
3: $X_{adv} \leftarrow \{\}$
4: **for** $class \in$ Unique Classes$(X)$ **do**
5: $\quad X_{adv}[class] \leftarrow []$
6: $\quad$ **for** $x_i \in$ Get Images$(X, class)$ **do**
7: $\quad\quad x_{adv_i} \leftarrow x_i$, $m \leftarrow 0$
8: $\quad\quad \text{PSNR-HVS-M}_{\text{max}} \leftarrow \text{PSNR-HVS-M}(x_i, x_{adv_i})$
9: $\quad\quad$ **while** $\text{PSNR-HVS-M}(x_i, x_{adv_i}) > \text{PSNR}_{\text{min}}$ **do**
10: $\quad\quad\quad x_{trans_i} \leftarrow \text{Transform}(x_{adv_i}, \sigma, \text{kernel size})$
11: $\quad\quad\quad f_{adv_i} \leftarrow \text{Feature Extractor}(x_{trans_i})$
12: $\quad\quad\quad L_{sim} \leftarrow -\text{Cosine Similarity}(f_{adv_i}, F_{target}[class])$
13: $\quad\quad\quad L_{psnr} \leftarrow \text{PSNR-HVS-M}_{\text{max}} - \text{PSNR-HVS-M}(x_i, x_{adv_i})$
14: $\quad\quad\quad Loss \leftarrow L_{sim} + L_{psnr}$
15: $\quad\quad\quad g \leftarrow \text{Compute Gradients}(Loss, x_{adv_i})$
16: $\quad\quad\quad m \leftarrow \mu m + g$
17: $\quad\quad\quad x_{adv_i} \leftarrow \text{Clip}(x_{adv_i} + \alpha \cdot \text{sign}(m), x_i - \epsilon, x_i + \epsilon)$
18: $\quad\quad$ **end while**
19: $\quad\quad X_{adv}[class].\text{append}(x_{adv_i})$
20: $\quad$ **end for**
21: **end for**
22: **return** $X_{adv}$

where $g$ represents the loss gradients (computed in line 15) and $\mu$ controls the influence of the history of previous gradients.

Then, update the image (line 17) as

$$x_{adv_i} = \text{Clip}(x_{adv_i} + \alpha \cdot \text{sign}(m), x_i - \epsilon, x_i + \epsilon),$$

where

- $\alpha$ is the step size that controls the magnitude of updates;
- $\epsilon$ defines the maximum allowed perturbation from the original image, constrained by the PSNR threshold, which ensures perturbations remain visually imperceptible by halting updates once the perceptual quality degradation exceeds $\text{PSNR}_{\text{min}}$;
- The *Clip* operation ensures perturbations remain within $[-\epsilon, \epsilon]$ of the original image; using $sign(m)$ rather than $m$ helps normalize the update magnitude across all dimensions.

Once the PSNR-HVS-M between $x_i$ and $x_{adv_i}$ falls below $\text{PSNR}_{\text{min}}$, the final adversarial image $x_{adv_i}$ is added to $X_{adv}$ (line 19), and the process is repeated for all images across all classes in $X$.

## 3.3 Empirical results

This section comprehensively evaluates PRISMA's effectiveness in protecting user privacy while maintaining image quality. We assess PRISMA against state-of-the-art methods across multiple dimensions: privacy protection effectiveness, resistance to reverse engineering, perceptual quality preservation, and computational efficiency.

All experiments reported in this thesis (PRISMA, EDI-Anon, and EDIALS) were conducted on a consistent hardware setup: an AMD Ryzen 5 3600 CPU (3.6 GHz base and max speed) with 32 GB of RAM (2x16 GB, 2667 MHz) and an NVIDIA GeForce RTX 4070 GPU (12 GB VRAM). Our implementation uses PyTorch 2.2.2 and is available (along with the datasets used) in the Git repository `https://github.com/yk0803/PRISMA`.

### 3.3.1 Evaluation dataset

We employed the Multi-Task Face (MTF) dataset (Haffar, Sánchez, and Domingo-Ferrer, 2025), a manually curated collection of 5,246 face images, representing 240 distinct individuals. Compared with other well-known datasets such as CelebA (Liu et al., 2015), which contains low-resolution images ($218 \times 178$), MTF offers high-resolution ($1024 \times 1024$) images that enable an accurate and fine-grained evaluation of perceptual quality metrics (PSNR-HVS-M depends on image resolution). Moreover, automatically labeled datasets such as the aforementioned CelebA contain significant labeling inaccuracies (Lingenfelter, Davis, and Hand (2022) reports that one-third of the images in CelebA are mislabeled). Finally, because CelebA has been automatically compiled by crawling the Web, it may incur copyright infringement, which raises concerns about its long-term availability (Haffar, Sánchez, and Domingo-Ferrer, 2025). In contrast, the MTF dataset includes manual annotations for a variety of attributes –identity, age, gender, and race–, and follows legally and ethically compliant compilation and curation practices that ensure its long-term availability and reproducibility.

However, noting the absence of images of minors in the MTF dataset, which limits the evaluation of privacy protection mechanisms across all age groups, we contribute a complementary collection of images of underage public figures. Following similar ethical and legal guidelines proposed in Haffar, Sánchez, and Domingo-Ferrer (2025), we developed this extension through four main stages:

- **Data collection:** Using IMDB and other platforms, we identified underage public figures through specific search strings (*e.g.*, "Best child stars", "Top Indian child stars", etc.). The icrawler[1] library and Bing's[2] advanced filtering capabilities enabled the collection of 127,527 commercially licensed images representing 512 identities.
- **Face detection:** Application of the Haar cascades (Viola and Jones, 2001) method through OpenCV[3] yielded 8,641 cropped face images from 130 identities.
- **Expert evaluation:** Two experienced evaluators verified each image for subject identity and apparent age (16 and under), removing uncertain cases to maintain dataset integrity.
- **Final processing:** We employed automated [4] and manual techniques to remove duplicates, excluded identities that were already present in the MTF dataset, and enforced a minimum requirement of 5 images per identity. This resulted in our final collection of 416 high-quality images representing 53 distinct underage public figures.

[1] `https://pypi.org/project/icrawler/`
[2] `https://www.bing.com/`
[3] `http://opencv.org`
[4] `https://www.duplicatephotosfixer.com`

The combined dataset comprises 5,662 images (5,246 MTF plus 416 new images of minors). We allocated 80% images as the *Original Dataset* and reserved 20% as *Probe Images* for independent model evaluation. From the Original Dataset, 10% was used for validation during model fine-tuning.

### 3.3.2 Parameter configuration and FR models

PRISMA's implementation used empirically determined parameters following standard practices in adversarial ML. For target class selection in Stage 1, we set NTIC to 2 based on the empirical evaluation we report in Section 3.3.9. For optimization in Stage 2, we employed step size $\alpha = 2/255$ and maximum perturbation $\epsilon = 8/255$ following (Dong et al., 2018). The Gaussian smoothing parameters were optimized through visual analysis, with kernel size 5 and $\sigma = 1.5$. The $\text{PSNR}_{\text{min}}$ threshold was set to 36 dB (decibels) to ensure our adversarial images maintain PSNR-HVS-M scores above 35 dB, which is considered the threshold for imperceptible image distortion in the literature (Ponomarenko et al., 2015).

Following the evolution of DL in FR from early breakthroughs like DeepFace to recent advances with ArcFace achieving near-human performance, we employed ConvNeXt as our primary FR model. ConvNeXt represents the current state-of-the-art in CNN architectures, incorporating modern design principles from both CNNs and transformers to achieve superior performance on FR tasks (Liu et al., 2022). For consistency across our privacy protection framework, ConvNeXt is used as the primary FR model in

all three works presented in this thesis (PRISMA, EDI-Anon, and EDIALS).

To rigorously assess PRISMA's transferability across the diverse landscape of FR systems described in the background, we selected four additional models representing different architectural paradigms and evolutionary stages of DL for FR. ResNet101 (Zhang, 2022) uses residual connections to enable very deep architectures, while VGG16 (Simonyan and Zisserman, 2014) represents traditional deep CNNs with sequential layer stacking; EfficientNet (Tan and Le, 2019) employs compound scaling for balanced network depth and width, and MobileNet-V3-Large (Howard, 2017) is designed for mobile devices using depth-wise separable convolutions.

This architectural diversity enables comprehensive evaluation across varying network depths (from 16 layers in VGG to 101 in ResNet), different computational building blocks (residual, traditional convolution, depth-wise separable), and contrasting design priorities that span from accuracy-optimized (ConvNeXt, ResNet) to efficiency-focused (MobileNet) approaches. Additionally, we evaluated PRISMA against Microsoft Azure's Face API (Application Programming Interface), a commercial black-box system representing the industrial-scale FR services discussed in the background, whose proprietary architecture, training data, and procedures remain unknown but are likely to incorporate state-of-the-art techniques similar to those used in ArcFace and other modern FR systems.

### 3.3.3 Experimental setup

In this section, we detail the configuration and implementation specifics of our evaluation setup, including both the model training and the baseline methods used for comparison.

#### Model training configuration

All FR models used in our evaluation were initialized with pre-trained weights and then fine-tuned on our dataset. We maintained consistent training parameters across all architectures. The models were trained using batches of 16 images with cross-entropy loss (Zhang and Sabuncu, 2018) as the optimization criterion. For optimization, we employed Adam (Kingma and Ba, 2015) with a learning rate of $1 \times 10^{-4}$. To prevent overfitting and ensure model convergence, we implemented early stopping with a patience of 3 epochs, where training was terminated if no improvement in validation loss was observed for three consecutive epochs. This standardized training configuration was applied across all models: ConvNeXt, ResNet101, VGG16, EfficientNet, and MobileNet-v3-Large.

#### Baseline methods

We compared PRISMA against two state-of-the-art methods for protecting facial images: Fawkes and Lowkey. For reproducibility, we next detail their specific implementations and configurations. We used the official implementations of both Fawkes[5] and

[5] `https://pypi.org/project/fawkes/`

Lowkey[6]. Lowkey's implementation includes an ensemble of ArcFace and CosFace FR systems, each implemented with ResNet-50, ResNet-152, IR-50, and IR-152 backbones trained on MS-Celeb-1M dataset (over five million images from 85,000+ identities) (Guo et al., 2016).

To enable a fair comparison, PRISMA, Fawkes, and Lowkey were set to produce similarly imperceptible images. For this, we aimed at achieving a PSNR-HVS-M above 35 dB for all three methods. For Fawkes, we achieved a PSNR-HVS-M around 35 dB with its 'low' privacy setting (see details in Section 3.3.7). However, for Lowkey we were unable to exceed 33 dB despite testing various parameter configurations including increased LPIPS restrictions. This was probably caused by the significant downscaling that Lowkey applies to the images, which significantly reduces the image details. Following the authors' recommendations for optimal image quality, we configured Lowkey with increased LPIPS restrictions (LPIPS=0.2) to achieve its best possible perceptual quality. For Fawkes, we employed their pip-installed library with the provided feature extractors (DenseNet-121 (Huang et al., 2017) and InceptionResNet V2 (Szegedy et al., 2017)).

### 3.3.4 Privacy protection assessment

We evaluated PRISMA's privacy protection in two complementary scenarios: protection of all classes to assess general effectiveness, and protection of underage identities to validate performance on vulnerable populations. Both scenarios present their

[6] Available under supplementary material at `https://openreview.net/forum?id=hJmtwocEqzc`

unique challenges and requirements for privacy protection.

**Protecting all classes**

We first evaluated PRISMA's effectiveness when protecting all identities in the dataset. Table 3.1 shows the performance of different FR models trained on PRISMA-protected images. Our models achieve accuracies of 44.94%-81.04% across models and 86.13% for Azure, which is in line with results reported on the MTF dataset of 240 classes by the original paper (Haffar, Sánchez, and Domingo-Ferrer, 2025) (ConvNeXt: 79.87%, ResNet: 58.39%, VGG16: 34.66%, MobileNet: 57.35%), albeit we work with an expanded set of 293 identities, each having approximately 14 images on average. This ability to maintain convergence on protected training data indicates that models can successfully learn from the perturbed images during training, which is in agreement with previous findings on the convergence of neural networks on modified training data (Mądry et al., 2017).

However, the performance drops significantly when attempting to recognize clean probe images of the same identities (22.30%-34.64% across models, 55.02% for Azure). This sharp decline is particularly notable for ConvNeXt which, despite achieving 81.04% accuracy on protected images, only manages 26.28% on clean images. This behavior demonstrates that PRISMA successfully prevents the models from learning generalizable features.

In contrast, examining the results for Lowkey and Fawkes reveals that they provide virtually no privacy protection. For Lowkey, models achieve even higher accuracies on clean probe images

TABLE 3.1: Privacy protection comparison against different FR models trained on images protected with PRISMA, Lowkey, and Fawkes. Accuracy is reported on both protected training images (Protected) and clean probe images (Clean). Lower accuracy on clean images indicates better privacy protection.

| Approach | Model | Accuracy (%) | |
|---|---|---|---|
| | | Protected | Clean |
| **PRISMA (Ours)** | ConvNext Base | 81.04 | 26.28 |
| | ResNet101 | 53.51 | 31.34 |
| | VGG16 | 44.94 | 34.64 |
| | EfficientNet | 60.00 | 22.30 |
| | MobileNet-v3-Large | 55.58 | 24.83 |
| | Azure | 86.13 | 55.02 |
| **Lowkey** | ConvNext Base | 68.83 | 74.64 |
| | ResNet101 | 42.34 | 47.13 |
| | VGG16 | 34.81 | 41.30 |
| | EfficientNet | 49.61 | 56.17 |
| | MobileNet-v3-Large | 47.79 | 51.49 |
| | Azure | 78.43 | 82.11 |
| **Fawkes** | ConvNext Base | 60.52 | 68.05 |
| | ResNet101 | 47.79 | 49.04 |
| | VGG16 | 39.48 | 41.84 |
| | EfficientNet | 45.97 | 50.27 |
| | MobileNet-v3-Large | 37.92 | 40.00 |
| | Azure | 70.00 | 79.30 |

(41.30%-74.64%, 82.11% for Azure) than on protected training images (34.81%-68.83%, 78.43% for Azure). Similarly, Fawkes shows higher accuracies (40.00%-68.05%, 79.30% for Azure) than training accuracies (37.92%-60.52%, 70.00% for Azure). This indicates that these methods' perturbations may be acting as regularizers

during training (Zhang et al., 2019; Ilyas et al., 2019), thereby improving rather than hampering model generalization. As shown in Table 3.1, this allows models to maintain high recognition rates on clean probe images, thereby providing no privacy protection at all.

PRISMA also shows strong attack transferability, as it reduces accuracies both for lightweight models like MobileNet, sophisticated ones like ConvNeXt, and commercial systems like Azure. The model performances on protected images align with or even improve upon the original performances reported in MTF (Haffar, Sánchez, and Domingo-Ferrer, 2025), despite our dataset being an expanded version of MTF. In contrast, on probe images the recognition accuracies significantly decrease for all models. This uniform degradation across architecturally diverse models indicates that PRISMA targets fundamental features rather than model-specific vulnerabilities.

**Protecting underage identities**

Protecting the privacy of minors is particularly critical given their inability to consent to or control their digital presence. With research indicating that 89% of children have online images by age 2 (Nominet, 2016), this extensive early-life exposure creates significant privacy risks through FR systems that can enable long-term tracking and profiling (Keeley and Little, 2017). This vulnerability is explicitly recognized by privacy regulations like the GDPR in its Article 8 and COPPA (Kosta, 2020; Federal Trade Commission, Commission, et al., 1998), which establish special protections

for minors' data. Additionally, minors represent a uniquely challenging case for FR systems due to their rapidly changing facial appearance over short time periods, making them an especially important group to be separately evaluated.

To evaluate the effectiveness of protection for this vulnerable group, we created specialized training sets where only the images of underage identities (53 individuals, with approximately 6 images per identity) were protected using each method, whereas adult images remained unmodified. The accuracies reported in Table 3.2 focus specifically on the recognition rates for underage identities, both during training (on protected images) and testing (on clean probe images), while excluding the performance for adult classes.

All methods achieve better protection compared to their performance on the full dataset, primarily because the perturbations are now targeted at a smaller subset of classes (53 minors vs. 240 total identities) and because minors' rapidly evolving facial features make them more challenging for FR models to learn consistently. Anyway, the focused protection allows for more distinct feature space manipulation for the protected identities.

