LICENTIATE THESIS

Anonymizing Faces without Destroying Information

Felix Rosberg

Halmstad University | School of Information Technology
LicentiateThesis | Halmstad University Dissertations no. 111

Anonymizing Faces without Destroying Information

Felix Rosberg

LICENTIATE THESIS | Halmstad University Dissertations no. 111

Anonymizing Faces without Destroying Information

Halmstad University Dissertations no. 111
ISBN 978-91-89587-36-6 (printed)
ISBN 978-91-89587-35-9 (pdf)
Publisher: Halmstad University Press, 2024 | www.hh.se/hup
Printer: Media-Tryck, Lund

Abstract

Anonymization is a broad term. Meaning that personal data, or rather data that identifies a person, is redacted or obscured. In the context of video and image data, the most palpable information is the face. Faces barely change compared to other aspect of a person, such as cloths, and we as people already have a strong sense of recognizing faces. Computers are also adroit at recognizing faces, with facial recognition models being exceptionally powerful at identifying and comparing faces. Therefore it is generally considered important to obscure the faces in video and image when aiming for keeping it anonymized. Traditionally this is simply done through blurring or masking. But this destroys useful information such as eye gaze, pose, expression and the fact that it is a face. This is an especial issue, as today our society is data-driven in many aspects. One obvious such aspect is autonomous driving and driver monitoring, where necessary algorithms such as object-detectors rely on deep learning to function. Due to the data hunger of deep learning in conjunction with society's call for privacy and integrity through regulations such as the General Data Protection Regularization (GDPR), anonymization that preserve useful information becomes important.

This Thesis investigates the potential and possible limitation of anonymizing faces without destroying the aforementioned useful information. The base approach to achieve this is through face swapping and face manipulation, where the current research focus on changing the face (or identity) while keeping the original attribute information. All while being incorporated and consistent in an image and/or video. Specifically, will this Thesis demonstrate how target-oriented and subject-agnostic face swapping methodologies can be utilized for realistic anonymization that preserves attributes. Thru this, this Thesis points out several approaches that is: 1) controllable, meaning the proposed models do not naively changes the identity. Meaning that what kind of change of identity and magnitude is adjustable, thus also tunable to guarantee anonymization. 2) subject-agnostic, meaning that the models can handle any identity. 3) fast, meaning that the models is able to run efficiently. Thus having the potential of running in real-time. The end product consist of an anonymizer that achieved state-of-the-art performance on identity transfer, pose retention and expression retention while providing a realism.

Apart of identity manipulation, the Thesis demonstrate potential security is-

sues. Specifically reconstruction attacks, where a bad-actor model learns convolutional traces/patterns in the anonymized images in such a way that it is able to completely reconstruct the original identity. The bad-actor networks is able to do this with simple black-box access of the anonymization model by constructing a pair-wise dataset of unanonymized and anonymized faces. To alleviate this issue, different defense measures that disrupts the traces in the anonymized image was investigated. The main take away from this, is that naively using what qualitatively looks convincing of hiding an identity is not necessary the case at all. Making robust quantitative evaluations important.

Acknowledgements

First of, I would like to express my gratitude to my principal supervisor Cristofer Englund for or providing the opportunity of my Ph.D study position in the industry. Cristofer has always made sure my progress been steady, whether that be navigating the academic bureaucracy, identifying and ensuring me of possible next paths or simply providing me with encouraging words. Sincerely thank you, your support have been invaluable.

I would like to thank my co-supervisor Eren Erdal Aksoy for providing excellent technical feedback and discussions, and for his contributions in rewriting and proof reading our papers. It has improved both in structuring my experiments and in my dissemination, whether it being through text or figures.

I also would like to thank my co-supervisor Fernando Alonso-Fernandez for our discussions and your provided perspective and knowledge in biometrics.

My sincere thanks to Engage Studios, Vinnova and Smart Industry Sweden for financing my research and my colleagues for their support. The majority of my colleagues does not work with the kind of subjects I tackled throughout, but they have always showed interest in my work and how it works.

A special thanks to Halmstad University for their provided resources, support and for providing means for me, an industrial Ph.D student, to engage with the people of the academic side. I also want to extend my thanks to my fellow Ph.D student colleagues at Halmstad University for our discussions and your welcoming nature, thank you Kevin Hernández Diaz, Kunru Chen, Anna Vettoruzzo and Tiago Fernandes Cortinhal.

I also would like to thank Martin Torstensson, my fellow industrial Ph.D student at RISE for our collaborations and discussions.

I am profoundly grateful to my family and friends for their support and continuous expressed interest of my progress. Finally I would like to express my sincerest thanks to my fiancee, Sara for her undying support and her cheering for my success! Sara you are my absolute anchor.

List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: **Towards Privacy Aware Data collection in Traffic: A Proposed Method for Measuring Facial Anonymity [1]**
Felix Rosberg, Cristofer Englund, Martin Torstensson, Boris Durán. **In Future Active Safety Technology: Towards zero traffic accidents**, (2021).

PAPER II: **Comparing Facial Expressions for Face Swapping Evaluation with Supervised Contrastive Representation Learning [2]**
Felix Rosberg, Cristofer Englund. **In 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), pages 01–05**, (2021).

PAPER III: **FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping [3]**
Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, Cristofer Englund. **In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 3454–3463**, (2023).

PAPER IV: **FIVA: Facial Image and Video Anonymization and Anonymization Defense [4]**
Felix Rosberg, Eren Erdal Aksoy, Cristofer Englund, Fernando Alonso-Fernandez. **In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, pages 362-371**, (2023).

Contents

Abstract **i**

Acknowledgements **iii**

List of Papers **v**

Abbreviations **ix**

List of Figures **xi**

List of Tables **xiii**

1 Introduction **1**
1.1 Motivation 1
1.2 Research Questions 2
1.2.1 RQ1: What measures can be employed to ensure effective anonymization of personal details within video frames and how can their performance be evaluated? . 3
1.2.2 RQ 2: What approaches can be employed to ensure the generation of unique and distinct faces in consecutive instances when capturing the same person in different video sequences? And, RQ 3: What techniques can be employed to ensure consistent anonymization of faces across frames within a video sequence? 3
1.2.3 RQ 4: What types of attacks pose vulnerabilities to anonymization methods, and what are the effective mitigation strategies to counteract them? 4
1.3 Structure of this Thesis 4

2 Facial Recognition **7**
2.1 Face Detection and Alignment 7
2.2 Embedding Faces 10

2.3 Evaluation 11

3 Face Swapping **15**
3.1 Source-Oriented 16
3.2 Target-Oriented 16

4 Privacy-Aware Machine Learning **19**
4.1 Federated Learning 19
4.2 Manipulating Identity 20

5 Summary of Papers **23**
5.1 PAPER I: Towards Privacy Aware Data collection in Traffic: A Proposed Method for Measuring Facial Anonymity 23
5.1.1 Summary and Purpose 23
5.1.2 Results and Contribution to the Licentiate Thesis . . . 24
5.1.3 Summary of Contributions 26
5.2 PAPER II: Comparing Facial Expressions for Face Swapping Evaluation with Supervised Contrastive Representation Learning 26
5.2.1 Summary and Purpose 26
5.2.2 Results and Contribution to the Licentiate Thesis . . . 27
5.2.3 Summary of Contributions 30
5.3 PAPER III: FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping 31
5.3.1 Summary and Purpose 31
5.3.2 Results and Contribution to the Licentiate Thesis . . . 33
5.3.3 Summary of Contributions 37
5.4 PAPER IV: Facial Image and Video Anonymization and Anonymization Defense 38
5.4.1 Summary and Purpose 38
5.4.2 Results and Contribution to the Licentiate Thesis . . . 39
5.4.3 Summary of Contributions 46

6 Conclusions **47**

References **45**

Appendix **51**
A PAPER I 51
B PAPER II 58
C PAPER III 64
D PAPER IV 79

Abbreviations

¬ID	Refers to the negated identity retrieval metric, a ratio of retrieving the correct identity in the data set after one of the embeddings has been multiplied by -1
3DMM	3D Morphable Model
c2s	Change to Source (Refers to comparing a changed face to a source face, essential same as s2c)
c2t	Change to Target (Refers to comparing a changed face to a target face, essential same as t2c)
s2c	Source to Change (Refers to comparing a source face to a changed face, e.g. a face swap)
s2t	Source to Target (Refers to comparing a source face to a target face)
t2c	Target to Change (Refers to comparing a target face to a changed face, e.g. a face swap)
AD	Autonomous Driving
ADAS	Advanced Driver-Assistance Systems
AFFA	Adaptive Feature Fusion Attention
CCPA	California Consumer Privacy Act
CSL	Cybersecurity Law of the People's Republic of China
EER	Equal Error Rate
FAR	False Acceptance Rate
FID	Fréchet Inception Distance
FPN	Feature Pyramid Network
FQR	Future Research Question
FRR	False Rejection Rate
GAN	Generative Adversarial Network

GDPR	General Data Protection Regulation
ID	Refers to the identity retrieval metric, a ratio of retrieving the correct identity in the data set
IFSR	Interpreted Feature Similarity Regularization
IJB-C	IARPA Janus Benchmark-C
L2	Euclidean norm or distance
LFW	Labeled Faces in the Wild
MLP	Multi-Layered Perceptron
RA	Reconstruction Attack, refers to the reconstruction attack identity retrieval metric, a ratio of retrieving the correct identity in the data set after a reconstruction attack
RQ	Research Question
TAR	True Acceptance Rate
TTC	Time To Collision

List of Figures

1.1 Publication timeline, which research questions (RQ) they address and planned future research questions (FRQ). 2

2.1 Illustration of face detection and alignment. 8
2.2 Illustration of inverse affine transformation. 10
2.3 Cosine distance distribution of random negative samples (Two different identities) and random positive samples (Two of the same identities). The blue line indicates equal error rate. . . . 12
3.1 Demonstration of face swapping and its naming conventions (Images corresponds to output from FaceDancer [3], PAPER III in this Thesis). 15

5.1 Illustration of full-body facial anonymization image used in the survey from PAPER I. Left two images are without facial manipulation, while the right two are with. In this case FS-GAN [5] was used to change the faces. 25
5.2 Normalized confusion matrix results on the AffectNet validation dataset. 28
5.3 Face swapping results generated by FaceDancer. 31
5.4 Overview of the architecture and training procedure of FaceDancer. For more in depth details and description of components, please refer to PAPER III and its appendix. The appendix also covers the main differences between different baselines and ablation derivatives of FaceDancer. 32
5.5 Comparing FaceDancer with SimSwap [6], FaceShifter [7], HifiFace [8], and FaceController [9]. 34
5.6 Qualitative comparison on low resolution images. 34
5.7 Illustration of the impact of IFSR. Config A given in the 3rd column here shows results once IFSR is omitted during training. 36

5.8 Cosine similarity between intermediate features between changed and target faces (c2t), changed and source faces (c2s), and different identities (Negative Samples). (a) Distances between features from first block of ArcFace. (b) Distances between features from final block of ArcFace. (c) Equal error rates (EER) between the distance distributions for intermediate features in every block. 38
5.9 Overview of the proposed anonymization pipeline and an illustration of the implication of reconstruction attacks. 40
5.10 Qualitative results of reconstruction attack, different defenses and anonymization using FIVA. 43
5.11 Qualitative comparison between CIAGAN [10], CFA-NET [11], Gafni et al. [12], DeepPrivacy [13] and FIVA. 43
5.12 Qualitative temporal comparison between CIAGAN [10], DeepPrivacy [13] and FIVA. Note we used ITM for tracking the identity for CIAGAN. 44
5.13 Qualitative comparison between FIVA and FaceDancer for gender and ethnicity retention. 44
5.14 Illustration of matching a desired anchor. The red lines illustrates matches to a desired anchor based on desired approximate distance from the target vector (black line). The green line illustrates the match that would occur when sampling for FIVA. Blue circle illustrates the desired distance. 45

List of Tables

5.1 Accuracy on AffectNet validation data set for 8 emotions. . . . 29

5.2 Comparison between our approach and 2D landmark approach for comparing expression for face swaps. Left to right column: Method, euclidean distance error between target face to source face and target face to changed face, mean euclidean distance for target face to changed face and ratio of $t2c < s2c$. 30

5.3 Quantitative experiments on FaceForensics++ [14]. See PAPER III for further details about configurations. 33

5.4 Ablative analysis together with the runtime performance. Inference time is given in millisecond and memory usage in GB. All models in this table were trained for 300k iterations. 35

5.5 Ablative analysis using AFLW2000-3D [15] as target and FaceForensics++ [14] as source. 36

5.6 Quantitative experiments on FaceForensics++ [14]. Evaluated with identity retrieval (ID), negated identity retrieval (¬ID, searching for a match with $-z_{id}$), and reconstruction attack (RA) identity retrieval. Temporal identity consistency $\mathcal{M}_{tc}$ calculated using 10 frames per video. The divide in the table separates inpainting-based methods from target-oriented ones. The $\times$ indicates that RA is not applicable to the corresponding method. +Sampling means we used the anchor sampling method to assign anonymized identities (See PAPER IV for details), while +ITM indicates both the anchor sampling and tracking (See PAPER IV and Figure 5.9). The $\downarrow$ indicates lower is better. 41

5.7 Quantitative identity retrieval experiments on LFW [16]. CFA-Net [11] and Gafni et al. [12] demonstrate the true positive rate for a false acceptance rate of 0.001 using FaceNet [17] as the facial recognition model. We evaluate the remaining methods with CosFace and a threshold of 0.63 (Cosine *distance*), for a false acceptance rate of 0.001. The $\downarrow$ indicates lower is better. . 42

5.8 Defense against reconstruction attack in FIVA, evaluated on FaceForensics++ [14]. Adversarial Defense in the form of a fast sign gradient method. Noise Defense just adds regular uniform noise to the image. Parameter Noise means adding a small Gaussian noise to the parameters. We report the fraction of successful retrievals of the original identity after applying the reconstruction attack. ε highlights how much the noise was scaled. The $\downarrow$ indicates lower is better. Black-box means it does not need access to the reconstruction attack model. . . 42

1. Introduction

1.1 Motivation

Machine learning and deep learning are areas of research and application that are widely used to solve complex problems in domains such as natural language and computer vision. In the context of computer vision, one such problem is object detection, which enables a wide range of technologies. Autonomous driving (AD) and advanced driver-assistance systems (ADAS) is an emerging technology that relies in many cases on deep learning, and object detection is an obvious example of this. Although deep learning can achieve impressive results in complex tasks, they are limited by their data hungriness [18; 19]. With recent progress with transformers, they have also been shown to push performance further as data access and parameters scale up [20]. It becomes clear in current machine learning research that data access and data *collection* is both important and necessary to build high performing models. However, this naturally clashes with integrity issues derived from collecting a massive amount of data. Arguably, the most important indicator of this concern is the General Data Protection Regulation (GDPR) [21].

