

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/338451749>

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

Conference Paper · August 2019

DOI: 10.1109/PST47121.2019.8949040

CITATIONS

7

READS

465

6 authors, including:



Jimmy Tekli

BMW Group

3 PUBLICATIONS 11 CITATIONS

[SEE PROFILE](#)



Bechara AL Bouna

Université Antonine

61 PUBLICATIONS 257 CITATIONS

[SEE PROFILE](#)



Raphaël Couturier

University Bourgogne Franche-Comté

301 PUBLICATIONS 2,837 CITATIONS

[SEE PROFILE](#)



Gilbert Tekli

University of Balamand

22 PUBLICATIONS 131 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Bib number detection and recognition: deep learning approach [View project](#)



Wireless Sensor Network [View project](#)

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

Jimmy Tekli ^{*1,2}, Bechara Al Bouna ^{†3}, Raphaël Couturier^{‡2},
Gilbert Tekli^{§4}, Zeinab Al Zein³, and Marc Kamradt¹

¹BMW Group, Munich, Germany

²Université de Franche Comté, Belfort, France

³ TICKET Lab., Antonine University, Baabda, Lebanon

⁴University of Balamand, Lebanon

Abstract

Computer vision applications such as object detection and recognition, allow machines to visualize and perceive their environments. Nevertheless, these applications are guided by learning-based methods that require capturing, storing and processing large amounts of images thus rendering privacy and anonymity a major concern. In return, image obfuscation techniques (i.e., pixelating, blurring, and masking) have been developed to protect the sensitive information in images. In this paper, we propose a framework to evaluate and recommend the most robust obfuscation techniques in a specific domain of application. The proposed framework reconstructs obfuscated faces via deep learning-assisted attacks and assesses the reconstructions using structural/identity-based metrics. To evaluate and validate our approach, we conduct our experiments on a publicly available celebrity faces dataset. The obfuscation techniques considered are pixelating, blurring and masking. We evaluate the faces reconstructions against five deep learning-assisted privacy attackers. The most resilient obfuscation technique is recommended with regard to structural and identity-based metrics.

Keywords : data privacy, face obfuscation, deep learning, image transformation

*jimmy.tekli@bmw.de

†bechara.albouna@ua.edu.lb

‡raphael.couturier@univ-fcomte.fr

§gilbert.tekli@balamand.edu.lb

1 Introduction

Manufacturing and automotive companies around the world are increasingly integrating computer vision applications such as object recognition[1], object detection[2] and segmentation[3] to increase their business value. In this direction, several use cases have been developed (e.g., autonomous driving and supply chain optimization), which require taking images that might contain sensitive information such as individuals identities, workers belongings, or name tags. Due to regulations, these companies must guarantee a level of anonymization¹ that requires “irreversibility preventing identification of the data subject”, taking into account all the means “reasonably likely to be used” for identification. Now, to preserve these sensitive information, several obfuscation techniques like pixelating (also known as mosaicking)[4], blurring (Gaussian/motion)[4, 5] and masking can be used (c.f. figure 1). More specifically, obfuscation is done by altering/removing features from the images to hide sensitive information but, at the same time, retaining some visual features to keep the image suitable for processing. However, these visual features can still be used to identify/reconstruct the protected objects via different attackers which we classify as *recognition-based*[6, 7, 8] and *restoration-based* attackers[9, 10].

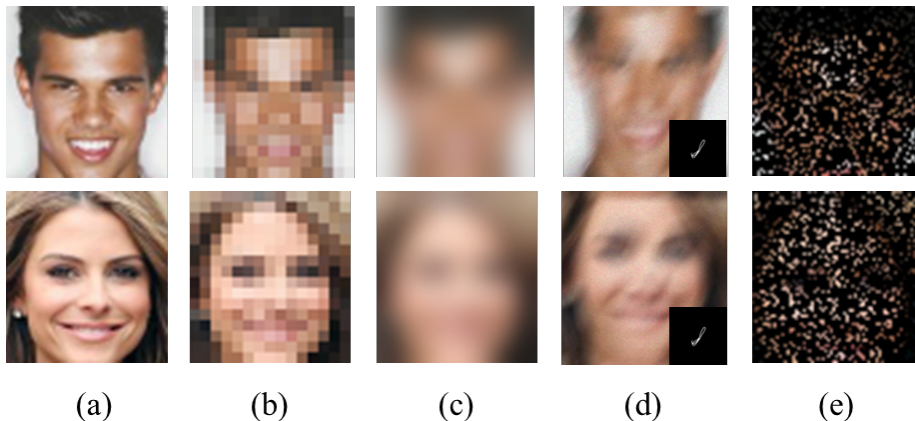


Figure 1: Obfuscation techniques left to right , (a) Original plain image, (b) pixelated image (4x) ,(c) Gaussian Blurred Image($\sigma = 5$), (d) Motion blurred via [5] and (e) masking by adding random black pixels.

On the one hand, *recognition-based* attackers breach the images privacy and anonymity by training learning-based algorithms to perform recognition tasks on obfuscated information. For instance, the authors in [6] demonstrate that obfuscated faces, objects and digits can be *recognized* by artificial neural networks trained via a supervised manner on obfuscated images. On the other hand,

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

restoration-based attackers de-anonymize privacy-protected images by trying to *restore/reconstruct* the plain original features of the obfuscated information. The authors in [9] cancel the impact of pixelating, blurring and masking with regard to face recognition algorithms by applying ad-hoc traditional image reconstruction techniques (e.g., bicubic interpolation[11]).