Although FR models achieve moderate success in recognizing PRISMA-protected underage images during training (18.87% – 60.38% for our models, and 72.85% for Azure), they clearly fail to recognize clean probe images of these identities. Testing accuracies drop dramatically to 0.00% – 9.43% for our models and 12.68% for Azure. PRISMA's attack transferability is evident here, as it achieves uniform protection across different architectures. In contrast, Lowkey and Fawkes offer significantly weaker protection: recognition rates on clean probe images range within 15.09%–37.74%

TABLE 3.2: Privacy protection comparison when protecting only minors.

| Approach | Model | Accuracy (%) | |
|---|---|---|---|
| | | Protected | Clean |
| **PRISMA (Ours)** | ConvNext Base | 60.38 | 0.00 |
| | ResNet101 | 35.85 | 5.66 |
| | VGG16 | 18.87 | 0.00 |
| | EfficientNet | 37.74 | 1.89 |
| | MobileNet-v3-Large | 39.62 | 9.43 |
| | Azure | 72.85 | 12.68 |
| **Lowkey** | ConvNext Base | 52.83 | 37.74 |
| | ResNet101 | 33.96 | 28.30 |
| | VGG16 | 24.53 | 26.42 |
| | EfficientNet | 41.51 | 28.30 |
| | MobileNet-v3-Large | 26.42 | 15.09 |
| | Azure | 64.12 | 16.82 |
| **Fawkes** | ConvNext Base | 52.83 | 18.87 |
| | ResNet101 | 28.30 | 26.42 |
| | VGG16 | 24.53 | 7.55 |
| | EfficientNet | 33.96 | 20.75 |
| | MobileNet-v3-Large | 32.08 | 24.53 |
| | Azure | 67.73 | 43.76 |

for Lowkey and within 7.55%–26.42% for Fawkes across our models. Interestingly, Lowkey achieved better protection than Fawkes on Azure, with a recognition rate of 16.82% compared to Fawkes' 43.76%.

### 3.3.5 Impact of unprotected training images

An important consideration for privacy protection methods is their resilience when an adversary has access to some unprotected images of the target individuals. To evaluate this scenario, we focused specifically on protecting the images of minors, as this represents a critical use case where privacy protection is especially important. We used ConvNeXt Base for these experiments because it demonstrated the most consistent performance across all three protection methods.

For this experiment, we started with all minors' images protected using each method, representing the ideal baseline for privacy protection. We then systematically replaced varying proportions of protected images with their unprotected versions to simulate data leakage scenarios. Table 3.3 shows how the protection degrades with increasing proportions of unprotected training data. PRISMA's protection degrades as unprotected images are introduced, rising to 12.53% accuracy with 5% unprotected images and nearly doubling to 24.36% with 15% unprotected images. Lowkey starts from a higher baseline accuracy of 37.74% and shows a more gradual degradation, reaching 39.33% at 15% unprotected images. Fawkes begins with better protection than Lowkey (18.87% accuracy) but degrades more quickly, reaching 32.08% accuracy at 15% unprotected images. As the proportion of unprotected images increases to 80%, all three methods converge to accuracy rates around 43%, indicating a severe compromise of their attained privacy.

These results highlight a critical insight: for privacy protection

TABLE 3.3: Impact of unprotected training data on the protection afforded by PRISMA, Lowkey, and Fawkes using ConvNeXt Base.

| Unprotected Images | Accuracy on Clean Probes (%) | | |
|---|---|---|---|
| | PRISMA | Lowkey | Fawkes |
| 0%* | 0.00 | 37.74 | 18.87 |
| 5% | 12.53 | 38.04 | 30.19 |
| 15% | 24.36 | 39.33 | 32.08 |
| 20% | 30.19 | 39.62 | 36.08 |
| 50% | 38.40 | 41.51 | 39.62 |
| 80% | 42.82 | 43.53 | 43.40 |

*Baseline results when 100% of training images are protected by each method. These baselines are used as reference points to evaluate how protection degrades when unprotected images are introduced into the training set.

methods to be truly effective, they should be applied comprehensively to *all* of an individual's images that may appear online.

### 3.3.6 Robustness against reverse attacks

In addition to the weaker privacy protection attained by Fawkes and Lowkey, we now show that, through a reverse attack, their protection can be circumvented by an adversary who applies the same protection technique to the probe images. The intuition behind this attack is straightforward: since these methods rely on simple feature space manipulation through adversarial perturbations, applying the same perturbation technique to probe images pushes them toward the same feature space as the protected (*i.e.*, training) images, thereby increasing the FR accuracy.

For these experiments, we focus on the minors' dataset comprising 53 identities (Section 3.3.4), as it was the scenario in which Lowkey and Fawkes offered some protection. Even if reverse attacks require knowledge of the protection parameters, these are straightforward to obtain: Fawkes operates with just three predefined protection levels ('low', 'mid', 'high'), while Lowkey's protection is controlled by its LPIPS value. Through simple experimentation with these limited parameter options, an adversary can readily determine their values for the reverse attack. Furthermore, both methods rely on general models pre-trained on public datasets, rather than on the private images to be protected (as PRISMA does). Since these general models must be shared to apply these methods, they can be straightforwardly leveraged by adversaries to facilitate reverse attacks.

Table 3.4 highlights this vulnerability. For Lowkey, although accuracies on clean probe images are reduced compared to unprotected images (*e.g.*, 37.74% vs 52.83% for ConvNext Base), applying the reverse attack allows accuracies to closely match or approach the protected accuracies across all models. For instance, with ConvNext Base, the reverse attack achieves 50.94% accuracy, nearly matching the protected accuracy of 52.83%. The effect is particularly evident with Azure, where the reverse attack achieves 63.88% accuracy compared to the protected accuracy of 64.12%.

Similarly, for Fawkes, while clean accuracies show significant reductions (*e.g.*, 18.87% vs 52.83% for ConvNext Base), the reverse attack substantially recovers the recognition capability. In some cases, like Azure, the reverse attack exceeds the protected accuracy (69.56% vs 67.73%). Even for models where the reversal does not fully match the protected accuracy, it still significantly

improves over the clean accuracy: for example, ConvNext Base improves from 18.87% to 41.51%.

TABLE 3.4: Vulnerability to reverse attacks in the minors' classes scenario (53 identities).

| Approach | Model | Accuracy (%) | | |
|---|---|---|---|---|
| | | Protected | Clean | Reverse |
| **Lowkey** | ConvNext Base | 52.83 | 37.74 | 50.94 |
| | ResNet101 | 33.96 | 28.30 | 35.85 |
| | VGG16 | 24.53 | 26.42 | 26.42 |
| | EfficientNet | 41.51 | 28.30 | 32.08 |
| | MobileNet-v3-Large | 26.42 | 15.09 | 28.30 |
| | Azure | 64.12 | 16.82 | 63.88 |
| **Fawkes** | ConvNext Base | 52.83 | 18.87 | 41.51 |
| | ResNet101 | 28.30 | 26.42 | 35.85 |
| | VGG16 | 24.53 | 7.55 | 20.75 |
| | EfficientNet | 33.96 | 20.75 | 32.09 |
| | MobileNet-v3-Large | 32.08 | 24.53 | 32.08 |
| | Azure | 67.73 | 43.76 | 69.56 |

Reverse attacks on PRISMA would aim to modify a probe image so that its feature representation aligns with that of a PRISMA-protected image. However, the architecture of PRISMA is data-driven, which substantially differs from the simple feature space manipulations employed by existing methods. Even if an adversary were to recreate similar model architectures, PRISMA's dependence on private local data makes it challenging to reverse its protection. On the other hand, if an adversary had access to the private data needed to replicate PRISMA and its protections, there would be nothing left to protect. In fact, can prove the following:

**Proposition 1.** *PRISMA resists reverse attacks if the attacker has no access to the training data.*

*Proof.* To mount a reverse attack for a probe image $p$, the attacker needs to recreate the feature extractor used to protect the training data, and use it to protect $p$. However, this feature extractor can only be obtained if the training data are available to the attacker and, by assumption, this is not the case. □

### 3.3.7 Perceptual quality

We next evaluate the perceptual quality using PSNR-HVS-M, which improves upon previous metrics, as it considers the human visual system's contrast sensitivity function and provides a comprehensive assessment by accounting for complex masking effects, where local features influence artifact perception in neighboring regions (Ponomarenko et al., 2015).

As discussed in Section 3.3.3, the parameters of all three methods were carefully tuned, aiming to achieve imperceptible perturbations, which typically require PSNR-HVS-M values of 35 dB or above. Whereas exactly matching PSNR-HVS-M scores across methods would be impractical due to their different optimization objectives, we configured PRISMA to target a PSNR-HVS-M of 36 dB to minimize visual artifacts. Even with this constraint limiting the magnitude of the perturbations allowed, PRISMA managed to achieve better privacy protection than the baselines.

The empirical results confirm this: PRISMA achieves a PSNR-HVS-M score of 35.82 dB, followed by Fawkes at 35.08 dB, while Lowkey reaches only 33.03 dB. It is worth noting that despite using the most conservative LPIPS parameter value, Lowkey falls short of the 35 dB imperceptibility threshold. This lower score can

be explained by the downscaling applied by Lowkey, which produces 112x112 images, whereas Fawkes and PRISMA preserve the original 1024x1024 resolution.

Figure 3.4 depicts a sample of images protected under the different methods. A closer inspection reveals characteristic differences among the methods: PRISMA (Row 2) adds subtle localized perturbations, whereas Lowkey (Row 3) blurs the images due to the aggressive downscaling and Fawkes (Row 4) perturbs some relevant facial features (*e.g.*, around the lips, the eyes or the face contour). Nevertheless, the overall perturbation of all methods is quite negligible (assuming that Lowkey downscaling is acceptable), as it might be mistaken by image compression artifacts usually found in images posted online.

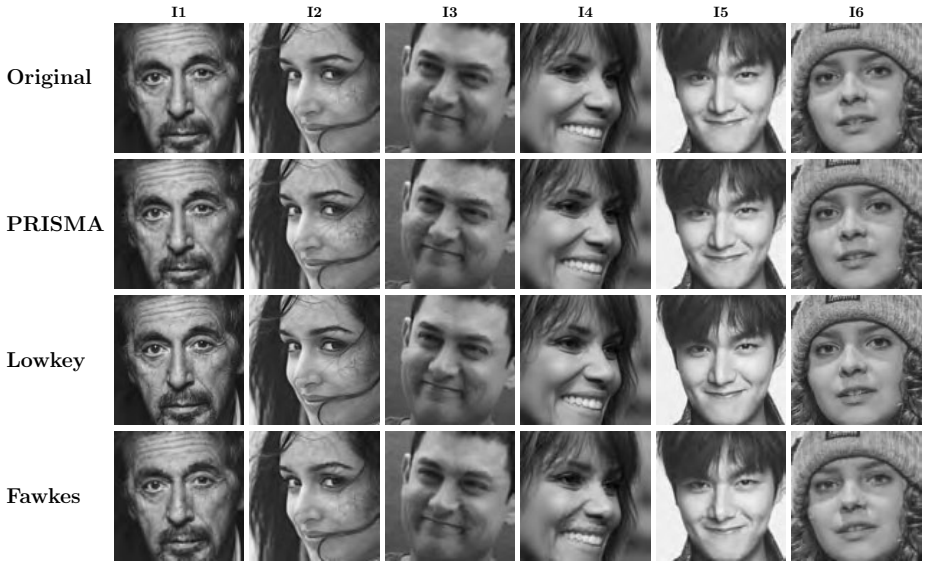



FIGURE 3.4: Visual comparison of privacy protection methods: original images (top row) and their protected versions using PRISMA, Fawkes, and Lowkey (subsequent rows)

### 3.3.8 Runtime analysis

For a comprehensive runtime comparison, PRISMA was run with a batch size of 16, while Lowkey and Fawkes used batch sizes of 8 and 10 respectively. The smaller batch sizes for Lowkey and Fawkes were needed because of their higher GPU memory requirements. To measure run-time, we counted the complete processing pipeline for each method, including model training, feature extraction, and adversarial image generation.

PRISMA achieved the best computational efficiency, requiring an average of 2.84 seconds per image for its complete pipeline, including model training, target feature extraction, and momentum-based iterative adversarial image generation. In comparison, Lowkey employed 3.24 seconds per image, though this excludes the unreported training time for its model ensemble. Fawkes took longest at 3.78 seconds per image, also excluding its unreported feature extractors' training time.

Lowkey's increased runtime is likely caused by two factors: its ensemble-based approach, which processes each image through four models, and its LPIPS-based perceptual similarity evaluation, which uses a fifth model. Despite operating on smaller 112×112 images, these computational requirements result in a longer processing time than that of PRISMA. Fawkes takes even longer, as it also uses two large-scale models as feature extractors. Also, unlike Lowkey, Fawkes operates on the original 1024×1024 images, which significantly increases both feature extraction and perturbation calculation times. Its runtime is further increased by its multi-step iterative optimization for each image.

These figures show that PRISMA's two-stage architecture offers better computational efficiency than existing methods, despite its comprehensive approach.

### 3.3.9 Impact of PRISMA parameters

In this section, we analyze the impact of $\alpha$, $\epsilon$, and NTIC on PRISMA when protecting the whole dataset.

Tables 3.5 and 3.6 show the results using ConvNeXt Base as a test model.

TABLE 3.5: Impact of varying PRISMA parameters. Results shown for ConvNeXt.

| Configuration | $\alpha$ | $\epsilon$ | Steps | Accuracy (%) | |
|---|---|---|---|---|---|
| | | | | Protected | Clean |
| **Recommended** | **2/255** | **8/255** | **172** | **81.04** | **26.28** |
| Small Step | 1/255 | 8/255 | 246 | 83.12 | 26.15 |
| Large Step | 4/255 | 8/255 | 121 | 83.78 | 34.56 |
| Small Bound | 2/255 | 4/255 | 294 | 83.64 | 43.45 |
| Large Bound | 2/255 | 12/255 | 161 | 80.52 | 41.61 |

Step size $\alpha$ significantly affects both computational efficiency and privacy protection. A reduction to $\alpha = 1/255$ increases iterations by 43% (246 steps) while marginally improving protection (26.15% clean accuracy). Increasing $\alpha$ to 4/255 reduces the computational cost by 30% (121 steps) but compromises privacy protection (34.56% clean accuracy). These experiments indicate that $\alpha = 2/255$ provides the best trade-off between computational cost and privacy protection.

Maximum perturbation $\epsilon$ exhibits distinct behavior patterns. Whereas larger $\epsilon$ values theoretically enable stronger perturbations, our fixed PSNR threshold of 36 dB effectively limits their impact. Halving $\epsilon$ to $4/255$ increases iterations by 71% (294 steps) while significantly reducing protection (43.45% clean accuracy). Increasing $\epsilon$ to $12/255$ offers minimal computational benefit (6% fewer steps) while also compromising protection (41.61% clean accuracy), as the PSNR constraint prevents making full use of the larger perturbation bound. Our experiments show that $\epsilon = 8/255$ provides optimal privacy protection, achieving the lowest clean accuracy (26.28%) while maintaining good protected accuracy (81.04%).

Our analysis of NTIC (Table 3.6) shows that using varying target images per target class, NTIC=2 achieves the strongest privacy protection (26.28% clean accuracy), significantly below NTIC=1 (34.48%) and NTIC=5 (33.79%). The model converges on the protected training data at 81.04% but fails to generalize on clean images at 26.28%. Larger NTIC values disperse target feature mappings – reducing perturbation effectiveness–, while smaller values limit feature diversity and adversarial robustness.

In contrast, experiments with a fixed number (5) of target images per class yield less effective protection: although NTIC=2 still performs best, its clean accuracy rises to 29.41%, compared to 32.33% and 31.95% of NTIC=1 and NTIC=5, respectively. This highlights the advantage of leveraging all available target images during feature computation to achieve stronger privacy protection.

These findings validate the recommended configuration we employed in the previous experiments ($\alpha = 2/255$, $\epsilon = 8/255$,

TABLE 3.6: Impact of NTIC and target class image count on PRISMA's performance using ConvNeXt.

| Configuration | NTIC | Accuracy (%) | |
|---|---|---|---|
| | | Protected | Clean |
| Varying Target Images per Target Class | 1 | 83.90 | 34.48 |
| | 2 | **81.04** | **26.28** |
| | 5 | 83.38 | 33.79 |
| Fixed (5) Target Images per Target Class | 1 | 82.12 | 32.33 |
| | 2 | **84.68** | **29.41** |
| | 5 | 83.12 | 31.95 |

NTIC=2 with varying target images), as it provides the best balance between privacy protection, computational efficiency, and protected accuracy.