GDPR and other data regulation laws such as Cybersecurity Law of the People's Republic of China (CSL) [22] and the California Consumer Privacy Act (CCPA) [23], highlight the importance of data privacy around the world. In short, when data collection is done, the data usages become limited unless all personal connections are either removed and/or anonymized. There are several situations in which people's faces are an involved point of personal information that is easily recognized by facial recognition [24–29]. The area of autonomous driving and traffic safety is an excellent example in which detection and avoidance of people plays a crucial role. Thus, the data in the image space naturally include faces. Some of the applications done to remove the personal data of faces is blurring, black boxing or pixelizing. The draw-back with this is that it directly destroys important information such as eye gaze, pose, facial expression, and the fact that it is a face. Furthermore, it risks affecting the data distribution in such a way that it deteriorates performance. Ren et al. studied the influence of pedestrian eye contact on drivers [30]. It has an important im-

pact on the driver's behavior in such a way that it can significantly increase the time to collision (TTC). This suggests that the face contains important potential markers that could improve traffic safety. Driver monitoring systems that analyze the face of the driver are another usage of facial information for traffic safety [31]. Especially considering that approximately 1.3 million people die each year from traffic injuries and between 20 and 50 million people suffer non-fatal injuries [32].

This Thesis addresses the capabilities of facial anonymization in such a way that we keep important information such as eye gaze and facial expression, but replaces the identity. Throughout this work, attribute information will represent all of the aforementioned information, while identity information will represent the identity.

1.2 Research Questions

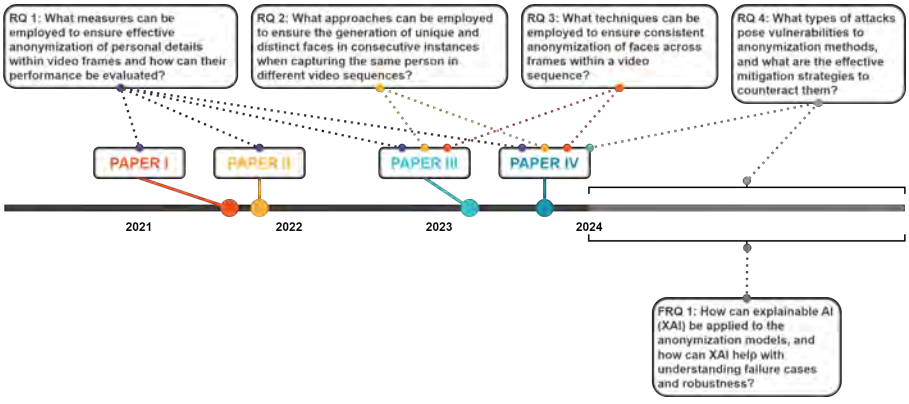

Figure 1.1: Publication timeline, which research questions (RQ) they address and planned future research questions (FRQ).

The overarching research questions that this Thesis aims to answer is as follows:

- What measures can be employed to ensure effective anonymization of personal details within video frames and how can their performance be evaluated?
- What approaches can be employed to ensure the generation of unique and distinct faces in consecutive instances when capturing the same person in different video sequences?

- What techniques can be employed to ensure consistent anonymization of faces across frames within a video sequence?
- What types of attacks pose vulnerabilities to anonymization methods, and what are the effective mitigation strategies to counteract them?

The timeline of the presented papers and their related research questions are illustrated in Figure 1.1. This figure also shows future research questions that are relevant as of the writing of this Thesis.

1.2.1 RQ1: What measures can be employed to ensure effective anonymization of personal details within video frames and how can their performance be evaluated?

The formulation of this question revolves around the identification of necessary tools and methods for achieving successful anonymization, incorporating both qualitative and quantitative measures of performance. To provide clarity, when multiple faces are anonymized within a frame, it is essential to ascertain that the anonymization process does not inadvertently leak personal information. Additionally, during the implementation of the anonymization method, it becomes imperative to evaluate its effectiveness.

This line of inquiry is of considerable significance as it encompasses several critical aspects that require evaluation. These aspects include performance in identity manipulation, retention of attributes, temporal consistency, potential leakage of identities, and more. As depicted in Figure 1.1 of the Thesis timeline, each presented paper within this Thesis addresses these aspects either directly or indirectly.

1.2.2 RQ 2: What approaches can be employed to ensure the generation of unique and distinct faces in consecutive instances when capturing the same person in different video sequences? And, RQ 3: What techniques can be employed to ensure consistent anonymization of faces across frames within a video sequence?

In the context of anonymization, various scenarios need to be carefully considered to ensure realistic outcomes. Let us envision a video sequence featuring a bustling environment with seven to twelve individuals moving around. These individuals frequently cross paths, occasionally exiting the frame only to reappear later. So, how can we effectively handle such a scenario while maintaining a plausible sense of realism in the anonymization process?

Several factors come into play in this regard. Firstly, it is crucial to ensure that a specific person within the video sequence consistently bears the same fabricated identity throughout. Introducing a new fake face for this individual in each frame would not make sense. Secondly, let us think on the situation where we encounter the same person for a second time a week later. Should we provide them with the same fake face as before, or should we assign them a new identity? The challenge lies in facilitating this functionality while maintaining control over the anonymization process.

Striking the right balance between realism and anonymity calls for creative solutions and meticulous engineering. By addressing these intricacies, we can achieve both a robust approach to anonymization, empowering us to navigate complex video scenarios while preserving the integrity of individuals' identities.

1.2.3 RQ 4: What types of attacks pose vulnerabilities to anonymization methods, and what are the effective mitigation strategies to counteract them?

While evaluating anonymization models based on predefined metrics provides valuable insights into their performance, relying on these metrics alone may not be sufficient. It is imperative to proactively search for, evaluate, and mitigate potential vulnerabilities within the models. Drawing inspiration from research conducted in the realm of federated learning [33–37], we find compelling parallels that highlight the need for vulnerability analysis.

One notable vulnerability explored in federated learning is reconstruction attacks [34; 35], where bad actors intercept communication lines to reconstruct sensitive data. Similar security concerns are pertinent to anonymization models, as elucidated in PAPER IV. Consequently, we must contend with the unknown unknowns, which entails addressing edge cases that could potentially compromise both integrity and privacy.

Navigating these intricacies requires a holistic approach that blends scientific rigor with a touch of creativity. By embracing the challenge of identifying and mitigating vulnerabilities, we ensure the robustness and efficacy of anonymization methods.

1.3 Structure of this Thesis

This chapter aims to provide a concise overview of the Thesis structure, facilitating your navigation through the information presented.

The subsequent three chapters delve into background areas of research. In the Facial Recognition chapter (Chapter 2), we unveil the foundations of face analysis and manipulation. The Thesis covers interplay between theory and practice, accompanied by an exploration of related works within the literature. This chapter also serves as a repository of general methodologies, offering valuable insights into how we work with faces in the context of facial anonymization research.

Moving forward, the Face Swapping chapter (Chapter 3) invites you to delve into the gist of facial manipulation. Here, the Thesis provides an overview of intricacies of seamlessly replacing one face with another. Throughout this chapter, we briefly cover relevant literature, allowing us to contextualize our work within the broader research landscape of anonymization.

Lastly, the Privacy-Aware Machine Learning chapter (Chapter 4) presents literature about preserving privacy and identity concealment in the context of both faces and general machine learning. The Thesis dive into the challenges of preserving privacy. This chapter serves as a valuable resource for understanding the methodologies employed and briefly covering previous scholarly works.

As we progress to the chapter that follows—Summary of Papers (Chapter 5). Here, The Thesis provide a more focused chapter, presenting specific summaries and results from our research. You will find concise summaries of each paper, accompanied by an overview of the addressed problems, contributions and results. This includes the highlighting figures and tables from each paper. You will also find a brief summary of contribution for each paper.

Finally, you can find the summarized conclusion of this Thesis work in Chapter 6. Get ready to explore the landscape of facial analysis, manipulation, and privacy preservation.

2. Facial Recognition

2.1 Face Detection and Alignment

Before working with and making in-depth analysis of faces in both research and in-practice deployment, it is necessary to detect faces and align them. It is rarely the case that faces appear perfectly in frame. Face detection, which can be seen as a task-specific object detection, is a long standing research topic. The task is considered challenging due to illumination, various poses, occlusions and varied scales [38; 39]. Recent work deal with this through feature pyramid networks (FPN) [40], multi-task training and anchor boxes [41–45]. FPN deals with several scales by aggregating features maps from several resolutions produced by the backbone. Example of backbones being ResNet50 [46], EfficientNet [47] or ConvNeXt [48] etc. Multi-task learning extends the detection objective of box classification and regression to include prediction of more tasks such as landmark regression, projected 3D face vertices and intersection-over-union prediction, and improves the performance of the face detector [41; 42]. Anchor boxes is what allow face detectors to deal with several faces in the frame robustly. The idea is to generate several (thousands) pre-determined anchor boxes across each scale produced by the FPN in different sizes. Predictions is then outputted for each anchor boxes. For classification, the prediction for each anchor box is if there exist a face or not at that anchor. For bounding-box regression, the detector regresses the offset from the anchor-box, yielding a more refined position than just the anchor box itself. For landmark regression, the detector regresses the points of the landmarks. At least one anchor is generated per pixel for each scale. As an example, a feature map of shape $B \times C \times 160 \times 160$, where B and C is the batch size and number of channels respectively, and where 160×160 is the resolution. If we choose to generate 2 anchor boxes per pixel, we would get $160 * 160 * 2 = 51200$ anchor boxes.

One state-of-the-art face detector, RetinaFace [41], provides a 5-point landmark prediction consisting of left eye, right eye, nose, left-part of the mouth and the right-part of the mouth (Figure 2.1). This is especially useful, because

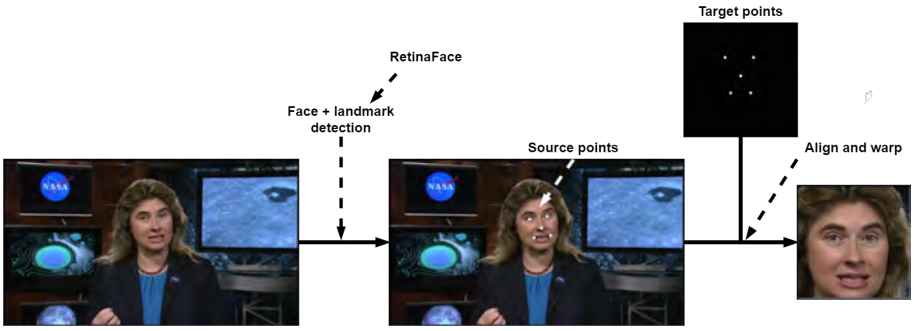

Figure 2.1: Illustration of face detection and alignment.

to strengthen performance of models that analyses the face one should align the face. Figure 2.1 demonstrates how RetinaFace is used to detect landmark *source points*, how these points is used to to produce an affine transformation that aligns the face according to the *target points*. The affine transformation matrix parameters is found by calculating the parameters in such way that achieves the least mean squared error between the *source points* and the *target points* [49]. Let the *target points* be a matrix X of shape (5, 2) and let the *source points* be a matrix Y of the same shape. Then we can calculate the transformation parameters of those points in the following steps from [49]:

$$\mu_x = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{2.1}$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} ||X_i - \mu_x||^2 \tag{2.2}$$

where μ and σ^2 is calculated along the first axis, yielding a vector of means and variances. Then we calculate the covariance matrix $\mathcal{A}$ of *target points* X and *source points* Y as

$$\mathcal{A} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu_y)(X_i - \mu_x)^T \tag{2.3}$$

where the mean for both X and Y are calculated along the first axis as in Equation 2.1. Next we establish a vector d as

$$d = \begin{cases} [1,1], & \text{if } det(\mathcal{A}) \geq 0 \\ [1,-1], & \text{if } det(\mathcal{A}) < 0 \end{cases} \tag{2.4}$$

where the resulting vector depend on the determinant of $\mathcal{A}$. Then we factorize the covariance matrix $\mathcal{A}$ as

$$U, S, V = svd(\mathcal{A}) \tag{2.5}$$

where $svd()$ is the singular value decomposition function. Now we can do the final calculations to construct the transformation matrix T. First we estimate the scaling factor c as

$$c = \frac{S \cdot d}{\sum \sigma_x^2}. \tag{2.6}$$

Following equation

$$R = U \cdot diag(d) \cdot V \tag{2.7}$$

establishes the matrix R of shape (2, 2) and where $diag(d)$ operation generates a zero matrix with d along the diagonal. Next equation

$$K = (\mu_y - c * R \cdot \mu_x^T) \tag{2.8}$$

establishes the vector K of shape (2,). Using R, K and c, we construct the affine transformation matrix T as

$$T = \begin{bmatrix} R_{0,0} * c & R_{0,1} * c & K_0 \\ R_{1,0} * c & R_{1,1} * c & K_1 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.9}$$

that allow for *scale*, *rotation* and *translation* of the image. The detected face is warped as illustrated in Figure 2.1 using this transformation matrix T. In practice (and in this work), only the two first rows of T is used as

$$M = \begin{bmatrix} R_{0,0} * c & R_{0,1} * c & K_0 \\ R_{1,0} * c & R_{1,1} * c & K_1 \end{bmatrix} \tag{2.10}$$

denoted as M. For simplification and clarity we can describe all these steps as

$$M = estimate\_tm(Y, X) \tag{2.11}$$

This alignment operation, also sometimes called normalization, allow for better facial manipulation and analysis. Facial recognition models such as ArcFace [24] uses this as preprocessing before embedding the identity. Recent face swapping models such as SimSwap [6] and work presented in this Thesis (PAPER III [3] and PAPER IV) deploy this alignment as well. The inverse operation for blending the manipulated image with the frame is trivial. Given *source points* Y we define a matrix Y_l as

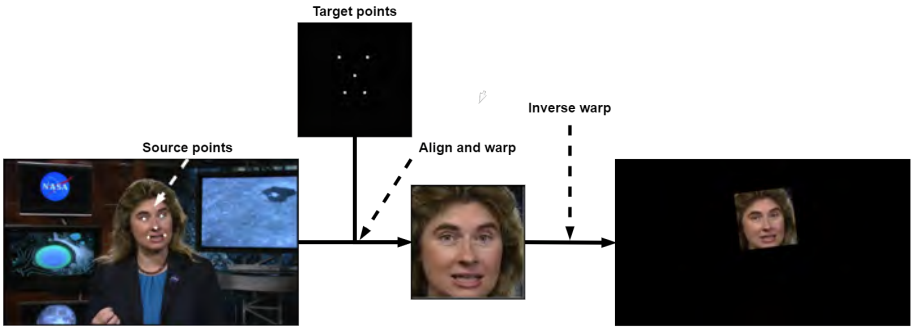

Figure 2.2: Illustration of inverse affine transformation.

$$Y_l = \begin{bmatrix} X_{0,0} & X_{0,1} & 1 \\ X_{1,0} & X_{1,1} & 1 \\ X_{2,0} & X_{2,1} & 1 \\ X_{3,0} & X_{3,1} & 1 \\ X_{4,0} & X_{4,1} & 1 \end{bmatrix} \tag{2.12}$$

which is the Y matrix with a inserted column of ones. The *inverse points* Z is calculated as

$$Z = Y_l \cdot M^T \tag{2.13}$$

where M is the transformation matrix from Equation 2.10. Then the same calculation described in Equation 2.3 through Equation 2.10 is done, except Y is treated as the *new target points* and Z is treated as the *new source points*. Using operation from Equation 2.11 as follows

$$M_i = estimate\_tm(Z,Y) \tag{2.14}$$

gives us a new transformation matrix M_i that let us invert the alignment. This operation warps the aligned face back to its original position in the frame (Figure 2.2). This is particular useful for facial anonymization, as it allow for a process that focus on one face at a time and utilizes the performance boost that is provided with alignment [50]. It also allows for working with faces in the same way that facial recognition models expects, making identity embedding in conjunction with face manipulation simple and straight forward.