Several studies showed that Deep Neural Networks are capable of restoring finer details of the obfuscated information[12]. Hence, from a privacy perspective, these Deep Learning-based (DL) techniques are highly nominated as tough *restoration-based* attackers. For instance, as its name indicates, SRResNet[13] is a ResNet-based neural network [1] designed to upscale images by a factor of 4, i.e. *image super resolution*. The authors in [14] proposed a technique that generates the missing content in an image by conditioning on the available pixels (DSLinpaint), i.e. *image inpainting*. In addition, the authors in [15] proposed a multi-scale convolutional neural network that restores sharper images from a motion blurred input (DSF_deblur), i.e. *image deblurring*. All these techniques prove the effectiveness of deep learning for image transformation tasks.

In this paper, we propose a recommendation framework that evaluates the robustness of image obfuscation techniques and recommend the most resilient obfuscation against DL-assisted attackers. Our approach is based on an iterative 4-layered workflow:

1. Detecting and obfuscating the sensitive information in an image.
2. Applying DL-assisted attacks to reconstruct the obfuscated information.
3. Evaluating the reconstruction results via multiple reconstruction metrics.
4. Recommending the most resilient obfuscation technique with regard to each metric.

Although our proposed 4-layered workflow can be adapted to different sensitive information, we only consider in this study individual’s faces. Hence, we detect/obfuscate faces via the following techniques: pixelating[4], blurring (Gaussian/motion)[4, 5] and masking. In addition, we employ DL-assisted attackers trained to reconstruct obfuscated facial features [13, 14, 15, 16]. Furthermore, we evaluate the reconstruction results via structural and identity-based metrics [17, 19] in order to recommend the most resilient obfuscation technique with regard to each evaluation metric.

The remainder of this paper is organized as follows. In Section 2, we review different obfuscation techniques and evaluation metrics. In Section 3, we present the recommendation framework. Section 4 evaluates different faces obfuscation techniques via the proposed framework. In Section 5, we investigate works related to privacy attacks in the context of images.

2 Background and Preliminaries

Before introducing our framework, it is essential to shed some light on the main obfuscation techniques and evaluation metrics used in the literature.

2.1 Obfuscation techniques

Obfuscation techniques fall mainly under three categories: pixelating, blurring and masking.

2.1.1 Pixelating

Pixelating (a.k.a. mosaicking) is widely adopted as an obfuscation technique. The sensitive information 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 [4]. 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 [6], 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 4 is equivalent to applying a pixel box of size 4x4.(c.f. figure 1.b).

2.1.2 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[4]. It removes details from an image by applying a Gaussian kernel. The blurriness level is controlled by the standard deviation σ . A motion blur alters the details of an image by generating the effect of a synthetic camera motion blur[5]. The level of blurriness is affected by the length and the angle of the synthesized motion (c.f. figure 1(c-d)).

2.1.3 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 (c.f. figure 1.e).

2.2 Evaluation metrics

Two main evaluation metrics are used to measure the similarities between original and reconstructed faces in the context of image transformation tasks: SSIM and Identity distance.

2.2.1 SSIM

The Structural Similarity Index (SSIM)[17] quantifies image quality modifications (enhancement/degradations). It is widely used in the literature as an evaluation metric for image transformation techniques[13, 14, 15, 18]. It computes

a holistic similarity value between two images, the reference and the modified image. The SSIM value ranges between 0 and 1 with 1 indicating identical images.

2.2.2 Identity distance

OpenFace[19] is a Python and Torch implementation of face recognition with deep neural networks[20]. OpenFace directly learns a mapping from face images to a compact euclidean space where distances directly correspond to a measure of face similarity. A distance of 0.0 means the faces are identical and 4.0 corresponds to the opposite spectrum i.e. two different identities. All distance values below a threshold of 1.1 can be considered as the same individual.

3 Proposed Framework

In this section, we present the proposed recommendation framework. Its main objective is to attack obfuscation techniques by restoring hidden facial features, evaluate the reconstruction and suggest the most resilient obfuscation. Inspired by the KDD process[21], the proposed framework is composed of four units: (a) a data preparation unit, (b) a reconstruction unit, (c) an evaluation unit and (d) an interpretation unit (c.f. figure 2).

The data preparation unit detects and obfuscates the sensitive information in an image. The reconstruction unit groups DL-assisted attackers in order to reconstruct the obfuscated information. The evaluation unit measures the reconstruction ratio via different quality metrics. Last but not least, the recommendation unit proposes the most resilient obfuscation technique with regard to each metric based on the intra/inter-attacker comparisons.

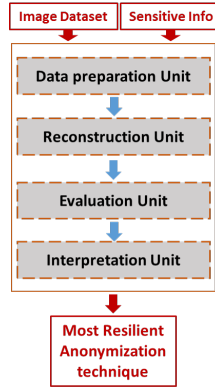


Figure 2: The generic recommendation framework

3.1 Data preparation unit

The data preparation unit takes as inputs an image dataset along with the sensitive information. It is divided into two modules: (a) *sensitive information (SI) detector* and (b) *Anonymizer* (c.f. figure 3.a). As its name indicates, the *SI detector* localizes and detects the sensitive information in the image, crops it and sends it to the *Anonymizer*. As stated before, we consider in this study the faces as sensitive information. Hence, the *SI detector* employs the OpenFace toolbox[19] to detect faces in an image, crop and forward it to the *Anonymizer*. In this study, the *Anonymizer* obfuscates the sensitive information via: pixelating, blurring (Gaussian/motion) and masking techniques and sends the anonymized images to the reconstruction unit.