## 3.4 Conclusion

This chapter introduced PRISMA, a novel method for protecting face images against unauthorized FR systems. Through comprehensive evaluations, we have shown PRISMA's benefits in three key areas. First, our two-stage architecture achieves substantially better privacy protection than existing methods while maintaining high perceptual quality. Specifically, while FR models trained on PRISMA-modified images appear to learn successfully (that is, they converge with a high accuracy on the training set), they basically fail to generalize (that is, they are unable to correctly classify clean, unseen images of the same individuals). This inability to extract meaningful features for ML training naturally provides protection against unauthorized use, thereby providing intrinsic

privacy protection. Second, a critical contribution of our work is the identification of a fundamental vulnerability in existing protection methods: they can be compromised by simple reverse attacks. In contrast, PRISMA is inherently resistant to such attacks. Third, our method exhibits strong transferability across diverse model architectures and commercial systems.

# Chapter 4

# Explainability-driven incremental image anonymization (EDI-Anon)

The challenge of facial privacy protection has grown increasingly critical in today's digital landscape, where modern FR systems can easily identify individuals. Balancing privacy protection with utility preservation represents a fundamental trade-off in image anonymization, as conventional methods often unnecessarily degrade analytical utility. As highlighted in Chapter 1, there is a pressing need for techniques that selectively target only identity-revealing features. This chapter introduces Explainability-Driven Incremental Image Anonymization (EDI-Anon), which leverages explainability techniques to detect and perturb only the most privacy-sensitive pixels in facial images, achieving superior privacy protection while maintaining significantly higher utility than conventional methods.

## 4.1 Contributions and plan of this chapter

This chapter presents EDI-Anon, a novel image anonymization framework that provides a better balance between privacy protection and utility preservation than existing methods. Rather than using a fixed masking criterion, EDI-Anon is disclosure risk-aware and it leverages explainability techniques to detect and target identity-revealing pixels within the images. As a result, and in contrast to conventional methods that uniformly perturb image pixels, our framework focuses on perturbing only those pixels that contribute the most to disclosure. Moreover, pixel perturbation is enforced incrementally, and it is driven by the observed residual risk in order to add the minimal distortion required to attain the desired level of privacy protection. Thus, our anonymization framework might be dubbed "utility-first", in that it strives to preserve as much utility as possible.

The empirical results we report show that our approach is able to encompass a wide variety of image perturbation techniques, and provides a significantly better privacy/utility balance than the state of the art, even when using the simplest perturbation techniques.

The remainder of this chapter is organized as follows. Section 4.2 presents our image anonymization framework. Section 4.3 reports and discusses experimental results. Finally, Section 4.4 gathers the conclusions of the chapter.

## 4.2 EDI-Anon

The conventional anonymization techniques discussed in Section 2, such as pixelation and blurring, operate in a one-shot and untargeted manner: they take an input image and apply a uniform perturbation across all pixels according to a predefined privacy parameter to produce an anonymized image in a single step. Facial masking also operates in a one-shot manner, but confines perturbation to the eyes or the mouth of the individual. Still, the perturbation applied to those facial features –which need to be defined beforehand– is uniform.

We argue that these conventional approaches incur an unnecessary loss of utility due to a fundamental intuition: not every pixel contributes the same to re-identification and, therefore, would not require the same level of perturbation (if any). This insight led us to develop a targeted anonymization framework that aims to perturb or obfuscate only the identity-revealing pixels, and up to a degree proportional to their contribution to re-identification. The goal of this disclosure risk-aware solution is to reduce the re-identification risk (RIR) of images while preserving as much analytical utility as possible. It is noteworthy that our framework neither proposes nor is tied to a specific image perturbation technique, but can enforce existing techniques –including pixelation, blurring, noise addition, and masking–, under this new image anonymization paradigm. Disclosure awareness is enabled in our approach by using an ML FR model trained on the images to be anonymized.

Then, an explainable AI method is applied to the FR model on the images to be protected in order to detect the image pixels

that contributed the most to successful re-identifications. Only the most disclosive pixels will be subjected to protection by applying existing perturbation techniques, such as those detailed in Section 2. Moreover, pixel perturbation is applied incrementally, and it is guided by a continuous re-evaluation of the RIR through the FR model in order to introduce the minimum perturbation needed and, therefore, maximize utility preservation. The workflow of our framework, termed *Explainability-Driven Incremental Anonymization* (EDI-Anon), is shown in Fig. 4.1.

### 4.2.1 EDI-Anon modules

EDI-Anon has a modular design as most of its components are replaceable, e.g., to adapt to specific needs or to incorporate new developments that may arise. In the following, we describe the modules of EDI-Anon in detail.

#### FR model

The *face recognition model* is trained on the *original dataset* of images that require protection, and provides the risk awareness that drives EDI-Anon. Specifically, once trained, the FR model is applied to each image from the same dataset it was trained on in order to generate probabilities for each class, indicating the likelihood of each *input image* belonging to each class. The class with the highest probability is considered the predicted label (i.e., identity) for the input image.

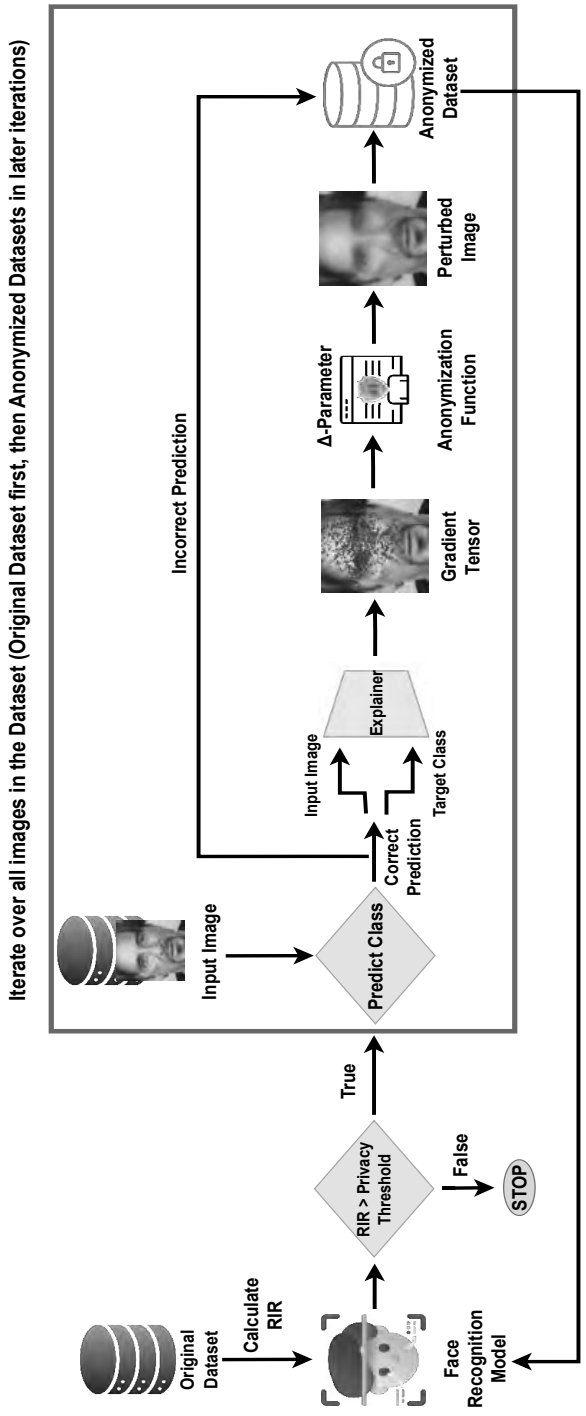



FIGURE 4.1: Workflow of EDI-Anon, illustrating the iterative process of detecting identity-revealing pixels through explainability techniques, applying targeted perturbations, and continuously re-evaluating RIR until the privacy threshold is reached.

On the one hand, the FR model is used to quantify the RIR of the (preliminary) anonymized dataset that results from each iteration. RIR is calculated as the ratio of correctly identified images to the total number of images, bounded below by the random guessing probability:

$$RIR = \max\left\{\frac{Correct_Predictions}{Total_Predictions}, \frac{1}{C}\right\} \tag{4.1}$$

where $C$ is the number of identities in the dataset. Since an adversary can always randomly guess among $C$ identities with probability $\frac{1}{C}$, this represents the theoretical minimum RIR.

On the other hand, the FR model is also the basis for detecting the pixels that contributed the most to successful re-identification, as described in Section 4.2.1.

#### Explainer

The *explainer* provides information about the impact of each individual pixel on the FR model's predictions, thereby allowing us to detect the pixels that contributed the most to re-identification, which we call *identity-disclosing pixels set* (IDPS). We opted to use the explainer introduced in Haffar et al. (2021), which leverages adversarial examples to alter the model predictions. This method provides a comprehensive weighted list that reflects the influence of each pixel on the model's prediction. In contrast, other explainers like LIME Ribeiro, Singh, and Guestrin (2016) employ linear models to interpret the model's predictions. However, they categorize the image pixels into only two labels, making them less suitable to guide the anonymization process.

Specifically, the explainer takes each input image and its second-most probable class predicted by the FR model. The latter is treated as the target class of the adversarial example for that image; thus, it corresponds to forcing the minimum possible misclassification. Then, we calculate the loss between the correct prediction of the FR model and that incorrect target class, and backpropagation computes the gradients of this loss w.r.t. each pixel in the input image. Finally, the explainer provides us with a tensor of these gradients as output.

These gradients represent the sensitivity of the loss to changes in individual pixels. Higher magnitude gradients indicate pixels that have a greater influence on the model's prediction, especially for misclassification (a sample of such pixels is depicted in Fig. 4.1 with a blue shade on the image to be protected). Therefore, by analyzing the gradients, we can detect which pixels, if changed, would increase the likelihood of the model misclassifying the input; or, in other words, which pixels, if obfuscated, will result in an incorrectly re-identified (that is, anonymized) image. These pixels, which consequently contribute the most to re-identification, constitute the IDPS, and will be subjected to perturbation.

**Anonymization function**

The *anonymization function* in our framework can be based on any of the conventional image perturbation techniques detailed in Section 2. It takes the input image and adds a chosen pixel-level obfuscation (i.e., pixelation, blurring, noise-addition, or masking) to

the IDPS obtained in the previous step. In addition to being targeted, pixel perturbation is also applied *incrementally*: instead of perturbing pixels in a single shot, we perturb them incrementally over several times, by taking each time a *weak* value of the method's privacy parameter (e.g., the pixel block size for pixelation or the kernel size for Gaussian noise), that is, a value that achieves only a small privacy increment (and presumably a small utility decrement). We call the privacy parameter used in this manner the $\Delta$-parameter. As we will describe in Section 4.2.2, by iteratively applying perturbations to each image in small increments, we can re-evaluate the RIR of each previously perturbed image through the FR model and apply (if needed) further perturbations on a freshly re-assessed IDPS. The motivation underlying this targeted incremental approach is to minimize the amount of perturbation needed to reach a desired level of protection, and thereby maximize utility preservation.

**Privacy threshold**

It is important to note that, whereas the (privacy) parameter employed in conventional methods directly influences the attained level of protection, in our approach, the $\Delta$-parameter described above only sets the amount of perturbation added per iteration. Larger increments will result in fewer iterations and faster anonymization, but also coarser (and therefore more harmful) perturbations. Inversely, smaller increments will result in more iterations and slower anonymization but more fine-tuned (i.e., more utility-preserving) perturbations.

In our approach, the desired level of privacy is set as a *privacy threshold* that defines the maximum RIR allowed for the anonymized dataset. RIR is recomputed with Equation (4.1) at the end of each iteration by giving the perturbed image dataset resulting from that last iteration as input to the FR model. The resulting RIR value is used as the stopping criterion: the process halts when the observed RIR falls below the user-defined privacy threshold, as depicted in Fig. 4.1. Defining the desired privacy level in terms of RIR is also more intuitive and aligns better with real-world privacy requirements than employing the abstract and non-privacy-grounded parameters of conventional methods (such as pixel block or kernel sizes).

### 4.2.2 EDI-Anon workflow and algorithm

Next, we describe in detail the workflow of EDI-Anon in detail, which is also formalized in Algorithm 3.

First, the FR model is trained on the original image dataset (line 3 of Algorithm 3). Then, the anonymized dataset is initialized to the original data set, and the RIR of the anonymized dataset is assessed through the FR model (line 5). If it does not fulfill the user-defined privacy threshold (line 6), each input image in the anonymized dataset is individually assessed for re-identification using the FR model (line 8). Each correctly re-identified input image (line 9) then undergoes the following process in order to be anonymized:

1. Calculate its gradient tensor using the explainer to quantify the contribution of each pixel toward re-identification as detailed in Section 4.2.1 (line 10).

2. Select the IDPS most disclosive from the gradient tensor based on a predefined IDPS percentage (line 11).

3. Apply the user-selected anonymization function with the $\Delta$ parameter to perturb the IDPS. The output of this function is a perturbed image (line 12).

4. Replace the input image with the perturbed image in the anonymized dataset (line 13).

Once all the images have been assessed for re-identification (and perturbed, if necessary), the RIR of the resulting (anonymized) dataset is re-evaluated through the FR model (line 16). The process is repeated until the RIR falls below the user-defined privacy threshold. Through this iterative process, each perturbation increment is applied on a freshly re-assessed IDPS, which may significantly change from one iteration to the other due to the incrementally added perturbations. This results in more targeted and less unnecessary perturbations, thereby reducing utility loss.

In addition to the input parameters that define the type of anonymization function to be used (i.e., pixelation, blurring, noise addition, or masking) and the desired privacy threshold for the RIR, the algorithm relies on two interrelated constants for the anonymization function: the IDPS percentage and the $\Delta$-parameter. Both control the amount of distortion added to the images at each iteration: whereas the IDPS percentage defines the number of pixels to be distorted, the $\Delta$-parameter controls the amount of distortion added to those pixels. Their values thus influence i) the speed of the process (i.e., the larger they are, the fewer iterations are required to reach the desired level of privacy) and ii) the granularity of the added perturbation (i.e., the smaller the parameter

values, the more targeted the added perturbation is, because more re-assessments of the IDPS are performed). The nature of the $\Delta$-parameter depends on the type of anonymization function. For instance, for pixelation the $\Delta$-parameter is block size, for blurring it is kernel size, and for DP-based noise addition it is $\epsilon$. Even though the values of those parameters are, in principle, fixed for a given anonymization function, advanced users might want to tune them in order to encompass specific utility/performance needs. The influence of such tuning will be evaluated in Section 4.3.4.

It is important to highlight that once its parameters are configured, EDI-Anon operates seamlessly without the need for human intervention, and thus it is a fully automated framework.

## 4.3 Empirical results

In this section, we report empirical results from conventional image anonymization techniques, along with the outcomes of their application to the EDI-Anon framework under different privacy thresholds. For comparison, we also include the results of DeepPrivacy2, a state-of-the-art GAN-based image anonymization approach described in Section 2. All methods have been evaluated and compared in terms of privacy protection, utility preservation, and runtime. In the following, we describe the evaluation dataset and the privacy and utility metrics employed for the evaluation. The code associated with the reported results is available in the Git repository `https://github.com/RamiHaf/EDI-Anon`

All experiments were carried out on the same system specifications described in Chapter 3.

**Algorithm 3** Explainability-Driven Incremental Anonymization

**Constants:** Δ-parameter, IDPS Percentage
**Inputs:** Original Dataset, Anonymization Function, Privacy Threshold
Model.Train(Original Dataset)
Anonymized Dataset ← Original Dataset
RIR ← Model.Evaluate(Anonymized Dataset)
**while** RIR > Privacy Threshold **do**
  **for** (Input Image, Label) **in** Anonymized Dataset **do**
    Prediction ← Model.Predict(Input Image)
    **if** Prediction = Identity label **then**
      Gradient Tensor ← Explainer (Input Image, Target Class)
      IDPS ← Disclosive Pixels(Gradient Tensor, IDPS Percentage)
      Perturbed Image ← Anonymization Function(Input Image, Δ-parameter, IDPS)
      Replace.Image(Anonymized Dataset, Input Image, Perturbed Image)
    **end if**
  **end for**
  RIR ← Model.Evaluate(Anonymized Dataset)
**end while**
**return** Anonymized Dataset

### 4.3.1 Evaluation dataset

As described in Chapter 3, the MTF dataset (Haffar, Sánchez, and Domingo-Ferrer, 2025) provides high-resolution ($1024 \times 1024$) images with manually verified annotations, avoiding the low resolution and labeling inaccuracies of automatically compiled datasets like CelebA. For this work, we employ the original MTF collection of 5,246 images depicting 240 individuals (without the additional underage celebrity images contributed in Chapter 3), as it provides the comprehensive multi-label annotations necessary for evaluating both privacy protection and utility preservation across identity, age, gender, and race classification tasks. Table 4.1 reports the distribution of annotation labels in all four tasks.