2.2 Embedding Faces

Modern facial recognition models aims to represent a face image in an embedding space $z_{id} \in \mathcal{R}^D$, where D is the dimensionality of the embedding vector.

The common practice is to use deep learning models such as convolutional neural networks or vision transformers to embed images to the embedding space $\mathcal{R}^D$, which is connected to a classification head for classifying the identity [24–27]. Or using contrastive methods such as triplet-loss [17] to minimize distance between embeddings of the same identity and maximize distance between embeddings of different identity. Usually this is done in such a way so the geometrical representation of z_{id} is cleverly restricted. FaceNet [17], which is trained with the aforementioned triplet-loss, does this by normalizing the embeddings by the L2-norm (restricting the embeddings to a unit-sphere). State-of-the-art models such as ArcFace [24] and CosFace [25], which are trained through classification, uses clever geometrical regularization- and loss functions to produce strong representative embeddings. Resulting embeddings are not necessarily restricted to a unit-sphere, but they are specifically designed to lend itself well when normalized with the L2-norm. Thus the common practice for actual face comparison and verification is done using the cosine distance (or cosine similarity). The equation for cosine similarity is as follow,

$$d_s(u,v) = \frac{u \cdot v}{||u||_2||v||_2} \tag{2.15}$$

where, $u \in \mathcal{R}^D$ and $v \in \mathcal{R}^D$ is two vectors. $d_s(.,.)$ results in a scalar which is bound between -1 and 1 which equals to $cos(\theta)$, where θ is the angle between u and v. -1 indicates dissimilarity and 1 indicates complete similarity. Conversion to cosine distance is straight forward using:

$$d_d(u,v) = 1 - d_s(u,v) \tag{2.16}$$

resulting in a scalar bound between 0 and 2 instead. Other distance metrics such the L2-distance can be used as well. Recent publication and work all focus on the cosine distance as metric [24–29; 51]. As of today, the performance on facial recognition reaches around 98% to 99% accuracy for recognition and verification on several benchmark datasets such as MegaFace [52], Labeled Faces in the Wild (LFW) [16] and IARPA Janus Benchmark-C (IJB-C) [53]. Therefore the representation and embeddings of identity information is exceptionally powerful. Facial recognition models is not only useful for verification, but as discussed in Chapter 3 an integral part for face manipulation, face anonymization and evaluating the ideas and work presented in this Thesis. This is further detailed in PAPER III and PAPER IV.

2.3 Evaluation

In practice, facial recognition models do not make classification of an identity directly. As mentioned in the section above (Chapter 2.2), identification

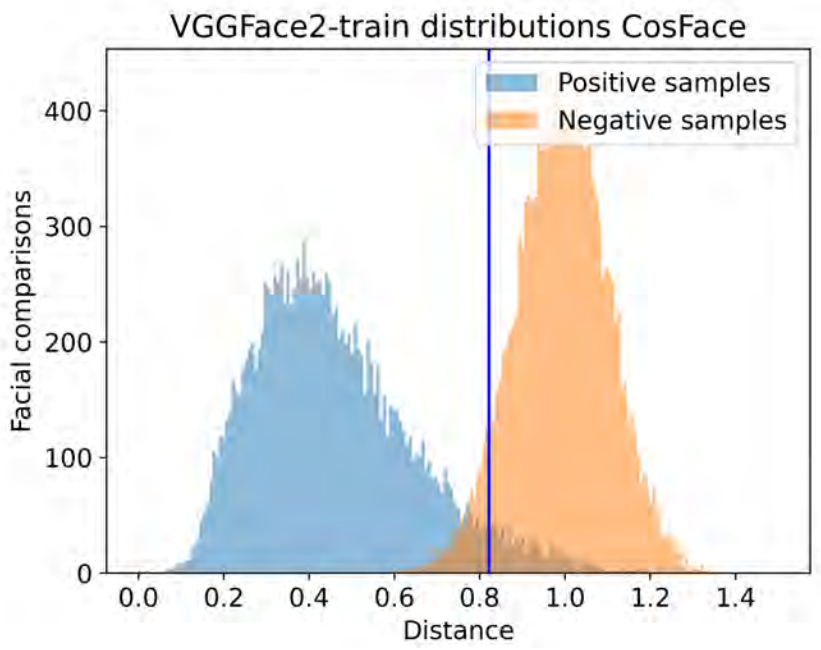

Figure 2.3: Cosine distance distribution of random negative samples (Two different identities) and random positive samples (Two of the same identities). The blue line indicates equal error rate.

and verification is done by comparing an unknown face with a database of known faces using a distance metric. We consider a match between a known identity and an unknown face when the distance is lower than a predetermined threshold. To determine a threshold, we specify an allowed false acceptance rate (FAR). Facial recognition models is usually evaluated and compared using several FAR values such as 0.0001 and 0.00001, with 0.00001 being the common baseline [24; 26; 29; 51; 54]. For clarification, during evaluation we report the true acceptance rate (TAR) for a specific FAR.

Figure 2.3 demonstrate two distributions of the cosine distance when comparing two different identities (Negative samples) and two of the same identities (Positive samples) using CosFace [25] as facial recognition model and the *train* data of the VGGFace2 dataset [55]. The figure also demonstrates the equal error rate (EER) as the vertical blue line, which is the threshold where false rejection rate (FRR) equals to FAR. Determining the EER for two overlapping distributions is simple. Algorithm 1 describes how to find the EER and the threshold for said EER. *eer_idx* corresponds to the index of all the

Algorithm 1: Algorithm for finding the equal error rate and the equal error rate threshold between a set of imposter pair distances and a set of genuine pair distances.

EER n_dis, p_dis, $start\_th$, $stop\_th$, $step\_size$;
Input : Distances between imposter pairs n_dis, distances between genuine pairs p_dis, start threshold $start\_th$, stop threshold $stop\_th$, step size for adjusting the threshold $step\_size$
Output: Equal error rate eer and equal error rate threshold eer_th
th = $start\_th$
far_curve = []
frr_curve = []
th_curve = []
while th <= $stop\_th$ **do**
    far = sum(where(n_dis < th, 1, 0)) / sum(ones_like(n_dis))
    frr = sum(where(p_dis >= th, 1, 0)) / sum(ones_like(p_dis))
    far_curve.append(far)
    frr_curve.append(frr)
    th_curve.append(th)
    th += $step\_size$
end
eer_idx = argmin(abs(far_curve - frr_curve))
eer = (far_curve[eer_idx] + frr_curve[eer_idx]) / 2
eer_th = th_curve[eer_idx]
return eer, eer_th;

thresholds tested that represent the EER threshold. *eer* is simply the EER.

The threshold for EER is usually not desirable, as most biometric system demands a low FAR to avoid unauthorized access or verification. Algorithm 2 describes how we can search for a threshold that satisfies a specific FAR. It starts from a initial guess threshold *init_th* and in small steps lowers it by *step_size* for each iteration. Once the desired FAR is estimated, it terminates and return the corresponding threshold. Now that the threshold for a specified FAR is found, we can start evaluating both facial recognition models and anonymization models identity retrieval performance. Face swapping models falls under this category as well, but the identity retrieval evaluation does not require a threshold as face swapping tries to convince that the transfer identity is after the face swap most similar to the correct identity [3; 6–8]. For anonymization, the threshold is necessary because the closest identity may be the original identity, but the distance is still large. For example a face that

is anonymized matches with its original identity. However, if for example the threshold for a FAR equal to 0.001 is 0.74, but the distance between the anonymized face and the match is 0.83. In this case we can not consider this as an actual match for the specified FAR of 0.001. For further details regarding evaluation of anonymization methods, see Chapter 5.4 and PAPER IV, which includes description of other relevant metrics. For further clarification in difference of evaluation between face swapping and anonymization, see Chapter 5.3 and PAPER III.

Algorithm 2: Algorithm for finding the threshold for a desired false acceptance rate.

FIND FAR n_dis, $init\_th$, $step\_size$, far_t;
Input : Distances between imposter pairs n_dis, initial threshold $init\_th$, step size for adjusting the threshold $step\_size$, target FAR we want to find a threshold for far_t
Output: Threshold th for a desired FAR of far_t
$th = init\_th$
$far = 1$
while $far > far_t$ **do**
    far = sum(where($n\_dis < th$, 1, 0)) / sum(ones_like(n_dis))
    th -= $step\_size$
end
return th;

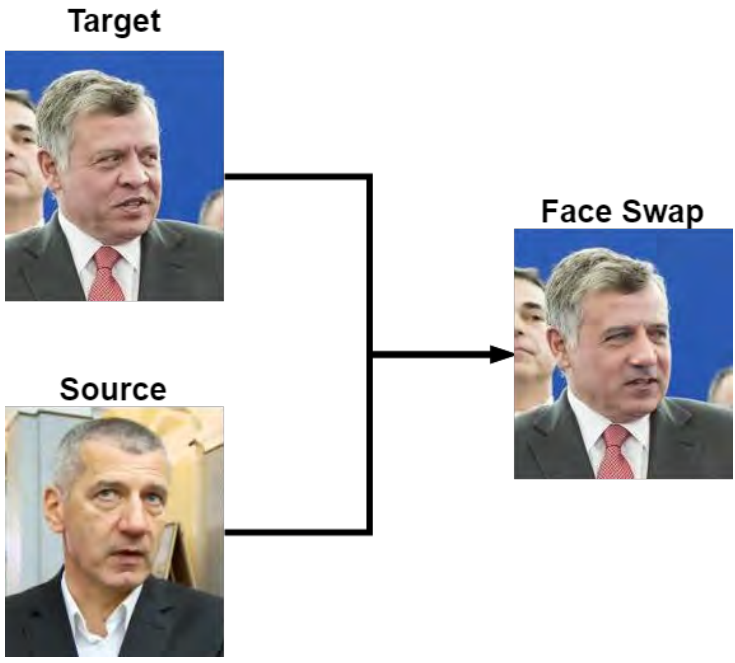

Figure 3.1: Demonstration of face swapping and its naming conventions (Images corresponds to output from FaceDancer [3], PAPER III in this Thesis).

3. Face Swapping

There exist two mainstream approaches for face synthesis-based face swapping: *source-oriented* and *target-oriented* methods. By target we mean the face that is to be manipulated, and by source we mean the face whose identity is being imposed in the target image (See Figure 3.1). This chapter will briefly cover the two approaches and how they differ. The Thesis will focus on, and cover target-oriented methods in more detail. The reason being that most recent work focuses on target-oriented methods, including the ones investigated and developed within this Thesis. Furthermore, source-oriented approach has a couple of issues, which is described below.

3.1 Source-Oriented

Source-oriented initially synthesize or manipulate the source face to match the attributes captured on the target face, followed by a blending step that replaces the target face with the source face [5; 56–58]. These approaches tend to struggle with lighting, occlusion, and complexity. One of the earliest approaches is The Digital Emily project [56], which performs face swapping with expensive and time-consuming 3D scanning of a single actor. Getting a face ready for manipulation in this way takes months. Banz et al. [59] adopted 3D morphable models (3DMM) [60] to generate source faces with target attributes matching. One drawback in their work is that the subject's hair must be carefully marked. Nirken et al. [57] also utilized 3DMM to extract attribute information, which is then used to reconstruct the source face with these attributes. The blending is then performed in combination with a face segmentation network. This method struggled with textures and lighting conditions. FSGAN and FSGANv2 [5; 58] introduced a reenactment network, designed to reenact the source face based of the target face's 68 point landmarks. The blending is further automated with a segmentation network and an inpainting network. Similar to the previous mentioned method, FSGAN and FSGANv2 struggles with lightning conditions. More importantly, due to relying on the target landmarks for reenactment, the reenacted source falls short in having effective identity transfer. In the context of facial anonymization, using source faces based of target landmarks risks leaking identity information. Another issue in this regard, with source-oriented approach, is the need to generate and track fake identities in image space. Which is severely costly in comparison with target-oriented approaches.

3.2 Target-Oriented

The second approach for face swapping, target-oriented, directly convert the identity of the target face into the source's identity [6–9; 52; 61]. These methods rely on Generative Adversarial Networks (GAN) in a one-stage optimization setting. This helps preserve attribute information such as pose and lighting, without requiring any additional processing step. For example, by learning perceptual and deep features directly in the training stage [6; 7; 62–65]. To clarify, target-oriented methods utilizes generative models to manipulate features of an encoded target face in conjunction with semi-supervised loss function or a regularization method to preserve attributes while shifting the identity. Most of these methods, including those introduced in PAPER III and PAPER IV, utilize facial recognition models to extract identity information. Details on this can be found above in Chapter 2.2. The extracted identity information

is generally used to condition the target-oriented network and to calculate an identity loss.

FaceShifter [7] maintains attributes with strong identity transfer by having an attribute encoder-decoder model, which is trained semi-supervised. The encoder-decoder is coupled with a generator that is conditioned on the source identity information and adaptively learns to gate between identity conditioned feature maps and attribute conditioned feature maps. FaceShifter also introduced a secondary stage for occlusion error correction. SimSwap [6] uses an encoder-decoder that condition the high level features in the bottleneck on the source identity information. For attribute retention, SimSwap uses a modified version of the feature matching loss from pix2pixHD [65]. SimSwap achieved state-of-the-art performance on pose retention with an arguably large trade-off in identity transferability. HifiFace [8] utilizes 3DMMs to achieve state-of-the-art identity transfer and shape performance. Although HifiFace produces high resolution photo-realistic face swaps and is not only conditioned on identity vectors, but also the 3DMM coefficients, it seems not to improve the pose considerably and performs worse than SimSwap in this regard. The identity performance quantitatively surpassed by work presented in this Thesis through work in PAPER III and PAPER IV. The Thesis will cover PAPER III and PAPER IV, their results and attribute retention in Chapter 5.3 and 5.4. You can also find mentions and comparisons of other work other than the ones mentioned above.

4. Privacy-Aware Machine Learning

There are two primary methodologies in privacy-aware machine learning that have garnered significant attention: federated learning and direct manipulation of identity information within the data. The research in this Thesis centers around the latter approach, which can be further categorized into identity masking methods and identity manipulation methods. However, there are lessons to be learned and inspired by in the subject of federated learning. Therefore, we will briefly give an overview of federated learning's role in privacy-aware machine learning, and how its problems can be translated into facial manipulation. Face-swapping techniques have emerged as a particularly promising avenue for achieving facial anonymization in recent years, serving as the bedrock for this Thesis.

4.1 Federated Learning

Federated learning approaches offer a privacy-preserving paradigm by leveraging data from multiple partners without directly sharing it [33–35]. This collaborative framework typically involves each partner training a globally shared model that receives updates from local participants [36]. Consequently, a collective model is generated, trained on diverse data sources while ensuring privacy awareness by avoiding direct data sharing. However, like any approach, it has its limitations and vulnerabilities that warrant consideration.

Firstly, the sharing of a global model among participants exposes it to reconstruction attacks, wherein unauthorized entities intercept communication channels to reconstruct private data [34; 35]. In an eye-opening study, Wang et al. [34] successfully demonstrated the reconstruction of faces that appeared eerily similar to the original ones. This underscores the need for robust security measures in both the federated learning ecosystem and facial manipulation-based anonymization.

Secondly, it is important to acknowledge that this methodology does not ad-

dress the underlying data collection process itself. It assumes that the utilized data already contains identity information, which may introduce biases or privacy concerns. Hence, it is vital to consider the broader context and implications of data collection while implementing federated learning approaches.