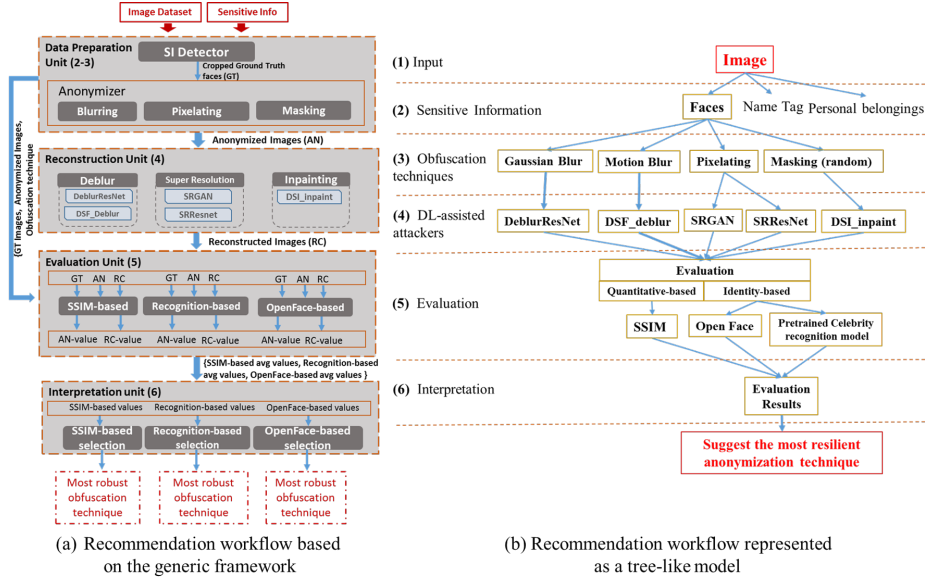


Figure 3: Adapting the proposed generic framework to images with faces as sensitive information

3.2 Reconstruction unit

The reconstruction unit receives the obfuscated cropped faces and the obfuscation technique used. As shown in figure 3.a, it is divided into three modules, one per obfuscation category: (a) the *super-resolution module* (for pixelating), (b) the *deblur module* (for Gaussian and motion blurring) and (c) the *inpainting module* (for masking). Each module contains one or multiple attackers. An attacker is a DL-based technique trying to reconstruct and recover the obscured faces. Moreover, we assume that each attacker has access to publicly available faces datasets and knows the exact algorithm used for obfuscation. At the

end of the reconstruction process, each DL-assisted attacker outputs a dataset where each face image has three derivatives: the unobfuscated, obfuscated and reconstructed face as shown in figure 4.

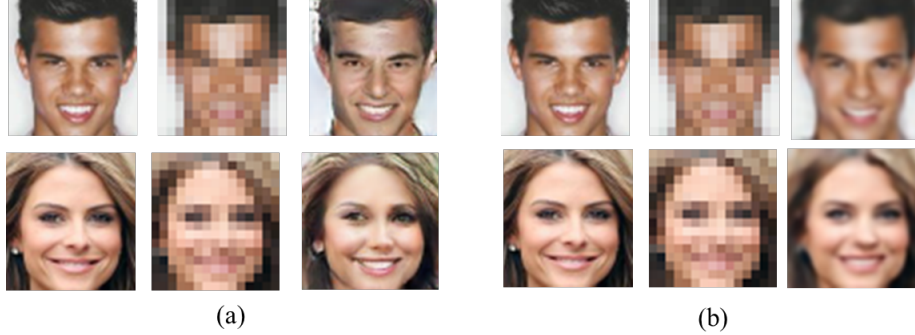


Figure 4: (a) Ground truth, Anonymized and Reconstructed images outputted via the SRGAN network. (b) Ground truth, Anonymized and Reconstructed image outputted via the SRResNet network.

3.3 Evaluation unit

The evaluation unit assesses the face restoration based on one *structural* and two *identity*-based metrics. Hence, the evaluation unit is divided into three evaluation modules: (a) *SSIM-based*, (b) *OpenFace-based* and (c) *Recognition-based evaluation module* as shown in figure 3.a. Each evaluation module receives as input three images: a unobfuscated face image (GT), an obfuscated face image (AN) and a reconstructed face image (RC). Furthermore, each evaluation module outputs two metric-based values, *AN-value* and *RC-value*(c.f. figure 3.a).

- The *SSIM-based evaluation module* measures the holistic similarity between the plain image (GT) and the obfuscated (AN) image and between the plain image (GT) and the reconstructed (RC) image [17]. For normalization purposes, the SSIM-based module computes the SSIM’s complement, i.e. $1 - \text{SSIM}$. Hence, the output values are between 0 and 1 where 0 means the two images are identical:

$$AN\text{-value} = 1 - \text{SSIM}(GT, AN) \quad (1)$$

$$RC\text{-value} = 1 - \text{SSIM}(GT, RC) \quad (2)$$

- The *OpenFace-based evaluation module* computes the identity distance via the OpenFace toolbox between the unobscured face images (GT) and both the obscured (AN) and the reconstructed face image (RC) [19]. As stated in section 2.2.2, OpenFace maps the two input faces to an identity distance between 0 and 4. The *Openface-based evaluation module* normalizes the

values to 0 and 1 where 0 value means that the two faces are identical, hence:

$$AN\text{-value} = \text{Normalized}(\text{OpenFace}(GT, AN)) \quad (3)$$

$$RC\text{-value} = \text{Normalized}(\text{OpenFace}(GT, RC)) \quad (4)$$

The normalization function, *Normalized*, is defined as follow:

$$\text{Normalized}(x_i) = \frac{x_i - \min(x)}{\max(x) - \min(x)}$$

where, $x = (x_i, \dots, x_n)$, x_i represents the identity distance $\text{OpenFace}(\text{image}_1, \text{image}_2)$ value bounded by 0 and 4, and $\text{Normalized}(x_i)$ is the corresponding normalized i th value between 0 and 1. In our case, $\min(x) = 0$ and $\max(x) = 4$.