The data set is divided into 70% training data, while the remaining 30% is allocated to validate and test the four tasks.

TABLE 4.1: Number of individuals for the various tasks and labels in the MTF dataset

| **Task** | **Number of individuals** | | | |
|---|---|---|---|---|
| **FR** | Identities = 240 | | | |
| **Age** | Young = 190 | | Old = 50 | |
| **Gender** | Males = 130 | | Females = 110 | |
| **Race** | Asian (Chin. & Kor.) = 80 | Asian (Indian) = 49 | Black = 35 | White = 76 |

### 4.3.2 Evaluation metrics

To evaluate privacy, we employed the RIR metric defined in Equation (4.1), which is calculated by using the FR model trained on

the original FR data from the MTF dataset. Our evaluation considers a worst-case scenario in which the adversary is assumed to have complete access to all images of the individuals. The privacy figures we report are thus on the conservative side, and better privacy will be attained in more plausible attack settings, in which adversaries have access to smaller sets of images.

For the FR model, we used the same ConvNeXt architecture described in Chapter 3, which is also recommended by the authors of the MTF dataset, adapting the last layer to the number of classes (i.e., number of individuals) in the MTF dataset. Recall that RIR measures the accuracy of the predictions without making a distinction between false positives (FP) and false negatives (FN), and thereby captures the overall correctness of the FR task.

Since our goal is to evaluate the performance of anonymization methods at protecting a set of images (or, in other words, to assess the re-identifiability of the actual anonymized images), we measured the RIR of the anonymized images in the training dataset through the FR model trained on the original (unanonymized) versions of those very same images. As a reference, the RIR of the original FR training images was 100%. This means that the RIR of the anonymized images for each of the methods considered in the following can be interpreted in absolute terms.

To assess utility preservation, we focused on the other three tasks supported by the MTF dataset: age, gender, and race classification. Following the methodology outlined in the MTF paper (Haffar, Sánchez, and Domingo-Ferrer, 2025), we trained three corresponding ML models on the original (non-anonymized) images of the training dataset by using the ConvNeXt pre-trained Base model. The hyper-parameters for those three models were

the same as those used in the MTF paper, with a batch size of 32 images.

Utility preservation was measured as the performance of the three models on the anonymized images. The models' performance was quantified via the $F_1$ score, which is the harmonic mean of precision and recall (Rijsbergen, 1979), and is calculated as

$$F_1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

where *Precision* is the ratio of true positives (TP) to the sum of TP and FP, i.e. $\frac{\text{TP}}{\text{TP+FP}}$, and *Recall* is the ratio of TP to the sum of TP and FN i.e. $\frac{\text{TP}}{\text{TP+FN}}$.

Considering both precision and recall, the $F_1$ score provides a comprehensive evaluation of model performance (and, hence, of data utility), in that it captures both the ability to correctly identify instances and to avoid FP and FN. This metric is particularly useful for ML tasks with a low number of classes (as is the case for the three utility-oriented tasks) because it is less affected by potential class imbalance. As a reference and upper bound for the utility figures reported in the next sections, the baseline $F_1$ scores on the original non-anonymized images were: 97.93% for age classification, 98.77% for gender classification, and 97.32% for race classification.

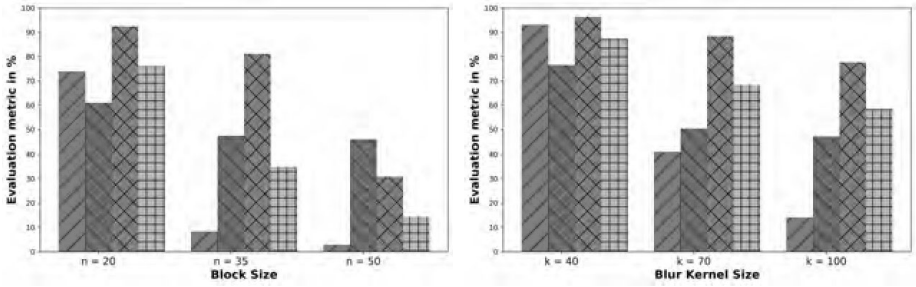
### 4.3.3 Conventional methods

For pixelation and blurring, the privacy parameters used were $n$ (block size) and $k$ (blur kernel size), respectively. The proposed values for these parameters in Yuan et al. (2022) and Schuler et al. (2015) –which did not evaluate empirical re-identification– were

$n \in [4, 10]$ and $k = 25$. However, according to our privacy evaluation, these resulted in negligible reductions in RIR. Therefore, we decided to use larger parameter values to produce observable differences, as shown in Figs. 4.2a and 4.2b. For DP-Pix and DP-Blur, Fan (2019) employed $\epsilon \in \{0.1, 0.5, 1\}$ and $m = 16$ neighboring pixels. Furthermore, for DP-Blur, a kernel size ($k$) of 99 was recommended. We used the same parameter values in our tests, even though $m = 16$ only accounts for 0.0015% pixels in our images. Considering that in face images, a substantial proportion of pixels contribute to re-identification, it is clear that such a small set of pixels would not result in any meaningful DP guarantee. When no *ex ante* privacy guarantees hold, the attained level of privacy can only be evaluated empirically (Blanco-Justicia et al., 2023), as we do in this section. As a matter of fact, setting larger –and safer– $m$ (ideally, the whole image, $m = 1024 \times 1024$) produced random pixels for any value $\epsilon$, even for extremely large ones.

For masking, we simulated the effect of wearing a face mask and/or sunglasses. We used Haar-cascade (Santana et al., 2008) for mouth and eye detection via the OpenCV library (Bradski, 2000), and applied large enough round and square masks to hide the detected facial features.

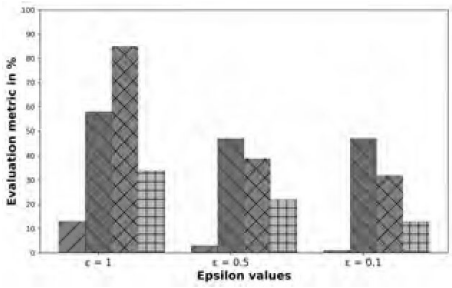
As shown in Fig. 4.2e, the least effective technique w.r.t. RIR was mouth masking, due to the mouth region lacking identity-revealing features. Although eye masking proved more effective at concealing the individual's identity, it still yielded an RIR larger than most other methods, while combining mouth and eyes masking produced a minor decrease in RIR over just masking the eyes. However, this minor decrease in RIR came at the cost of a small reduction in utility for the classification of age and gender, and a



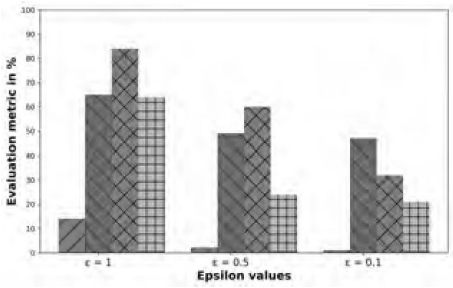

(A) Pixelation



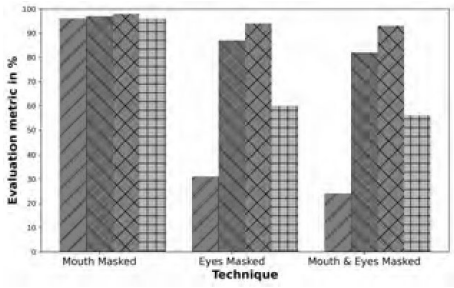

(B) Blurring




(C) DP-Pix




(D) DP-Blur




(E) Mouth and eye masking

FIGURE 4.2: Evaluation results of conventional anonymization methods for different privacy parameters. Within each group, from left to right, the bars correspond to RIR, $F_1$-Age, $F_1$-Gender, and $F_1$-Race.

more noticeable reduction for the classification of races. The latter can be explained due to the importance of the shape of the eyes in recognizing Asian faces.

Pixelation and blurring, which apply a uniform perturbation across all pixels, achieved stronger privacy (i.e., lower RIR) for $n > 20$ and $k > 40$. However, both methods significantly hampered classification by age and race, as shown in Figs. 4.2a and 4.2b. The accuracy of the age classification model was particularly affected by the misclassification of the *Old* class because pixelation and blurring can hardly preserve facial wrinkles that are a distinctive feature of elderly individuals. On the other hand, the performance of the race classification model was hampered by the loss of finer details, including the shapes of the eyes, lips, and noses, which are key for identifying the *Asians*. In particular, blurring performed better to preserve the *black* class due to the unchanged facial contours and skin tones.

Although DP-based methods lack formal privacy guarantees for face images due to the small value of $m$ employed, both DP-Pix and DP-Blur achieved the strongest empirical privacy, as reported in Figs. 4.2c and 4.2d. Lower $\epsilon$ values resulted in more effective anonymization for both methods. DP-Pix exhibited the lowest utility preservation, whereas DP-Blur offered a slightly better privacy/utility trade-off. The relatively better performance of DP-Blur can be attributed to the application of Gaussian blur, which improved image quality by smoothing the images and reducing the effect of the added (colored) noise. This was especially useful for race classification. Nevertheless, the utility reached for both methods with $\epsilon < 1$ approached the random guess for all tasks.

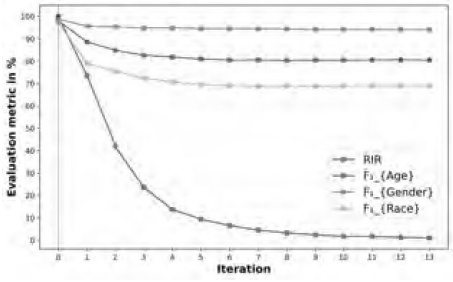
### 4.3.4 EDI-Anon

As explained in Section 4.2.1, EDI-Anon is enforced via an anonymization function that can implement a variety of conventional methods. In this section, we evaluate the application of the conventional methods considered in the previous section within the EDI-Anon framework. We refer to them by preceding their names by "EDI-". To accommodate the iterative nature of EDI-Anon, we took weak values of the privacy parameters of the different methods (i.e., what we refer to as $\Delta$-*parameter*, as explained in Section 4.2.1). Specifically, for EDI-Pix and EDI-Blur, their respective $\Delta$-parameters were set to $n = 20$ and $k = 40$ (i.e., the weakest values employed in the previous experiments), whereas for EDI-DP-Pix and EDI-DP-Blur we used a notably higher $\epsilon = 100$ while keeping $m$ and $k$ unchanged from their conventional counterparts with one shot. As discussed in Section 4.3.3, these loose parameters do not provide *ex ante* privacy guarantees for EDI-DP, but neither did the one-shot counterparts.

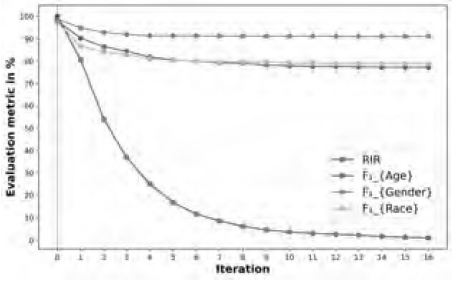
Facial masking was also adapted to the targeted nature of EDI-Anon: instead of applying fixed-size masks to predefined facial features, the EDI-Masking version iteratively added small-sized masks of 25-pixel radius over each IDPS pixel detected at each iteration.

We set the IDPS percentage to 10% for the first four techniques, whereas for EDI-Masking –which blacks out the masked pixels and, therefore, causes more utility loss at each application– we used a smaller value of 1%.

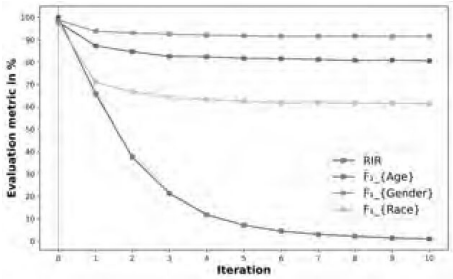
The privacy threshold for the EDI-Anon methods was set to the lowest RIR achieved by any conventional method, that is, the



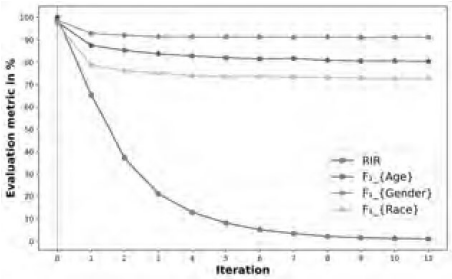

(A) EDI-Pix



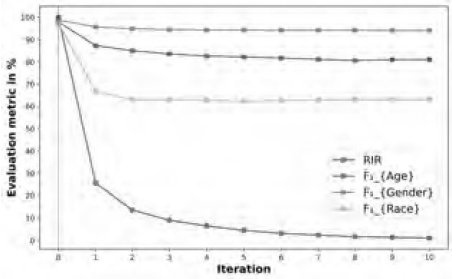

(B) EDI-Blur




(C) EDI-DP-Pix




(D) EDI-DP-Blur




(E) EDI-Masking

FIGURE 4.3: Evolution of RIR and utilities for EDI-Anon across iterations with different anonymization functions until reaching a 1% privacy threshold

1% achieved by DP-Blur. In Fig. 4.3, we report the evolution of the RIR and the utility metrics across iterations until that privacy threshold is reached.

We can see that, regardless of the underlying anonymization function, EDI-Anon was able to achieve the target privacy threshold with a similar number of iterations. Despite their differences, all EDI-Anon methods significantly lowered the RIR within the initial iterations, and most iterations were devoted to lower the RIR from 10% to 1%. The sharp decline in RIR was accompanied by a milder decrease in the utility metrics. Differences across methods manifest in some utilities being slightly better preserved than others. In general, all EDI-Anon methods affected all three utilities proportionally to their number of classes; that is, the more classes, the less utility because of having to deal with a more difficult classification problem. Age classification was similarly preserved across all methods, indicating that age-related facial features were minimally affected by the EDI-Anon process. On the other hand, EDI-DP and EDI-Masking hampered race classification the most due to the loss of race-related facial features (e.g., those depicting eye shapes, facial contours, and skin tones).

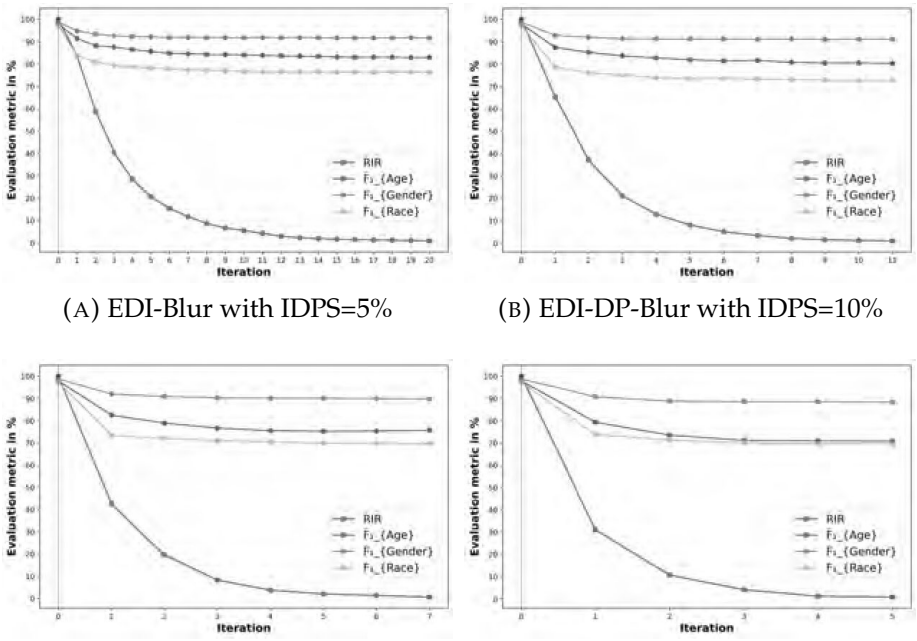
**On EDI-Anon constants**

As explained in Section 4.2.2, in addition to the parameters that define the type of anonymization function to be used by EDI-Anon and the desired privacy threshold, two other values also influence the behavior of the algorithm: the IDPS percentage and the $\Delta$-parameter of the chosen anonymization function. In the following, we analyze the impact of these values on EDI-DP-Blur,

which provided a good balance between RIR reduction and utility preservation.

