

# A Framework for Evaluating Image Obfuscation under Deep Learning-Assisted Privacy Attacks

Jimmy Tekli<sup>1,2\*</sup>, Bechara Al Bouna<sup>3†</sup>, Gilbert Tekli<sup>4†</sup>  
and Raphaël Couturier<sup>2</sup>

<sup>1</sup>Tech Office Munich, BMW Group, Germany.

<sup>2</sup>FEMTO-ST, Univ. Bourgogne France-Comté, France.

<sup>3</sup>TICKET Lab, Antonine University, Lebanon.

<sup>4</sup>University of Balamand, Lebanon.

\*Corresponding author(s). E-mail(s): [jimmy.tekli@bmw.de](mailto:jimmy.tekli@bmw.de);

†These authors contributed equally to this work.

## Abstract

Image obfuscation techniques (e.g., pixelation, blurring and masking,...) have been developed to protect sensitive information in images (e.g. individuals' faces). In a previous work, we designed a recommendation framework that evaluates the robustness of image obfuscation techniques and recommends the most resilient obfuscation against Deep-Learning assisted attacks. In this paper, we extend the framework due to two main reasons. First, to the best of our knowledge there is not a standardized evaluation methodology nor a defined model for adversaries when evaluating the robustness of image obfuscation and more specifically face obfuscation techniques. Therefore, we adapt a three-components adversary model (goal, knowledge and capabilities) to our application domain (i.e., facial features obfuscations) and embed it in our framework. Second, considering several attacking scenarios is vital when evaluating the robustness of image obfuscation techniques. Hence, we define three threat levels and explore new aspects of an adversary and its capabilities by extending the background knowledge to include the obfuscation technique along with its hyper-parameters and the identities of the target individuals. We conduct three sets of experiments on a publicly available celebrity faces dataset. Throughout the first experiment, we implement and

## 2 *Evaluating Image Obfuscation under DL-Assisted Privacy Attacks*

evaluate the recommendation framework by considering four adversaries attacking obfuscation techniques (e.g. pixelating, Gaussian/motion blur and masking) via *restoration*-based attacks. Throughout the second and third experiments, we demonstrate how the adversary's attacking capabilities (*recognition*-based and *Restoration & Recognition*-based attacks) scale with its background knowledge and how it increases the potential risk of breaching the identities of blurred faces.

**Keywords:** Face obfuscation, Deep learning-assisted attacks, adversary model, background knowledge, image transformation, privacy-preserving techniques

### *Acknowledgement*

The authors especially thank Mr. Marc Kamradt for providing the GPU hardware available at the BMW TechOffice located in Munich to conduct all the experiments.

## Declarations

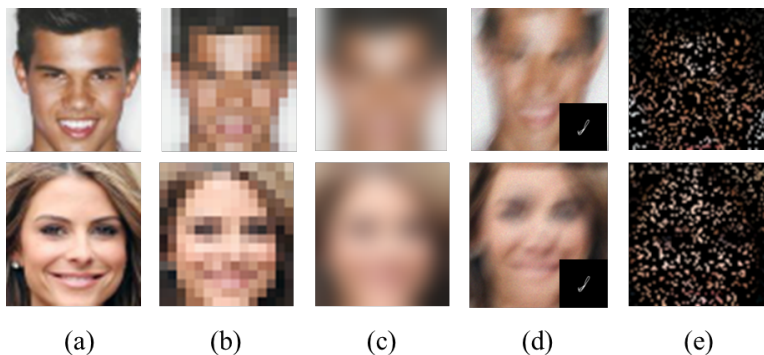
- **Funding:** Not applicable.
- **Conflict of interest:** On behalf of all authors, the corresponding author states that there is no conflict of interest.
- **Ethics approval:** Not Applicable.
- **Consent to participate:** Not Applicable
- **Consent for publication:** The authors consent that this paper can be published in case of acceptance.
- **Availability of data and materials:** The dataset used throughout our experiments is publicly available.
- **Code availability:** The code that we implemented is publicly available as GitHub repositories and is referenced.
- **Authors' contributions:** Both the second and the third author contributed equally to this study.

## 1 Introduction

From supply chain optimization to autonomous driving vehicles, manufacturing units and automotive companies are increasingly integrating Computer Vision (CV) applications to further improve the efficiency and the quality of their processes and products. These CV applications (e.g. object recognition [1, 2], object detection [3] and segmentation [4]) rely on learning-based techniques which require capturing images that might contain Sensitive Information (SI) such as individuals faces, workers belongings, or name tags. Due

to privacy regulations, these companies must guarantee a level of anonymization<sup>1</sup> that requires “irreversibility preventing identification of the data subject” by taking into account all the means “reasonably likely to be used” for identification. In order to preserve these SI, several obfuscation techniques like pixelation (also known as mosaicking) [5], blurring (Gaussian/motion) [5, 6] and masking can be used (c.f. Figure 1).

In a nutshell, obfuscation is done by altering/removing features from the images to hide SI while, at the same time, retaining some visual features to keep the image suitable for processing. However, these visual features can be used to identify/reconstruct the obfuscated SI via different attacks that can be classified as *recognition*-based [7–9] and *restoration*-based attacks [10–12].



**Fig. 1:** Obfuscation techniques left to right , (a) Original clear image, (b) pixelated image (4x), (c) Gaussian blurred Image, (d) motion blurred and (e) masking by adding random black pixels [15]

*Recognition*-based attacks breach the images privacy and anonymity by training learning-based algorithms to perform recognition tasks on obfuscated information [7]. *Restoration*-based attacks de-anonymize privacy-protected images by trying to restore/reconstruct the clear original features of the obfuscated information [10–12]. Last but not least, *Restoration & Recognition*-based (*R&R*) attacks combine both techniques in order to recognize restored features of the obfuscated information. Several studies showed that Deep Neural Networks outperform traditional learning-based approaches for image restoration and recognition tasks [1, 13, 14]. Hence, from a privacy perspective, these Deep Learning-based (DL) techniques are highly nominated as strong *recognition*-based and *restoration*-based attacks [7, 15–17].

As defined in [18] within the field of security, an adversary refers to an attacker, often with malicious intents, that undertakes an attack on a secure system to prevent or disrupt its proper operation. The authors in [18] formalized the adversary as a three-components model having a **goal**, **knowledge**

<sup>1</sup>Throughout the rest of this paper, we will use the terms obfuscation and anonymization interchangeably.

and **capabilities**. Let us consider an adversary who has access to a dataset of obfuscated faces belonging to certain individuals and its goal is to recover their identities. On the one hand, the adversary is capable of performing a *recognition*-based, a *restoration*-based or an *R&R*-based attack in order to extract the needed information from the anonymized faces. On the other hand, undertaking these attacks depends heavily on the adversary's knowledge with regard to the anonymized dataset, more specifically its background knowledge. For instance, the adversary should only be aware of the obfuscation technique employed in the anonymized dataset when performing a *restoration*-based attack. Whereas, she/he is capable of performing an identity *recognition*-based or an *R&R*-based attack only when equipped with knowledge related to the identities *present*<sup>2</sup> in the target anonymized dataset [16].

As stated in a recent review [19], existing obfuscation techniques in the context of images do not come with formal provable privacy guarantees. Hence, several evaluation frameworks have been proposed in the literature to evaluate empirically the obfuscation techniques in the context of images/videos. Some frameworks rely on human observers [20] whereas others [21–23] employ quantitative metrics, e.g. structural similarity metrics SSIM [24], recognition algorithms, etc. As classified by the authors in [19], these evaluations consider either (i) the efficiency of the privacy enhancement, (ii) the biometric utility preserved after privacy enhancement or (iii) the robustness to attempts to reverse the obfuscation techniques. In a previous work, we designed a quantitative recommendation framework that evaluates the robustness of image obfuscation techniques and recommends the most resilient obfuscation against DL-assisted attacks [15]. We assumed that the background knowledge of the adversary comprises the obfuscation technique and its hyperparameters. Hence, we performed *restoration*-based attacks. In this work, we extend the recommendation framework proposed in [15] by:

1. Embedding and adapting the three-components adversary model inspired from [18] to the context of facial image obfuscation. To the best of our knowledge and as mentioned in a recent review [19], there is not a standardized evaluation methodology nor a defined model for adversaries when evaluating the robustness of image obfuscation [7, 25] and more specifically face obfuscation techniques [15, 16].
2. Defining several threat levels with regard to the adversary's background knowledge which constitutes the obfuscation technique employed, its hyperparameters and the identities *present* in the target dataset (c.f. Table 2). As stated by the authors in [19], considering several attacking scenarios and exploring new aspects of the adversary is critical when evaluating the robustness of image obfuscation techniques.
3. Supporting *restoration*-based and *recognition*-based as well as *R&R*-based attacks (c.f. Figure 3).

---

<sup>2</sup>That possess obfuscated face images

We consider that the adversary’s **goal** is to recover the identity of the obfuscated faces while its **capabilities** (i.e., *restoration*-based, *recognition*-based or *R&R*-based attacks) depend heavily on its **background knowledge** (i.e., obfuscation technique and identities *present* in the target dataset). Inspired by Shannon’s Maxim<sup>3</sup>, we defined three threat levels  $T_1$ ,  $T_2$  and  $T_3$  with regard to the adversary’s background knowledge (i.e., “*our system*”). In  $T_1$ , we assume an adversary aware of the obfuscation technique used to obfuscate the target dataset along with its hyperparameters. In  $T_2$ , we assume an adversary aware (i) of the identities *present* in the target dataset and (ii) of the obfuscation technique used along with its hyperparameters. As for  $T_3$ , we assume an adversary aware (i) of the identities *present* in the target dataset and (ii) of the obfuscation technique used but not of its hyperparameters. We conduct three sets of experiments on a publicly available celebrity faces dataset [26]. Throughout the first experiment, we implement and evaluate the recommendation framework by considering four adversaries in  $T_1$  against four obfuscation techniques (e.g. pixelating, Gaussian/motion blur and masking). Throughout the second experiment, we demonstrate how the adversary’s attacking capabilities vary and scale with its knowledge in  $T_2$  and how it increases the potential risk of breaching the identities of blurred face images. *For instance, we re-identify 692 anonymized individuals out of 854 (81%) when simulating our strongest adversary in  $T_2$ .* Throughout the third experiment, we study the possible privacy breaches and the attack range of an adversary in  $T_3$  against face images blurred with different kernels. *For instance, we demonstrate that an adversary has the widest attack range against target datasets blurred with different blurring kernels when she/he prepares the training dataset of the recognition-based attacks with a blurring kernel of (37,37).*

**Table 1:** Acronyms

<i>CV</i>	Computer Vision
<i>SI</i>	Sensitive Information
<i>SR</i>	Super Resolution
<i>DL</i>	Deep Learning
<i>R&amp;R</i> -based attack	<i>Restoration &amp; Recognition</i> -based attack
<i>GAN</i>	Generative Adversarial Network
<i>GT</i>	Ground-truth image
<i>AN</i>	Anonymized image
<i>RC</i>	Reconstructed image
<i>BK</i>	Background Knowledge

The remainder of this paper is organized as follows. In Section 2, we review different obfuscation techniques. In Section 3, we extend the recommendation framework by adapting the adversary model to our application domain and considering three threat levels. Section 4 evaluates different face obfuscation techniques via the proposed framework and studies the effect of the background

---

<sup>3</sup>“*the enemy knows the system*”, i.e., “*one ought to design systems under the assumption that the enemy will immediately gain full familiarity with them.*”

knowledge on the adversary’s capabilities. In Section 5, we present how our recommendation framework can be extended to other SI and scaled to include different adversaries, DL-assisted attacks and evaluation metrics. In Section 6, we investigate works related to privacy attacks in the context of images and to evaluation frameworks. Last but not least, we conclude by discussing our study’s limitations and possible future work.