- *The recognition-based evaluation module* employs a pre-trained celebrity recognition model. It uses the inferences² over the three received images in order to compute two average relative error values. By definition, the average relative error is the absolute difference between the exact "theoretical" value and its measured counterpart, divided by the exact value. We consider the inference over the unobscured image (GT) as the "exact" value whereas the prediction over the anonymized (AN) and the reconstructed (RC) as the measured values. The average relative error ranges between 0 and 1. We denote the confidence probability returned by the celebrity recognition model as *conf*.

$$AN\text{-value} = \frac{|\text{conf}(GT) - \text{conf}(AN)|}{\text{conf}(GT)} \quad (5)$$

$$RC\text{-value} = \frac{|\text{conf}(GT) - \text{conf}(RC)|}{\text{conf}(GT)} \quad (6)$$

In (5) and (6), both confidences in the numerator belong to the same *predicted celebrity name*. In other words, in case the five inferences, of the celebrity recognition model, over the obfuscated *AN* image do not contain the celebrity name predicted over its *GT* counterpart, the *AN*-value would be 1.

AN-values and *RC*-values, outputted by the three evaluation-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 evaluation module computes the average *AN*-values and the *RC*-values over the entire obfuscated/restored dataset received from each DL-assisted attacker. As a result, we can visualize the evaluation unit's output as shown in table 1.

We used in our evaluation unit the SSIM metric and the identity distance metric, measured via the OpenFace toolbox[19], because they are widely used in

²The pretrained celebrity model receives an image and infers 5 predictions, each containing the *detected celebrity name* alongside the *confidence probability*.

Table 1: Output of the three modules in the Evaluation unit

Module	Reconstruction technique	Evaluation values	SSIM-based	OpenFace-based	Recognition-based
Gaussian Blur ($\sigma=5$)	DeblurResNet	AN-value	0.548	1	1
		RC-value	0.167	0.272	0.972
		AN-value	0.530	0.434	0.998
Motion Blur	DSF deblur	RC-value	0.350	0.117	0.846
		AN-value	0.334	0.987	1
Pixelating (4x4)	SRGAN	RC-value	0.306	0.221	0.990
		AN-value	0.334	0.987	1
	SRresNet	RC-value	0.190	0.206	0.946
		AN-value	0.926	1	1
Masking (Random)	DSI inpaint	RC-value	0.172	0.285	0.920

the literature to evaluate the reconstruction of degraded faces in the context of image transformation tasks[15, 18]. Furthermore, we used a pre-trained celebrity recognition model as a third metric in order to simulate a human evaluator. A *human-based evaluation module* could be added in the future to evaluate the reconstruction qualitatively based on the human visual system.

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 (c.f. table 1). The interpretation unit is divided into three *selection modules*, one per evaluation metric: (a) *SSIM-based*, (b) *recognition-based* and (c) *OpenFace-based selection module*. Each module performs a 2-steps comparison in order to select the most resilient obfuscation technique: *intra-attacker* and *inter-attacker* comparison(e.g., the *SSIM-based selection module* selects the most resilient obfuscation with regard to the SSIM metric whereas the *Openface-based selection module* selects the most resilient obfuscation with regard to the identity distance metric).

As a first step, the *intra-attacker* comparison allows us to identify the toughest DL-assisted attacker against each obfuscation technique with regard to each evaluation metric. In other words, the DL-based reconstruction technique that restored the most of the obfuscated image. Once the toughest DL-assisted attacker is identified against each obfuscation, the *inter-attacker* comparison chooses the most resilient obfuscation against the selected DL-assisted attackers.

Following the example shown in table 1, we can see the corresponding results of the interpretation unit after the intra/inter-attacker comparisons in figure 5. The *OpenFace-based selection module* picks, via the intra-attacker comparison, the SRResNet as the toughest attacker against “pixelating” since it produced the minimum identity distance value after reconstruction, compared to SRGAN. The inter-attacker comparison, then, selects “masking” as the most robust obfuscation technique against the reconstruction attempts since it maintained the highest identity distance value compared to all the other obfuscation techniques.

In short, after selecting the toughest DL-assisted attackers against each obfuscation technique, the *OpenFace-based selection module* recommends the most resilient one, which is "Masking".

Module	Reconstruction technique	Evaluation values	OpenFace-based
Gaussian Blur ($\sigma=5$)	DeResNet	AN-value	1
		RC-value	0.272
		AN-value	0.434
Motion Blur	DSF deblur	RC-value	0.117
		AN-value	0.987
		RC-value	0.221
Pixelating (4x4)	SRGAN	AN-value	0.987
		RC-value	0.206
		AN-value	1
Masking (Random)	DSI inpaint	RC-value	0.285

(a)Evaluation results of the Openface-based evaluation module

Module	Reconstruction technique	Evaluation values	OpenFace-based
Gaussian Blur ($\sigma=5$)	DeResNet	AN-value	1
		RC-value	0.272
		AN-value	0.434
Motion Blur	DSF deblur	RC-value	0.117
		AN-value	0.987
		RC-value	0.206
Pixelating (4x4)	SRresNet	AN-value	1
		RC-value	0.285
		AN-value	0.285

(b) Intra-attacker comparison selects the toughest DL-based attackers per obfuscation technique

Module	Reconstruction technique	Evaluation values	OpenFace-based
Masking (Random)	DSI inpaint	AN-value	1
		RC-value	0.285

(c) Inter-attacker comparison selects Masking(random) as the most resilient obfuscating technique based on the Identity distance metric

Figure 5: Inter/Intra-attacker comparison

In section 3, we showed how the proposed framework recommends the most resilient obfuscation technique via the 4-layered iterative workflow: (a) detecting/obfuscating the sensitive information, (b) reconstructing via the DL-assisted attackers, (c) evaluating the reconstructions and (d) selecting the most robust obfuscation based on the inter/intra-attacker comparisons.