It is worth noting that these challenges not only impact data privacy but also limit the viewing, demonstration, and remote work capabilities associated with the collected data. Addressing these issues requires a delicate balance between safeguarding privacy and enabling efficient utilization of data in various applications. So, there are arguments for focusing on identity replacement based methods such as face manipulation.

4.2 Manipulating Identity

The manipulation of identity involves concealing or altering identity information to protect privacy. In the domain of image and video faces, traditional masking techniques like blurring or applying black boxes have been commonly used. While these methods effectively safeguard privacy, they often eliminate valuable details, such as eye gaze, which may impact the underlying data distribution. Naively training models on such distorted data can lead to model dependence on the introduced distortions. To address this challenge, researchers have turned to the possibilities offered by generative models to replace identities with realistic faces [4; 11–13; 66–70].

Several works, including those by Ma et al. [11], Li et al. [68], Li and Han [69], and Ren et al. [70], have explored direct face modification techniques. However, the current evaluation methodologies vary significantly, and none of these works specifically focus on realism within a spatio-temporal context. Gafni et al. [12] introduce a face modification autoencoder network with a strong emphasis on spatio-temporal consistency. Their approach ensures consistent operations across frames, generating a learned occlusion-aware mask. However, there is a clear lacking of automatic in-the-wild anonymization. Specifically, their method does not allow for automatic control of the anonymization, assuming either manual specifying which treatment each face gets. Furthermore, a quantitative evaluation of temporal consistency is not evaluated.

DeepPrivacy and DeepPrivacy2 [13; 67] employ a U-net-based model trained to inpaint removed faces, conditioned on pose information to maintain pose consistency. However, complete removal of the face results in the loss of crucial information, including facial expressions and eye gaze. Moreover, temporal consistency is disregarded, leading to the generation of new faces for

frames that exhibit minor differences.

The work of Çiftçi et al. [66], who employ the face swapping model SimSwap [6], utilize a gender and ethnicity-based analysis, to make sure that they sample fake identities that match those attributes. To clarify why, face-swapping models usually transfers the source gender and ethnicity. Later on in Chapter 5, this Thesis will elaborate on vulnerabilities in target-oriented face swapping models such as SimSwap. Meaning that the naivety briefly mentioned in Chapter 1.2.3 is already present in the current literature and research.

5. Summary of Papers

5.1 PAPER I: Towards Privacy Aware Data collection in Traffic: A Proposed Method for Measuring Facial Anonymity

5.1.1 Summary and Purpose

I want to start of by noting that this was an early investigation into measuring anonymity. Some insights has been made redundant by later studies and experiments that show stronger robustness (E.g. see PAPER IV). But I do want to underscore there are still important pieces of information to be considered in the paper. In this research, we investigate approaches for measuring anonymity in facial recognition systems. Traditionally, facial recognition involves identifying a face based on its similarity to known identities. However, we invert this task by considering a larger distance between embeddings as indicating a higher degree of anonymity. To protect identities, we anonymize faces and then measure the distance between the original and anonymized versions. If the distance is significantly large, we discard the original image and retain the anonymized one. This is as of the writing of this Thesis, for lacking of a better term, made redundant by current comparison and evaluation protocols for facial recognition model.

To perform face extraction and alignment, we explore two methods: MTCNN [38] and a five-point similarity transformation approach (See Chapter 2). MTCNN offers automatic cropping and alignment, while the five-point approach calculates transformation parameters based on target coordinates. This ensures the anonymized face closely matches the desired specifications.

For identity swaps, we utilize open-source code for FSGAN [5] and an implementation of the FaceShifter [7] model trained on the FFHQ dataset [71]. To demonstrate the effectiveness of our approach, we conduct various experiments. Initially, we compare the distance distributions between faces of the same identity and different identities using different distance metrics (L1, L2, squared L2, and cosine similarity). We leverage the VGGFace2 dataset [55], which contains diverse poses, lighting conditions, ages, and samples for each

identity. By analyzing the intersection between the distributions, we determine the optimal identity encoder and discriminating distance metric. This also helps us establish an acceptable threshold for identity anonymization.

In addition to computational analysis, we compare our results with a human survey. The survey includes images of well-known celebrities with anonymized faces, and participants are asked to recognize the individuals, provide their guesses, indicate their certainty, and identify distinguishing attributes. We divide the survey into sections, including full-body images (See Figure 5.1) with anonymized faces, close-up facial images, unaltered full-body images, and unaltered close-up facial images. This comparison helps us understand the importance of facial anonymity relative to other identifying attributes. We further compare survey results with the distances between celebrities' faces and their anonymized counterparts using ArcFace [24].

By combining computational analysis and human feedback, our results offers valuable insights into the measurement of facial anonymity and its implications for facial recognition systems.

5.1.2 Results and Contribution to the Licentiate Thesis

In this study, we examine the performance of two facial recognition models, ArcFace [24] and FaceNet [17], using 30,000 negative and positive image pairs. The results show that FaceNet exhibits a better intersection area between positive and negative sample distributions when using the cosine similarity metric. However, ArcFace allows for a higher threshold and provides a more reliable method for securing a sufficient distance after anonymization. We recommend using ArcFace with RetinaFace [43] alignment for optimal anonymization determination.

To assess the effectiveness of our proposed metric, we compare the distance distributions between target identities, source identities, and anonymized targets. The results indicate that using FaceShifter for anonymization maintains a good identity transfer, although random identity swaps can result in distances below the mentioned threshold. The slight leftward shift in the anonymized distribution suggests some data leakage or the preservation of background information by FaceShifter. To add retroactively to the discussion, it is more likely that a random swap could choose a similar face.

We also conducted a survey involving 14 participants to understand human

Figure 5.1: Illustration of full-body facial anonymization image used in the survey from PAPER I. Left two images are without facial manipulation, while the right two are with. In this case FSGAN [5] was used to change the faces.

identity recognition behavior. The recognition ratios for full body images (See Figure 5.1) and facial images were 35.7% and 40% respectively, with varying levels of certainty. The true positive rates for identifying the person were high, indicating a reasonable capability to recognize celebrities. The average recognition rates for selected celebrities were 70% for full body images and 61.4% for facial images, suggesting a potentially higher recognition rate for anonymized identities.

Furthermore, we compared the cosine similarity between anonymized celebrity faces and their unaltered counterparts. Using the ArcFace + RetinaFace embedding and alignment, the average distance exceeded the threshold for determining identity. However, the minimum distance fell below the threshold, indicating some leakage of identifying information. The limited sample size of 15 image pairs restricts the generalizability of these findings. Similar measurements using FaceShifter yielded a mean distance of 0.95, suggesting instances of low distances during random identity swaps.

Overall, our study provides valuable insights into facial recognition models, anonymization recognition, human identity recognition behavior, and the effectiveness of distance measurements in comparing anonymized faces. In conclusion, our study evaluate an approach for evaluating anonymization through the inversion of facial recognition tasks. By considering further advancements

and incorporating additional attributes and survey comparisons, we can gain deeper insights into the effectiveness of anonymization procedures and their alignment with human perception of recognizability. We also recognize the bias introduced when presenting the survey participants with celebrities, who are common public figures with distinct styles.

5.1.3 Summary of Contributions

In terms of this Thesis, the main contribution of PAPER I is the insight to important identifying aspects of people outside of the actual face. Our human survey showed that other parts such as the hair or eyes. Meaning that we need to be mindful of how people actually identify other people in practice, instead of just changing the face. This survey did have an introduced bias of informing the participants that the images were of celebrities. This would limit the true reflection of people being able to recognizing individuals based of other attributes than the face, as one can argue that the 'search' space becomes severally constrained to the celebrities.

5.2 PAPER II: Comparing Facial Expressions for Face Swapping Evaluation with Supervised Contrastive Representation Learning

5.2.1 Summary and Purpose

Privacy-aware data collection in traffic safety has become a burgeoning field of research, particularly in the task of anonymizing image data while retaining important information. In the context of traffic video data collection, where numerous individuals are present in each frame, safeguarding data security necessitates concealing identities while preserving facial expressions and eye gaze. This approach ensures the maintenance of realistic behaviors among road users even after anonymization. To deploy, enhance, or introduce new methods for achieving this objective, evaluating the anonymization process becomes crucial, encompassing aspects such as the effectiveness of identity obfuscation, preservation of eye gaze, and facial expression fidelity. In this study, we focus on two primary tasks: (a) representing facial expressions and (b) quantifying the preservation of facial expressions after the anonymization process. To accomplish this, we aim to develop robust facial expression embeddings that can be utilized to calculate distances between expression embeddings.

The use of contrastive loss has demonstrated considerable potential in extracting informative embeddings [72]. We leverage supervised contrastive loss to

address class imbalance in facial expression recognition. This loss function has also exhibited the ability to handle classes with subtle differences, which is particularly relevant in our case, as all images are facial images with minor variations in the main facial region.

For training and evaluation, we employed the AffectNet dataset [73], which comprises an extensive collection of 287,651 cropped and aligned facial images. The dataset includes annotations for eight distinct emotions, as well as valence and arousal values. Valence and arousal provide continuous representations of emotions using two dimensions. Valence describes the comfort associated with an emotion, where lower values represent emotions such as anger and disgust, while higher values correspond to happiness. Arousal signifies the intensity of an emotion, with lower values denoting emotions like sadness and calmness, and higher values indicating anger and excitement. The dataset encompasses discrete emotions, including neutral, happy, sad, surprise, fear, disgust, anger, and contempt. In total, the dataset comprises 440,000 facial images. For training identity swaps, we utilized the FFHQ dataset [71] and employed the FaceShifter method as our chosen approach.

5.2.2 Results and Contribution to the Licentiate Thesis

To investigate the representational power of supervised contrastive representation learning as a pretraining method, we present the top achieved accuracy obtained by training with weighted cross-entropy using a classification head instead of a projection head. The results are summarized in Table 5.1, where we achieve a remarkable accuracy of 59.58% by employing a frozen EfficientNetB0 encoder pretrained with weighted contrastive loss. Notably, EfficientNetB0 outperforms the ResNet50V2 baseline, improving performance by 1.07 percentage units. I refer to PAPER II for implementation and loss details ($\mathcal{L}_{wtot}$, $\mathcal{L}_{tot}$, $\mathcal{L}_{wc}$).

Comparing our approach to recent works that do not employ additional training data beyond AffectNet, our method exhibits a slight improvement in performance, as depicted in Table 5.1. We also attempted to train the classification network end-to-end using the same configuration but adjusting the learning rate. However, the results were unsatisfactory, with the accuracy reaching only 15.15%. Additionally, in the *same* configuration as EfficientNetB0 + $\mathcal{L}_{wtot}$ (Table 5.1), we experimented with semi-supervised contrastive learning using siamese representation learning (SimSiam) [80] and a simple autoencoder [81] built upon the original encoder. Unfortunately, both methods yielded poor performance and failed to learn meaningful representations. We suspect that siamese representation learning's reliance on heavy augmen-

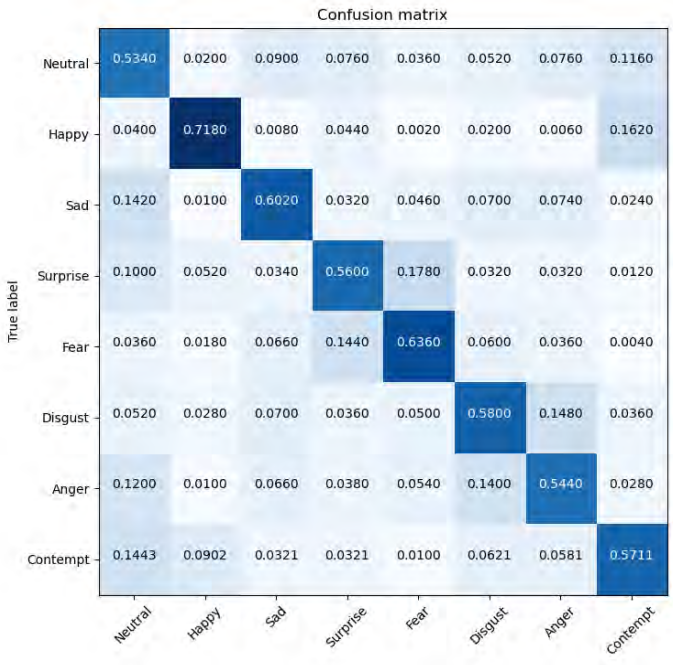

Figure 5.2: Normalized confusion matrix results on the AffectNet validation dataset.

tation, such as random crop, zoom, and translation, hinders its effectiveness. Aligned images are generally preferred for facial recognition and facial expression recognition, and heavy augmentation may compromise the alignment preprocessing.

In the ablation study, we conducted several steps: downgrading the backbone from EfficientNetB0 to ResNetV2, removing the weighting of the contrastive loss, and finally eliminating the multi-task prediction head for arousal and valence. The resulting accuracy are displayed in Table 5.1. Collectively, these additional components boosted the performance from 48.96% to 59.58%, with the multi-task component providing the most significant improvement.

Figure 5.2 presents a confusion matrix depicting the classification performance among different classes. Shi et al. [77] also reported a confusion matrix in their work. While they achieved an overall better accuracy by utilizing additional training data from the RAF-DB dataset [82], the per-class accuracy in their results appears to be more varied. Notably, our approach maintains a 57.11% accuracy for the contempt expression, while Shi et al. achieved only 39.00%

Table 5.1: Accuracy on AffectNet validation data set for 8 emotions.

Methods	Accuracy	Extra data
Schoneveld et. al (Multimodial) [74]	**61.60%**	*yes*
Savchenko et. al (Multi-task) [75]	61.32%	*yes*
Vo et. al (PSR) [76]	60.68%	*yes*
Shi et. al (ARM) [77]	59.75%	*yes*
Wang et. al (RAN) [78]	59.50%	*yes*
Ours (EfficientNetB0) + $\mathcal{L}_{wtot}$	**59.58%**	*no*
Ours (ResNet50V2) + $\mathcal{L}_{wtot}$	58.51%	*no*
Ours (ResNet50V2) + $\mathcal{L}_{tot}$	57.76%	*no*
Ours (ResNet50V2) + $\mathcal{L}_{wc}$	48.96%	*no*
Siqueira et. al [79]	59.30%	*no*
Mollahosseini et. al [73]	58.00%	*no*
End-to-end classification (EfficientNetB0)*	15.15%	*no*
SimSiam (EfficientNetB0)*	12.50%	*no*
Autoencoder (EfficientNetB0)*	12.50%	*no*

* Trained with the same configuration and hyper-parameters as our best method.

accuracy for the same class. It is worth mentioning that contempt is often regarded as a challenging expression to analyze and is frequently excluded from assessments.

As suggested, the representation network can be employed to evaluate the extent to which face/identity swapping methods maintain facial expressions by operating on the embedding vectors. We compare our approach with a 2D landmark baseline approach used to measure expression preservation for FSGAN [5]. The comparison is conducted using all 68 landmarks from dlib [83] and, alternatively, 51 landmarks excluding those around the face. We evaluate the normalized Euclidean distance ($L2$) error between the source face and target face ($s2t$), the source face and the changed face ($s2c$), as well as the mean distance from the target face to the changed face ($t2c$). Additionally, we report the ratio of $t2c$ being less than the distance between the source face and the changed face ($s2c$). These values were obtained by leveraging a pretrained FaceShifter [7] to swap faces between different expression classes in the AffectNet validation set. A total of 3000 samples were generated, divided equally between randomly assigned different expression labels and the same expression labels. The results are presented in Table 5.2 for comparisons within the

Table 5.2: Comparison between our approach and 2D landmark approach for comparing expression for face swaps. Left to right column: Method, euclidean distance error between target face to source face and target face to changed face, mean euclidean distance for target face to changed face and ratio of $t2c < s2c$.