First, we investigate the effect of modifying the IDPS percentage by selecting four different values (5%, 10%, 20%, and 30%), while keeping all other parameters as in the previous experiment. The results are reported in Fig. 4.4. As expected, higher IDPS percentages resulted in fewer iterations needed to achieve the privacy threshold, as more pixels were distorted per iteration. However, the less targeted distortion also resulted in diminished utilities. This shows the benefit of adding distortion to only the most re-identifying pixels, and the need for re-evaluating those pixels after each iteration, since they may significantly change from one iteration to the next as a result of the added perturbation. The reduction in the number of iterations was not linear in the increase of the IDPS percentage. This suggests that the less important pixels considered when enlarging the IDPS set had a smaller impact on the prediction of the FR model. On the other hand, perturbing them had a noticeable impact on the three utilities. This illustrates that the lower the IDPS percentage, the more accurate and targeted the added distortion is; yet, one should take into account the added runtime resulting from the additional iterations (see the runtime analysis in Section 4.3.5).

Second, we examined the influence of the $\Delta$-parameter on EDI-DP-Blur by setting $\epsilon \in (1, 10, 50, 100)$ while keeping all other parameters unchanged. The findings are presented in Fig. 4.5. The results show a similar tendency as when modifying the IDPS percentage: the stricter the $\Delta$-parameter (i.e., the lower $\epsilon$ is), the fewer iterations are needed to attain the privacy threshold, but, at the same time, the more affected the utilities are. In fact, the strictest



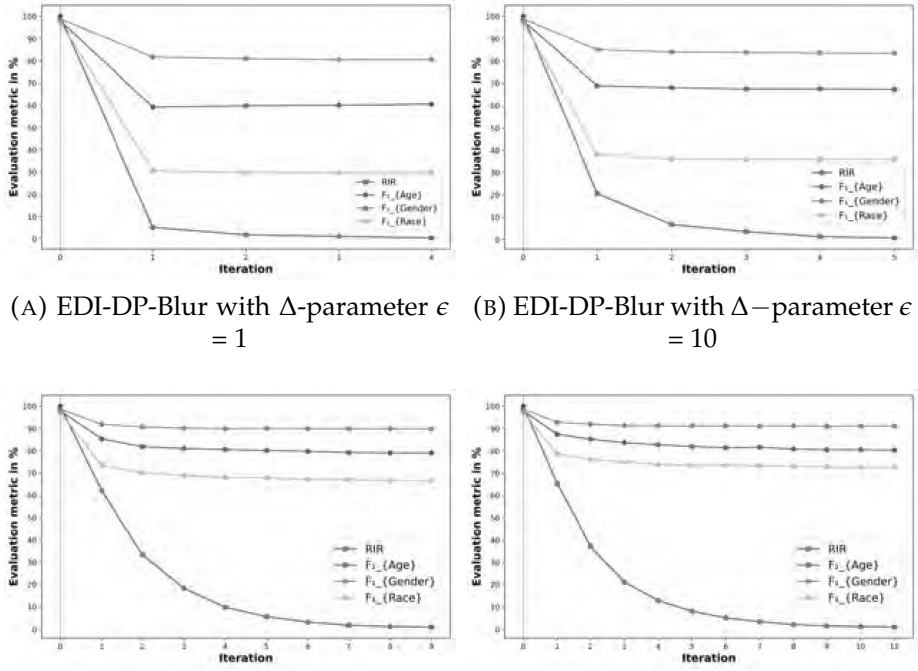

(A) EDI-Blur with IDPS=5%

(B) EDI-DP-Blur with IDPS=10%

(C) EDI-DP-Blur with IDPS=20%

(D) EDI-DP-Blur with IDPS=30%

FIGURE 4.4: Evolution of RIR and utilities for EDI-DP-Blur across iterations with different IDPS Percentages until reaching a 1% privacy threshold




(A) EDI-DP-Blur with Δ-parameter $\epsilon$ = 1

(B) EDI-DP-Blur with Δ−parameter $\epsilon$ = 10

(C) EDI-DP-Blur with Δ-parameter $\epsilon$ = 50

(D) EDI-DP-Blur with Δ-parameter $\epsilon$ = 100

FIGURE 4.5: Evolution of RIR and utilities for EDI-DP-Blur across iterations with different Δ-parameter values until reaching a 1% privacy threshold

Δ-parameter produced a sharp decline of all three utilities, which barely changed in subsequent iterations. This indicates that the affected pixels were almost randomized due to the large perturbation added by DP in just one iteration. Thus, it is important to choose a relatively weak Δ-parameter, which will allow keeping the added perturbation per iteration small enough for the affected pixels to still retain (some of) their utility.

### 4.3.5 Comparisons

To gain a better understanding of the advantages offered by EDI-Anon over existing methods, in this section, we compare conventional techniques against their EDI-Anon counterparts under equivalent conditions. Furthermore, we report results of a state-of-the-art GAN-based image anonymization method (DeepPrivacy2). Finally, we report and discuss the run-time incurred by each method.

#### Comparison with conventional methods

Whereas EDI-Anon can be set to reach a target RIR/privacy threshold, the RIR attained by the conventional one-shot methods depends on the values of their parameters and is hard to fix *ex ante*. To enable a fair comparison, in the following, we report EDI-Anon utilities for the same RIR achieved by its conventional counterpart with the privacy parameter that resulted in the lowest RIR (see Fig. 4.2). To facilitate the comparison, Fig. 4.6 reports the average of the three utilities considered for the different methods.

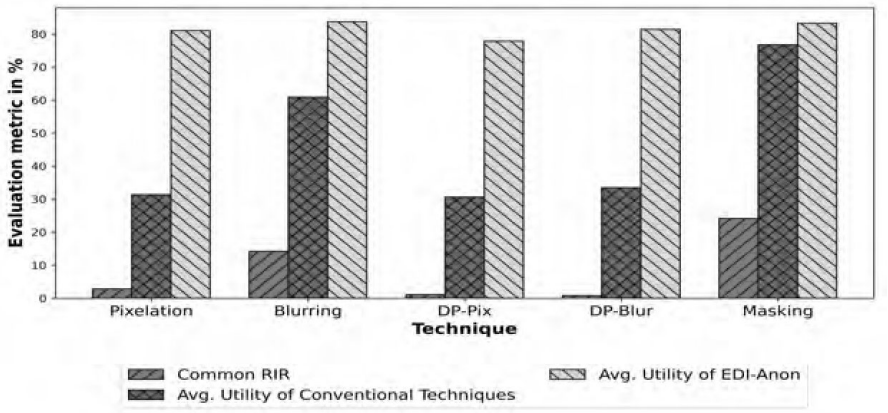



FIGURE 4.6: Average utility comparison between EDI-Anon and its conventional counterparts at equal RIRs

We observe significant disparities between the methods. In instances where conventional methods yielded lower privacy levels/higher RIR (such as masking and blurring), the difference in utility between EDI-Anon and its conventional counterpart was minor. On the other hand, when conventional methods attained higher privacy levels (as was the case for Pixelation, DP-Pix and DP-Blur), utilities were significantly compromised due to the uniform and random nature of the (large) perturbation added. For those methods, their EDI-Anon counterparts more than doubled the average utility due to the targeted and incremental distortion. We can, therefore, conclude that EDI-Anon is especially beneficial when high privacy levels are needed.

**Comparison with DeepPrivacy2**

As discussed in Section 2, GANs are employed in state-of-the-art image anonymization methods due to their ability to create synthetic images that conceal identifiable information while retaining relevant facial features. In this section, we compare EDI-Anon with DeepPrivacy2 (Hukkelås and Lindseth, 2023), a state-of-the-art cGAN-based approach.

DeepPrivacy2 has been trained on the FDF dataset, which was specifically curated for that purpose. The FDF dataset encompasses an extensive collection of over a million human faces with a wide spectrum of poses, diverse age groups, genders, and racial identities, complete with bounding box and some automated keypoint annotations. We employed the implementation offered by the authors of DeepPrivacy2 [1], which includes the trained GAN model with 46 million parameters. To generate the anonymized images, we adhered to the recommended batch size of 1.

Applying DeepPrivacy2 to the MTF dataset, we obtained an anonymized dataset with a RIR = 16.27%. To allow a fair comparison with EDI-Anon, in Table 4.2 we report the average utility obtained by DeepPrivacy2 and all versions of EDI-Anon when setting the privacy threshold of the latter to 16%.

We can see that all EDI-Anon versions outperformed DeepPrivacy2 by a significant margin. Moreover, even though the average utility for DeepPrivacy2 was –apparently– not that low, the synthetic nature of the images it produces may harm tasks requiring truthful images and/or realistic human faces.

[1] Github URL of DeepPrivacy2: `https://github.com/hukkelas/deep_privacy2`

TABLE 4.2: Average utility comparison between DeepPrivacy2 and EDI-Anon methods for equal RIR $\approx 16\%$

| Technique | Average Utility |
|---|---|
| DeepPrivacy2 | 73.89% |
| EDI-Pix | 82.46% |
| EDI-Blur | 83.75% |
| EDI-DP-Pix | 79.23% |
| EDI-DP-Blur | 82.64% |
| EDI-Masking | 81.00% |

**Runtime comparison**

In this section, we report the average runtime per image required to anonymize the MTF dataset under the various methods considered so far. This has been calculated by dividing the total runtime required to anonymize the whole dataset by the number of images.

For each method, we picked the instantiation that resulted in the best privacy (i.e., lowest RIR). Even if the level of privacy attained does not affect the runtime of conventional methods applied in one shot, it does affect the runtime of EDI-Anon methods. Thus, for the latter, we report the time of their longest run. The results are shown in Table 4.3. To put EDI-Anon figures in context, we also specify the number of iterations needed to reach the privacy threshold.

As expected, the simplest methods, such as pixelation and blurring, resulted in the lowest runtimes. DP-based methods were costlier due to the additional effort required to compute noise distributions. In comparison, their EDI-Anon counterparts took 3-17 times longer runtimes, a magnitude that is proportional to the

TABLE 4.3: Average runtime per image for conventional methods, their EDI-Anon counterparts and DeepPrivacy2

| Method | Conventional | EDI-Anon (iterations) |
|---|---|---|
| **Pixelation** | 0.04s | 0.65s (13) |
| **Blurring** | 0.04s | 0.7s (16) |
| **DP-Pix** | 0.13s | 0.4s (7) |
| **DP-Blur** | 0.17s | 0.44s (8) |
| **Masking** | 0.09s | 0.5s (10) |
| **DeepPrivacy2** | 2.29s | NA |

number of iterations required to reach the desired privacy threshold. In contrast, DeepPrivacy2 was, by far, the slowest in generating anonymized images, which can be attributed to the complexity of the GAN model used.

For both DeepPrivacy2 and EDI-Anon, we should consider the added cost of training the GAN and FR models, respectively. Although this is a one-time cost for the whole set of images, for EDI-Anon on the MTF dataset, this accounted for 12,104.4 seconds. Unfortunately, the authors of DeepPrivacy2 did not provide information on the GAN model training time. Nevertheless, considering the size of the training data they used (over 1 million images) and the complexity of the GAN model (46 million parameters), we can assume it to be orders of magnitude larger than that of EDI-Anon's FR model training.

In spite of its longer runtime, EDI-Anon has an intrinsic advantage over conventional one-shot methods. Whereas the latter requires re-running the whole process from scratch to create anonymizations with different privacy parameters, using a strict privacy threshold with EDI-Anon immediately yields not only the

results corresponding to that threshold but any of the (weaker) anonymizations resulting from intermediate iterations. This is quite convenient for practitioners who aim to optimize the privacy/utility trade-off *ex post*, a task that is usually implemented as a trial and error process for conventional methods.

## 4.4 Conclusion

We have presented Explainability-Driven Incremental Image Anonymization (EDI-Anon), a novel framework that proposes an iterative anonymization process guided by the observed RIR. Our proposal addresses the limitations of conventional anonymization techniques, which enforce one-shot uniform pixel perturbations. EDI-Anon leverages explainability techniques to identify and perturb only pixels that hold identity-disclosing features. Due to its incremental nature, EDI-Anon can automatically achieve any level of privacy protection while preserving image features more accurately than conventional methods. Our framework also offers flexibility and customization, allowing end users to choose from a variety of anonymization techniques, and to select *ex ante* the desired privacy threshold in terms of maximum allowed RIR.

EDI-Anon has empirically demonstrated its superiority over conventional and GAN-based techniques. It excels at providing higher levels of privacy while preserving more utility-critical pixels.

# Chapter 5

# Explainability-driven image anonymization in latent space (EDIALS)

As stated in the previous chapter, conventional image anonymization methods apply uniform distortion across all pixels, unnecessarily degrading both privacy and utility. Chapter 4 introduced EDI-Anon, which improved upon these conventional methods by using explainability techniques to selectively target identity-revealing pixels. Building upon this approach, EDIALS (Explainability-Driven Image Anonymization in Latent Space) addresses the same challenge, but by operating in the latent feature space where facial attributes are represented in a structured manner. Unlike EDI-Anon, which modifies pixel values directly, EDIALS uses explainability techniques to identify and selectively distort only those latent features most responsible for identification.

## 5.1 Contributions and plan of this chapter

The main contributions of this chapter are threefold. First, we adapt the ABELE framework (Guidotti et al., 2020) for face image anonymization by employing specialized encoder and decoder architectures designed for facial image processing. Second, we introduce a novel feature distortion mechanism that selectively modifies identity-revealing features in the latent space based on explainability-derived decision rules. Third, we demonstrate through extensive evaluation that our approach achieves a better privacy-utility trade-off than both conventional anonymization methods and state-of-the-art GAN-based techniques.

The remainder of this chapter is organized as follows. Section 5.2 describes the EDIALS methodology. Section 5.3 presents our experimental setup, evaluation framework, and the analysis and discussion of the empirical results. Section 5.4 reports the conclusions of the work.

## 5.2 EDIALS

EDIALS selectively targets identity-revealing features while preserving characteristics essential for analytical tasks such as age, gender, and race classification. In this way, EDIALS enables researchers to train models on anonymized images that can later generalize well to non-anonymized images in real-world applications.

### 5.2.1 Image explainer

Our work builds on the foundational framework introduced in the ABELE explainer (Guidotti et al., 2020). ABELE offers a model-agnostic approach for explaining black-box decisions for image classification by leveraging the latent feature space learned through an AAE. The key strength of ABELE lies in its ability to identify features that are deemed important by any black-box model ($b$) without requiring access to its internal structure. This capability is particularly valuable for our anonymization task, as it enables us to precisely target the features that contribute most significantly to identity recognition. We leverage ABELE's methodology for encoding images into the latent space ($z$) and extracting decision rules. We then extend it to a distortion-based mechanism tailored to protect face images.

We chose ABELE over explainers like LIME (Ribeiro, Singh, and Guestrin, 2016) because LIME fails to account for pixel relationships, which produces explanations that uniformly cover most of the image and make them unsuitable for targeted anonymization. ABELE, by contrast, leverages the latent feature space to generate meaningful exemplars that respect feature relationships. Empirical evidence by Guidotti et al. (2020) shows that ABELE outperforms LIME in fidelity (mimicking black-box behavior), coherence, and stability, making it ideal for precisely identifying identity-revealing features.

### 5.2.2 Adapting ABELE to facial recognition

Since FR represents a significantly more complex task than the general image classification problems addressed in the original

ABELE implementation, we have adapted its architecture components while preserving ABELE's core model-agnostic methodology.

For the AAE, we employ a truncated ConvNeXt-Base model (Liu et al., 2022) pre-trained on ImageNet as the encoder ($\mathcal{E}$), removing the classification layers to capture rich visual features at multiple scales. The encoder outputs a 1024-dimensional bottleneck representation, compressing the image information while preserving semantic details.

The decoder ($\mathcal{D}$) uses a progressive upsampling architecture featuring five transposed convolutional stages that gradually increase the spatial resolution from 8×8 to 256×256 pixels. Each upsampling stage incorporates a transposed convolution followed by LayerNorm (Ba, Kiros, and Hinton, 2016) and GELU activation (Hendrycks and Gimpel, 2016) for stable training. The final layer employs a tanh activation function to generate normalized red green blue (RGB) images within the [-1,1] range.

The discriminator ($D$) consists of convolutional layers that process the latent vectors, followed by fully connected layers that determine whether a latent representation belongs to the prior distribution. To ensure stable adversarial training, we employ hinge loss (Lim and Ye, 2017) with gradient penalty regularization (Gulrajani et al., 2017).