4 Experiments

To evaluate and validate our approach, we implemented and tested our framework on the CelebA dataset[22] with the celebrity faces being the sensitive information. To do so, we assessed the reconstruction of the obfuscation techniques based on structural and identity-based metrics. Then, we recommended the most robust obfuscation technique, with regard to each evaluation metric, based on the intra/inter attacker comparisons.

4.1 Dataset

We used the CelebA dataset for training and evaluating the DL-assisted attackers. CelebA dataset contains 202,599 face images with 10,177 identities[22]. In order to prepare our evaluation set, we select³ 370 images from the CelebA test set. To normalize our experimental setup, we use the same images to evaluate all our DL-assisted attackers. The training sets vary between the DL-based

³We first selected 1307 images from the celebA test set, then we filtered out, via the pretrained celebrity recognition model, the faces that were wrongly recognized or correctly recognized with a probability lower than 0.7 . This step was done for the *recognition-based evaluation module*.

techniques, however, no images from the test set were included throughout the training of any of the DL-based networks.

4.2 Obfuscation techniques

We used in this study four obfuscation techniques: (1) Pixelation (a.k.a. mosaicking), (2) Gaussian blurring, (3) Motion Blur and (4) Masking. We specified for each obfuscation technique a fixed parameter as shown in table 2. Regarding the pixelation, we simply downsampled the images by a factor of 4. For the Gaussian blur, we applied a Gaussian filter with a kernel size (31x31) and standard deviation of 5. As for the motion blur, we synthesized a motion blur kernel from random 3D camera trajectories [5]. Regarding the masking technique, we replaced random pixels all over the image by black pixels.

We chose these fixed parameters values for the pixelating and the gaussian blurring because they are considered the average intensity values regarding each obfuscation technique in the literature and they guarantee the anonymity of the individual as shown in figure 1(b-c). For instance, the pixelating technique is usually adapted, in the context of privacy, with a pixel box of size 2x2, 4x4 and 8x8. Nevertheless, in our study we chose a pixel box of size 4x4. As for the motion blur and the masking techniques, we generated multiple kernels however we chose the ones which guarantee the most the anonymity of the individual as seen in figure 1(d-e).

4.3 DL-assisted attackers

As mentioned previously, we grouped DL-based techniques for image reconstruction tasks as attackers against each obfuscation technique.












For the super resolution task, we chose two models: SRResNet and SRGAN. On the one hand, the SRResNet is a ResNet-based architecture[1] and is considered a benchmark when it comes to SR algorithms [12, 13]. Moreover, SRResNet is a generic SR-network applicable to our faces dataset⁴. On the other hand, SRGAN is a GAN-based super resolution model implemented by [16] similar to [24]. The model was developed specifically for faces. Both networks were trained from scratch. We generated the training pairs by downsampling the unobfuscated (GT) images by a factor of 4.

For the deblur task, we chose two DL-based techniques. Regarding the Gaussian blur, we adapted the SRResNet architecture by modifying the input size of the network implemented in [23]. In addition, we generated the training pairs by applying Gaussian blur to the unobfuscated (GT) images. As for the motion blur, we used the implementation and the pre-trained model provided by the authors [25].

Last but not least, we applied the deep generative model proposed in [14] and the implementation in [26] for the image inpainting task. We trained the

⁴The implementation [23] 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 2: Technical details regarding the obfuscation techniques and the implementations of the DL-assisted attackers

Ground Truth face	Obfuscating technique	Parameters	Parameter values	Obfuscated faces	DL-based reconstruction	Implementation/ Framework	Results
	Pixelation	Pixel Box Size	4		SRresNet [13]	TensorFlow [23] <i>Trained from scratch</i>	
					SRGAN [16]	TensorFlow [16] <i>Trained from scratch</i>	
	Gaussian Blur	Standard deviation σ	5		DeblurResNet[13]	Tensorflow [23] <i>Trained from scratch</i>	
	Motion Blur	Length and angle of the motion			DSF_deblur [15]	Matcaffe / Matlab [25] <i>Pre-trained model</i>	
	Masking	Location of the black pixels	Random		DSI_inpaint [14]	TensorFlow [26] <i>Trained from scratch</i>	

DCGAN network on our dataset from scratch.

As stated in section 3.3, our framework provides structural and identity-based evaluations regarding the different DL-based techniques over our test set. Each evaluation module in our framework computes two metric-based values for each unobfuscated image: (a) *AN*-value and (b) *RC*-value. In the following sections, we report the average values over the entire test set.