## 2 Obfuscation Techniques

Numerous obfuscation techniques have been proposed in the literature to hide/protect SI in images such as (i) the traditional techniques (e.g. pixelating, blurring, masking), (ii) the k-same methods [27] or (iii) the inpainting approaches [28, 29]. Nowadays, the majority of social media platforms, news agencies and publicly available research datasets still use the traditional techniques such as pixelating or blurring: *for instance, Google Maps [30] as well as the large-scale dataset nuScenes [31] published in 2019 for autonomous driving still employ blurring kernels to obfuscate individuals’ faces/homes or vehicle plates*. Therefore, we focus in this study on the following three obfuscation techniques: pixelating, blurring and masking.

### 2.1 Pixelating

Pixelating (a.k.a. mosaicking) is widely adopted as an obfuscation technique. The SI to be obfuscated is divided into a square grid, a.k.a. “a pixel box”. Each pixel box will have one color after averaging the values of the grouped pixels in it [5]. The size of the pixel box can be modified depending on the needed level of privacy. The larger the box, the more pixels will be averaged together, the higher the level of privacy. As stated in [7], although the size of the image stays the same, pixelating can be thought of as reducing the obfuscated section’s resolution. For instance, downscaling an image by a factor of four is equivalent to applying a pixel box of size 4x4 (c.f. Figure 1.b).

### 2.2 Blurring

Blurring is also a degradation technique utilized in image processing. It can be generated by a Gaussian kernel or via a camera motion effect, a.k.a. motion blur. A Gaussian like blur kernel is used extensively as an obfuscation technique [5]. It removes details from an image by applying a Gaussian kernel. A motion blur alters the details of an image by generating the effect of a synthetic camera motion blur [6]. The level of blurriness is affected by the length and the angle of the synthesized motion (c.f. Figure 1(c-d)).

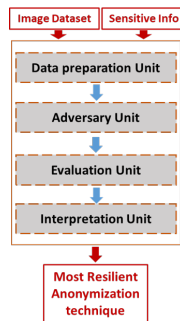
### 2.3 Masking

Masking removes details from an image by replacing the original pixels by black pixels. The masking technique can have multiple derivatives depending mainly on the color intensity and location of the altered pixels. For instance, if

an individual's face is considered sensitive, pixels can be modified around the eyes and nose or at random points of the face. The level of privacy depends on the amount, location and color intensity of the modified pixels. In our case, we consider random black pixels around the entire face (c.f. Figure 1.e).

### 3 Proposed Framework

In this section, we extend the recommendation framework proposed in [15] by defining a three-components adversary model with three threat levels and supporting *recognition*-based, *restoration*-based and *R&R*-based attacks. The framework attacks obfuscation techniques by restoring/recognizing hidden facial features, evaluates the reconstruction/recognition and suggests the most resilient obfuscation. It is mainly composed of four units: (a) a data preparation unit, (b) an adversary unit, (c) an evaluation unit and (d) an interpretation unit (c.f. Figure 2).



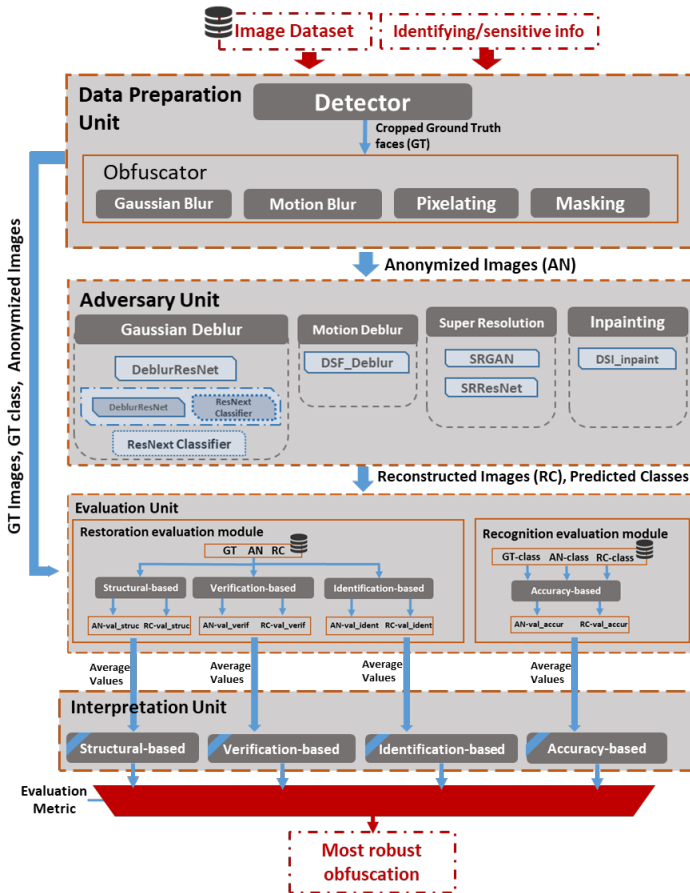
**Fig. 2:** The generic recommendation framework [15]

#### 3.1 Data preparation unit

The data preparation unit takes as inputs an image dataset along with the SI. It is divided into two modules: (a) *SI detector* and (b) *Anonymizer* (c.f. Figure 3). As its name indicates, the *SI detector* localizes and detects the SI in the image, crops it and sends it to the *Anonymizer*. As stated before, we consider in this study the faces as SI. Hence, the *SI detector* employs the OpenFace toolbox[32] to detect faces in an image, crop and forward it to the *Anonymizer*. In this study, the *Anonymizer* obfuscates the SI via: pixelating, blurring (Gaussian/motion) and masking techniques and sends the anonymized images to the adversary unit.

#### 3.2 Adversary unit

The adversary unit receives the obfuscated cropped face images. As shown in Figure 3, it is divided into four modules, one per obfuscation category: (a) the



**Fig. 3:** Extending the proposed generic framework in [15] to include additional adversaries, DL-assisted attacks and evaluation metrics

*super-resolution module* (for pixelating), (b) the *Gaussian deblur module* (for Gaussian blurring), (c) the *motion deblur module* (for motion blurring) and (d) the *inpainting module* (for masking). Each module contains one or more adversaries.

### 3.2.1 Adversary Model

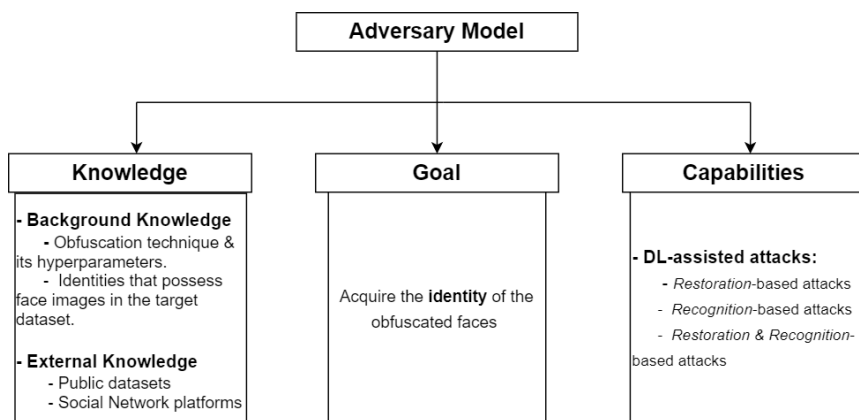
In our domain of application, an adversary undertakes an attack on obfuscated face images in order to extract particular information from the hidden facial features. Inspired by the authors in [18, 33–35], we define our adversary as a three-components model (c.f. Figure 4):

- **Adversary’s goal:** It refers to the adversary’s intentions and to the particular information she/he attempts to obtain/extract from the target



anonymized dataset. In our work, the adversary’s goal is to acquire the identity of the obfuscated faces.

- **Adversary’s knowledge:** The *background knowledge* is any sort of information regarding the anonymized dataset itself. In our work, (i) the obfuscation technique used to anonymize the target face images along with its hyperparameters and (ii) the identities *present* in the target dataset constitutes the background knowledge.
- **Adversary’s capabilities:** It represents to what extent can the adversary act in order to reach its goal, i.e. the *adversary’s abilities*. It depends on the adversary’s background knowledge. In our work, we consider that the adversary can perform a *restoration*-based, a *recognition*-based or a *R&R*-based attack.



**Fig. 4:** Adapting the three-components adversary model to the face obfuscation scenario

### 3.2.2 Threat Levels

In addition, we consider three threat levels based on the adversary’s background knowledge (c.f. Table 2):

- **Threat level T<sub>1</sub>:** assumes an adversary aware of the obfuscation technique used to protect the target dataset along with its hyperparameters.
- **Threat level T<sub>2</sub>:** assumes an adversary with full/partial knowledge about the identities *present* in the target dataset in addition to the obfuscation technique used and its hyperparameters.
- **Threat level T<sub>3</sub>:** assumes an adversary with full/partial knowledge about the identities *present* in the target dataset in addition to the obfuscation technique used without being aware of its hyperparameters.

**Table 2:** Comparing the adversary’s capabilities and knowledge with regard to the three threat levels

Adversary’s Components		Threat Levels		$T_1$	$T_2$	$T_3$
		External Knowledge	Public Datasets	✓	✓	✓
Goal		Identity/recover the identity of the obfuscated faces				
Knowledge	Background Knowledge	Obfuscation technique	✓	✓	✓	
		Obfuscation technique’s hyperparameters	✓	✓	✗	
		Identities present in the target dataset	✗	✓	✓	
Capabilities	DL-assisted attacks	Restoration-based attack	✓	✓	✓	
		Recognition-based attack	✗	✓	✓	
		Restoration & Recognition-based attack	✗	✓	✓	

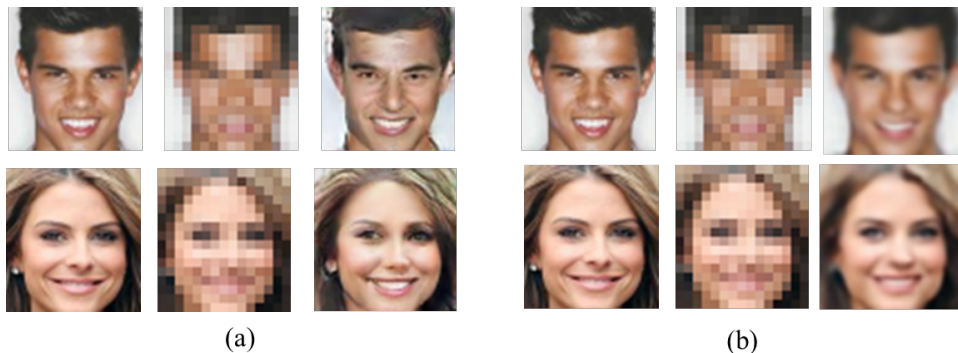
As we mentioned before, the adversary’s capabilities scale with regard to its background knowledge. Hence, the attacks that the adversary can perform vary between  $T_1$ ,  $T_2$  and  $T_3$ . We consider three attacks: (i) *restoration*-based, (ii) *recognition*-based and (iii) *Restoration & Recognition*-based (*R&R*-based) attacks.

- **Restoration-based attack:** de-anonymize obfuscated faces by trying to reconstruct the clear original features of the anonymized information. Training a Deep Neural Network to perform a *restoration*-based attack requires randomly gathering pairs of original/obfuscated face images. Hence, the adversary is capable of performing this sort of attack in  $T_1$ ,  $T_2$  and  $T_3$ <sup>4</sup>.
- **Recognition-based attack:** The adversary breaches the images privacy and anonymity by training learning-based algorithms to perform recognition tasks on obfuscated faces. An identity *recognition*-based attack requires gathering obfuscated face images for specific identities. Hence, the adversary can perform this attack in  $T_2$  and  $T_3$ .
- **Restoration & Recognition-based attack:** The adversary attempts to defeat the obfuscation technique via a two-steps attack: (1) reconstructing the hidden features of an obfuscated face and (2) trying to associate it with an identity by training an identity recognition model on clear face images. Therefore, only the adversaries in  $T_2$  and  $T_3$  can perform this two-steps process because it requires knowledge of the identities.