Together, these components enable high-fidelity reconstruction of facial images from their latent representations while maintaining the semantic details that are essential for both identity protection and utility preservation. Our method employs AAEs rather than conventional autoencoders because they combine GANs

with autoencoder learning, ensuring that generated samples follow the same distribution as original images. The AAE architecture allows us to work in a lower-dimensional latent space ($k \ll n$), where features are more structured and semantically meaningful, though not completely disentangled. While perfect feature disentanglement remains an active research challenge, as demonstrated by recent work (Mienye and Swart, 2025), the AAE's latent space provides sufficient organization of semantic concepts to allow targeted manipulation of identity-related features.

### 5.2.3 EDIALS design

Thanks to ABELE's model-agnostic nature, EDIALS is designed with a modular architecture that decouples the encoding-decoding mechanisms from the core anonymization logic. This enables seamless integration of future advancements in representation learning: as more sophisticated autoencoder variants emerge with improved feature structuring capabilities, they can be seamlessly incorporated into our algorithm without redesigning the fundamental anonymization procedure. Similarly, for the black-box model ($b$) that EDIALS aims to anonymize against, we utilize a ConvNeXt architecture described in Chapter 3. This choice reflects the practical reality of modern FR systems, ensuring evaluation against a realistic adversary.

The EDIALS workflow is illustrated in Figure 5.1, which shows the progress from initial image encoding to final anonymization, and it is formalized in Algorithm 4. It begins with initialization (line 1) of an empty set to store anonymized images ($X_{\text{anon}}$). For

each image in the original dataset ($X$) (line 2), the algorithm encodes it into its latent representation using the encoder $\mathcal{E}$ (line 3). This transformation is crucial as it enables manipulation in a lower-dimensional space where semantic features are more distinctly represented.

Following ABELE's approach, EDIALS generates a neighborhood of latent vectors ($H$) around the original encoding through the *neighgen* module (line 4). This neighborhood includes instances with both decisions, i.e., $H = H_{=} \cup H_{\neq}$ where instances $h \in H_{=}$ are such that $b(\tilde{h}) = b(x)$, and $h \in H_{\neq}$ are such that $b(\tilde{h}) \neq b(x)$. This balanced neighborhood is crucial as it enables the decision tree to accurately learn the local behavior of the black-box model.

To ensure that the generated latent vectors correspond to plausible images, the algorithm filters them using $D$ (line 5), retaining only valid vectors ($\tilde{H}$) for further processing. These valid latent vectors are then decoded back into the image space using $\mathcal{D}$ (line 6), producing a set of images ($\tilde{X}$) that represent valid variations of the original. By querying the black-box model with the decoded images (line 7), EDIALS obtains their classifications ($Y$) to train a decision tree ($c$) alongside the corresponding latent vectors (line 8). This interpretable model approximates the decision boundary of the black-box model in the latent space. From this tree, the algorithm extracts a decision rule ($r$) and identifies the set of features $Z_r = \{z_1, z_2, \ldots, z_m\}$ that significantly influence the classification outcome (line 9). These identity-revealing features become the target for the image distortion process.

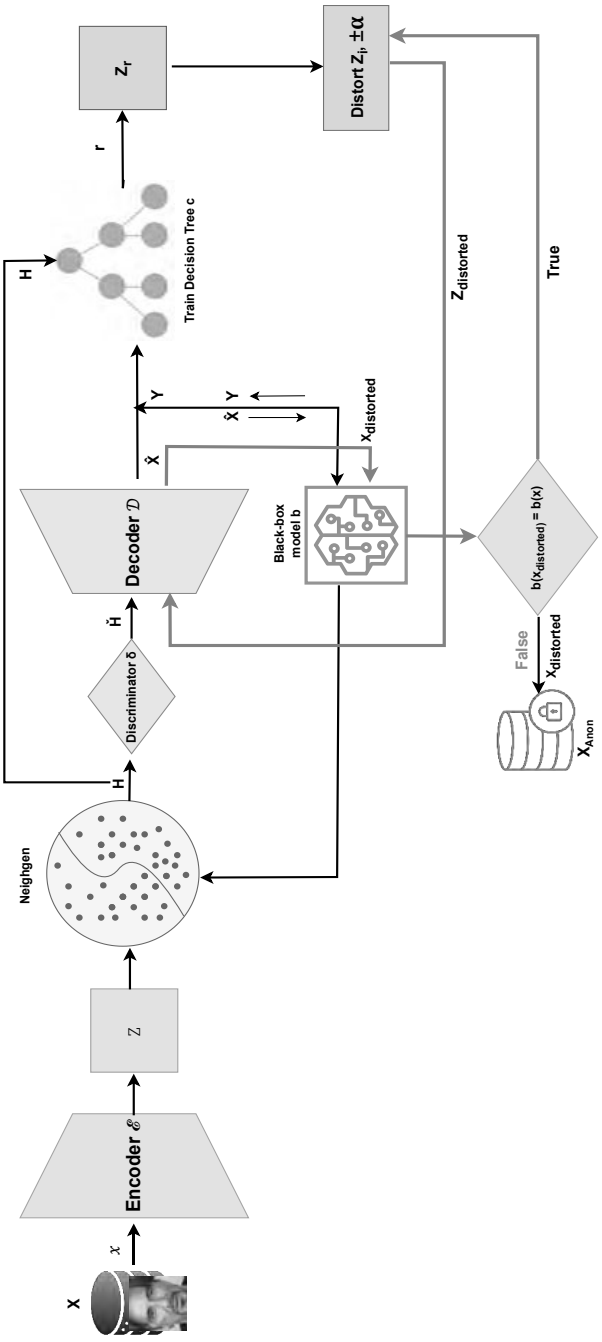



FIGURE 5.1: EDIALS workflow: Input $x$ from dataset $X$ is encoded by $\mathcal{E}$ into latent representation $z$, then processed through *neighgen* to create neighborhood $H$. After validation by $\sigma$ and decoding by $\mathcal{D}$, a decision tree $c$ extracts rule $r$ and features $Z_r$ for targeted distortion. The green pathways illustrate how $z_{\text{distorted}}$ is incrementally and iteratively modified by $\alpha$ and decoded until $b$ misclassifies (i.e., fails to re-identify) the resulting $x_{\text{distorted}}$, which is then added to $X_{\text{anon}}$.

At this point, EDIALS extends ABELE's methodology by implementing a novel, incremental distortion strategy. Whereas ABELE generates both factual and counter-factual rules to explain black-box decisions through exemplars and counter-exemplars, EDIALS deliberately focuses only on the factual rules. This design choice is motivated by our anonymization objective: factual rules directly capture the features that most strongly support the current identity classification, making them ideal targets for privacy-preserving distortion. Counter-factual rules, while valuable for explanations, could lead to unpredictable identity transitions rather than targeted anonymization. By concentrating solely on factual rules, we achieve more controlled and smaller distortions. The core distortion phase initializes a copy of the original latent vector ($z_{\text{distorted}}$) and establishes a feature counter ($i$) to track the feature currently being modified (line 10). The algorithm then enters an iterative distortion loop that continues until the black-box model misclassifies (i.e., fails to re-identify) the distorted image (line 12), indicating successful anonymization.

Within each iteration, the algorithm modifies one feature at a time by adding or subtracting the distortion factor $\alpha$ (line 13). The direction of the distortion is determined by the condition in the decision rule: if a rule condition specifies that $z_i < \mathit{threshold}$, then $\alpha$ is added to increase the value and violate this condition; conversely, if a rule condition states $z_i \geq \mathit{threshold}$, then $\alpha$ is subtracted to push the value below the threshold. This targeted approach ensures that the distortion directly counters the specific conditions that lead to the correct identity classification. After each feature modification, the distorted latent vector is decoded to produce an image ($x_{\text{distorted}}$) (line 14), which is evaluated by the

black-box model to check if misclassification has occurred. The distortion process systematically progresses through the identified features by incrementing the feature counter after each modification (line 15). When all features have been processed once without achieving misclassification, the counter is reset (lines 16-17), allowing multiple passes through the feature set with cumulative distortions.

This targeted and incremental approach ensures minimal image distortion by modifying only the features necessary for preventing re-identification while preserving characteristics contributing to utility-oriented tasks. The key advantage of incremental distortion is that it adapts to each feature's varying sensitivity to perturbation. By testing after each increment, we apply precisely the necessary distortion per feature, avoiding over-distortion. Once an image is successfully misclassified, it is added to the anonymized dataset (line 20). The algorithm continues this process for all images in the original dataset and ultimately returns the complete anonymized collection (line 22).

## 5.3 Empirical results

We next report a comprehensive evaluation of EDIALS and compare it to conventional anonymization methods and state-of-the-art approaches. Our analysis focuses on both privacy protection and utility preservation metrics.

All experiments were conducted on the same system specifications described in Chapter 3. Our implementation uses PyTorch 2.2.2 and is available, along with the datasets used, at `https://github.com/yk0803/EDIALS`.

**Algorithm 4** EDIALS

**Require:** Dataset $X$, Encoder $\mathcal{E}$, Decoder $\mathcal{D}$, Discriminator $D$, Black-box model $b$, Distortion factor $\alpha$
**Ensure:** Anonymized dataset $X_{\text{anon}}$

Initialize $X_{\text{anon}} \leftarrow \emptyset$
**for** each image $x \in X$ **do**
  $z \leftarrow \mathcal{E}(x)$
  $H \leftarrow \text{neighgen}(z, b)$
  $\tilde{H} \leftarrow \{h \in H \mid \delta(h) = \text{valid}\}$
  $\tilde{X} \leftarrow \{\mathcal{D}(h) \mid h \in \tilde{H}\}$
  $Y \leftarrow \{b(\tilde{x}) \mid \tilde{x} \in \tilde{X}\}$
  $c \leftarrow \text{DecisionTree}.Train(H, Y)$
  Extract rule $r$ & features $Z_r = \{z_1, z_2, \ldots, z_m\}$ from $c$
  $z_{\text{distorted}} \leftarrow z$, $i \leftarrow 1$
  $x_{\text{distorted}} \leftarrow \mathcal{D}(z_{\text{distorted}})$
  **while** $b(x_{\text{distorted}}) = b(x)$ **do**
    $z_{\text{distorted}}[z_i] \leftarrow z_{\text{distorted}}[z_i] \pm \alpha$
    $x_{\text{distorted}} \leftarrow \mathcal{D}(z_{\text{distorted}})$
    Increment $i$
    **if** $i > m$ **then**
      Reset $i \leftarrow 1$
    **end if**
  **end while**
  $X_{\text{anon}} \leftarrow X_{\text{anon}} \cup \{x_{\text{distorted}}\}$
**end for**
**return** $X_{\text{anon}}$

### 5.3.1 Evaluation dataset and metrics

We utilize the MTF dataset as described in Chapter 4. We employ the same evaluation metrics as defined in Chapter 4: RIR for privacy protection and $F_1$ scores for utility preservation across age, gender, and race classification tasks.

However, the utility evaluation methodology differs from Chapter 4 in an important way. Whereas EDI-Anon evaluated utility by applying models trained on original images to anonymized images (due to the hundreds of iterations making re-training impractical), EDIALS follows a more practical approach: we re-train the age, gender, and race classification models directly on the anonymized datasets (with identity labels removed). This methodology better reflects real-world scenarios where researchers would train their models on already anonymized data, providing a more realistic assessment of the practical utility of the anonymized data. The configurations of the FR and utility models remain the same as in the previous chapter.

### 5.3.2 Conventional anonymization methods

We begin our empirical evaluation by analyzing the conventional image anonymization techniques introduced in Chapter 2: pixelation, Gaussian blurring, motion blurring, facial feature masking, DP-Pix, and DP-Blur. These methods serve as baselines for further comparison. We evaluate conventional methods using the retrained model methodology described in Section 5.3.1, providing realistic baselines for comparing EDIALS performance.

Parameter values were selected following the same methodology established in Chapter 4, where we demonstrated that standard recommended parameters resulted in negligible privacy protection. Since our objective was to reduce RIR to levels that approach random guessing while maintaining utility metrics well

above baseline random performance (50% for binary classifications such as age and gender, and 25% for the four-class race classification), we expanded the parameter ranges to achieve meaningful anonymization.

For DP-based methods, we used the same parameter configurations as in Chapter 4 ($\epsilon \in \{0.1, 0.5, 1\}$ with $m = 16$), which, as previously discussed, provide only empirical privacy protection rather than theoretical DP guarantees.

As shown in Table 5.1, increasing the block size in pixelation improves privacy protection but reduces utility across all classification tasks. The race classification task, which requires distinguishing between four different classes, suffers the most significant degradation, dropping from 87.33% at $n = 20$ (and very large RIR) to just 18.25% at $n = 50$ (with reasonably low RIR). This occurs because pixelation obscures distinctive facial features such as eye shape, nose structure, and skin tone variations that are essential for recognizing different racial characteristics. Age classification performance also declines substantially as pixelation removes crucial age-related markers, particularly affecting the ability to identify older individuals who form the minority class in our dataset. Gender classification remains more resilient to pixelation compared to the other tasks, likely because gender recognition can leverage broader facial proportions and contours that remain partially discernible even at lower resolutions.

Gaussian blurring follows a different pattern: while it preserves some of the general facial structure, it significantly degrades fine details and texture information. Age classification is particularly affected by blurring, as this task relies heavily on facial wrinkles, skin texture, and other subtle age-related features that

are smoothed out even by moderate levels of Gaussian blur. Race classification also suffers, as blurring homogenizes facial features crucial for distinguishing between racial groups. Gender classification maintains better performance under blurring because it can still utilize the preserved facial contours and proportions that remain visible after the blurring process.

Motion blurring creates directional distortion that has a unique impact pattern. At moderate levels, it disrupts identifying features while maintaining the overall facial structure, achieving better privacy-utility trade-offs than equivalent levels of Gaussian blurring. However, at high intensities (with $k = 300$), motion blur severely distorts all facial features, causing substantial degradation across all classification tasks. Unlike pixelation and Gaussian blur, extreme motion blur particularly impacts gender classification because the directional smearing distorts the overall facial shape and proportions that gender classification typically relies upon, dropping below a random guess (41.67%) at $k = 300$.

We also evaluated masking techniques by simulating the effect of wearing face masks and/or sunglasses. Using Haar-cascade (Santana et al., 2008) for detection via the OpenCV library (Bradski, 2000), we applied masks to hide specific facial features. As shown in Table 5.1, mouth masking alone proved least effective, with a RIR as large as 94.12%, because mouth regions generally lack strong identity-revealing features. However, utility metrics remained high with only minimal degradation from the original values. Eye masking achieved better privacy protection with RIR reduced to 30.21%, while combining both eye and mouth masking further reduced RIR to 27.27%. The eye masking approach

had a moderate impact on utility, particularly affecting race classification (dropping to 61.25%) since eye shape is a key feature for racial identification. Combined eye and mouth masking had more pronounced effects on all utility metrics, though gender classification remained relatively robust at 83.11%.

DP-based methods achieve stronger privacy protection than previous methods, but at a higher cost to the utility, as depicted in Table 5.1. DP-Pix and DP-Blur both add random RGB noise to the images, which not only obscures identifying features but also introduces strong artifacts that do not allow classification models to generalize to unseen data. At $\epsilon = 0.1$, both methods achieve RIR values close to 1%, indicating near-perfect anonymization. DP-Blur shows slightly better utility preservation than DP-Pix at comparable privacy levels, particularly for gender classification (44.03% vs. 31.68% at $\epsilon = 0.1$). This occurs because the additional blurring step in DP-Blur creates a smoothing effect that mitigates some of the disruptive noise patterns introduced by the DP mechanism. However, at these high privacy levels, utility metrics for both methods approach random guessing, especially for race classification.