4.4 Structural-based evaluation

The evaluation unit (c.f. section 3.3) uses the *SSIM-based evaluation module* to compute the SSIM’s complement, (c.f. equation(1,2)). Hence, a value equal to 0 means that the images are identical. As shown in figure 6, the average RC-values for all the attackers is lower than the average AN-values since the reconstructed images are overall more similar to the plain ground truth images than the obfuscated ones in terms of SSIM. As mentioned in section 3.4, in order to select the most resilient obfuscation regarding the SSIM metric, the interpretation unit executes the intra/inter-attacker comparisons. In case an obfuscation technique has more than one DL-assisted attacker, the intra-attacker comparison selects the attacker that reconstructed the most obfuscated images, i.e. the toughest attacker. In our case, the pixelating technique has two DL-assisted attackers (a) SRGAN and (b) SRResNet. As a first step, the intra-attacker comparison selects the method with the minimum RC-value, SRResNet (c.f. figure 6.a). As a second step, the inter-attacker comparison selects the obfuscation technique that achieves the highest RC-value (the most resilient obfuscation technique against its toughest attacker). As seen in figure 6.b, the obfuscation technique “Motion Blur” has the highest RC-value which means it is the most resilient obfuscation with regard to the SSIM metric.

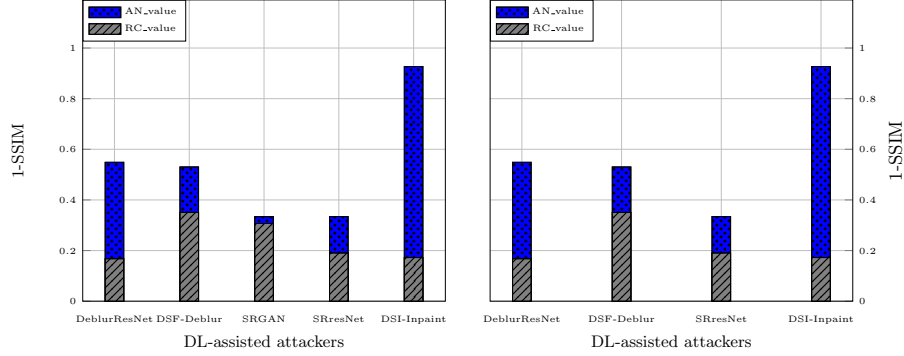


Figure 6: The *SSIM-based evaluation module* output before and after the intra-attacker comparisons

4.5 Identity-based evaluation

Our proposed framework measures the identity reconstruction of the obfuscated faces via two evaluation modules: (a) the *OpenFace-based* and (b) *Recognition-based evaluation modules*.

4.5.1 OpenFace-based evaluation

As its name indicates, the *OpenFace-based evaluation module* computes an identity distance measured by the Openface toolbox and normalize it to a value between 0 and 1. Two faces belong to the same identity if the distance is below a threshold of 0.275 (after normalization) [19]. Hence, we report in figure 7 the average values outputted by the evaluation module and in figure 8 the percentage of the values below the 0.275 threshold. Opposite to the values reported in figure 7, higher values in figure 8 signals higher level of privacy breaches. Based on the intra/inter-attacker comparisons, both results show “masking” as the most resilient obfuscation technique regarding the identity distance metric. The intra-attacker comparison first selects the obfuscation technique that achieved respectively the lowest and the highest RC-values in figure 7.a and figure 8.a for each obfuscation technique (the toughest DL-assisted attacker against each obfuscation technique). Followed by the inter-attacker comparison which selects the obfuscation technique that achieved the highest RC-value in figure 7.b and the lowest in figure 8.b, i.e. “masking” (the most resilient obfuscation techniques against the different DL-assisted attackers).

4.5.2 Recognition-based evaluation

The *recognition-based evaluation module* computes the average relative error based on the inferences of a pre-trained celebrity recognition model. All the plain ground-truth images in the test set are correctly recognized with a probability higher than 0.7. In the figures below, we report the average *AN*-values,

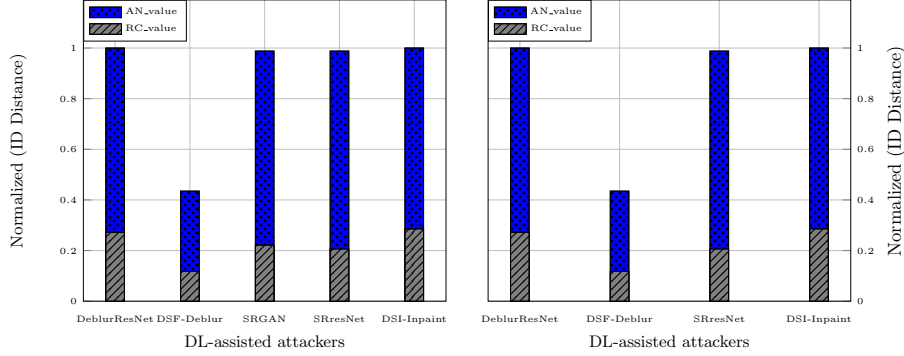


Figure 7: The *OpenFace-based evaluation module* output before and after the intra-attacker comparisons

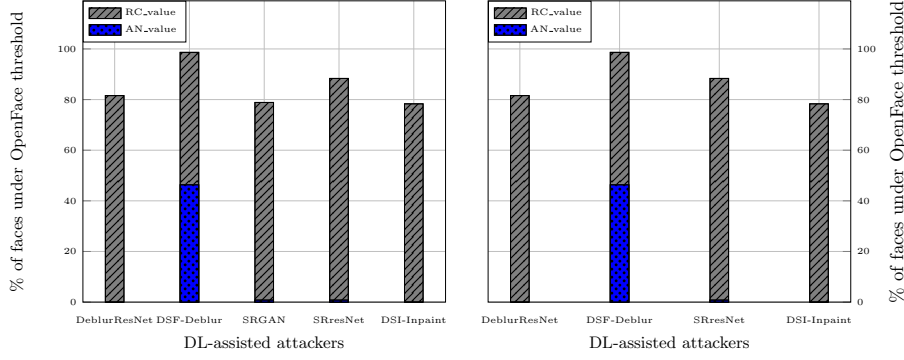


Figure 8: The Percentage of faces under the OpenFace ID threshold (0.275) before and after the intra-attacker comparison