At the end of each attack, the adversary outputs either a reconstructed face (in case she/he performed a *restoration*-based or a *R&R*-based attack) or a predicted class label/probability (in case she/he performed a *recognition*-based

<sup>4</sup>In  $T_3$  the *restoration*-based attack could be less dangerous compared to  $T_1$  and  $T_2$  because the adversary is not aware of the exact hyperparameters of the obfuscation technique.

or a  $R\mathcal{E}R$ -based attack). Both ways, each face image has three derivatives: the clear, obfuscated and reconstructed class/face as shown in Figure 5.



**Fig. 5:** (a) Ground truth, anonymized and reconstructed images output via the SRGAN network. (b) Ground truth, anonymized and reconstructed image outputted via the SRResNet network [15]

### 3.3 Evaluation unit

The evaluation unit is divided into two main modules: (1) the restoration and (2) the recognition evaluation modules (c.f. Figure 3). The former assesses the reconstruction ratio of the *restoration*-based attacks whereas the latter measures the accuracy of the *recognition*-based attacks. As for the  $R\mathcal{E}R$ -based attacks, both the restoration and the recognition evaluation modules are employed.

#### 3.3.1 Restoration evaluation module

The restoration evaluation module assesses the face restoration with regard to *structural*, *verification* and *identification*-based metrics. Each metric-based sub-module receives as input three images: a clear face image (GT), an obfuscated face image (AN) and a reconstructed face image (RC).

- *The structural-based evaluation sub-module* quantifies the image enhancement/degradation quality after reconstruction attempts. In this study, we measure the holistic similarity between the clear image (GT) and the obfuscated image (AN) and between the clear image (GT) and the reconstructed image (RC) via SSIM<sup>5</sup> [24]. For normalization purposes, the *structural-based sub-module* computes the SSIM’s complement, i.e. 1-SSIM. Hence, the output values are between 0 and 1 where 0 means the two images are

---

<sup>5</sup>The Structural Similarity Index (SSIM) measures image quality modifications (enhancement/degradations)

identical:

$$AN\_value\_struc = 1 - SSIM(GT, AN) \quad (1)$$

$$RC\_value\_struc = 1 - SSIM(GT, RC) \quad (2)$$

- *The verification-based evaluation sub-module* validates the identity of a target face with a reference image. It mainly tries to conduct a 1-to-1 matching. In this study, we compute the identity distance via the OpenFace<sup>6</sup> toolbox [32] between the reference clear face images (GT) and both the obfuscated face image (AN) and the reconstructed face image (RC). OpenFace maps the two input faces to an identity distance between 0 and 4. The *verification-based evaluation module* normalizes the values to 0 and 1 where 0 value means that the two faces are identical, hence:

$$AN\_value\_verif = Normalized(OpenFace(GT, AN)) \quad (3)$$

$$RC\_value\_verif = Normalized(OpenFace(GT, RC)) \quad (4)$$

We employed in this study SSIM [24] and the OpenFace toolbox [32] because they are both publicly available and widely used in the literature to evaluate the reconstruction of degraded faces in the context of image transformation tasks[37, 38].

- *The identification-based evaluation sub-module* attempts to recognize an identifying feature of an individual based on a single face image. It mainly tries to compare the face in question with many others by conducting a 1-to-many matching. In our study, we employed a DL-based identity recognition model (c.f. Section 4.1.4). The *identification-based sub-module* uses the inferences over the three received images in order to compute two average relative error values<sup>7</sup>. The average relative error ranges between 0 and 1. We denote the class probability returned by the DL-based recognition model as *conf*.

$$AN\_value\_ident = \frac{|conf(GT) - conf(AN)|}{conf(GT)} \quad (5)$$

$$RC\_value\_ident = \frac{|conf(GT) - conf(RC)|}{conf(GT)} \quad (6)$$

In (5) and (6), both confidences in the numerator belong to the same *predicted class label*. In other words, in case the inferences of the recognition model over the obfuscated or reconstructed (*AN* or *RC*) image do not contain the GT class name, the *AN\_value\_ident* or the *RC\_value\_ident* would be 1.

*AN*-values and *RC*-values, outputted by the three restoration evaluation sub-modules, ranges between 0 and 1 where 0 indicates that the individual’s privacy is completely breached whereas 1 means that it is intact. Each

---

<sup>6</sup>OpenFace is a Python and Torch implementation of face recognition with deep neural networks [36]. OpenFace directly learns a mapping from face images to a compact euclidean space where distances directly correspond to a measure of face similarity.

<sup>7</sup>By definition, the average relative error is the absolute difference between the “exact theoretical” value and its ”measured” counterpart, divided by the “exact theoretical” value. We consider the inference over the clear face image (GT) as the “exact” value whereas the prediction over the anonymized (AN) and the reconstructed (RC) face images as the ”measured” values.

restoration evaluation sub-module computes the average *AN*-values and the *RC*-values over the entire obfuscated/restored dataset received from each DL-assisted attack and forwards them to the corresponding module in the interpretation unit.

### 3.3.2 Recognition evaluation module

The recognition evaluation module assesses the face (obfuscated or restored) recognition ratio with regard to an *accuracy*-based metric. The *accuracy-based sub-module* receives as input the class names and probabilities predicted (by *recognition*-based or *R<sup>ℓ</sup>R*-based attacks) over the obfuscated face image (*AN*-class) and the reconstructed face image (*RC*-class) along with the ground-truth label (*GT*-class).

- *Accuracy-based evaluation sub-module* measures the Top-n accuracy of the DL-based recognition models employed as *recognition*-based attacks by the adversaries. For each face image, the *accuracy-based sub-module* determines if the *GT* class label (*GT*-class) is equal to one of the top n predicted class labels<sup>8</sup> over the obfuscated (*AN*-class) and the reconstructed (*RC*-class) faces. After analyzing the entire obfuscated/restored dataset, the sub-module outputs the *AN-value-accur* and the *RC-value-accur*. For normalization purposes, we compute the fraction and the complement of the Top-n accuracy. Hence, the output values are between 0 and 1 where 0 means that the recognition model used by the adversary was highly accurate<sup>9</sup>, i.e. the individual's anonymity is completely breached.

## 3.4 Interpretation unit


The interpretation unit selects the most robust obfuscation techniques per evaluation metric based on the results provided by the evaluation unit. As seen in Figure 3, the interpretation unit is divided into four *selection modules*, one per evaluation metric: (a) *structural-based*, (b) *identification-based*, (c) *verification-based* and (d) *accuracy-based selection module*. Each module performs a two-steps comparison in order to select the most resilient obfuscation technique: (1) *intra-attack* and (2) *inter-attack* comparisons (e.g., the *structural-based selection module* selects the most resilient obfuscation with regard to the SSIM metric whereas the *verification-based selection module* selects the most resilient obfuscation with regard to the Openface identity distance metric). As a first step, the *intra-attack* comparison allows us to identify the strongest DL-assisted attack against each obfuscation technique with regard to each evaluation metric. In other words, the attack that restored/recognized most of the obfuscated face images. As a second step, the *inter-attack* comparison chooses the most resilient obfuscation against the selected DL-assisted attacks. A detailed example is showcased in Section 4.1.4.

---

<sup>8</sup>Class labels with the Top n highest probabilities.

<sup>9</sup>Top-n accuracy is 100%.

**Table 3:** The different adversaries considered for the first experimental setup

Threat Levels			$T_i$			
			Adversary 1	Adversary 2	Adversary 3	Adversary 4
Adversary's Component			Identity/recover the <i>identity</i> of the obfuscated faces			
Goal			Identity/recover the <i>identity</i> of the obfuscated faces			
Knowledge	External Knowledge	Public Datasets	CelebA	CelebA	CelebA	CelebA
	Background Knowledge	Obfuscation Technique	Pixelation	Gaussian blurring	Motion blur	Masking
		Obfuscation Technique's hyperparameters	4x4	(31,31)		Random
		Identities known by the adversary	✗	✗	✗	✗
Capabilities	DL-assisted attacks	Restoration-based attack	SRGAN[36] SRResNet [25]	DeblurResNet[25]	DSF_Deblur[39]	DSI_inpaint[26]
		Recognition-based attack	✗	✗	✗	✗
		Restoration & Recognition-based attack	✗	✗	✗	✗

In this section, we described how our recommendation framework recommends the most resilient obfuscation technique via the 4-layered iterative workflow: (a) detecting/obfuscating the SI, (b) restoring/recognizing via the DL-assisted attacks performed by the adversaries, (c) evaluating the reconstruction/recognition and (d) selecting the most robust obfuscation based on the inter/intra-attack comparisons.

## 4 Experiments

To validate and assess our approach, we set up our experiments to (i) evaluate the recommendation framework (c.f. Section 4.1) and study thoroughly the effect of the background knowledge on the adversary's capabilities with regard to (ii) the identities *present* in the target dataset (c.f. Section 4.2) and (iii) the obfuscation technique (c.f. Section 4.3). Throughout the three experiments, we considered that the adversaries have the same goal: identify/recover the identities of the obfuscated faces.

### 4.1 Evaluating the recommendation framework

In this experimental setup, we evaluate our recommendation framework by considering four obfuscation techniques: pixelation, Gaussian blur, motion blur and masking.

#### 4.1.1 Input Dataset & SI

In order to prepare our evaluation test dataset, we select<sup>10</sup> 370 face images from the official CelebA test set. Our test set contains face images belonging to male and female celebrities of different races and different age (majority are

<sup>10</sup>We first selected 1307 images from the official CelebA test set, then we filtered out, via a pre-trained celebrity recognition model, the faces that were wrongly recognized or correctly recognized with a probability lower than 0.7

above 18). To normalize our experimental setup, we use the same face images to evaluate the different DL-assisted attacks. The training sets vary between DL-assisted attacks, however, no face images from the test set were included throughout the training of any of the DL models.

#### 4.1.2 Obfuscation techniques

We employed in this setup four obfuscation techniques: (1) pixelation (a.k.a. mosaicking), (2) Gaussian blur, (3) motion blur and (4) masking. We specified for each obfuscation technique a fixed parameter as shown in Table 3. Regarding the pixelation, we simply downscaled the face images by a factor of 4. For the Gaussian blur, we applied a Gaussian filter with a kernel size (31,31) and standard deviation of 5. As for the motion blur, we synthesized a motion blur kernel from random 3D camera trajectories [6]. Regarding the masking technique, we replaced random pixels all over the image by black pixels. As seen in Figure 1, the different obfuscation techniques guarantee "visually" the anonymity of the target identities.

#### 4.1.3 Adversaries & DL-assisted attacks

We simulated four adversaries in  $T_i$  who perform 5 *restoration*-based attacks against the four obfuscation techniques (c.f. Table 3).

For the Super Resolution (SR) task, we considered that the adversary performed two *restoration*-based attacks against pixelation: SRResNet and SRGAN. On the one hand, the SRResNet is a ResNet-based architecture [2] and is considered a benchmark when it comes to SR algorithms [13, 39]. Moreover, SRResNet is a generic SR-network applicable to our faces dataset<sup>11</sup>. On the other hand, SRGAN is a GAN-based super resolution model implemented by [40] similar to [41]. The model was developed specifically for faces. We generated the training pairs by downsampling the unobfuscated (GT) face images by a factor of four and trained both networks from scratch.












For the deblur task, we considered two distinct adversaries against two distinct blurring techniques. Regarding the Gaussian blur, we adapted the SRResNet architecture by modifying the input size of the network implemented in [42] (i.e., DeblurResNet). In addition, we generated the training pairs by applying Gaussian blur to the unobfuscated (GT) face images and trained the network from scratch. As for the motion blur, we used the implementation and the pre-trained model provided by the authors (i.e., DSF\_Deblur) [43].