Method	$L2$ error ↓	Mean $t2c$ $L2$ ↓	Ratio ↑
Ours*	**0.07**	**0.17**	**0.73**
68 2D landmarks*	0.29	0.32	0.39
51 2D landmarks*	0.28	0.31	0.49
Ours$^+$	**0.07**	**0.17**	**0.70**
68 2D landmarks$^+$	0.31	0.38	0.39
51 2D landmarks$^+$	0.35	0.35	0.47

* Different class comparison. $^+$ Same class comparison.

same expression class and different expression classes.

In conclusion, our study demonstrates the effectiveness of supervised contrastive representation learning for facial expression recognition. The ablation study highlights the importance of various components in improving performance. Furthermore, our evaluation of identity swapping methods provides insights into their ability to preserve facial expressions, offering a quantitative analysis using the proposed metrics. In future work, specifically in PAPER III, our expression embedder is used to evaluate the expression performance. I refer to PAPER II for further details such as detailed methodology, T-SNE plot of embeddings and network structure.

5.2.3 Summary of Contributions

PAPER II core idea is to learn a rich embedding for facial expressions. We demonstrate that the supervised contrastive loss in conjunction with a class weight is able to learn rich embeddings. It turns out to be significantly more stable than classical softmax cross entropy loss, achieving at the time competitive expression classification accuracy *without* using any extra data on AffectNet. Classification was done by training a multi-layered perceptron (MLP) to classify the extracted expression embeddings, highlighting the richness of the embeddings. Confusion matrix analysis also demonstrates the improved bias of the model, able to deal well with difficult classes such as *contempt*. In the context of facial anonymization, the idea is to use the model for estimating expression retention when anonymizing or face swapping faces.

5.3 PAPER III: FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping

5.3.1 Summary and Purpose

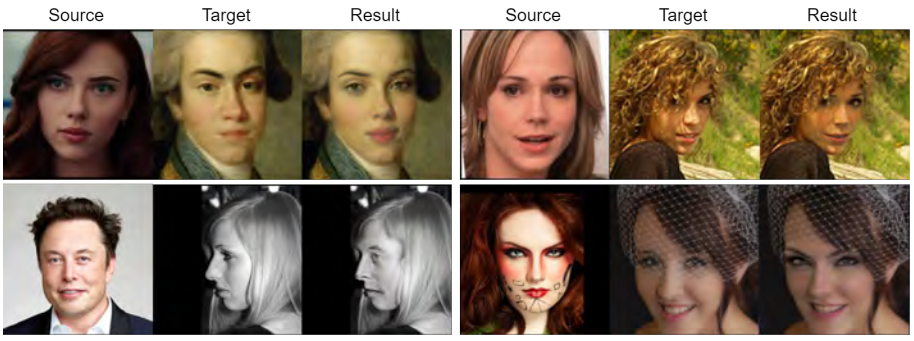

Figure 5.3: Face swapping results generated by FaceDancer.

Face swapping is a complex task that involves transferring the identity of a source face to a target face while preserving important facial attributes such as expression, pose, and lighting. This capability to generate non-existent face pairs finds applications in various industries, including film, gaming, and entertainment [56]. As a result, face swapping has gained considerable attention in the fields of computer vision and graphics.

The main challenge in face swapping lies in achieving a high-fidelity transfer of identity from the source face while ensuring consistency with the target face's attributes. In this work[1], we propose a novel and single-stage method called FaceDancer to address these challenges, including lighting variations, occlusion, pose differences, and semantic structure preservation (See Figure 5.3). FaceDancer stands out for its simplicity, speed, and accuracy.

Our contributions are twofold: Firstly, we introduce an Adaptive Feature Fusion Attention (AFFA) module that dynamically learns to produce attention masks during training. Inspired by recent methods [7; 8], the AFFA module is integrated into the decoder and learns attribute features without the need for additional facial segmentation. The AFFA module incorporates both conditioned features based on the source identity information and unconditioned features from the target information in the encoder's skip connection (See Fig-

[1]Work done within the Vinnova project MIDAS (2019-05873).r

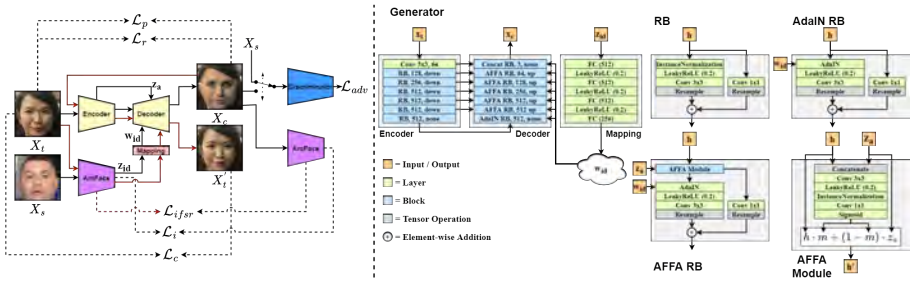

Figure 5.4: Overview of the architecture and training procedure of FaceDancer. For more in depth details and description of components, please refer to PAPER III and its appendix. The appendix also covers the main differences between different baselines and ablation derivatives of FaceDancer.

ure 5.4). It enables FaceDancer to determine which conditioned features (e.g., identity information) to discard and which unconditioned features (e.g., background information) to retain in the target face. Our experiments demonstrate that gating from the AFFA module significantly improves identity transfer.

Secondly, we propose the Interpreted Feature Similarity Regularization (IFSR) loss to enhance attribute preservation. IFSR acts as a regularization technique for FaceDancer, promoting the preservation of facial expression, head pose, and lighting while maintaining high-fidelity identity transfer. Specifically, IFSR explores the similarity between intermediate features in the identity encoder by comparing cosine distance distributions of these features in target, source, and generated face triplets, learned from a pretrained state-of-the-art identity encoder, ArcFace [24] (See Figure 5.4).

We conduct comprehensive quantitative and qualitative experiments on the FaceForensic++ [14] and AFLW2000-3D [15] datasets, demonstrating that FaceDancer outperforms existing face swapping frameworks in terms of identity transfer while exhibiting superior pose preservation compared to most previous methods. Furthermore, we address scalability concerns by applying FaceDancer to low-resolution images with severe distortions and show qualitative improvements in pose preservation compared to other methods.

Although this work primarily focuses on face swapping, it serves as a foundation for researching and contributing to the field of facial anonymization. This connection is highlighted in PAPER IV, which provides details on the usage of FaceDancer [3], SimSwap [6], and an enhanced version of [12] dubbed FIVA as anonymization models.

5.3.2 Results and Contribution to the Licentiate Thesis

We conducted a comprehensive quantitative evaluation of our proposed FaceDancer method on the FaceForensics++ dataset [14]. We compared FaceDancer with other state-of-the-art face swapping networks, including SimSwap [6], FaceShifter [7], HifiFace [8], and FaceController [9]. The evaluation metrics used were identity retrieval (ID), pose error, expression error, and Frechét Inception Distance (FID) [84].

For identity retrieval, we performed random swaps on each image in the test set and then used a secondary identity encoder, CosFace [25], to retrieve the correct identity. To assess pose, we utilized a pose estimator [85] and reported the average L2 error. While expression metrics are often omitted in comparisons due to limited accessibility of models, we employed the expression embedder discussed in PAPER II [2] and reported the average L2 error. FID was calculated between the swapped versions of the test set and the unaltered test set, providing insights into lighting, occlusion, visual quality, and posture issues.

Table 5.3: Quantitative experiments on FaceForensics++ [14]. See PAPER III for further details about configurations.

Method	ID↑	Pose↓	Exp↓	FID↓
FaceSwap [86]	54.19	2.51	N/A	N/A
FaceShifter [7]	97.38	2.96	N/A	N/A
MegaFS [52]	90.83	2.64	N/A	N/A
FaceController [9]	98.27	2.65	N/A	N/A
HifiFace [8]	98.48	2.63	N/A	N/A
SimSwap [6]	92.83	**1.53**	8.04	**11.76**
FaceDancer (Config B)	98.54	2.24	8.52	25.11
FaceDancer (Config C)	**98.84**	2.04	7.97	16.30
FaceDancer (Config D)	98.19	2.15	**5.70**	19.10

Similar to previous works [6–8], we sampled 10 frames from each video in the FaceForensics++ dataset, resulting in a test dataset of 10,000 images. Table 5.6 shows that our FaceDancer method outperforms all previous works in terms of identity retrieval. Regarding the pose metric, FaceDancer achieves the second-lowest pose error (2.04) after SimSwap [6].

For qualitative evaluation, we compared the performance of FaceDancer with the recent state-of-the-art works, including SimSwap [6], FaceShifter [7],

Figure 5.5: Comparing FaceDancer with SimSwap [6], FaceShifter [7], HifiFace [8], and FaceController [9].

HifiFace [8], and FaceController [9] (see Figure 5.5). We provided more in-depth comparisons with SimSwap due to its public accessibility, while qualitative results for other baseline models were limited to sample images reported in their respective works. Detailed comparisons can be found in PAPER III and its appendix.

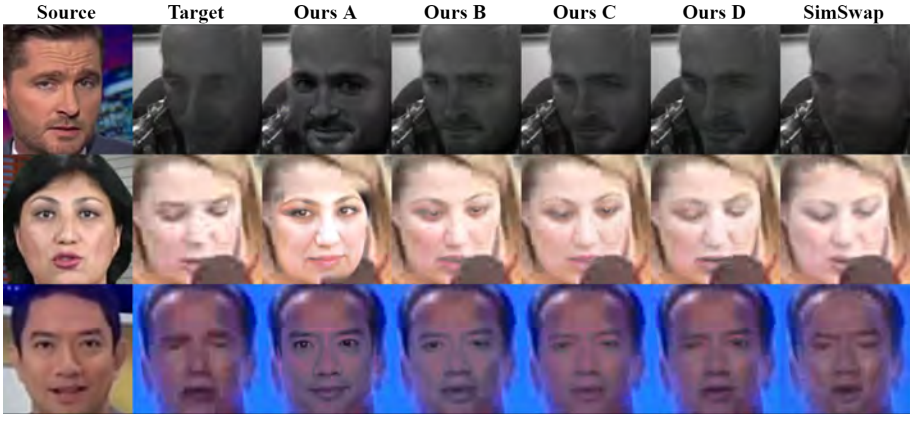

Figure 5.6: Qualitative comparison on low resolution images.

Figure 5.5 illustrates that the FaceDancer model exhibits similar behavior to SimSwap but noticeably improves identity transfer. FaceShifter demonstrates good identity transfer and the preservation of relevant attributes such as facial hair while maintaining occlusion and the identity face shape. However, FaceShifter struggles with lighting and gaze direction, heavily relying on the second-stage model. FaceController demonstrates good identity transferability and decent pose error but frequently fails in preserving gaze direction. Our approach effectively addresses these challenges. Lastly, HifiFace shows promising results, particularly in terms of facial shape preservation. While our

model falls slightly short in facial shape preservation compared to HifiFace, it outperforms quantitatively (see Table 5.6).

FaceDancer also showcases its ability to understand and preserve image distortions, such as maintaining pixelation artifacts (See Figure 5.6). It performs well on videos without considering temporal information, as demonstrated in the supplementary material of PAPER III. The supplementary material also includes results on higher resolution images, further comparisons, handling occlusion, challenging poses, extreme cases, and failure cases. Failures typically occur when the face poses away from the camera or when the face pose represents an uncommon angle not well-represented in the training data.

Table 5.4: Ablative analysis together with the runtime performance. Inference time is given in millisecond and memory usage in GB. All models in this table were trained for 300k iterations.

Config	IFSR	AFFA	Concat final skip*	6 skips	Mapping	ID↑	Pose↓	Exp↓	FID↓	Inference	Memory
Baseline 1	✓	-	-	-	✓	97.66	1.97	8.20	16.72	74.9	1.25
Baseline 2	✓	-	-	-	✓	92.61	**1.87**	7.97	13.51	70.2	1.25
A	-	✓	-	-	✓	98.14	3.61	9.82	31.63	75.8	**1.18**
B	✓	✓	-	-	✓	96.96	2.48	8.25	23.11	75.8	**1.18**
C	✓	✓	✓	-	✓	**98.57**	2.27	7.98	14.59	78.3	1.26
D	✓	✓	✓	✓	✓	97.53	2.04	7.76	**13.50**	78.2	1.27
E	✓	✓	✓	✓	-	97.38	2.07	**5.73**	14.68	**64.6**	1.21

* Concatenation instead of AFFA at resolution 256 + one extra AFFA modules at resolution 32. See supplementary materials for detailed figures for each configuration.

To highlight the impact of our contributions, we conducted ablation experiments by removing different components, such as the AFFA module and the IFSR loss, and compared the results with two baselines. Table 5.4 presents the evaluations on the FaceForensics++ dataset [14], while the ablations in Table 5.7 are performed on the AFLW2000-3D dataset [15]. For a detailed explanation of the differences in baselines and configurations, refer to PAPER III. The contribution of IFSR and AFFA becomes more evident when evaluating the pose-challenging AFLW2000-3D dataset (Table 5.7). PAPER III provides further analysis of the impact of AFFA, including its influence in different resolutions.

In this Thesis, we extensively discuss the IFSR loss, which constitutes the key ingredient and contribution enabling FaceDancer to achieve robust attribute retention alongside strong identity transferability. In the case of IFSR, we investigate intermediate features within the ArcFace ResNet50 backbone by comparing cosine distances between feature maps computed for the target

Table 5.5: Ablative analysis using AFLW2000-3D [15] as target and FaceForensics++ [14] as source.

Config	ID↑	Pose↓	Exp↓	FID↓
Baseline 1	89.10	**5.63**	5.34	19.26
Baseline 2	94.95	6.23	5.60	21.30
A	**98.50**	14.97	7.07	40.34
B	97.95	5.86	5.74	21.50
C	97.65	5.82	**4.13**	18.50
D	97.10	5.75	4.15	20.41
E	95.45	6.16	4.19	**18.13**

face, the source face, the changed face, and negative pairs using the VGGFace2 dataset [55]. This process is repeated for each residual block output in ArcFace.

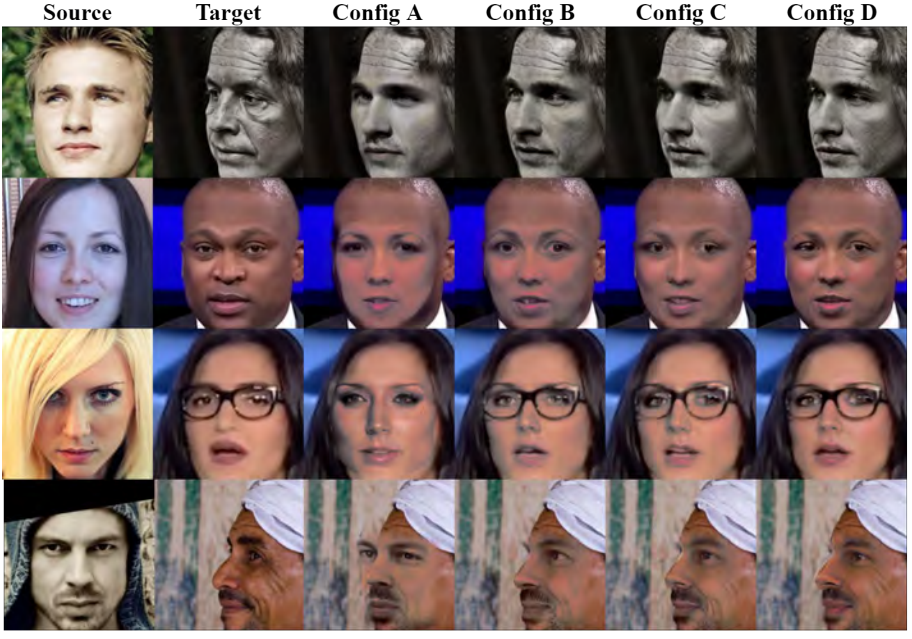

Figure 5.7: Illustration of the impact of IFSR. Config A given in the 3rd column here shows results once IFSR is omitted during training.