RC-values and the top-5 recognition accuracy⁵ for the different reconstruction techniques. Opposite to the values reported in figure 9, higher values in figure 10 signals higher level of privacy breaches. Based on the intra/inter-attacker comparisons, both results present “Gaussian blur ($\sigma = 5$)” as the most resilient obfuscation technique regarding the recognition-based metric. The intra-attacker comparison first selects the obfuscation technique that achieved the lowest *RC*-value in figure 9.a and the highest in figure 10.a for each obfuscation technique (the toughest DL-assisted attacker against each obfuscation technique). Then, the inter-attacker comparison chooses the obfuscation technique that achieved respectively the highest and the lowest *RC*-values in figure 9.b and figure 10.b, “Gaussian blur ($\sigma = 5$)” (the most resilient obfuscation against the different DL-assisted attackers).

⁵The top-5 recognition accuracy is the percentage of obfuscated/reconstructed images where one of the top 5 predicted celebrity names match the top predicted celebrity name of the ground-truth image.

In short, we can summarize our results as follows: the *SSIM-based evaluation module* selects “Motion Blur” as the most robust technique. Whereas the *OpenFace-based evaluation module* selects “masking” and the *recognition-based module* chooses “Gaussian blur”. We can pick the most resilient technique based on the evaluation metric desired. As a future step, we would like to add a *human-based evaluation module*. Therefore, we would have the possibility to select the most resilient obfuscation technique against reconstruction attempts with regard to human evaluators.

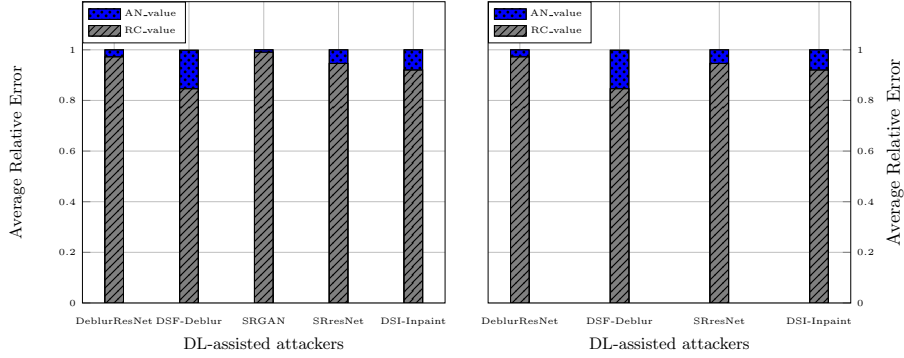


Figure 9: The *Recognition-based evaluation module* output before and after the intra-attacker comparisons

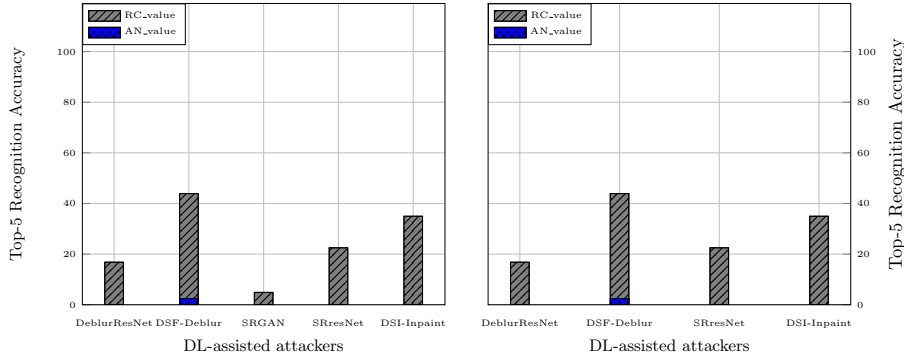


Figure 10: Top-5 recognition accuracy of the pre-trained celebrity recognition model

5 Related Works

Pixelating, blurring and masking are mainly the most obfuscation techniques used for identity protection in the context of images[7, 8]. In [8], Lander et

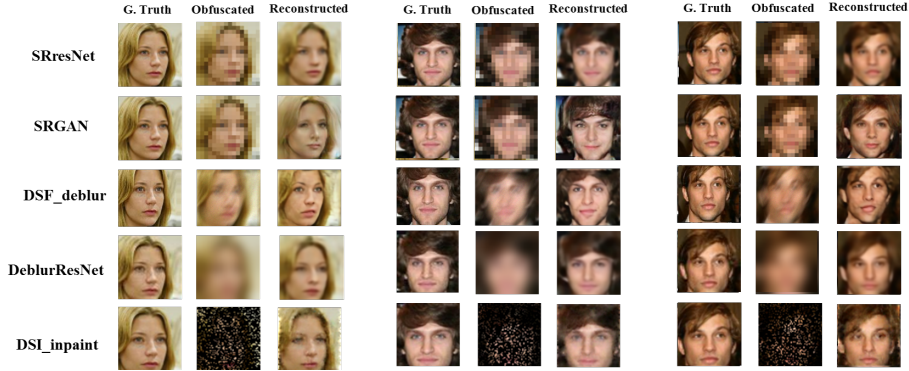


Figure 11: Comparison of the different reconstructions. Columns from left to right include Ground truth, Obfuscated and Reconstructed faces. Rows from top to bottom include the DL-assisted attackers