Last but not least, we considered that the adversary applied the deep generative model DCGAN proposed in [44] and the implementation in [45] to attack the masking technique (i.e., DSI\_inpaint). We trained the DCGAN network on our face dataset from scratch. Table 4 summarizes the technical details regarding each DL-assisted attack.

---

<sup>11</sup>The implementation [42] provided a network which upscales the input image by a factor of 2. Hence, we added an upscaling function and re-trained it from scratch for upscaling by a factor of 4.

**Table 4:** Technical details regarding the obfuscation techniques and the implementations of the DL-assisted attacks [15]

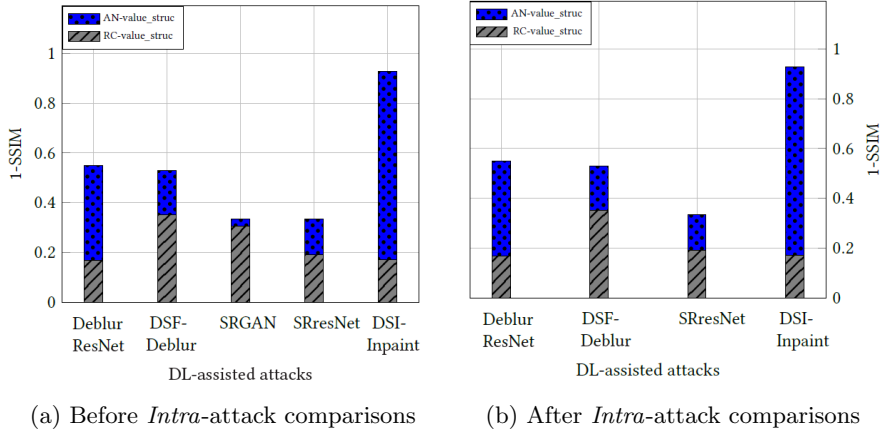
Ground Truth face	Obfuscating technique	hyperparameters	Parameter values	Obfuscated faces	Restoration-based attacks	Implementation/Framework	Results
	Pixelation	Pixel Box Size	4x4		SRResNet [39]	TensorFlow [42] Trained from scratch	
					SRGAN [40]	TensorFlow [40] Trained from scratch	
	Gaussian Blur	Kernel Size	(31,31)		DeblurResNet [39]	Tensorflow [42] Trained from scratch	
	Motion Blur	Length and angle of the motion			DSF_deblur [37]	Matcaffe / Matlab [43] Pre-trained model	
	Masking	Location of the black pixels	Random		DSL_inpaint [44]	TensorFlow [45] Trained from scratch	

#### 4.1.4 Evaluation & Interpretation

As stated in section 3.3, the framework provides (i) structural, (ii) verification and (iii) identification-based evaluations to assess the reconstruction ration of the *restoration*-based attacks. Each evaluation sub-module in our framework computes two metric-based values for each clear image: (a) *AN*-value and (b) *RC*-value. These values range between 0 and 1 where 0 indicates that the individual’s privacy is completely breached. In the following sections, we report the average values over the entire test set.

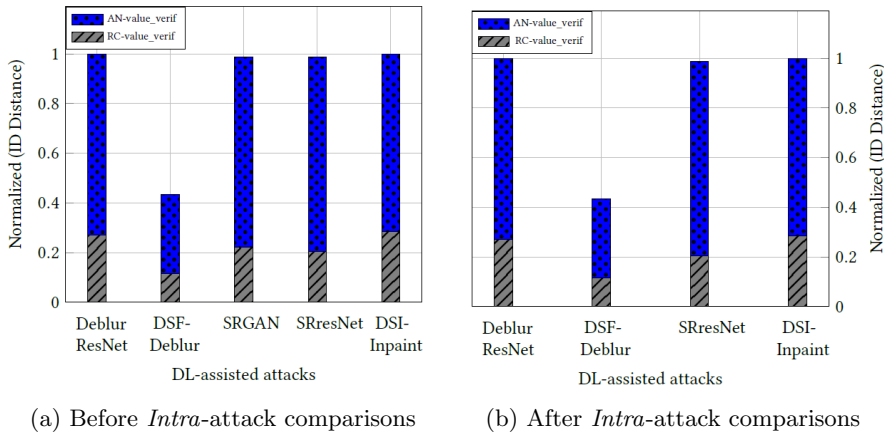
- *Structural-based evaluation (c.f. Section 3.3.1)*: As shown in Figure 6.(a), the average *RC\_values\_struct* of all the DL-assisted attacks are lower than the average *AN\_values\_struct* since the reconstructed RC face images are overall more similar to the clear GT face images than the obfuscated AN face images in terms of SSIM. As mentioned in Section 3.4, the interpretation unit executes the intra/inter-attack comparisons in order to select the most resilient obfuscation. First, the intra-attack comparison selects the strongest DL-assisted attack against each obfuscation technique. For instance in our case, all adversaries performed a single DL-assisted attack against each obfuscation except “Adversary 1” (c.f. Table 3) which performed two DL-assisted attacks against the pixelation technique: (a) SRGAN and (b) SRResNet attacks. Therefore, the intra-attack comparison selects the attack that caused the highest privacy breach against pixelation, i.e. the SRResNet attack because it resulted in the lowest *RC\_value\_struct* as seen in Figure 6.(a) when compared to SRGAN. Furthermore, the inter-attack comparison selects the most resilient obfuscation technique, i.e. the obfuscation whose DL-assisted attack records the highest *RC\_value\_struct*. As seen in Figure 6.(b), the “DSF-Deblur” attack records the highest *RC\_value\_struct*, as such, “motion blur” is the most resilient obfuscation with regard to the SSIM metric.





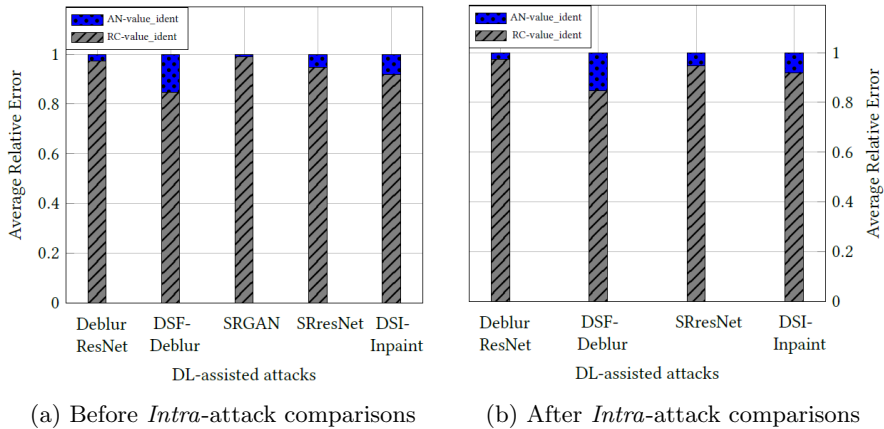
**Fig. 6:** The *Structural-based evaluation sub-module* output before and after the intra-attack comparisons [15]

- *Verification-based evaluation* In Figures 7.(a,b), we report the average *RC\_val\_verif* and *AN\_val\_verif* values. The intra/inter-attack comparisons select "masking" as the most resilient obfuscation technique with regard to the identity distance metric because the corresponding "DSI\_Inpaint" attack recorded the highest *RC-value\_verif* in Figure 7.(b).



**Fig. 7:** The *Verification-based evaluation sub-module* output before and after the intra-attack comparisons [15]

- *Identification-based evaluation* In Figures 8.(a,b), we report the average *AN\_values\_ident* and *RC\_values\_ident*. The intra/inter-attack comparisons select “Gaussian blur” as the most resilient obfuscation technique with regard to the identification-based metric because the “DeblurResNet” attack recorded the highest *RC\_values\_ident* in Figure 8.(b).



**Fig. 8:** The *Identification-based evaluation sub-module* output before and after the intra-attack comparisons [15]

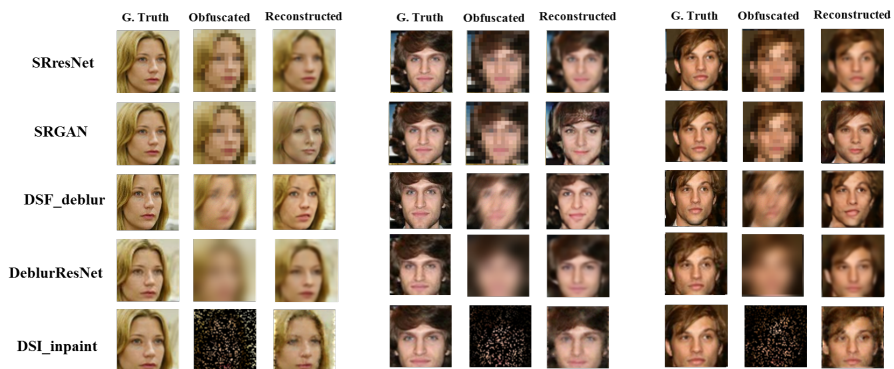
The “masking” technique would be the most robust obfuscation with regard to the different metrics, even after reconstruction attempts, if we were to block the entire face image with black pixels. In this study, we are masking the face image by randomly placing black pixels and leaving some original pixels intact (c.f. Figure 1.(e)), hence our results.

#### 4.1.5 Comparison with other evaluation frameworks

On a different note, the authors in [20] considered a human-based evaluation where they showed each participant an obfuscated face image and a couple of clear face images in order for her/him to match them up and guess the obfuscated identity. If we were to apply the same scenario to our four obfuscations, it would be more difficult for a human to re-identify an identity masked with random black pixels in comparison to the pixelated or blurred identities (as seen via the obfuscated face images in Figure 9). Nevertheless, after reconstructing the obfuscated faces via *restoration*-based attacks, the “masking” technique becomes also vulnerable to the human visual system (as seen via the reconstructed face images in Figure 9). Similarly, if we were only to evaluate the obfuscation techniques like the authors did in [21–23] where they compared the obfuscated image to the original image via quantitative metrics, then “masking” would be the most resilient obfuscation with regard to

the different evaluation metrics as we notice when observing the AN-values in Figures 6,7,8 (e.g. the AN-values of the "masking" technique are always 1). However, we notice that it is not always the case after performing the *restoration*-based attacks: for instance when observing Figure 6.(b), we notice that the *RC-value\_struct* of "masking" is lower than "pixelation" and "motion blur", i.e. the reconstruction of the masked face images was better with regard to the SSIM metric compared to the pixelated and motion blurred face images therefore making it more vulnerable. These two observations stress the importance of employing *restoration*-based attacks when evaluating the robustness of an obfuscation technique.

As a future step, we would like to add a *human-based evaluation sub-module* similar to what the authors did in [20]. Therefore, we would have the possibility to select the most resilient obfuscation technique after reconstruction attempts with regard to the human visual system as well.



**Fig. 9:** Comparison of the different reconstructions. Columns from left to right include Ground truth, Obfuscated and Reconstructed faces. Rows from top to bottom include the DL-assisted attacks [15]

## 4.2 Studying the effect of the background knowledge regarding the *known* individuals

In this experimental setup, we show how the adversary's knowledge regarding the identities *present* in the target dataset affects its capabilities. In the earlier section, we identified "masking", "motion blur" and "Gaussian blur" as the most robust obfuscations with regard to the different evaluation metrics in our framework. In this section, we focus on the "Gaussian blur" obfuscation technique as it is still the most widely used technique for privacy preservation purposes [30, 31, 46].

### 4.2.1 Data Preparation

In this setup, we needed to train identity recognition models in order to perform *recognition*-based and *R&R*-based attacks. Hence, we had to gather face images for each identity. Although the CelebA dataset is not designated for identity recognition tasks, we used it for training and evaluating the DL-assisted attacks<sup>12</sup>[26]. We selected<sup>13</sup> 854 identities from the CelebA dataset and we gathered 60 face images for each celebrity<sup>14</sup>. Out of these 60 images, 5 were left for testing and the remaining 55 were used for training purposes. Therefore, our test set contained 4270 face images (854 selected individuals x 5 test images) which are not part of the official CelebA test set (58% are female and 42% are male). We resized all the images to 64x64 and then applied the blurring function with a kernel size (31x31) and standard deviation of 5 (c.f. Figure 1.(c)).