Figure 5.8 demonstrates that the changed face shares significantly more similar features with the target face than with the source face in the early layers of ArcFace, while this behavior diminishes in the final residual block.

This suggests that the identity encoder contains important information, such as pose, expression, and occlusions, in the earlier layers, whereas the final blocks predominantly store identity information. To quantify the separability of the changed-to-target (c2t) and changed-to-source (c2s) distributions, we calculate the equal error rate (EER) between these distributions. As depicted in Figure 5.8, the c2t and c2s distributions remain completely separable until block 14. The qualitative impact of our proposed IFSR method is demonstrated in Figure 5.7, where the lack of IFSR results in more apparent pasted face effects and compromised expression preservation.

It is worth noting that the layers and information used in IFSR are obtained from a frozen identity encoder. Therefore, any pretrained face swap framework can be employed to calculate the IFSR margins. IFSR itself does not contain any learnable parameters. It serves as a means to gain an interpretable insight into the information contained in different layers (e.g., expression, pose, color, lighting, identity) and define appropriate margins (see PAPER III for margin details) for IFSR. Furthermore, in PAPER IV, we demonstrate the successful training of a robust anonymization model using IFSR, wherein the margins are completely omitted.

To conclude PAPER III in the context of this Thesis, we introduce FaceDancer, a single-stage face swapping model that quantitatively reached state-of-the-art. One of its strongest contribution is the IFSR loss that utilizes intermediate features to preserve attributes such as pose, facial expression, and occlusion. Furthermore, IFSR in the context of this Thesis, provide a strong contribution that enables facial anonymization. I demonstrate strong facial anonymization using FaceDancer in PAPER IV. PAPER IV also introduces a specialized model that is specialized in anonymization, which was trained using IFSR. See PAPER IV or below for further details.

5.3.3 Summary of Contributions

The work in PAPER III does not explicit address facial anonymization for this Thesis. However it implicit do so as its contributions are useful for further work such as in PAPER IV. The core of PAPER III is the FaceDancer face swapping framework, reaching state-of-the-art performance. The main contribution that allowed for this is the introduced IFSR, which is a regularizing loss that forces the network to keep important attributes. Secondly, we introduced the Adaptive Feature Fusion Attention (AFFA) module, which adaptively allow the network to fuse feature maps conditioned on identity information with

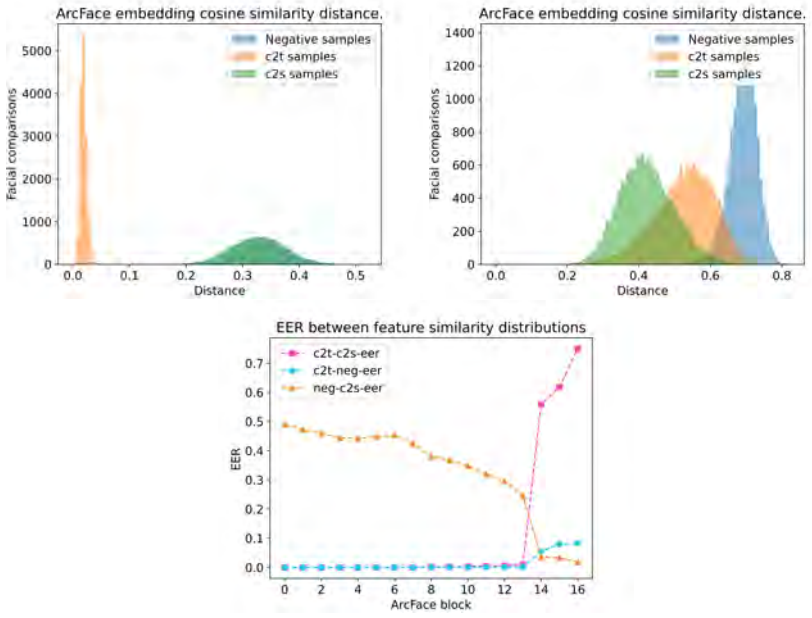

Figure 5.8: Cosine similarity between intermediate features between changed and target faces (c2t), changed and source faces (c2s), and different identities (Negative Samples). (a) Distances between features from first block of ArcFace. (b) Distances between features from final block of ArcFace. (c) Equal error rates (EER) between the distance distributions for intermediate features in every block.

unconditioned feature maps. Allowing for better trade-off between attribute retention and identity transfer metrics. Furthermore, FaceDancer and IFSR are used to derive the results and contributions in PAPER IV.

5.4 PAPER IV: Facial Image and Video Anonymization and Anonymization Defense

5.4.1 Summary and Purpose

Privacy plays a crucial role in numerous domains, including data collection and storage, and is further emphasized by the implementation of regulations like the General Data Protection Regulation (GDPR) [21]. As the demand for data and interest in privacy increase, the need for data anonymization becomes imperative. Anonymization techniques aim to conceal, remove, or replace identity information with arbitrary pseudo-identities while preserving essential attribute information. However, direct manipulation of the data distribution to obscure or remove identity information often leads to the loss of

significant attributes. Traditional methods, such as blurring faces or replacing them with black boxes, eliminate crucial details like eye gaze, pose, and expressions. In contrast, replacement-oriented approaches focus on preserving essential attributes while altering individual identities.

This study (PAPER IV) specifically concentrates on replacement-oriented approaches for anonymizing faces in both images and videos. We propose a novel method that utilizes target-oriented face swapping models for anonymization purposes (See Figure 5.9). We investigate two target-oriented face swapping models, namely FaceDancer (PAPER III) and SimSwap [6], along with a third model based on the work in [12]. Notably, our proposed method can be implemented with any existing target-oriented face manipulation model that leverages identity embeddings [3; 6–8] for facial manipulation. These models ensure strong consistency across video frames when the employed identity embedding remains stable. To ensure realistic and robust anonymization, we introduce a simple and efficient approach for identity tracking and sampling of fake identities, thus enabling consistency across frames.

Moreover, this research addresses a critical security aspect concerning the vulnerability to reconstruction attacks, which is prevalent in federated learning [34; 35] (Briefly mentioned in Chapter 4) but has received limited exploration within the context of facial anonymization. In a reconstruction attack, an adversarial model attempts to translate the anonymized face back to its original identity. We hypothesize and provide compelling evidence that target-oriented models leave traces in the images that can be exploited for successful reconstruction attacks. To mitigate and evaluate this threat, we investigate the effectiveness of various noise types, including adversarial noise, uniform noise, and parameter noise, in disrupting the reconstruction attack.
Finally, we emphasize that maximizing the distance between identities can be potentially detrimental to privacy. This is primarily due to the fact that state-of-the-art facial recognition models constrain embeddings to a hyper-unit-sphere, allowing the original identity to be easily identified by negating one of the embeddings.

5.4.2 Results and Contribution to the Licentiate Thesis

We conducted a comprehensive series of experiments to effectively demonstrate the capabilities of target-oriented face swapping models, specifically FaceDancer and SimSwap, as well as our proposed model, FIVA. Since certain previous works lack availability and detailed information, we compared the performance on multiple datasets. In Table 5.6, we present quantitative results

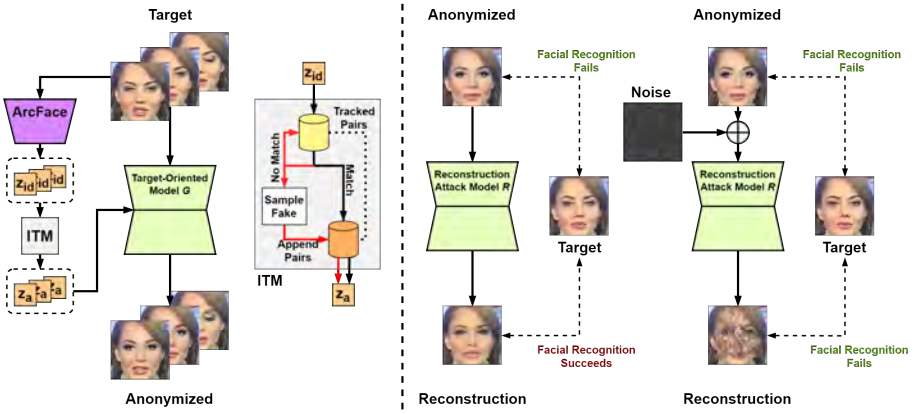

Figure 5.9: Overview of the proposed anonymization pipeline and an illustration of the implication of reconstruction attacks.

obtained from the FaceForensic++ dataset [14] and compare them with previous works. The evaluation metrics include identity retrieval (ID), reconstruction attack identity retrieval (RA), negated identity retrieval (¬ID), and temporal consistency (for detailed information on temporal consistency, refer to PAPER IV). As depicted in Table 5.6, both FIVA, SimSwap, and FaceDancer demonstrate a high level of face anonymization. The proposed Identity Tracking Module (ITM) contributes to robust temporal consistency. Due to the nature of FIVA being trained to drive identity away from the target face, it guarantees a large cosine distance but are susceptible to have the original identity verified if during the verification process the extracted identity embedding z_{id} extracted from the anonymized image is negated to $-z_{id}$. In short, this is due to the hyper sphere constrained properties of cosine similarity. This Thesis covers the details of this behaviour later in the chapter and more details can be found in PAPER IV.

Furthermore, we assert that FIVA and other target-oriented approaches leave discernible traces in the image, enabling an adversarial network to learn the reconstruction of the original identity (see Figure 5.10 and refer to PAPER IV). To reinforce this claim and explore potential defense strategies, we illustrate in Table 5.8 and Figure 5.10 that the output of the reconstruction attack model can be disrupted by introducing perturbations.
Next, we compare the anonymization performance of FIVA, SimSwap, and FaceDancer with previous works using the Labeled Faces in the Wild (LFW) benchmark [16] in Table 5.7. It should be noted that we employ a more powerful facial recognition model, CosFace [25], for identity retrieval, ensuring a fair comparison with the prior works.

Table 5.6: Quantitative experiments on FaceForensics++ [14]. Evaluated with identity retrieval (ID), negated identity retrieval (¬ID, searching for a match with $-z_{id}$), and reconstruction attack (RA) identity retrieval. Temporal identity consistency $\mathcal{M}_{tc}$ calculated using 10 frames per video. The divide in the table separates inpainting-based methods from target-oriented ones. The $\times$ indicates that RA is not applicable to the corresponding method. +Sampling means we used the anchor sampling method to assign anonymized identities (See PAPER IV for details), while +ITM indicates both the anchor sampling and tracking (See PAPER IV and Figure 5.9). The $\downarrow$ indicates lower is better.

Method	ID$\downarrow$	¬ID$\downarrow$	RA$\downarrow$	$\mathcal{M}_{tc}^{\mu}\downarrow$	$\mathcal{M}_{tc}^{\sigma}\downarrow$
Real Data	-	-	-	0.150	0.074
CIAGAN [10]	0.035	**0.000**	$\times$	0.521	0.220
CIAGAN [10] + ITM	0.030	**0.000**	$\times$	0.300	0.151
DeepPrivacy [13]	0.004	**0.000**	$\times$	0.359	0.184
CFA-Net [11]	0.012	N/A	N/A	N/A	N/A
SimSwap [6] + Sampling	0.002	**0.000**	**0.994**	0.607	0.345
SimSwap [6] + ITM	0.002	**0.000**	**0.994**	0.084	0.051
FaceDancer [3] + Sampling	**0.000**	**0.000**	0.999	0.556	0.314
FaceDancer [3] + ITM	**0.000**	**0.000**	0.999	0.186	0.141
FIVA	**0.000**	0.966	0.998	0.227	0.101
FIVA + Sampling	**0.000**	**0.000**	0.996	0.550	0.310
FIVA + ITM	**0.000**	**0.000**	0.996	**0.075**	**0.041**

For qualitative evaluation, we compare the output of FIVA together with previous work (Figure 5.11 and Figure 5.12). In depth comparisons are done with available models such as CIAGAN and DeepPrivacy. Analysing the images in Figure 5.11 and Figure 5.12, we see that CFA-NET struggles with maintaining the color and eye-gaze. CIAGAN struggles with the resolution of the image and general output quality. DeepPrivacy struggles with eye-gaze, expression and often producing artifacts. Gafni et al. together with CFA-NET and FIVA, is the only approach that has demonstrated successful results on video. CFA-NET does use similar identity control as target-oriented face swapping and FIVA, which means that they need to manually assign identity embedding for each face in a video. For video results, we refer to the supplementary material in PAPER IV.

However, a qualitative analysis reveals an intriguing aspect concerning

Table 5.7: Quantitative identity retrieval experiments on LFW [16]. CFA-Net [11] and Gafni et al. [12] demonstrate the true positive rate for a false acceptance rate of 0.001 using FaceNet [17] as the facial recognition model. We evaluate the remaining methods with CosFace and a threshold of 0.63 (Cosine *distance*), for a false acceptance rate of 0.001. The ↓ indicates lower is better.

Method	ID↓
Gafni et al. [12]	0.035
CIAGAN [10]	0.034
CFA-Net [11]	0.012
DeepPrivacy [13]	0.002
FaceDancer [3] + ITM	0.002
SimSwap [6] + ITM	0.001
FIVA + ITM	**0.000**

FIVA's preservation of gender and ethnicity during face swaps, even when confronted with scenarios involving dissimilar target and source attributes. This phenomenon is visually evident in Figure 5.13. In comparison to FaceDancer, FIVA consistently exhibits a tendency to retain the gender and ethnicity attributes during face swaps. Although FIVA outperforms other methods quantitatively in terms of identity transfer, it falls short qualitatively in fully capturing the desired facial transformations. Nevertheless, this distinctive behavior is of value in other applications, such as anonymization, eliminating the necessity of sampling identities based on gender. Despite this behavior, FIVA achieves remarkable performance on the identity retrieval metric employed to evaluate

Table 5.8: Defense against reconstruction attack in FIVA, evaluated on FaceForensics++ [14]. Adversarial Defense in the form of a fast sign gradient method. Noise Defense just adds regular uniform noise to the image. Parameter Noise means adding a small Gaussian noise to the parameters. We report the fraction of successful retrievals of the original identity after applying the reconstruction attack. ε highlights how much the noise was scaled. The ↓ indicates lower is better. Black-box means it does not need access to the reconstruction attack model.