### 5.3.3 EDIALS results

Our EDIALS approach was evaluated using the methodology described in Section 5.2. For the neighborhood generation phase, we ensured adequate coverage of the latent space surrounding each image encoding by generating a balanced set of samples representing both positive and negative classifications. This strategy

TABLE 5.1: Performance metrics across different privacy-preserving transformations

| Technique | Parameter | Values in % | | | |
|---|---|---|---|---|---|
| | | RIR | $F_1$Age | $F_1$Gender | $F_1$Race |
| Original images | – | 100.00 | 93.09 | 98.88 | 94.65 |
| Pixelation | $n = 20$ | 84.00 | 96.95 | 97.43 | 87.33 |
| | $n = 35$ | 13.87 | 53.31 | 54.25 | 34.59 |
| | $n = 50$ | 3.01 | 47.35 | 49.17 | 18.25 |
| Blurring | $k = 50$ | 77.08 | 78.19 | 96.95 | 82.84 |
| | $k = 100$ | 13.02 | 63.30 | 74.71 | 29.67 |
| | $k = 200$ | 2.50 | 36.81 | 54.57 | 23.48 |
| Motion Blur | $k = 100$ | 61.74 | 75.85 | 95.60 | 79.74 |
| | $k = 200$ | 19.25 | 34.77 | 52.35 | 42.14 |
| | $k = 300$ | 7.95 | 30.29 | 41.67 | 26.11 |
| DP-Pix | $\epsilon = 1$ | 9.83 | 65.81 | 91.46 | 58.51 |
| | $\epsilon = 0.5$ | 6.34 | 52.93 | 44.94 | 18.16 |
| | $\epsilon = 0.1$ | 1.09 | 34.77 | 31.68 | 23.55 |
| DP-Blur | $\epsilon = 1$ | 24.03 | 68.97 | 91.47 | 49.17 |
| | $\epsilon = 0.5$ | 9.85 | 42.13 | 61.62 | 16.19 |
| | $\epsilon = 0.1$ | 1.01 | 29.12 | 44.03 | 13.69 |
| Mouth Masked | – | 96.12 | 90.11 | 95.28 | 91.65 |
| Eyes Masked | – | 30.21 | 80.13 | 85.81 | 61.25 |
| Eyes & Mouth Masked | – | 27.27 | 76.15 | 83.11 | 59.15 |

allowed our explainability component to identify decision boundaries with sufficient precision while maintaining computational efficiency. For the distortion component, we applied a balanced factor of $\alpha = 0.5$ to the identity-revealing features identified in the latent space (see Section 5.3.6 for additional results with other distortion factors). Because EDIALS requires both the AAE and

black-box models to remain on the GPU, our hardware resources only allowed us to run EDIALS with reasonable runtimes on images with a maximum resolution of 256×256 pixels. Therefore, MTF images were downscaled to that resolution (see Section 5.3.7 for an analysis of the influence of image resolution on performance and runtime).

TABLE 5.2: EDIALS performance (with $\alpha = 0.5$) vs. original images.

| **Metric** | **EDIALS (%)** | **Original images (%)** |
|---|---|---|
| RIR | 0.42 | 100.00 |
| $F_1$-Age | 84.66 | 93.09 |
| $F_1$-Gender | 97.91 | 98.88 |
| $F_1$-Race | 82.61 | 94.65 |

As shown in Table 5.2, EDIALS achieved the theoretically higher levels of anonymization with RIR of 0.42%, equivalent to random guessing. Simultaneously, it largely preserved utility across all classification tasks, with $F_1$ scores of 84.66% for age classification, 97.91% for gender classification, and 82.61% for race classification.

The effectiveness of EDIALS comes from its targeted distortion in the latent feature space. By operating in this transformed space rather than directly on pixel values, EDIALS can exploit the structural organization of facial attributes that occurs in well-trained autoencoders. Our explainability component identifies which specific latent dimensions most strongly influence identity recognition, allowing us to modify only those features while leaving other characteristics largely intact.

This selective modification is key to preserving utility. For instance, gender classification remains highly accurate (97.91%)

because the latent features that influence gender information appear to be sufficiently distinct from those that strongly influence identity recognition. The explainability-guided approach ensures these gender-defining features remain undisturbed during the anonymization process. Similarly, age-related and race-related features are substantially preserved, allowing for effective demographic analysis even after complete anonymization.

The ability to maintain high $F_1$ scores for race classification (82.61%) is particularly noteworthy, as this task typically requires the preservation of subtle facial features that conventional methods often destroy when achieving strong privacy protection. By targeting only identity-revealing components in the latent space, EDIALS preserves these critical features that contribute to race classification tasks, such as certain aspects of facial structure and skin tone.

These results demonstrate that the conventional trade-off between privacy and utility can be substantially mitigated through our targeted approach to feature distortion in the latent space.

### 5.3.4 Comparison with DeepPrivacy2

GAN-based approaches like DeepPrivacy2 represent the current state-of-the-art in image anonymization. We employed the implementation provided by the authors of DeepPrivacy2[1], which includes a pre-trained GAN model with 46 million parameters.

[1]Github URL of DeepPrivacy2: `https://github.com/hukkelas/deep_privacy2`

TABLE 5.3: Performance comparison between DeepPrivacy2 and EDIALS at comparable RIR levels

| **Method** | $F_1$ **Score (%)** | | | |
|---|---|---|---|---|
| | **Age** | **Gender** | **Race** | RIR (%) |
| DeepPrivacy2 | 78.32 | 82.58 | 73.32 | 16.27 |
| EDIALS | 89.32 | 97.98 | 87.14 | 15.68 |
| Original images | 93.09 | 98.88 | 94.65 | 100.00 |

Our evaluation revealed that DeepPrivacy2 achieved a RIR of 16.27%, indicating weaker than desired privacy protection. To enable a fair comparison at similar RIR levels, we configured EDIALS with $\alpha = 0.5$ and set an iteration limit of 60 (as 68 iterations achieved a RIR of 0.42%). With this configuration, EDIALS achieved a RIR of 15.68%, closely matching DeepPrivacy2's privacy level.

At these comparable privacy levels, EDIALS significantly outperforms DeepPrivacy2 across all utility metrics: age classification (89.32% vs. 78.32%), gender classification (97.98% vs. 82.58%), and race classification (87.14% vs. 73.32%). This performance advantage demonstrates that our targeted approach to identity disruption better preserves features essential for demographic analysis.

The synthetic nature of GAN-generated images introduces artifacts that impair classification models. While DeepPrivacy2's images may appear more detailed than conventional anonymization methods, they exhibit unnaturally synthetic characteristics that affect classification accuracy. Additionally, DeepPrivacy2's complete replacement of facial features can unintentionally alter

demographic characteristics essential for utility tasks. EDIALS avoids these issues through its more selective modification approach that targets identity-revealing features while minimizing distortion to characteristics important for demographic analysis.

### 5.3.5 Runtime analysis

Computational efficiency is an important consideration when processing large image collections. As detailed in Chapter 4, conventional anonymization methods vary in runtime from 0.04 seconds per image (for pixelation and blurring) to 2.31 seconds per image for DeepPrivacy2.

EDIALS incurs a higher processing time, as it takes 33.00 seconds (on average) to process an image. The time reported also includes the complete pipeline: AAE training, black-box model training, encoding, neighborhood generation, decision tree learning, feature distortion, and decoding. The variant that matches DeepPrivacy2's RIR requires 2 fewer seconds as it takes fewer iterations (60 vs. 68) due to fewer decode-evaluate cycles during incremental distortion.

The AAE and black-box models must both remain on the GPU for the technique to work properly, which precludes batch processing with the current hardware. This technical requirement is intrinsic to the operation of the method.

### 5.3.6 Impact of the distortion factor

This section analyzes how the distortion factor employed at each iteration ($\alpha$) affects computational efficiency and utility preservation in our approach. Table 5.4 presents experimental results for

different $\alpha$ values.

TABLE 5.4: Impact of distortion factor $\alpha$ on efficiency and utility preservation

| Distortion Factor $\alpha$ | Average Iterations | Theoretical Total Distortion* | $F_1$ Score (%) | | |
|---|---|---|---|---|---|
| | | | Age | Gender | Race |
| 1.0 | 36 | 3.60 | 82.22 | 95.09 | 80.23 |
| 0.7 | 51 | 3.57 | 83.15 | 96.22 | 81.32 |
| 0.6 | 59 | 3.54 | 83.89 | 97.05 | 81.86 |
| **0.5** | **68** | **3.40** | **84.66** | **97.91** | **82.61** |
| 0.2 | 168 | 3.36 | 85.78 | 97.91 | 83.75 |
| 0.1 | 328 | 3.28 | 85.80 | 97.92 | 83.76 |

*Theoretical Total Distortion calculated as $(iterations \times \alpha)/10$, where 10 is the average number of features.

All $\alpha$ values achieved anonymization at the level of random guessing (0.42% RIR). As expected, smaller $\alpha$ values required more iterations to reach the same privacy threshold, increasing from 36 iterations at $\alpha = 1.0$ to 328 iterations at $\alpha = 0.1$.

The results show that as $\alpha$ decreases from 1.0 to 0.1, utility improves across all tasks. Gender classification $F_1$ scores increase from 95.09% to 97.92%, age classification from 82.22% to 85.80%, and race classification from 80.23% to 83.76%. This improvement aligns with decreased total distortion, which drops from 3.60 at $\alpha = 1.0$ to 3.28 at $\alpha = 0.1$.

As introduced above, we used $\alpha = 0.5$ in the main experiments as it provides a good balance between utility and efficiency. Although utility barely increases for $\alpha < 0.5$, the computational cost significantly increases. At $\alpha = 0.5$, we achieve $F_1$ scores of 84.66% (age), 97.91% (gender), and 82.61% (race) with only 68 iterations, compared to 328 iterations at $\alpha = 0.1$ for marginally better scores. This makes $\alpha = 0.5$ the optimal choice for practical applications

where both utility preservation and computational efficiency are important.

### 5.3.7 Impact of image resolution

To evaluate EDIALS' scalability, we investigated the effect of reducing image resolution on utility preservation and computational efficiency. We re-trained our AAE at 128×128 and 64×64 resolutions and compared the results with the standard implementation employed so far, which uses 256×256 images.

TABLE 5.5: Impact of image resolution on runtime and utility at 0.42% RIR

| | | | $F_1$ Score (%) | | |
|---|---|---|---|---|---|
| **Resolution** | **Batch size** | **Runtime (s)** | **Age** | **Gender** | **Race** |
| 256×256 | 1 | 33.0 | 84.66 | 97.91 | 82.61 |
| 128×128 | 8 | 9.0 | 81.24 | 94.53 | 78.42 |
| 64×64 | 16 | 3.2 | 72.38 | 88.65 | 70.14 |

Reducing resolution offers significant computational benefits with varying impacts on utility preservation. Downsampling to 128×128 enabled batch processing of 8 images simultaneously, reducing processing time by 73% (from 33.0s to 9.0s per image). Further reduction to 64×64 allowed for batches of 16 images, decreasing processing time by 90% (to 3.2s per image). These improvements stem from the quadratic reduction in memory requirements, which enables more efficient parallel processing.

The impact on utility varies by classification task. Gender classification demonstrated remarkable resilience to resolution reduction, with only a minor degradation at 128×128 (3.38%) and moderate impact at 64×64 (9.26%). Age and race classifications showed

greater sensitivity, with modest degradation at 128×128 (3.42% and 4.19%, respectively) but more substantial declines at 64×64 (12.28% and 12.47%). As observed in Section 5.3.3, gender classification generally maintains better performance under image modifications because it relies on broader facial structure.

The decline in $F_1$ scores at lower resolutions reflects the compound effect of information loss in our pipeline. While high-frequency details persist at higher resolutions, they become increasingly compromised at lower resolutions (Patel et al., 2025). When we apply targeted distortion to these already resolution-compromised features, the effect particularly impacts tasks requiring fine detail discrimination, especially at 64×64 resolution, where age and race classification performance deteriorates significantly. These findings indicate that 128×128 resolution offers a good compromise between efficiency and utility preservation. When time constraints are not critical, the 256×256 resolution offers better utility preservation.

Importantly, when considering inference time alone, the 128×128 resolution version achieves processing times comparable to DeepPrivacy2 while maintaining better RIR values and utility metrics.

## 5.4 Conclusion

We have presented EDIALS, a novel approach to facial image anonymization that operates in the latent space using explainability techniques to selectively target identity-revealing features. Our empirical evaluation confirms that EDIALS achieves a superior balance between privacy and utility compared to conventional methods and GAN-based approaches, achieving anonymization

at the theoretical minimum of random guessing (RIR = 0.42%) while maintaining high utility for demographic classification tasks (scores $F_1$: 84.66% for age, 97.91% for gender and 82.61% for race).

# Chapter 6

# Conclusions and future work

This thesis has addressed the critical challenge of protecting the privacy of facial images through three novel approaches: PRISMA, EDI-Anon, and EDIALS. We have shown how AI techniques can be repurposed to provide robust privacy protection, using adversarial examples, explainability, and latent space manipulation to protect images against ML-based threats.

Each method successfully addresses the specific objectives outlined in Chapter 1: PRISMA tackles the social media sharing context by preventing unauthorized FR models from generalizing to clean images while maintaining visual quality for sharing; EDI-Anon and EDIALS address the data anonymization context by selectively targeting identity-revealing features while preserving utility for analytical tasks such as age, gender, and race classification.

Our comprehensive evaluation demonstrates that these AI-driven approaches achieve superior privacy-utility trade-offs compared

to conventional methods, establishing new benchmarks for facial privacy protection across both social and analytical contexts.

## 6.1 Contributions and publications

Before starting the actual thesis work, a detailed review of the security and privacy issues in machine learning was performed. This review covered attacks and defenses, with a special focus on federated learning.

This work was published as a journal paper:

- Younas Khan, David Sánchez, and Josep Domingo-Ferrer. "Federated learning-based natural language processing: a systematic literature review." *Artificial Intelligence Review*, vol. 57, art. no. 20, 2024. Impact factor: 10.7 (first decile).

Chapter 3 introduced PRISMA, a two-stage approach that addresses privacy challenges in social media image sharing. PRISMA identifies robust target features and uses momentum-based optimization to generate visually imperceptible perturbations. Its key innovation produces protected images that enable FR models to converge during training but fail to recognize clean images of the same individuals. PRISMA significantly reduces FR accuracy while maintaining high perceptual quality and demonstrates strong transferability across diverse model architectures.

This work has been submitted as a journal paper:

- Younas Khan, David Sánchez, and Josep Domingo-Ferrer. "Perceptually Robust Iterative Similarity Momentum Attack." *Engineering Applications of Artificial Intelligence* (submitted). Impact factor: 7.5 (first decile).

Chapter 4 presented EDI-Anon, a disclosure risk-aware approach enabling privacy-preserving sharing for research purposes. It leverages explainability techniques to identify identity-revealing pixels and applies incremental perturbation guided by continuous re-evaluation of re-identification risk. EDI-Anon significantly improves the privacy/utility trade-off compared to conventional methods.

This work is accepted for publication as a journal paper:

- Rami Haffar, David Sánchez, Younas Khan, and Josep Domingo-Ferrer. "Explainability-Driven Incremental Image Anonymization." *Transactions on Data Privacy* (to appear). Impact factor: 0.9 (third quartile).

Chapter 5 described EDIALS, which operates in the latent feature space to achieve better data anonymization. It employs specialized architectures and implements novel feature distortion mechanisms that selectively modify identity-revealing latent features. EDIALS achieves random guessing-level re-identification while maintaining remarkably high utility across all classification tasks.

This work has been submitted as a conference paper:

- Younas Khan, Anna Monreale, Carlo Metta, David Sánchez, and Josep Domingo-Ferrer. "Explainability-Driven Image Anonymization in Latent Space." *Knowledge Discovery and Data Mining* (submitted). CORE ranking: A.

## 6.2 Limitations and future work

While our methods provide significant advances, they can be improved through the following lines of future work.

- **PRISMA:** PRISMA's effectiveness can be compromised by background knowledge attacks, where adversaries have access to unprotected images of the individuals to be protected. Our empirical analysis showed that protection begins to deteriorate when an adversary has access to 15% or more unprotected images in their training data. This limitation affects all privacy protection methods and highlights the need for comprehensive deployment strategies. Additionally, PRISMA's architecture requires first training an FR model on the images to be protected before computing robust target features for generating adversarial perturbations, creating computational overhead that appears significant for single-image protection.

  Future work addressing these limitations includes:

  - **Extending to non-facial regions** by adapting PRISMA's two-stage architecture to identify and protect other biometric identifiers such as iris recognition, fingerprint matching, body posture, gait patterns, or clothing style that could enable identification even when faces are protected. This would involve training the FR model on full-body features and extending the momentum-based optimization to these regions.
  - **Adapting the architecture to handle multiple faces** by developing efficient processing pipelines that can simultaneously compute target features for multiple individuals in a single image while maintaining protection effectiveness for each person.

  - **Reducing computational overhead** by developing pre-trained models that can generate effective target features without requiring user-specific training, and implementing batch processing strategies that allow users to protect multiple images efficiently by training once and applying to many images. This addresses the computational overhead by amortizing the model training cost across multiple faces within the same image.