### 4.2.2 Incremental Background Knowledge

In order to simulate the adversary in  $T_2$ , we designed an adversary with an incremental background knowledge regarding the number of identities *present* in the target dataset. We denoted as  $N$  the set of identities known by the adversary. We varied  $|N|$  between 0 (no knowledge about the identities *present* in the target dataset, i.e.  $T_1$ ) and 854 (Full knowledge, i.e.  $T_2$ )<sup>15</sup>. In total, we considered 10 distinct values for  $|N| = \{0, 100, 200, \dots, 800, 854\}$ .

### 4.2.3 Adversary & DL-assisted attacks

As mentioned in Section 3.2.2, the adversary in  $T_2$ , is capable of executing either a (i) *R&R*-based or a (ii) *recognition*-based attack.

On the one hand, the *R&R*-based attack is a combination of a DL restoration model followed by a DL identity recognition model. Regarding the restoration model, we trained the same DeblurResNet network [39, 42] as in section 4.1.3. The only difference is that we included in the training set 10 pairs of clear/obfuscated face images for each identity in  $N$ . As for the recognition model, the adversary uses the remaining 45 clear face images of each identity in  $N$  to train a SEResNext101<sup>16</sup> classifier with  $|N| + 1$  classes<sup>17</sup> and attempt to recognize reconstructed faces [71, 72].

On the other hand, the *recognition*-based attack tries to associate each anonymized face image with an identity (bypassing the reconstruction process). Therefore, the adversary obfuscates the 55 face images of each identity in  $N$  and train a SEResNext101-based classifier with  $|N| + 1$  classes in order to

---

<sup>12</sup>For instance, we did not employ the FaceScrub dataset [60], which is designated for identity recognition tasks, because the number of identities is limited to 530 whereas it is 10,177 in the CelebA dataset.

<sup>13</sup>for additional details regarding the data preparation process, please contact jimmytekli@hotmail.com

<sup>14</sup>we mined images from google via *google-images-download* as well.

<sup>15</sup>854 being the maximum number of individuals in our test set

<sup>16</sup><https://github.com/BMW-InnovationLab/BMW-Classification-Training-GUI>

<sup>17</sup>In addition to the classes regarding the individuals in  $N$ , we also added an additional class to our classifier entitled “others” which grouped 800 images that belong to other individuals

**Table 5:** Technical details regarding the DL-based models employed as *restoration*, *recognition* and *R&R*-based attacks

Adversary's Component		Threat Levels	$T_1$	$T_2$				
Goal		Identity/recover the <i>identity</i> of the obfuscated faces						
Knowledge	External Knowledge	Public Datasets	CelebA & Google Images which resulted to 55 face images for each identity in N					
	Background Knowledge	Obfuscation Technique & hyper-parameters	Gaussian Blurring (31,31)					
		Set of identities N known by the adversary	0	100	200	...	854	
Capabilities	DL-assisted attacks	<i>Restoration-based attack</i>	DeblurResNet [18] trained on randomly chosen face images from CelebA dataset.	✗	✗	...	✗	
		<i>R&amp;R-based attack</i>	<i>Restoration task</i>	✗	DeblurResNet trained with a dataset that includes 10 face images of each of the 100 identities in N	DeblurResNet trained with a dataset that includes 10 face images of each of the 200 identities in N	...	DeblurResNet trained with a dataset that includes 10 face images of each of the 854 identities in N
			<i>Recognition task</i>	✗	SEResNext101 Classifier with 101 classes where the first 100 classes contain 45 clear face images of each identity in N and the additional class contains 800 face images of other identities	SEResNext101 Classifier with 201 classes where the first 200 classes contain 45 clear face images of each identity in N and the additional class contains 800 face images of other identities	...	SEResNext101 Classifier with 854 classes where the first 854 classes contain 45 clear face images of each identity in N and the additional class contains 800 face images of other identities
		<i>Recognition-based attack</i>	✗	SEResNext101 Classifier with 101 classes where the first 100 classes contain 55 obfuscated face images of each identity in N and the additional class contains 800 face images of other identities	SEResNext101 Classifier with 201 classes where the first 200 classes contain 55 obfuscated face images of each identity in N and the additional class contains 800 face images of other identities	...	SEResNext101 Classifier with 854 classes where the first 854 classes contain 55 obfuscated face images of each identity in N and the additional class contains 800 face images of other identities	

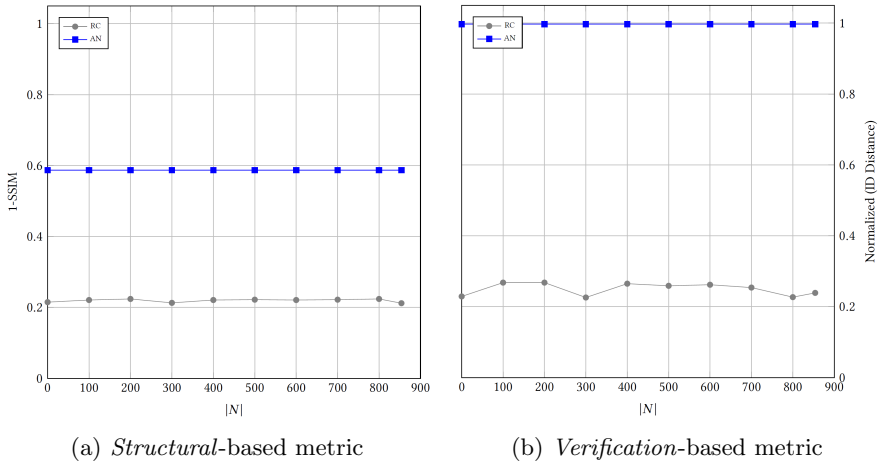
recognize obfuscated face images. We applied transfer learning to our classifier network by employing ImageNet pre-trained weights and also augmented our training datasets by randomly flipping, resizing and adding noise (e.g. color variations and saturation) to the face images.

We simulated for each value of  $|N|$  the corresponding attacking capabilities hence we trained 1 *restoration*-based, 9 *R&R*-based and 9 *recognition*-based attacks as seen in Table 5.

#### 4.2.4 Results and Interpretations

We show how the incremental background knowledge with regard to the identities *present* in the target dataset affect the adversary's capabilities. Our results show that:

- The incremental background knowledge does not affect the reconstruction accuracy of the restoration models in the *R&R*-based attacks.
- The incremental background knowledge increases the accuracy of the recognition models in both the *R&R*-based and the *recognition*-based attacks, i.e. increases the privacy breaches.
- The adversary is more dangerous when performing *recognition*-based attacks compared to *R&R*-based attacks.



**Fig. 10:** Effect of the background knowledge on the reconstruction quality with regard to the structural and verification-based evaluation sub-modules

As stated in Section 3, our framework provides structural and verification-based evaluations regarding the *restoration*-based attacks. In the following part, we measure the AN-values and the RC-values of the restoration models in the *restoration*-based and *RER*-based attacks for each value of  $|N|$ .

- **The incremental background knowledge does not affect the reconstruction accuracy of the restoration models in the *RER*-based attacks:** We notice in Figure 10 for the different values of  $|N|$  that (a) the *RC\_values\_struct* and (b) *RC\_values\_verif* values are stable with minor fluctuations. This demonstrate that increasing  $|N|$  does not increase nor affect the reconstruction accuracy of the restoration models with regard to the SSIM [24] and the Openface [32] evaluation metrics. In other words, even if the adversary knows the identity of a particular individual in the target dataset, adding face images of this particular individual to the training set of the restoration model (DeblurResNet [42] in our case) does not affect its reconstruction accuracy. This behavior should be further investigated when employing other restoration models (e.g. [40, 44]) as DL-assited attacks. Also, another experiment could study the effect of modifying the number of face images per individual in the training set instead of considering a fixed number as we did (10 face images per individual).

In the following part, we count the number of individuals who were re-identified and whose anonymity was breached. As we mentioned before, our test set contains 5 anonymized face images per identity. Hence, we consider that an individual is re-identified if  $L$  face images out of 5 are correctly recognized (Top-1 recognition) where  $0 < L \leq 5$ . In the following, we report the values for  $L = 2$ .

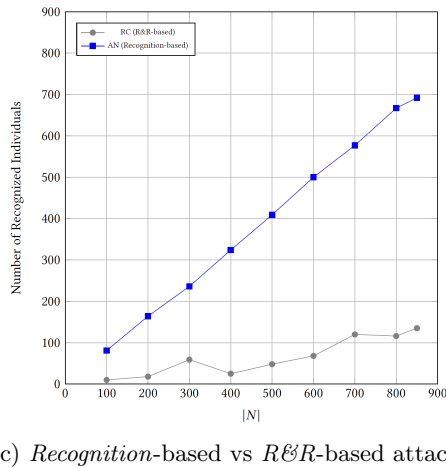
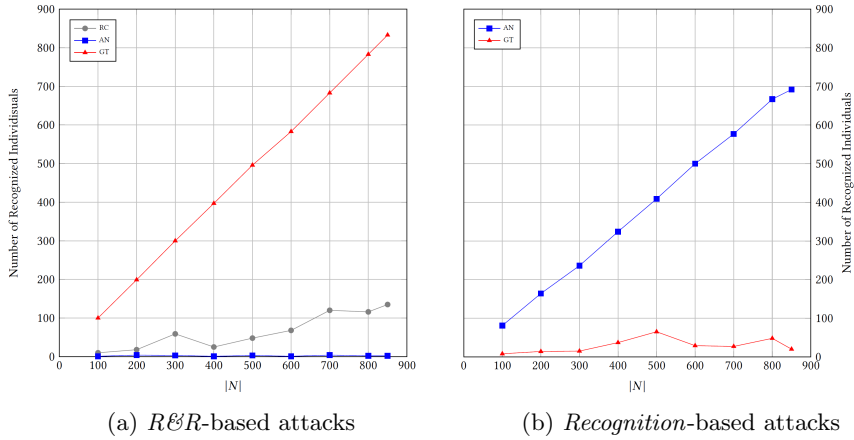
- **The incremental background knowledge increases the accuracy of the recognition models in both the  $R\mathcal{E}R$ -based and the *recognition*-based attacks, i.e., increases the privacy breaches:** In Figure 11.(a), we count the number of individuals re-identified by the  $R\mathcal{E}R$ -based attacks. Because the recognition model is trained on clear face images, the GT curve serves as a reference. We report that the number of re-identified individuals with regard to the RC images increased along with the background knowledge of the adversary. For  $|N|=100$ , the adversary re-identified, via the  $R\mathcal{E}R$ -based attack, 10 out of 854 (1.2%) individuals after reconstruction whereas at  $|N|=854$  she/he recognized 135 individuals (15.8%). The increase in the privacy breach is mainly due to the increase of the recognition models' accuracy because as we notice in Figure 10, the reconstruction accuracy of the restoration-based attacks does not change despite the incremental BK.

In addition, in Figure 11.(b) we report the number of individuals re-identified by the *recognition*-based attacks. We notice a steady increase in the number of re-identified individuals with regard to the AN face images along with the background knowledge. At  $|N|=854$ , the adversary re-identified 692 individuals out of 854, i.e., almost 81% of the anonymized individuals. The *recognition*-based attacks demonstrate poor results when inferring over clear (GT) face images in Figure 11.(b) because the identity recognition models are trained via obfuscated face images.