Method	ε	ID↓	Black-box
Parameter Noise	0.10	0.442	**yes**
Adversarial Defense	0.15	**0.002**	no
Noise Defense	0.15	0.004	**yes**

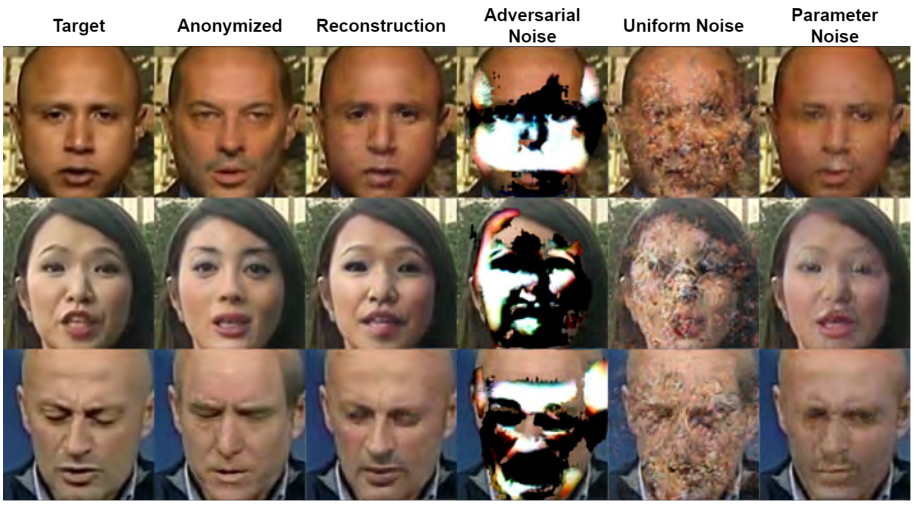

Figure 5.10: Qualitative results of reconstruction attack, different defenses and anonymization using FIVA.

Figure 5.11: Qualitative comparison between CIAGAN [10], CFA-NET [11], Gafni et al. [12], DeepPrivacy [13] and FIVA.

face swapping methods, as demonstrated in PAPER III and [6–9; 52]. For detailed results, refer to PAPER IV.

Lastly, it is important to address a minor security concern associated with the utilization of target-oriented methods, specifically those employing identity vectors that are significantly distant. In PAPER IV, we delve into the concept of employing anchor identities within the Identity Tracking Module, sampled using a defined margin. Figure 5.14 illustrates the relationship between anchor matches and varying margin values, wherein the green line represents the match achieved with a margin of $m = 0.7$. As FIVA's training objective involves driving the generated identities away from the source, it becomes essential to identify an anchor identity that is in close proximity. Conversely, when employing target-oriented face swapping methods that aim to align the

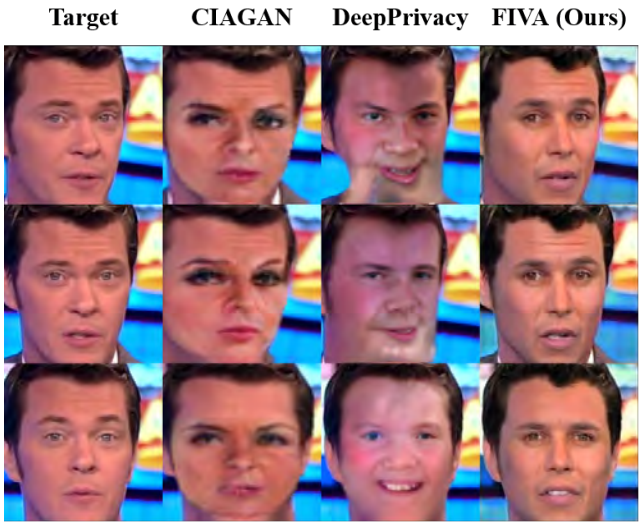

Figure 5.12: Qualitative temporal comparison between CIAGAN [10], DeepPrivacy [13] and FIVA. Note we used ITM for tracking the identity for CIAGAN.

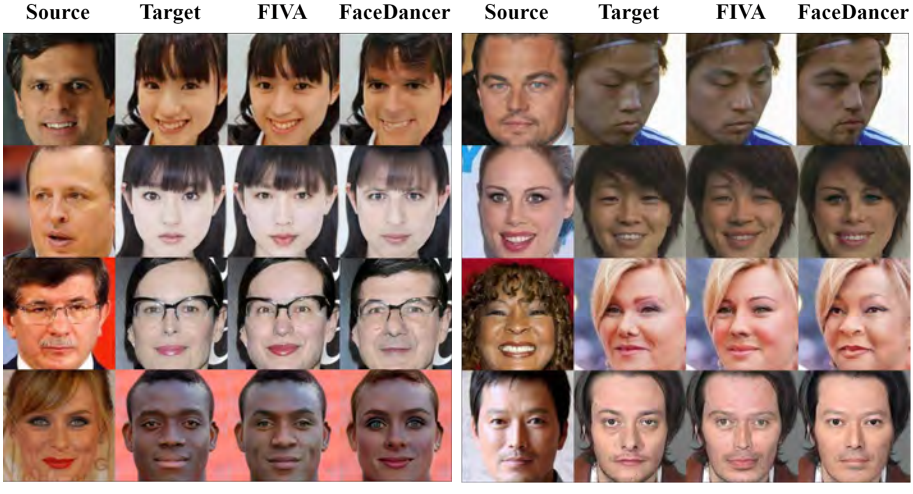

Figure 5.13: Qualitative comparison between FIVA and FaceDancer for gender and ethnicity retention.

generated identity with the source, it becomes necessary to sample identities that are distant from the original target identity.

To elucidate further, our findings indicate that selecting the furthest iden-

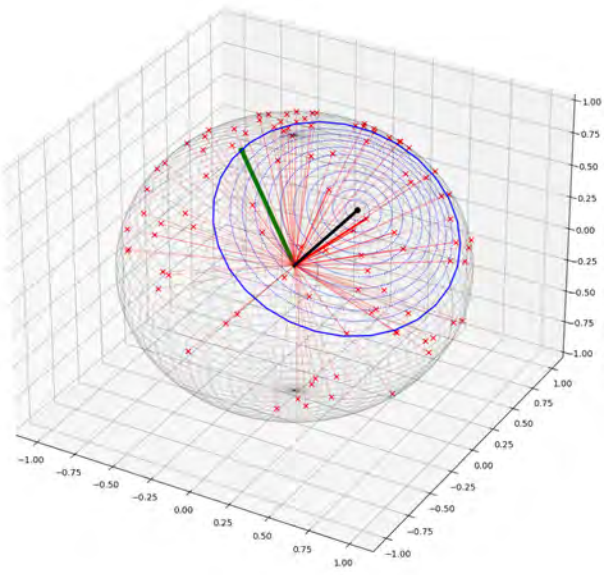

Figure 5.14: Illustration of matching a desired anchor. The red lines illustrates matches to a desired anchor based on desired approximate distance from the target vector (black line). The green line illustrates the match that would occur when sampling for FIVA. Blue circle illustrates the desired distance.

tity (or the closest in the case of FIVA, owing to its counterfactual training scheme) allows facial recognition systems to consistently identify the original identity by negating one of the vectors with a value of -1. As an illustrative example, suppose we aim to anonymize face X_t with an identity vector of z_{id}. By employing a face swapping model such as FaceDancer or SimSwap and inputting $G(X_t, -z_{id})$, where G denotes the face swapping model, we theoretically achieve the maximum possible difference in identities. However, the drawback of this approach lies in the high likelihood of successfully matching the resulting face with $-z_{id}$. Therefore, to ensure robust anonymization, it becomes imperative to ensure that the resulting face is distanced significantly from both z_{id} and $-z_{id}$. This is precisely the purpose served by the use of anchor identities.

To conclude this section. PAPER IV is the most exhaustive contribution to this Thesis. We demonstrated that target-oriented face manipulation models are excellent candidates for facial anonymization. They are fast, efficient and controllable. However, as one of the first to our knowledge, the paper

demonstrated some serious security concerns that could be of a problem if these methods are naively used.

5.4.3 Summary of Contributions

PAPER IV contributes significantly to the Thesis through several core ideas related to facial anonymization. Firstly, the ITM allow for better temporal consistency, in-the-wild anonymization, and prevents the issue of leaking the identity when the cosine distance is to large. ITM in conjunction of FIVA allow for a 0 true positives for a FAR of 0.001. While Todt et al. [87] did investigate traditional anonymization method (such as blurring and pixelation), along with inpainting-based methods DeepPrivacy [13; 67] and CIA-GAN [10], this work is first to our knowledge that show that reconstruction attack is not only possible for target-oriented models, but also surprisingly simple. We empirically demonstrate this further by disrupting the reconstruction attacks with a small noise added to the images. The reconstruction attack model can also be used for deep fake detection, but is as of now not model agnostic. FIVA is able to do zero-shot face swapping, reaching state-of-the-art performance in identity retrieval metric even if the the image it self is not perceptually convincingly a face swap.

6. Conclusions

To summarize this Thesis, it has investigated target-oriented face manipulation for facial anonymization, provided exhaustive evaluation, identified its shortcomings and its potential for privacy-aware data collection, data storage, data sharing and machine learning. The first two appended publications (PAPER I and PAPER II) focus directly on the evaluation step, addressing Research Question 1 (See Chapter 1.2). Later work (PAPER III and IV) improved upon the anonymization evaluation that PAPER I sought to study. However, PAPER I provided insight into what we humans use for identification, demonstrating that other parts than the face can leak crucial identity information. Through PAPER II, the Thesis demonstrates a novel method for embedding facial expression. Which in turn, was used to measure expression retention when manipulating facial identity in PAPER III. However, I would like to highlight that the current literature as of writing this, choose to either omit evaluating expression retention or uses 3DMM [60] to regress expression coefficients. In PAPER IV we also chose to omit evaluating the expression and focus on identity and pose.

The Thesis aimed to answer how we can make sure anonymized faces are maintained properly both through time and with multiple people (Research Question 2 and 3, see Chapter 1.2). The Thesis address this through its related work studies and contribution in both PAPER III and PAPER IV. We achieved a fast, efficient and realistic method to address this. With these two publications, we also provide a novel contribution to methodologies for succeeding with facial manipulation (swapping, anonymization). For example IFSR for attribute retention. Exhaustive metrics were identified, and the work as of their publications achieved state-of-the-art performance.

PAPER IV contributed with not only improved methodologies in achieving facial anonymization, but also identified vulnerabilities that could potentially be detrimental to its goal of privacy (Research Question 4, see Chapter 1.2). We demonstrated that we could reconstruct original identities purely from input image and manipulated image pairs (black box attack), showing strong evidence of information traces hidden in the anonymized image. The start of addressing these issues are also highlighted.

For future work and research, investigating potential vulnerabilities in depth as per findings of PAPER IV (Research Question 3, see Chapter 1.2) and investigate the potential for explainability for addressing and understanding vulnerabilities in the model (Future Research Question 1, see Chapter 1.2).

References

[1] FELIX ROSBERG, CRISTOFER ENGLUND, MARTIN TORSTENSSON, AND BORIS DURÁN. **Towards Privacy Aware Data collection in Traffic**. In *FAST-zero - International Symposium on Future Active Safety Technology toward zero-traffic-accident*, 2021. v

[2] FELIX ROSBERG AND CRISTOFER ENGLUND. **Comparing Facial Expressions for Face Swapping Evaluation with Supervised Contrastive Representation Learning**. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–05, 2021. v, 33

[3] FELIX ROSBERG, EREN ERDAL AKSOY, FERNANDO ALONSO-FERNANDEZ, AND CRISTOFER ENGLUND. **FaceDancer: Pose- and Occlusion-Aware High Fidelity Face Swapping**. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3454–3463, January 2023. v, xi, 9, 13, 15, 32, 39, 41, 42

[4] FELIX ROSBERG, EREN ERDAL AKSOY, CRISTOFER ENGLUND, AND FERNANDO ALONSO-FERNANDEZ. **FIVA: Facial Image and Video Anonymization and Anonymization Defense**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 362–371, October 2023. v, 20

[5] YUVAL NIRKIN, YOSI KELLER, AND TAL HASSNER. **FSGAN: Subject Agnostic Face Swapping and Reenactment**. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7183–7192, 2019. xi, 16, 23, 25, 29

[6] RENWANG CHEN, XUANHONG CHEN, BINGBING NI, AND YANHAO GE. *SimSwap: An Efficient Framework For High Fidelity Face Swapping*, page 2003–2011. Association for Computing Machinery, New York, NY, USA, 2020. xi, 9, 13, 16, 17, 21, 32, 33, 34, 39, 41, 42, 43

[7] LINGZHI LI, JIANMIN BAO, HAO YANG, DONG CHEN, AND FANG WEN. **Advancing high fidelity identity swapping for forgery detection**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020. xi, 16, 17, 23, 29, 31, 33, 34

[8] YUHAN WANG, XU CHEN, JUNWEI ZHU, WENQING CHU, YING TAI, CHENGJIE WANG, JILIN LI, YONGJIAN WU, FEIYUE HUANG, AND RONGRONG JI. **HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping**. In ZHI-HUA ZHOU, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1136–1142. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. xi, 13, 17, 31, 33, 34, 39

[9] ZHILIANG XU, XIYU YU, ZHIBIN HONG, ZHEN ZHU, JUNYU HAN, JINGTUO LIU, ERRUI DING, AND XIANG BAI. **FaceController: Controllable Attribute Editing for Face in the Wild**. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**(4):3083–3091, May 2021. xi, 16, 33, 34, 43

[10] MAXIM MAXIMOV, ISMAIL ELEZI, AND LAURA LEAL-TAIXÉ. **Ciagan: Conditional identity anonymization generative adversarial networks**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020. xii, 41, 42, 43, 44, 46

[11] TIANXIANG MA, DONGZE LI, WEI WANG, AND JING DONG. **CFA-Net: Controllable Face Anonymization Network with Identity Representation Manipulation**. *arXiv preprint arXiv:2105.11137*, 2021. xii, xiii, 20, 41, 42, 43

[12] ORAN GAFNI, LIOR WOLF, AND YANIV TAIGMAN. **Live face de-identification in video**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9378–9387, 2019. xii, xiii, 20, 32, 39, 42, 43

[13] HÅKON HUKKELÅS, RUDOLF MESTER, AND FRANK LINDSETH. **DeepPrivacy: A Generative Adversarial Network for Face Anonymization**. In *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019. xii, 20, 41, 42, 43, 44, 46

[14] ANDREAS ROSSLER, DAVIDE COZZOLINO, LUISA VERDOLIVA, CHRISTIAN RIESS, JUSTUS THIES, AND MATTHIAS NIESSNER. **Faceforensics++: Learning to detect manipulated facial images**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019. xiii, xiv, 32, 33, 35, 36, 40, 41, 42

[15] XIANGYU ZHU, ZHEN LEI, XIAOMING LIU, HAILIN SHI, AND STAN Z LI. **Face alignment across large poses: A 3d solution**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016. xiii, 32, 35, 36

[16] GARY B. HUANG, MANU RAMESH, TAMARA BERG, AND ERIK LEARNED-MILLER. **Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments**. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. xiii, 11, 40, 42

[17] FLORIAN SCHROFF, DMITRY KALENICHENKO, AND JAMES PHILBIN. **Facenet: A unified embedding for face recognition and clustering**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. xiii, 11, 24, 42

[18] GARY MARCUS. **Deep learning: A critical appraisal**. *arXiv preprint arXiv:1801.00631*, 2018. 1

[19] AMINA ADADI. **A survey on data-efficient algorithms in big data era**. *Journal of Big Data*, **8**(1):24, 2021. 1

[20] JASON WEI, YI TAY, RISHI BOMMASANI, COLIN RAFFEL, BARRET ZOPH, SEBASTIAN BORGEAUD, DANI YOGATAMA, MAARTEN BOSMA, DENNY ZHOU, DONALD METZLER, ET AL. **Emergent abilities of large language models**. *arXiv preprint arXiv:2206.07682*, 2022. 1

[21] **General Data Protection Regulation**. Accessed 2023-05-09. 1, 38

[22] **Cybersecurity Law of the People's Republic of China**. Accessed 2023-05-09. 1

[23] **California Consumer Privacy Act**. Accessed 2023-05-09. 1

[24] JIANKANG DENG, JIA GUO, NIANNAN XUE, AND STEFANOS ZAFEIRIOU. **ArcFace: Additive Angular Margin Loss for Deep Face Recognition**. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 1, 9, 11, 12, 24, 32

[25] HAO WANG, YITONG WANG, ZHENG ZHOU, XING JI, DIHONG GONG, JINGCHAO ZHOU, ZHIFENG LI, AND WEI LIU. **Cosface: Large margin cosine loss for deep face recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 11, 12, 33, 40

[26] QIANG MENG, SHICHAO ZHAO, ZHIDA HUANG, AND FENG ZHOU. **Magface: A universal representation for face recognition and quality assessment**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021. 12

[27] WEIYANG LIU, YANDONG WEN, ZHIDING YU, MING LI, BHIKSHA RAJ, AND LE SONG. **Sphereface: Deep hypersphere embedding for face recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017. 11

[28] PHILIPP TERHÖRST, MALTE IHLEFELD, MARCO HUBER, NASER DAMER, FLORIAN KIRCHBUCHNER, KIRAN RAJA, AND ARJAN KUIJPER. **Qmagface: Simple and accurate quality-aware face recognition**. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3484–3494, 2023.