- **EDI-Anon:** EDI-Anon's modular framework enables integration of various perturbation techniques while maintaining targeted anonymization. Future enhancements will focus on expanding this flexibility:

  - **Integrating additional perturbation techniques** by extending the EDI-Anon to support motion blur, variable pixelation patterns, and other obfuscation methods that can be applied incrementally to identity-revealing pixels.
  - **Optimizing the iterative process** by developing adaptive algorithms that automatically adjust IDPS percentages and $\Delta$-parameters based on image characteristics and privacy requirements, reducing the need for manual parameter tuning.
  - **Evaluating emerging explainability techniques** by systematically testing new explainability methods as they become available to determine if they can provide more accurate identification of identity-revealing pixels, potentially improving EDI-Anon's targeting precision and

reducing the number of iterations required for effective anonymization.

- **Enhancing robustness evaluation** by testing EDI-Anon against sophisticated attacks and developing defensive mechanisms to counter potential reconstruction attempts.

- **EDIALS:** EDIALS' main limitation is its computational complexity, as it requires both the AAE and black-box models to remain on the GPU, precluding efficient batch processing at higher resolutions. When utility preservation is critical, this performance cost is justified, but can be reduced by processing at lower resolutions. Future work extending the method includes:

  - **Investigating disentangled representations** by developing autoencoder architectures that better separate identity features from demographic attributes in the latent space. Achieving better disentanglement would allow more precise targeting of identity-specific features while preserving demographic characteristics, potentially enabling processing at lower resolutions while maintaining utility due to cleaner feature separation.
  - **Developing theoretical privacy guarantees** by establishing formal bounds on the amount of identity information that can be recovered from anonymized latent representations. This would provide mathematical frameworks for quantifying privacy protection levels and establishing confidence intervals for anonymization effectiveness under different attack scenarios.

# Bibliography

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim (2018). "Sanity checks for saliency maps". In: *Advances in neural information processing systems* 31.

Ba, Jimmy Lei, Jamie Ryan Kiros, and Geoffrey E Hinton (2016). "Layer normalization". In: *arXiv preprint arXiv:1607.06450*.

Bakhshi, Saeideh, David A Shamma, and Eric Gilbert (2014). "Faces engage us: Photos with faces attract more likes and comments on instagram". In: *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 965–974.

Blanco-Justicia, Alberto, David Sánchez, Josep Domingo-Ferrer, and Krishnamurty Muralidhar (2023). "A Critical Review on the Use (and Misuse) of Differential Privacy in Machine Learning". In: *ACM Computing Surveys* 55.8, 160:1–160–16.

Bradski, Gary (2000). "The openCV library." In: *Dr. Dobb's Journal: Software Tools for the Professional Programmer* 25.11, pp. 120–123.

Chen, Jia-Wei, Li-Ju Chen, Chia-Mu Yu, and Chun-Shien Lu (2021). "Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6478–6487.

Cherepanova, Valeriia, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein (2021).

"Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition". In: *arXiv preprint arXiv:2101.07922*.

Deng, Jiankang, Jia Guo, Niannan Xue, and Stefanos Zafeiriou (2019). "Arcface: Additive angular margin loss for deep face recognition". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4690–4699.

Dong, Yinpeng, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li (2018). "Boosting adversarial attacks with momentum". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith (2006). "Calibrating noise to sensitivity in private data analysis". In: *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, pp. 265–284.

Egiazarian, Karen, Jaakko Astola, Nikolay Ponomarenko, Vladimir Lukin, Federica Battisti, and Marco Carli (2006). "New full-reference quality metrics based on HVS". In: *Proceedings of the second international workshop on video processing and quality metrics*. Vol. 4, p. 4.

Evtimov, Ivan, Pascal Sturmfels, and Tadayoshi Kohno (2020). "Foggysight: A scheme for facial lookup privacy". In: *arXiv preprint arXiv:2012.08588*.

Fan, Liyue (2018). "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*. Springer, pp. 148–162.

Fan, Liyue (2019). "Differential privacy for image publication". In: *Theory and Practice of Differential Privacy (TPDP) Workshop*. Vol. 1, p. 6.

Federal Trade Commission, M, Federal Trade Commission, et al. (1998). "Children's online privacy protection rule ("COPPA")". In: *Retrieved January* 19, p. 2018.

Fraser, Ian H, Greg L Craig, and Denis M Parker (1990). "Reaction time measures of feature saliency in schematic faces". In: *Perception* 19.5, pp. 661–673.

Gajic, Ana (2023). "Instagram Marketing Statistics". In: 99firms.com.

Ghorbani, Amirata, Abubakar Abid, and James Zou (2019). "Interpretation of neural networks is fragile". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 3681–3688.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2020). "Generative adversarial networks". In: *Communications of the ACM* 63.11, pp. 139–144.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy (2014). "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572*.

Guidotti, Riccardo, Anna Monreale, Stan Matwin, and Dino Pedreschi (2020). "Black box explanation by learning image exemplars in the latent feature space". In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I*. Springer, pp. 189–205.

Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville (2017). "Improved training of

wasserstein gans". In: *Advances in neural information processing systems* 30.

Guo, Yandong, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao (2016). "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, pp. 87–102.

Haffar, Rami, Najeeb Moharram Jebreel, Josep Domingo-Ferrer, and David Sánchez (2021). "Explaining image misclassification in deep learning via adversarial examples". In: *International Conference on Modeling Decisions for Artificial Intelligence*. Springer, pp. 323–334.

Haffar, Rami, David Sánchez, and Josep Domingo-Ferrer (2025). "Multi-Task Faces (MTF) Data Set: A Legally and Ethically Compliant Collection of Face Images for Various Classification Tasks". In: *IEEE Access* 13, pp. 63827–63840.

Haig, Nigel D (1986). "Exploring recognition with interchanged facial features". In: *Perception* 15.3, pp. 235–247.

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross Girshick (2017). "Mask r-cnn". In: *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969.

Hendrycks, Dan and Kevin Gimpel (2016). "Gaussian error linear units (gelus)". In: *arXiv preprint arXiv:1606.08415*.

Higgins, Irina, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner (2017). "beta-vae: Learning basic visual concepts with a constrained variational framework". In: *International Conference on Learning Representations*.

Howard, Andrew G (2017). "Mobilenets: Efficient convolutional neural networks for mobile vision applications". In: *arXiv preprint arXiv:1704.04861*.

Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger (2017). "Densely connected convolutional networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708.

Hukkelås, Håkon and Frank Lindseth (2023). "Deepprivacy2: Towards realistic full-body anonymization". In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1329–1338.

Hukkelås, Håkon, Rudolf Mester, and Frank Lindseth (2019). "Deepprivacy: A generative adversarial network for face anonymization". In: *International symposium on visual computing*. Springer, pp. 565–578.

Ilyas, Andrew, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry (2019). "Adversarial examples are not bugs, they are features". In: *Advances in neural information processing systems* 32.

Keeley, Brian and Céline Little (2017). *The State of the Worlds Children 2017: Children in a Digital World.* ERIC.

Khorzooghi, Seyyed Mohammad Sadegh Moosavi and Shirin Nilizadeh (2023). "Examining StyleGAN as a utility-preserving face de-identification method". In: *Proceedings on Privacy Enhancing Technologies*.

Kingma, Diederik P. and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9,*

*2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. URL: http://arxiv.org/abs/1412.6980.

Kosta, Eleni (2020). "Article 8 Conditions applicable to child's consent in relation to information society services". In: *The EU General Data Protection Regulation: A commentary*. Oxford University Press, pp. 355–364.

Kurakin, Alexey, Ian J Goodfellow, and Samy Bengio (2018). "Adversarial examples in the physical world". In: *Artificial intelligence safety and security*. Chapman and Hall/CRC, pp. 99–112.

Laugel, Thibault, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki (2018). "Defining locality for surrogates in post-hoc interpretablity". In: *arXiv preprint arXiv:1806.07498*.

Li, Jian, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Jilin Li, and Feiyue Huang (2019). "DSFD: dual shot face detector". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5060–5069.

Lim, Jae Hyun and Jong Chul Ye (2017). "Geometric gan". In: *arXiv preprint arXiv:1705.02894*.

Lingenfelter, Bryson, Sara R Davis, and Emily M Hand (2022). "A quantitative analysis of labeling issues in the celeba dataset". In: *International Symposium on Visual Computing*. Springer, pp. 129–141.

Liu, Terence (2023). *How We Store and Search 30 Billion Faces*. URL: https://www.clearview.ai/post/how-we-store-and-search-30-billion-faces.

Liu, Yanpei, Xinyun Chen, Chang Liu, and Dawn Song (2017). "Delving into transferable adversarial examples and black-box attacks". In: *International Conference on Learning Representations*.

Liu, Zhuang, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie (2022). "A ConvNet for the 2020s". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986.

Liu, Ziwei, Ping Luo, Xiaogang Wang, and Xiaoou Tang (2015). "Deep learning face attributes in the wild". In: *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738.

Makhzani, Alireza, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey (2015). "Adversarial autoencoders". In: *arXiv preprint arXiv:1511.05644*.

Maximov, Maxim, Ismail Elezi, and Laura Leal-Taixé (2020). "Ciagan: Conditional identity anonymization generative adversarial networks". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5447–5456.

Mienye, Ibomoiye Domor and Theo G Swart (2025). "Deep autoencoder neural networks: A comprehensive review and new perspectives". In: *Archives of computational methods in engineering*, pp. 1–20.

Mądry, Aleksander, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu (2017). "Towards deep learning models resistant to adversarial attacks". In: *stat* 1050.9.

Neustaedter, Carman, Saul Greenberg, and Michael Boyle (2006). "Blur filtration fails to preserve privacy for home-based video conferencing". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 13.1, pp. 1–36.

Neverova, Natalia, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi (2020). "Continuous

surface embeddings". In: *Advances in Neural Information Processing Systems* 33, pp. 17258–17270.

Newton, Elaine M, Latanya Sweeney, and Bradley Malin (2005). "Preserving privacy by de-identifying face images". In: *IEEE transactions on Knowledge and Data Engineering* 17.2, pp. 232–243.

Nominet, UK (2016). *Parents "Oversharing" Family Photos Online, But Lack Basic Privacy Know-How*.

Omeiza, Daniel, Skyler Speakman, Celia Cintas, and Komminist Weldermariam (2019). "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models". In: *arXiv preprint arXiv:1908.01224*.

Patel, Ashay, Michela Antonelli, Sebastien Ourselin, and M Jorge Cardoso (2025). "Resolution Invariant Autoencoder". In: *arXiv preprint arXiv:2503.09828*.

Ponomarenko, Nikolay, Vladimir Lukin, Jaakko Astola, and Karen Egiazarian (2015). "Analysis of HVS-metrics' properties using color image database TID2013". In: *Advanced Concepts for Intelligent Vision Systems: 16th International Conference, ACIVS 2015, Catania, Italy, October 26-29, 2015. Proceedings 16*. Springer, pp. 613–624.

Ponomarenko, Nikolay, Flavia Silvestri, Karen Egiazarian, Marco Carli, Jaakko Astola, and Vladimir Lukin (2007). "On between-coefficient contrast masking of DCT basis functions". In: *Proceedings of the third international workshop on video processing and quality metrics*. Vol. 4. Scottsdale USA.

Radford, Alec, Luke Metz, and Soumith Chintala (2016). "Unsupervised representation learning with deep convolutional generative adversarial networks". In: *International Conference on Learning Representations*.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin (2016). ""' Why should i trust you?" Explaining the predictions of any classifier". In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144.

Rijsbergen, C. J. van (1979). *Information Retrieval*. 2nd ed. London: Butterworths. ISBN: 0-408-70929-4.

Royer, Jessica, Caroline Blais, Isabelle Charbonneau, Karine Déry, Jessica Tardif, Brad Duchaine, Frédéric Gosselin, and Daniel Fiset (2018). "Greater reliance on the eye region predicts better face recognition ability". In: *Cognition* 181, pp. 12–20.

Sadr, Javid, Izzat Jarudi, and Pawan Sinha (2003). "The role of eyebrows in face recognition". In: *Perception* 32.3, pp. 285–293.

Santana, Modesto Castrillón, Oscar Déniz-Suárez, Luis Antón-Canalís, and Javier Lorenzo-Navarro (2008). "Face and Facial Feature Detection Evaluation - Performance Evaluation of Public Domain Haar Detectors for Face and Facial Feature Detection". In: *VISAPP 2008: Proceedings of the Third International Conference on Computer Vision Theory and Applications, Funchal, Madeira, Portugal, January 22-25, 2008 - Volume 2*. Ed. by Alpesh Ranchordas and Helder Araújo. Setúbal, Portugal: INSTICC - Institute for Systems, Technologies of Information, Control, and Communication, pp. 167–172.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). "FaceNet: A unified embedding for face recognition and clustering". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Schuler, Christian J, Michael Hirsch, Stefan Harmeling, and Bernhard Schölkopf (2015). "Learning to deblur". In: *IEEE transactions on pattern analysis and machine intelligence* 38.7, pp. 1439–1451.

Selvaraju, Ramprasaath R, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra (2017). "Grad-cam: Visual explanations from deep networks via gradient-based localization". In: *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Shan, Shawn, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao (2020). "Fawkes: Protecting privacy against unauthorized deep learning models". In: *29th USENIX security symposium (USENIX Security 20)*, pp. 1589–1604.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman (2014). "Deep inside convolutional networks: Visualising image classification models and saliency maps". In: *Workshop at International Conference on Learning Representations*.

Simonyan, Karen and Andrew Zisserman (2014). "Very deep convolutional networks for large-scale image recognition". In: *arXiv preprint arXiv:1409.1556*.

Sinha, Pawan, Benjamin Balas, Yuri Ostrovsky, and Richard Russell (2006). "Face recognition by humans: Nineteen results all computer vision researchers should know about". In: *Proceedings of the IEEE* 94.11, pp. 1948–1962.

Smilkov, Daniel, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg (2017). "SmoothGrad: removing noise by adding noise". In: *Workshop on Visualization for Deep Learning, ICML*.

Sundararajan, Mukund, Ankur Taly, and Qiqi Yan (2017). "Axiomatic attribution for deep networks". In: *International Conference on Machine Learning*. PMLR, pp. 3319–3328.

Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi (2017). "Inception-v4, inception-resnet and the impact of residual connections on learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus (2013). "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199*. DOI: arXiv:1312.6199. URL: https://doi.org/10.48550/arXiv.1312.6199.

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf (2014). "Deepface: Closing the gap to human-level performance in face verification". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1701–1708.

Tan, Mingxing and Quoc Le (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *International conference on machine learning*. PMLR, pp. 6105–6114.

Tramèr, Florian, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel (2017). "The space of transferable adversarial examples". In: *arXiv preprint arXiv:1704.03453*.

Viola, Paul and Michael Jones (2001). "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the*

*2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*. Vol. 1. Ieee, pp. I–I.

Wang, Zhou, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli (2004). "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4, pp. 600–612.

Wright, Mike and Natasha Bernal (2018). *Digital footprint starts in the womb, Children's Commissioner warns*. URL: https://www.telegraph.co.uk/technology/2018/11/08/emb-thurs-children-have-1000-pictures-posted-online-turn-13/.

Yuan, Lin, Linguo Liu, Xiao Pu, Zhao Li, Hongbo Li, and Xinbo Gao (2022). "PRO-Face: A Generic Framework for Privacy-preserving Recognizable Obfuscation of Face Images". In: *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1661–1669.

Zhang, Hongyang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan (2019). "Theoretically principled trade-off between robustness and accuracy". In: *International conference on machine learning*. PMLR, pp. 7472–7482.

Zhang, Qi (2022). "A novel ResNet101 model based on dense dilated convolution for image classification". In: *SN Applied Sciences* 4, pp. 1–13.

Zhang, Richard, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang (2018). "The unreasonable effectiveness of deep features as a perceptual metric". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595.

Zhang, Zhilu and Mert Sabuncu (2018). "Generalized cross entropy loss for training deep neural networks with noisy labels". In: *Advances in neural information processing systems* 31.

Zhao, Qiang Alex and John T Stasko (1998). "Evaluating image filtering based techniques in media space applications". In: *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pp. 11–18.

Zhuchkov, Aleksei (2021). "Analyzing the effectiveness of image augmentations for face recognition from limited data". In: *2021 International Conference" Nonlinearity, Information and Robotics"(NIR)*. IEEE, pp. 1–6.



UNIVERSITAT ROVIRA i VIRGILI