- **The adversary is more dangerous when performing *recognition*-based attacks compared to the  $R\mathcal{E}R$ -based attacks:**

When comparing the adversary's capabilities in Figure 11.(c), we notice that when equipped with  $|N|=100$  as background knowledge, the adversary re-identified 10 individuals when performing a  $R\mathcal{E}R$ -based attack whereas she/he re-identified 81 when performing a *recognition*-based attack. The same behavior persists throughout the incremental process of the background knowledge. When equipped with  $|N|=854$  as background knowledge, the adversary re-identified 135 out of 854 (15.8%) when performing an  $R\mathcal{E}R$ -based attack whereas she/he re-identified 692 (79.8%) when performing a *recognition*-based attack. We notice such a difference regarding the number of re-identified individuals because the recognition models of the  $R\mathcal{E}R$ -based attacks are trained on clear images while inferring over reconstructed images. Including obfuscated version of each face image in the training dataset of the recognition models might/should increase the recognition accuracy of the restored image as well as the privacy breach making the adversary more dangerous.

In this experimental setup, we demonstrate that it is not sufficient to protect face images in a target dataset by obfuscating them via certain techniques (e.g. blurring). However, it is also vital to protect any information (e.g. statistical information) about the target dataset that could be published, such as the identities of the faces present in the target dataset or information about the obfuscation technique employed, which could enforce the BK of the adversary








**Fig. 11:** Counting and comparing the number of recognized individuals in the test set when performing *R&R*-based and *recognition*-based attacks

and lead to higher privacy breaches. Furthermore, publishing quasi-identifying information (e.g. the gender or race distribution of the face images in a target dataset) could also lead to privacy breaches [47]. For instance, the adversary could perform DL-assisted attacks to recognize the gender or the race of the target individual instead of the full identity which might lead as well to potential privacy breaches when linked to other data sources (i.e., identity disclosure via linking attacks[19]).



**Table 6:** The seven target datasets blurred with distinct  $k_{test}$  values

Original	(19,19)	(25,25)	(31,31)	(37,37)	(43,43)	(49,49)	(55,55)
							

### 4.3 Studying the effect of the background knowledge regarding the obfuscation technique

In this experimental setup, we show how the background knowledge with regard to the obfuscation technique and its hyper-parameters affects the adversary’s capabilities. Similar to the previous section, we consider “Gaussian blur” as the obfuscation technique.

#### 4.3.1 Data Preparation

We selected randomly 100 identities from the dataset prepared in Section 4.2.1. For each identity, 55 face images were left for training purposes and 5 images were left for testing (i.e. 500 images in the test set). We prepared 7 different versions of the target dataset, each blurred with a kernel from  $k_{test} = \{19, 25, 31, 37, 43, 49, 55\}$  (c.f. Table 6).

#### 4.3.2 Background knowledge

We consider threat level  $T_s$ , i.e., the adversary is aware of the obfuscation technique employed in the test/target dataset (e.g., Gaussian blur) however not of its hyper-parameters (e.g., the blurring kernel’s size). In addition, we consider the adversary is aware of the identities in the target dataset (i.e.,  $N = 100$ ).

#### 4.3.3 Adversary & DL-assisted attacks

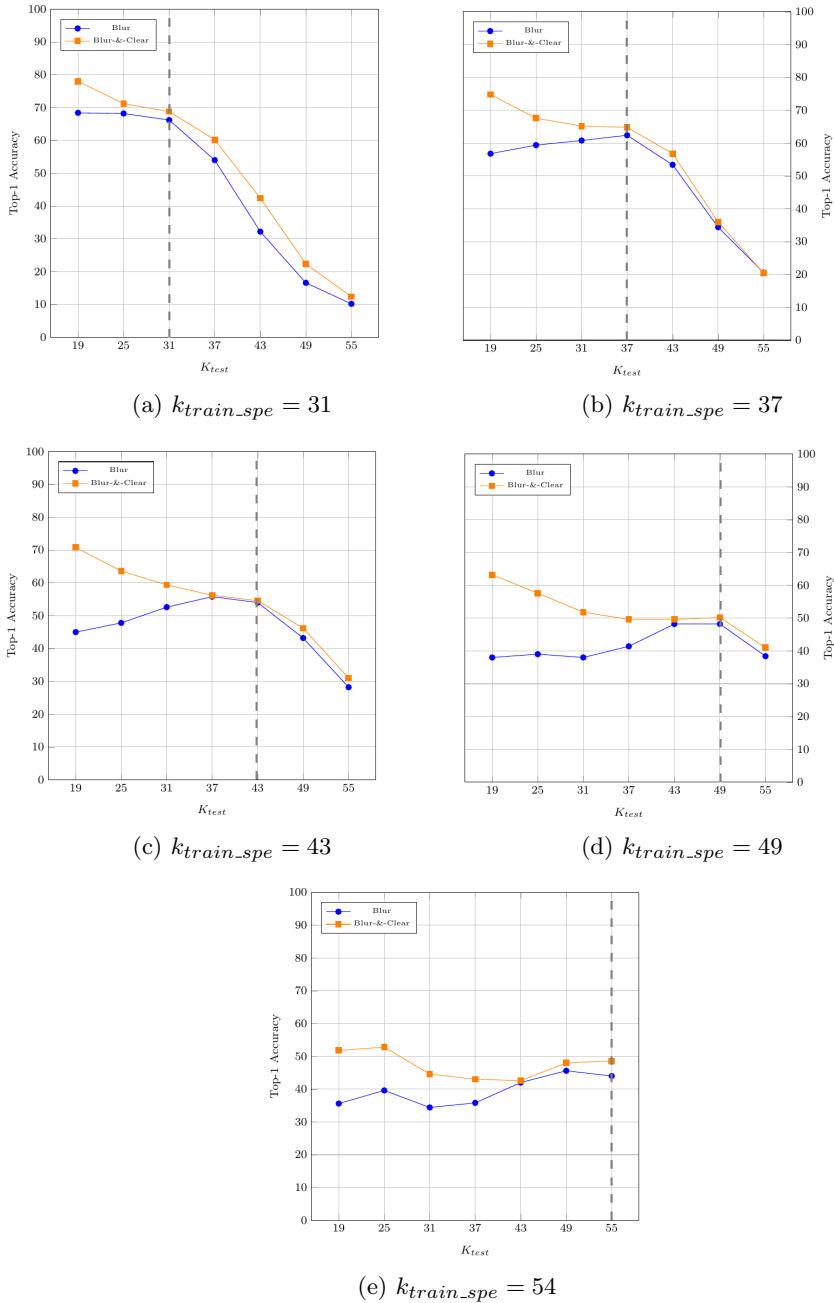
We perform *recognition*-based attacks as it is more dangerous compared to the *R&R*-based attacks as demonstrated in the earlier section (c.f. Section 4.2.4). We employ the same SEResNext101 classifier and training parameters used in section 4.2.3. In  $T_s$ , the adversary can choose any blurring kernel and prepare the training dataset accordingly because she/he is not aware of the blurring kernel used to obfuscate the target dataset. Hence, we trained 5 *recognition*-based attacks, each with a distinct kernel from  $k_{train} = \{31, 37, 43, 49, 55\}$ . We report the privacy breaches of each attack against the 7 target datasets blurred via  $k_{test}$ . Last but not least, we considered two training modes for each *recognition*-based attack: (i) “**blur-&-clear mode**” where the training set contains clear and blurred version of each face image and (ii) “**blur mode**” where the training set contains blurred face images only.

### 4.3.4 Results and Interpretations

In the following section, we show that:

- The adversary must not know the exact blurring kernel of a target dataset in order to breach its anonymity.
- The privacy breaches decrease steadily in a linear fashion when attacking face images blurred with kernels greater than the kernel chosen by the adversary while preparing its training dataset.
- Including both, clear and blurred images in the training datasets increases the recognition accuracy of the *recognition*-based attacks, specifically when the target dataset's blurring kernel is smaller than the training dataset's.
- Preparing the training dataset with blurring kernel (37,37) provides the widest attack range against the 7 target datasets.

Each subfigure in Figure 12 reports the Top-1 accuracy of the *recognition*-based attacks (**blur-&-clear** and **blur** modes) trained with a specific kernel  $k_{train\_spe}$  and attacking the 7 target datasets. For instance, Figure 12.(a) corresponds to the *recognition*-based attacks (**blur-&-clear** and **blur** modes) trained on face images blurred via  $k_{train\_spe}=(31,31)$ . We report the following observations:



**Fig. 12:** Top-1 accuracy of the *recognition*-based attacks (**blur-&-clear** and **blur** modes) trained with a specific kernel  $k_{train\_spe}$  and attacking the 7 target datasets.

- **The adversary must not know the exact blurring kernel of the target dataset in order to breach its anonymity.** For instance in Figure 12.(e), we notice that the adversary breached the faces' anonymity 46% (**blur-&-clear** mode) and 36% (**blur** mode) of the time on average against the 7 target datasets despite training the recognition models with kernel size  $k_{train\_spe}=(55,55)$ . Similar behavior is observed, although with different magnitudes, for the other  $k_{train\_spe}$  values as well.
- **The privacy breaches decrease steadily in a linear fashion when attacking face images blurred with kernels greater than  $k_{train\_spe}$ .** As we notice in Figure 12.(a), the Top-1 accuracies (for both **blur-&-clear** and **blur** modes) decrease steadily when the adversary attacks target datasets blurred with kernels greater than  $k_{train\_spe}=(31,31)$ . The same behavior clearly persists in Figures 12.(b) and 12.(c) for  $k_{train\_spe}=(37,37)$  and  $k_{train\_spe}=(43,43)$  respectively.
- **Including both, clear and blurred images in the training datasets increases the recognition accuracy of the *recognition*-based attacks, specifically when attacking a target dataset obfuscated with a kernel smaller than  $k_{train\_spe}$ .** When trained via the **blur** mode, the highest privacy breach occurs when the training and the target datasets are blurred with the same kernel, i.e.,  $k_{train\_spe}=k_{test}$ . Whereas, when trained via the **blur-&-clear** mode, the highest privacy breach occurs against the target dataset blurred with the smallest kernel, i.e.,  $k_{test}=(19,19)$ . Last but not least, we notice that both, **blur** and **blur-&-clear** modes report almost the same Top-1 accuracies for  $k_{test}=k_{train\_spe}$  when observing Figures 12.(b)-(c) and (d).

**Table 7:** The AUC values of the Top-1 accuracy curves for each kernel  $k_{train}$  measured against the different kernels  $k_{test}$ .

Mode \ $k_{train\_spe}$	(31,31)	(37,37)	(43,43)	(49,49)	(55,55)
<b>Blur</b>	276.5	<b>309.1</b>	290	253	237.2
<b>Blur-&amp;-Clear</b>	310.2	<b>338</b>	330.9	310.9	281.2

- **Using blurring kernel (37,37) provides the widest attack range and highest privacy breaches against the 7 target datasets.** To estimate the range of each attack (e.g. each kernel  $k_{train\_spe}$ ) against the 7 target datasets, we report in Table 7 the Area Under the Curve (AUC) of the Top-1 accuracy curves for each kernel in  $k_{train}$  (i.e., for each sub-figure in Figure 12). We notice that kernel size (37,37) reports the highest AUC values for both training modes, i.e. the highest privacy breaches and attack range against the 7 target datasets. The adversary does not need to blur its training datasets with the highest blurring kernel ( $k_{train\_spe}=55$  in our case) in order to cause the highest privacy breaches over the different target datasets. In other words, considering an adversary unaware of the target dataset's blurring kernel, the

most dangerous attack she/he could perform is a *recognition*-based attack trained via the **blur-&-clear** mode using kernel  $k_{train\_spe}=(37,37)$ .

As demonstrated in this experimental setup, an adversary can still breach the anonymity of face images even when they are obfuscated with a greater degree (e.g. higher blurring kernel) than its own training set. In other words, simply increasing the degree of obfuscation will not guarantee the privacy of the face images in the target dataset. Hence, it should not be the only strategy adopted when defending a target dataset against an adversary performing DL-assisted attacks. This in turn stresses the importance of investigating, developing and employing other approaches (e.g., using adversarial examples [48] to trick recognition-based attacks) for better and more robust defense mechanisms.