[29] XIANG AN, XUHAN ZHU, YUAN GAO, YANG XIAO, YONGLE ZHAO, ZIYONG FENG, LAN WU, BIN QIN, MING ZHANG, DEBING ZHANG, ET AL. **Partial fc: Training 10 million identities on a single machine**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1445–1449, 2021. 1, 11, 12

[30] ZEHENG REN, XIAOBEI JIANG, AND WUHONG WANG. **Analysis of the influence of pedestrians' eye contact on drivers' comfort boundary during the crossing conflict**. *Procedia engineering*, **137**:399–406, 2016. 1

[31] ABDALLAH MOUJAHID, MOUNIR ELARAKI TANTAOUI, MANOLO DULVA HINA, ASSIA SOUKANE, ANDREA ORTALDA, AHMED ELKHADIMI, AND AMAR RAMDANE-CHERIF. **Machine learning techniques in ADAS: a review**. In *2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, pages 235–242. IEEE, 2018. 2

[32] **Road traffic injuries**. Accessed 2023-05-09. 2

[33] TIAN LI, ANIT KUMAR SAHU, AMEET TALWALKAR, AND VIRGINIA SMITH. **Federated learning: Challenges, methods, and future directions**. *IEEE signal processing magazine*, **37**(3):50–60, 2020. 4, 19

[34] ZHIBO WANG, MENGKAI SONG, ZHIFEI ZHANG, YANG SONG, QIAN WANG, AND HAIRONG QI. **Beyond inferring class representatives: User-level privacy leakage from federated learning**. In *IEEE INFOCOM 2019-IEEE conference on computer communications*, pages 2512–2520. IEEE, 2019. 4, 19, 39

[35] MENGKAI SONG, ZHIBO WANG, ZHIFEI ZHANG, YANG SONG, QIAN WANG, JU REN, AND HAIRONG QI. **Analyzing user-level privacy attack against federated learning**. *IEEE Journal on Selected Areas in Communications*, **38**(10):2430–2444, 2020. 4, 19, 39

[36] BRENDAN MCMAHAN, EIDER MOORE, DANIEL RAMAGE, SETH HAMPSON, AND BLAISE AGUERA Y ARCAS. **Communication-efficient learning of deep networks from decentralized data**. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 19

[37] RUI SHAO, BOCHAO ZHANG, PONG C. YUEN, AND VISHAL M. PATEL. **Federated Test-Time Adaptive Face Presentation Attack Detection with Dual-Phase Privacy Preservation**. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, 2021. 4

[38] KAIPENG ZHANG, ZHANPENG ZHANG, ZHIFENG LI, AND YU QIAO. **Joint face detection and alignment using multitask cascaded convolutional networks**. *IEEE signal processing letters*, **23**(10):1499–1503, 2016. 7, 23

[39] JIANFENG WANG, YE YUAN, AND GANG YU. **Face attention network: An effective face detector for the occluded faces**. *arXiv preprint arXiv:1711.07246*, 2017. 7

[40] TSUNG-YI LIN, PIOTR DOLLÁR, ROSS GIRSHICK, KAIMING HE, BHARATH HARIHARAN, AND SERGE BELONGIE. **Feature pyramid networks for object detection**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 7

[41] JIANKANG DENG, JIA GUO, EVANGELOS VERVERAS, IRENE KOTSIA, AND STEFANOS ZAFEIRIOU. **Retinaface: Single-shot multi-level face localisation in the wild**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 7

[42] YANJIA ZHU, HONGXIANG CAI, SHUHAN ZHANG, CHENHAO WANG, AND YICHAO XIONG. **Tinaface: Strong but simple baseline for face detection**. *arXiv preprint arXiv:2011.13183*, 2020. 7

[43] SHIFENG ZHANG, CHENG CHI, ZHEN LEI, AND STAN Z LI. **Refineface: Refinement neural network for high performance face detection**. *IEEE transactions on pattern analysis and machine intelligence*, **43**(11):4008–4020, 2020. 24

[44] BIN ZHANG, JIAN LI, YABIAO WANG, YING TAI, CHENGJIE WANG, JILIN LI, FEIYUE HUANG, YILI XIA, WENJIANG PEI, AND RONGRONG JI. **Asfd: Automatic and scalable face detector**. *arXiv preprint arXiv:2003.11228*, 2020.

[45] YANG LIU, XU TANG, JUNYU HAN, JINGTUO LIU, DINGER RUI, AND XIANG WU. **HAMBox: Delving Into Mining High-Quality Anchors on Face Detection**. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13043–13051, 2020. 7

[46] KAIMING HE, XIANGYU ZHANG, SHAOQING REN, AND JIAN SUN. **Deep residual learning for image recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[47] MINGXING TAN AND QUOC LE. **Efficientnet: Rethinking model scaling for convolutional neural networks**. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 7

[48] ZHUANG LIU, HANZI MAO, CHAO-YUAN WU, CHRISTOPH FEICHTENHOFER, TREVOR DARRELL, AND SAINING XIE. **A convnet for the 2020s**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022. 7

[49] SHINJI UMEYAMA. **Least-squares estimation of transformation parameters between two point patterns**. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **13**(04):376–380, 1991. 8

[50] XIN JIN AND XIAOYANG TAN. **Face alignment in-the-wild: A survey**. *Computer Vision and Image Understanding*, **162**:1–22, 2017. 10

[51] YIFAN SUN, CHANGMAO CHENG, YUHAN ZHANG, CHI ZHANG, LIANG ZHENG, ZHONGDAO WANG, AND YICHEN WEI. **Circle loss: A unified perspective of pair similarity optimization**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020. 11, 12

[52] YUHAO ZHU, QI LI, JIAN WANG, CHENG-ZHONG XU, AND ZHENAN SUN. **One Shot Face Swapping on Megapixels**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4834–4844, June 2021. 11, 16, 33, 43

[53] BRIANNA MAZE, JOCELYN ADAMS, JAMES A DUNCAN, NATHAN KALKA, TIM MILLER, CHARLES OTTO, ANIL K JAIN, W TYLER NIGGEL, JANET ANDERSON, JORDAN CHENEY, ET AL. **Iarpa janus benchmark-c: Face dataset and protocol**. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018. 11

[54] BOXIAO LIU, SHENGHAN ZHANG, GUANGLU SONG, HAIHANG YOU, AND YU LIU. **Rectifying the Data Bias in Knowledge Distillation**. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1477–1486, 2021. 12

[55] QIONG CAO, LI SHEN, WEIDI XIE, OMKAR M PARKHI, AND ANDREW ZISSERMAN. **Vggface2: A dataset for recognising faces across pose and age**. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 12, 23, 36

[56] OLEG ALEXANDER, MIKE ROGERS, WILLIAM LAMBETH, MATT CHIANG, AND PAUL DEBEVEC. **Creating a Photoreal Digital Actor: The Digital Emily Project**. In *2009 Conference for Visual Media Production*, pages 176–187, 2009. 16, 31

[57] YUVAL NIRKIN, IACOPO MASI, ANH TRAN TUAN, TAL HASSNER, AND GERARD MEDIONI. **On Face Segmentation, Face Swapping, and Face Perception**. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 98–105, 2018. 16

[58] Y. NIRKIN, Y. KELLER, AND T. HASSNER. **FSGANv2: Improved Subject Agnostic Face Swapping and Reenactment**. *IEEE Transactions on Pattern Analysis amp; Machine Intelligence*, **45**(01):560–575, jan 2023. 16

[59] VOLKER BLANZ, KRISTINA SCHERBAUM, THOMAS VETTER, AND HANS-PETER SEIDEL. **Exchanging faces in images**. In *Computer Graphics Forum*, **23**, pages 669–676. Wiley Online Library, 2004. 16

[60] BERNHARD EGGER, WILLIAM AP SMITH, AYUSH TEWARI, STEFANIE WUHRER, MICHAEL ZOLLHOEFER, THABO BEELER, FLORIAN BERNARD, TIMO BOLKART, ADAM KORTYLEWSKI, SAMI ROMDHANI, ET AL. **3d morphable face models—past, present, and future**. *ACM Transactions on Graphics (TOG)*, **39**(5):1–38, 2020. 16, 47

[61] JIANMIN BAO, DONG CHEN, FANG WEN, HOUQIANG LI, AND GANG HUA. **Towards Open-Set Identity Preserving Face Synthesis**. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6713–6722, 2018. 16

[62] TAESUNG PARK, MING-YU LIU, TING-CHUN WANG, AND JUN-YAN ZHU. **Semantic Image Synthesis with Spatially-Adaptive Normalization**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 16

[63] JUSTIN JOHNSON, ALEXANDRE ALAHI, AND LI FEI-FEI. **Perceptual Losses for Real-Time Style Transfer and Super-Resolution**. In BASTIAN LEIBE, JIRI MATAS, NICU SEBE, AND MAX WELLING, editors, *Computer Vision – ECCV 2016*, pages 694–711, Cham, 2016. Springer International Publishing.

[64] RICHARD ZHANG, PHILLIP ISOLA, ALEXEI A EFROS, ELI SHECHTMAN, AND OLIVER WANG. **The unreasonable effectiveness of deep features as a perceptual metric**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[65] TING-CHUN WANG, MING-YU LIU, JUN-YAN ZHU, ANDREW TAO, JAN KAUTZ, AND BRYAN CATANZARO. **High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 16, 17

[66] UMUR A. ÇIFTÇI, GOKTURK YUKSEK, AND İLKE DEMIR. **My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization**. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1369–1379, January 2023. 20, 21

[67] HÅKON HUKKELÅS AND FRANK LINDSETH. **DeepPrivacy2: Towards Realistic Full-Body Anonymization**. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1329–1338, 2023. 20, 46

[68] JINGZHI LI, LUTONG HAN, HUA ZHANG, XIAOGUANG HAN, JINGGUO GE, AND XIAOCHUN CAO. **Learning disentangled representations for identity preserving surveillance face camouflage**. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9748–9755. IEEE, 2021. 20

[69] TAO LI AND LEI LIN. **Anonymousnet: Natural face de-identification with measurable privacy**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 20

[70] ZHONGZHENG REN, YONG JAE LEE, AND MICHAEL S RYOO. **Learning to anonymize faces for privacy preserving action detection**. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018. 20

[71] TERO KARRAS, SAMULI LAINE, AND TIMO AILA. **A style-based generator architecture for generative adversarial networks**. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 23, 27

[72] PRANNAY KHOSLA, PIOTR TETERWAK, CHEN WANG, AARON SARNA, YONGLONG TIAN, PHILLIP ISOLA, AARON MASCHINOT, CE LIU, AND DILIP KRISHNAN. **Supervised contrastive learning**. *arXiv preprint arXiv:2004.11362*, 2020. 26

[73] ALI MOLLAHOSSEINI, BEHZAD HASANI, AND MOHAMMAD H. MAHOOR. **AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild**. *IEEE Transactions on Affective Computing*, **10**(1):18–31, Jan 2019. 27, 29

[74] LIAM SCHONEVELD, ALICE OTHMANI, AND HAZEM ABDELKAWY. **Leveraging recent advances in deep learning for audio-Visual emotion recognition**. *Pattern Recognition Letters*, **146**:1–7, Jun 2021. 29

[75] ANDREY V. SAVCHENKO. **Facial expression and attributes recognition based on multi-task learning of lightweight neural networks**, 2021. 29

[76] THANH-HUNG VO, GUEE-SANG LEE, HYUNG-JEONG YANG, AND SOO-HYUNG KIM. **Pyramid With Super Resolution for In-the-Wild Facial Expression Recognition**. *IEEE Access*, **8**:131988–132001, 2020. 29

[77] JIAWEI SHI AND SONGHAO ZHU. **Learning to amend facial expression representation via de-albino and affinity**. *arXiv preprint arXiv:2103.10189*, 2021. 28, 29

[78] KAI WANG, XIAOJIANG PENG, JIANFEI YANG, DEBIN MENG, AND YU QIAO. **Region attention networks for pose and occlusion robust facial expression recognition**. *IEEE Transactions on Image Processing*, **29**:4057–4069, 2020. 29

[79] MINGXING TAN AND QUOC LE. **EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks**. In KAMALIKA CHAUDHURI AND RUSLAN SALAKHUTDINOV, editors, *Proceedings of the 36th International Conference on Machine Learning*, **97** of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 09–15 Jun 2019. 29

[80] XINLEI CHEN AND KAIMING HE. **Exploring Simple Siamese Representation Learning**. In *2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 27

[81] G. E. HINTON AND R. R. SALAKHUTDINOV. **Reducing the Dimensionality of Data with Neural Networks**. *Science*, **313**(5786):504–507, 2006. 27

[82] SHAN LI, WEIHONG DENG, AND JUNPING DU. **Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild**. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2584–2593, 2017. 28

[83] DAVIS E. KING. **Dlib-ml: A Machine Learning Toolkit**. *Journal of Machine Learning Research*, **10**(60):1755–1758, 2009. 29

[84] MARTIN HEUSEL, HUBERT RAMSAUER, THOMAS UNTERTHINER, BERNHARD NESSLER, AND SEPP HOCHREITER. **Gans trained by a two time-scale update rule converge to a local nash equilibrium**. *Advances in neural information processing systems*, **30**, 2017. 33

[85] NATANIEL RUIZ, EUNJI CHONG, AND JAMES M REHG. **Fine-grained head pose estimation without keypoints**. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 2074–2083, 2018. 33

[86] **FaceSwap**. Accessed 2022-02-18. 33

[87] JULIAN TODT, SIMON HANISCH, AND THORSTEN STRUFE. **Fant\^ omas: Evaluating Reversibility of Face Anonymizations Using a General Deep Learning Attacker**. *arXiv preprint arXiv:2210.10651*, 2022. 46

School of Information Technology

ISBN: 978-91-89587-36-6 (printed)
Halmstad University Dissertations, 2024

HALMSTAD
UNIVERSITY