## 5 Framework discussion

In this section, we describe briefly how our recommendation framework is generic and scalable.

First, our recommendation framework is generic because it can be adapted to other SI such as workers' badge name, workers' personal belongings or even the workers' silhouette (e.g., ReID scenarios [25]). Let us consider the worker's badge name as the SI instead of the worker's face. In other words, our goal is to recommend the most robust obfuscation techniques for the workers' badge names. In short, we need to:

- Change/train a detector in the data preparation unit (c.f. Section 3.1) to localize and detect the text/Badge Names in an image, e.g. OpenCV's scene text detector<sup>18</sup>.
- Train DL-assisted attacks to restore/recognize obfuscated characters in image by adding pairs of clear/obfuscated text images to their training sets.
- Change/adapt the evaluation metrics in the evaluation unit (c.f. Section 3.3): for instance, employing the Tesseract OCR library<sup>19</sup> in the verification-based sub-module in order to extract the text from the clear, obfuscated and restored images and compare the extracted results.

Second, our recommendation framework is scalable with regard to the:

- *Obfuscation techniques*: for instance we can evaluate the robustness of the inpainting method [28, 29] in the context of face images by implementing it in the anonymizer (data preparation unit) and train a DL-assisted attack accordingly.
- *Adversaries*: we can also consider adversaries with different threat levels, capable of performing more dangerous DL-assisted attacks either by considering additional knowledge, different neural network architectures [49–52], other training hyper-parameters or larger training datasets...

---

<sup>18</sup>[https://docs.opencv.org/master/da/d56/group\\_\\_text\\_\\_detect.html](https://docs.opencv.org/master/da/d56/group__text__detect.html)

<sup>19</sup><https://github.com/tesseract-ocr/tesseract>

- *Evaluation metrics*: including additional metrics provides 'redundancy' and 'diversity' for the evaluation process. For instance in the context of face images, we can consider human evaluators as in [20] or other identity-based metric to measure the identity distance between two faces alongside the OpenFace tool.

## 6 Related Works

In this section, we investigate works related to (i) adversaries attacking obfuscated images via *recognition/restoration*-based attacks and (ii) to evaluation frameworks.

### 6.1 Recognition-based attacks

In [8], Newton et al. designed an algorithmic attack to identify people from pixelated and blurred face images. The recognition rates increased after applying the same obfuscation to the probe and gallery set of the face recognition approach [53]. They showed that small pixel box (e.g., 2x-4x) and simple blurring cannot prevent identification attacks. In another study [54], Gopalan et al. presented a method to recognize faces obfuscated with non-uniform blurring by examining the blurred images. As a follow-up study [55], Punnappurath et al. applied blurring effects to images in the target gallery and measured the minimal distance between the gallery images and the blurred probe image. On another note, the authors in [7] demonstrated that modern image recognition approaches, based on artificial neural networks, can be employed as attacks to recover hidden information from obfuscated images. They focused on three forms of obfuscation: pixelating, blurring and P3 (an encryption-based method [56]). The adversary successfully identifies obfuscated faces and objects by training DL networks with obfuscated images (faces [57, 60], digits [58] and objects [59]). Also in a recent medical study [17], the authors performed DL-assisted attacks against a publicly available anonymized medical dataset [61] containing x-rays of patients with sensitive meta-data such as treatment history, clinical institution, diagnosis... They considered an adversary aware of the identities *present* in the target dataset. Therefore, she/he can perform *recognition*-based attacks and link the known identity to the anonymized x-rays in the target anonymized dataset in order to gain more sensitive data about the identity.

Similar to [17] and unlike the other studies, we assume a more realistic scenario where an adversary can perform a *recognition*-based attack only when equipped with the proper background knowledge. Additionally in our case, we study thoroughly how the background knowledge affects the *recognition*-based attacks. For instance, in  $T_2$  we show how the incremental background knowledge regarding the identities *present* in the target dataset intensifies the privacy breaches and increases the number of re-identified individuals. Whereas in  $T_3$ , we show how an adversary can perform a *recognition*-based

attack and breach the face's anonymity despite lacking knowledge regarding the hyper-parameters of the obfuscation technique used.

## 6.2 Restoration-based attacks

The authors in [10] tackled the privacy-preservation question in the context of obfuscated faces by restoring obfuscated features and evaluating the reconstruction with regard to face recognition. They considered three obfuscations: pixelating, blurring and masking. They used traditional image reconstruction techniques (i.e., reconstruction [62] and interpolation-based [12] techniques for super resolution). In addition, they evaluated the identity restoration using the same traditional face recognition techniques as in [23]. In our framework, we adopted DL-based techniques for both, face reconstruction and recognition because as stated in [32, 39], DL-based techniques demonstrate great superiority over traditional methods. Alternatively, the authors in [11] investigated the amount of obfuscation needed to guarantee patients anonymity. They applied CycleGAN [63] in order to reconstruct features from anonymized medical imaging. They considered two anonymization techniques: (a) blurring and (b) masking. They also compared the results qualitatively and quantitatively by computing correlation coefficients and SSIM between the original and reconstructed images as well as between the original and anonymized images. In our approach, we add a level of abstraction to the restoration and evaluation process, i.e. the intra/inter attack comparisons in the interpretation unit, in order to not only evaluate the reconstruction process but recommend the most robust obfuscation technique.

## 6.3 Background knowledge effect

In a similar study to ours, the authors in [16] evaluated the effectiveness of 8 obfuscation techniques by considering three threat levels based on the knowledge of the adversary with regard to the obfuscation technique employed along with its hyper-parameters. They considered that the weakest adversary has no knowledge about the obfuscation used whereas the strongest knows the exact one employed. In addition, they performed three types of attacks: an *recognition*-based, a *verification*-based and a *restoration*-based attack and they showed that the privacy breaches increase along with the background knowledge. In our work, we first defined the background knowledge of the adversary with regard to the identities *present* in the target dataset, not only the obfuscation technique employed. Second, we designed the adversary with an incremental background knowledge with regard to the number of identities known by the adversary. Whereas the authors in [16] considered a specific number of known identities when performing identification attacks and it was not part of the background knowledge. Third, in our work we considered the hyper-parameters of the obfuscation technique as part of the background knowledge. Last but not least, we adapted the three-components adversary

model (i.e., goal, knowledge and capability) to the image obfuscation application domain to clearly define and demonstrate how these different components (mainly knowledge and capabilities) affect one another.

## 6.4 Evaluation Frameworks

Several evaluation frameworks have been proposed in the literature to evaluate obfuscation techniques in the context of images/videos. Some frameworks rely on human participants [20] whereas others rely on quantitative metrics [21, 22] e.g. SSIM, recognition algorithms...

On the one hand, the authors in [20] conducted an online experiment with 271 participants to evaluate the effectiveness of different obfuscation techniques (e.g. blurring, pixelating, inpainting. . .) against human recognition and how they affect the viewing experience. In our study, we employ quantitative-based metrics however we can hybridize our framework by including either a human-based adversary that attempts to recognize obfuscated/restored faces or a human-based evaluation module that attempt to assess the reconstruction of the images. On the other hand, the authors in [22] propose a framework that evaluates the obfuscation techniques (pixelating, blurring, complete masking, cartooning) based on the privacy and utility aspects in the context of videos via quantitative-based metrics. They assess the privacy aspect by quantifying the appearance similarity between the original and the obfuscated image and assess the utility by quantifying the structural similarity. Also, the authors in [73] propose an adversarial framework to address the privacy preservation problem regarding action recognition in videos. The framework explicitly minimizes a hybrid loss function combining both, privacy and utility aspects in order to find an optimal level of privacy (anonymization) while maintaining a good level of utility. Our framework evaluates the robustness of obfuscation techniques by (i) simulating adversaries with different background knowledge, (ii) performing attacks (*recognition* or *restoration*-based) and (iii) evaluating these attacks via structural, verification, identification and accuracy-based metrics. In [21], the authors proposed a framework to verify the effectiveness of obfuscation techniques (pixelating, blurring, scrambling) by conducting recognition-based attacks via the PCA [53] and LDA [64] algorithms. Also, the authors in [23] investigated the privacy-intelligibility trade-off by proposing a framework for evaluation of privacy filters. They applied several privacy techniques to faces (e.g. blurring, pixelating and masking) with varying intensities. The accuracy of the face recognition algorithm was considered a measure of privacy (a specific person should not be identified). Whereas, the accuracy of the face detection algorithm was used as a measure of intelligibility (a face should be detected). Similar to [21], they applied traditional methods for face recognition such as PCA [53], LDA [64] and LBP [65]. They concluded that an increase in the strength of privacy filters leads to an increase in privacy and a decrease in intelligibility. Similarly, the framework proposed in [23] evaluates the best obfuscation technique regarding the privacy-intelligibility trade-off by varying the level of privacy and comparing the accuracy of both face detection



and recognition algorithms. Here and unlike [21, 23], (i) we employ DL-based approaches instead of traditional approaches, (ii) we add a level of abstraction to the framework to not only evaluate but recommend the most robust obfuscation technique and (iii) we study thoroughly how the background knowledge can limit/increase the adversary’s attacking capabilities (*restoration*-based or *recognition*-based attacks) and privacy breaches.

## 7 Conclusion

In this paper, we extend the recommendation framework proposed in [15] by (i) adapting a three-components adversary model (goal, knowledge and capabilities) to our application domain (i.e., facial features obfuscations), (ii) extending the background knowledge of the adversary to include both the obfuscation technique and the identities *present* in the target dataset and (iii) supporting *restoration*-based and *recognition*-based as well as *Restoration & Recognition*-based attacks. In addition, we conducted three sets of experiments on the CelebA dataset [26]. In the first experiment, we validated our approach by implementing and testing our framework on obfuscated faces. Throughout the second experiment, we demonstrated how the adversary’s attacking capabilities scale with its knowledge and how it increased the potential risk of breaching the identities of blurred face images. Throughout the third experiment, we studied the possible privacy breaches and the attack range of an adversary against blurred face images while lacking knowledge about the obfuscation’s hyper-parameters.

Prospects that we did not explore in this study, could be addressed in future work. First of all, other visual features such as an individual’s name tag, posture or personal belongings can be identifying and considered sensitive. In this work, we focused on individuals’ faces because they are the most revealing in the context of images. Second of all, the adversary’s background knowledge in this work covers the identities *present* in the target dataset, therefore she/he can mine images for each known identity and perform a DL-assisted attack to recognize and re-identify the identity of the obfuscated face images. Nevertheless, in other scenarios the adversary’s background knowledge could be limited to quasi-identifying information such as the individual’s race or gender. If that is the case, the adversary could perform DL-assisted attacks to recognize the gender or the race of the target individual [47] instead of the full identity which might lead as well to potential privacy breaches when linked to other data sources (i.e., identity disclosure via linking attacks [19]). Last but not least, different approaches have been proposed in the context of image classification and identity recognition to trick, ruin or corrupt DL models. Some approaches rely on designing adversarial examples by perturbing the query image at the inference phase either physically (e.g. the target individual wears special accessories, e.g. glasses or hats [66]) or quantitatively [48] (small perturbations are added on a pixel level which are not visible to the human visual system). Other approaches rely on modifying/corrupting the training dataset via data

poisoning (clean-label [67, 68] and dirty-label attacks [69]), to ruin the neural network’s weights and trick it into inferring incorrect labels when queried with non-perturbed images, i.e., breaking the DL models at training phase. These approaches can be employed in our scenario as a defense mechanism against the adversary’s attempts to breach the obfuscated faces’ anonymity [70]. Nevertheless, it requires a thorough examination and investigation therefore we leave the defender concept for a future study.

## References

- [1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, ImageNet Large Scale Visual Recognition Challenge *arXiv:1409.0575*, (2014)
- [2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, *CVPR* (2016)
- [3] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, SSD: Single shot multibox detector, *arXiv:1512.02325*, (2015)
- [4] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected *arXiv:1606.00915*, (2016)
- [5] S. Hill, Z. Zhou, L. Saul, and H. Shacham. On the in effectiveness of mosaicing and blurring as tools for document redaction, *PETS*, (2016)
- [6] G. Boracchi and A. Foi. Modeling the performance of image restoration from motion blur, *Image Processing, IEEE Transactions*, *21(8):3502–3517*, (2012)
- [7] R. McPherson, R. Shokri, and V. Shmatikov, Defeating image obfuscation with deep learning. *CoRR*, (2016)
- [8] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images, *IEEE transactions on Knowledge and Data Engineering*, (2005)
- [9] K. Lander, V. Bruce, and H. Hill. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces, *Applied Cognitive Psychology*, (2001)
- [10] Ruchaud, N. and Dugelay, J. L., Automatic Face Anonymization in Visual Data: Are we really well protected? *Electronic Imaging*, (2016)
- [11] D. Abramian and A. Eklund, Refacing: reconstructing anonymized facial features using GANs *COCR*, volume abs/1810.06455, (2018)

- [12] R. Keys. Cubic convolution interpolation for digital image processing, *Acoustics, Speech and Signal Processing, IEEE Transactions*, 29(6):1153–1160, (1981)
- [13] W. Yang , X. Zhang, Y. Tian, W. Wang, J.H. Xue, Deep learning for single image super-resolution: A brief review *arXiv preprint arXiv:1808.03344*, (2018)
- [14] Dargan, S., Kumar, M., Ayyagari, M.R. et al. A Survey of Deep Learning and Its Applications: A New Paradigm to Machine Learning Arch Computat Methods Eng 27, 1071–1092 , (2020)
- [15] J. Tekli, B. al Bouna, R. Couturier, G. Tekli, Z. al Zein and M. Kamradt, A Framework for Evaluating Image Obfuscation under Deep Learning-Assisted Privacy Attacks, 2019 17th International Conference on Privacy, Security and Trust (PST), Fredericton, NB, Canada, 2019, pp. 1-10 (2019)
- [16] H. Hao, D. Güera, J. Horváth, A. R. Reibman, E. J. Delp, Robustness analysis of face obscuration 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition, (2020)
- [17] K. Packhauser, S. Gündel, N. Münster, C. Syben, V. Christlein and A. Maier Is Medical Chest X-ray Data Anonymous? *arXiv:2103.08562:v1*, (2021)
- [18] Q.Do, B. Martini, K.-K. R. Choo, The role of the adversary model in applied security research, *Computers & Security* , (2018)
- [19] B. Meden and P. Rot and P. Terhörst and N. Damer and A. Kuijper and W. J. Scheirer and A. Ross and P. Peer and V. Struc Privacy-Enhancing Face Biometrics: A Comprehensive Survey *IEEE Trans. Inf. Forensics Secur.*, (2021)
- [20] Y.Li, N. Vishwamitra, B.P. Knijnenburg, H. Hu, K. Caine, Effectiveness and Users’ Experience of Obfuscation as a Privacy-Enhancing Technology for sharing Photos *In Proceedings of the ACM on Human-Computer Interaction* (2017)
- [21] F. Dufaux, T. Ebrahimi, A framework for the validation of privacy protection solutions in video surveillance *In IEEE International Conference on Multimedia and Expo* (2010)
- [22] T. Nawaz, A. Berg, J. Ferryman, J. Ahlberg, M. Felsberg Effective evaluation of privacy protection techniques in visible and thermal imagery *Journal Of Electroning Imaging* (2017)

- [23] P. Korshunov, A. Melle, J.-L. Dugelay, and T. Ebrahimi, Framework for objective evaluation of privacy filters. *In Proceedings of SPIE*, volume 8856, page 12, (2013)
- [24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment: From error measurement to structural similarity, *Image Processing, IEEE Transactions, vol. 13*, (2004)
- [25] A. Chattopadhyay, R. Ruska and L. Pfantz Determining the Robustness of Privacy Enhancing DeID Against the ReID Adversary: An Experimental Study The 16th International Conference on Availability, Reliability and Security, (2021)
- [26] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild In Proceedings of International Conference on Computer Vision (ICCV)(2015)
- [27] E. M. Newton, L. Sweeney, Preserving privacy by de-identifying face images. *In IEEE transactions on knowledge and Data Engineering* (2005)
- [28] H.Hao, D. Güera, A. R. Reibman and E.J. Delp, A utility-preserving gan for face obscuration. *Proceedings of the International Conference on Machine Learning, Synthetic Realities: Deep Learning for Detecting AudioVisual Fakes Workshop* (2019)
- [29] L.Jingzhi and H. Lutong and C. Ruoyu and Z. Hua and H. Bing and W. Lili and C. Xiaochun Identity-Preserving Face Anonymization via Adaptively Facial Attributes Obfuscation *Proceedings of the 29th ACM International Conference on Multimedia* (2021)
- [30] A. Frome, G. Cheung, A. Abdulkader, M. Zennaro, et al. Large-scale privacy protection in Google Street View IEEE 12th International Conference on Computer Vision, ICCV (2009)
- [31] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenec: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019)
- [32] B. Amos, B. Ludwiczuk, and M. Satyanarayanan. Openface: A general-purpose face recognition library with mobile applications, *Technical report, CMU School of Computer Science, CMU-CS-16-118*, (2016)
- [33] M. Bellare, and P. Rogaway Entity Authentication and Key Distribution, *Advances in Cryptology — CRYPTO’ 93, Lecture Notes in Computer Science D. R. Stinson, ed., pp. 232-249: Springer Berlin Heidelberg, 1993.*, (1993)

- [34] M. Bellare, and P. Rogaway Provably Secure Session Key Distribution: The Three Party Case, *in Proceedings of the 27th Annual ACM Symposium on Theory of Computing, Las Vegas, Nevada, USA, pp. 57-66.*, (1995)
- [35] M. Bellare, D. Pointcheval, and P. Rogaway, Authenticated Key Exchange Secure against Dictionary Attacks, *Advances in Cryptology — EURO-CRYPT 2000, Lecture Notes in Computer Science B. Preneel, ed., pp. 139-155: Springer Berlin Heidelberg*, (2000)
- [36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering, *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 815–823*, (2015)
- [37] Z. Shen, W. Lai, T. Xu, J. Kautz and SM. Yang, Deep semantic face deblurring *CVPR* pages 8260–8269, (2018)
- [38] Y. Li, S. Liu, J. Yang, and M.-H. Yang., Generative face completion, *arXiv preprint arXiv:1704.05838*, (2017)
- [39] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint arXiv:1609.04802*, (2016)
- [40] D. Garcia., srez: Adversarial super resolution, <http://github.com/david-gpu/srez>, (2016)
- [41] X. Yu and F. Porikli, Ultra-Resolving Face Images by Discriminative Generative Networks, Springer International Publishing, pages 318–333, (2016)
- [42] S. Majumdar Image Super Resolution, <https://github.com/titu1994/Image-Super-Resolution>, (2016)
- [43] Z. Shen Deep-Semantic-Face-Deblurring, <https://github.com/joanshen0508/Deep-Semantic-Face-Deblurring>, (2016)
- [44] R. A. Yeh, C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson and M. N. Do, Semantic image inpainting with deep generative models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5485–5493, (2017)
- [45] C.B. Jin, Semantic-image-inpainting, <https://github.com/ChengBinJin/semantic-image-inpainting>, (2018)
- [46] K. Yang, J. Yau, L. Fei-Fei, J. Deng, O. Russakovsky A Study of Face Obfuscation in ImageNet, *arXiv:2103.06191v2*, (2021)

- [47] Y. Linwei and L. Binglin and M. Noman and W. Yang and L. Jie Privacy-Preserving Age Estimation for Content Rating, 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), (2018)
- [48] I.J. Goodfellow, J. Shlens and C. Szegedy. Explaining and harnessing adversarial examples in ICLR, (2015)
- [49] K. Shaheed and A. Mao and I. Qureshi and M. Kumar and S. Hussain and I. Ullah and X. Zhang DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition Expert Systems with Applications (2022)
- [50] Zhang, Kai and Van Gool, Luc and Timofte, Radu Deep unfolding network for image super-resolution IEEE Conference on Computer Vision and Pattern Recognition, (2020)
- [51] Bansal M, Kumar M, Sachdeva M, Mittal A. Transfer learning for image classification using VGG19: Caltech-101 image data set J Ambient Intell Humaniz Comput. 2021 Sep 17:1-12, (2021)
- [52] H.N. Vu, M.H. Nguyen and C. Pham, Masked face recognition with convolutional neural networks and local binary patterns Applied Intelligence (2022)
- [53] M. Turk and A. Pentland, Face recognition using eigenfaces, *Computer Vision and Pattern Recognition, Proceedings CVPR '91., IEEE Computer Society Conference*, (1991)
- [54] R. Gopalan, S. Taheri, P. Turaga, and R. Chellappa. A blur-robust descriptor with applications to face recognition, *IEEE transactions on pattern analysis and machine intelligence*, (2012)
- [55] A. Punnappurath, A. N. Rajagopalan, S. Taheri, R. Chellappa, and G. Seetharaman. Face recognition across non-uniform motion blur, illumination, and pose, *IEEE Transactions on Image Processing*, (2015)
- [56] M.-R. Ra, R. Govindan, and A. Ortega. P3: Toward privacy-preserving photo sharing, *NSDI*, (2013)
- [57] Laboratories Cambridge. The database of faces, (1994)
- [58] Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, (1998)
- [59] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, (2009)

- [60] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets In *IEEE International Conference on Image Processing (ICIP)*, (2014)
- [61] X. Wang et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2097–2106, (2017)
- [62] W. Dong, L. Zhang, G. Shi, and X.Wu. Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization, *Image Processing, IEEE Transaction*, 20(7):1838–1857, (2011)
- [63] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks, *CoRR*, vol abs/1703.10593, (2017)
- [64] P. Belhumeur, J. Hespanha and D. Kriegman, Eigenfaces vs. sherfaces: recognition using class specific linear projection, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19(7), 711720, (1997)
- [65] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: Application to face recognition, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28(12), 20372041, (2006)
- [66] S. Komkov, A. Petiushko, Advhat: Real-world adversarial attack on arcfac face id system arXiv:1908.08705, (2019)
- [67] A. Shafahi, W.R. Huang, M. Najibi, O. Suci, C. Studer, T. Dumitras and T. Goldstein Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Proc. of NeurIPS*, (2018)
- [68] S. Shan, E. Wenger, J. Zhang, H. Li, H. Zheng and B.Y. Zhao Fawkes: Protecting personal privacy against Unauthorized Deep Learning Models arXiv:2002.08327, (2020)
- [69] B. Biggio, B. Nelson and P. Laskov. Poisoning attacks against support vector machines arxiv 1206.6389, (2012)
- [70] S. Rezaeifar and S. Voloshynovskiy and M. Asgari Jirhandeh and V. Kinakh Privacy-Preserving Image Template Sharing Using Contrastive Learning Entropy (2022)
- [71] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated residual transformations for deep neural networks, In *Computer Vision and Pattern Recognition (CVPR)*, 2017 IEEE Conference on, pp. 5987-5995. IEEE,

(2017)

- [72] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu Squeeze-and-excitation networks, arXiv:1709.01507 (2019)
- [73] Z. Wu, H. Wang, Z. Wang, H. Jin, Z. Wang Privacy-Preserving Deep Action Recognition: An Adversarial Learning Framework and A New Dataset arxiv:1906.05675v4 (2020)
- [74] Y. Chen, Y. Tai, X. Liu, C. Shen, and J. Yang. Fsrnet: End-to-end learning face super-resolution with facial priors. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,(2018)