al. asked participants to identify famous people in obfuscated movie clips and static images in order to evaluate the effectiveness of blurring and pixelation. They showed that celebrities can still be recognized after obfuscation. However, they reported higher identification rate for movie clips compared to static pictures. In [7], Newton et al. examined pixelating and masking as face obfuscation techniques. They first measured the similarity between obfuscated probe images and unaltered gallery images based on the PCA Eigenspace[27]. Then, they measured the minimal distance between the obfuscated probe and gallery images after re-training the PCA-based recognition on obfuscated images. They achieved a higher recognition rate. Nevertheless, they only considered obfuscating a small rectangle on the top part of the face, including the eyes and top of the nose. Inspired by [7], Gross et al. suggested a new obfuscation technique[28]. They also 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[27]. They showed that small pixelating windows (e.g., 2x-4x) and simple blurring would not prevent identification attacks. In addition, they tried to super resolve the image[29] but the recognition rates decreased. In another study [30], Gopalan et al. presented a method to recognize faces obfuscated with non-uniform blurring by examining the blurred images. As a follow-up study [31], 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. In a more recent study, the authors demonstrated that modern image recognition approaches, based on artificial neural networks, can be employed as attackers to recover hidden information from obfuscated protected images [6]. They focused on three forms of obfuscation: pixelating, blurring and P3 (an encryption-based method[32]). The attacker successfully identifies obfuscated faces and objects by training image recognition networks on obfuscated images (faces[33, 34],

digits[35] and objects[36]). While all the above studies strengthen the attacker by re-training image recognition approaches (PCA-based or NN-based) on obfuscated images, we focus in our work on reconstructing the obfuscated features and using face recognition methods as unbiased evaluators of the reconstruction.

The authors in [37] 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). The accuracy of the face detection algorithm was used as a measure of intelligibility (a face should be detected). They applied traditional methods for face recognition such as PCA[27], LDA[38] and LBP[39]. 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 [37] 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 the latter, we identify the most robust obfuscation technique against DL-assisted attackers by evaluating both, structural and identity-based metrics, which provide more profound insights.

In a similar study to ours, the authors in [9] tackled the privacy-preservation question in the context of obscured 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 [40] and interpolation-based [11] techniques for super resolution). In addition, they evaluated the identity restoration using the same traditional face recognition techniques as in [37]. In our framework, we adopted DL-based techniques for both face reconstruction and recognition because as stated in [13, 19], DL-based techniques demonstrate great superiority over traditional methods. In addition, our framework recommends with regard to structural/identity-based evaluation metric the most robust obfuscation. Alternatively, the authors in [10] investigated the amount of obfuscation needed to guarantee patients anonymity. They applied CycleGAN[41] in order to reconstruct features from anonymized medical imaging. They considered two anonymization techniques: (a) blurring and (b) masking. They also provided qualitative comparison by visually showing the results and a quantitative comparison in the form of correlation coefficients and SSIM between the original and reconstructed images as well as between the original and anonymized images. In our approach, we propose a generic and scalable framework. Generic because it can be adapted to different sensitive information (e.g. faces, badge names, etc.). Scalable because additional DL-assisted attacks can be added to the reconstruction unit (c.f. figure 3.b).

6 Conclusion

In this paper, we proposed a generic framework to evaluate and recommend the most robust obfuscation techniques for specific sensitive information, such as an individual’s face. The framework reconstructs obfuscated faces via DL-assisted attackers, evaluates the reconstructions via structural/identity based metrics and recommends the most robust obfuscation with regard to each metric. In order to validate the proposed approach, we conducted our experiments on the CelebA dataset. We considered four obfuscation techniques in the context of facial images and five DL-assisted attackers. “Motion Blur” was recommended as the most robust obfuscation with regard to the SSIM metric whereas “Gaussian Blur” and “Masking” were recommended respectively with regard to the OpenFace identity distance metric and the celebrity recognition model. In the near future, we will consider adding stronger DL-assisted attackers and using additional evaluation metrics (e.g., human evaluators) to evaluate more efficiently the robustness of the obfuscation techniques. We will also extend our framework to cover additional data types such as audio and textual data.

References

- [1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, *CVPR* (2016)
- [2] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, SSD: Single shot multibox detector, arXiv:1512.02325, (2015)
- [3] 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)
- [4] S. Hill, Z. Zhou, L. Saul, and H. Shacham. On the in effectiveness of mosaicing and blurring as tools for document redaction, *PETS*, (2016)
- [5] G. Boracchi and A. Foi. Modeling the performance of image restoration from motion blur, *Image Processing, IEEE Transactions*, 21(8):3502–3517, (2012)
- [6] R. McPherson, R. Shokri, and V. Shmatikov, Defeating image obfuscation with deep learning. *CoRR*, (2016)
- [7] E. M. Newton, L. Sweeney, and B. Malin. Preserving privacy by de-identifying face images, *IEEE transactions on Knowledge and Data Engineering*, (2005)
- [8] 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)
- [9] Ruchaud, N. and Dugelay, J. L., Automatic Face Anonymization in Visual Data: Are we really well protected? *Electronic Imaging*, (2016)
- [10] D. Abramian and A. Eklund, Refacing: reconstructing anonymized facial features using GANs *COCR*, volume abs/1810.06455, (2018)
- [11] R. Keys. Cubic convolution interpolation for digital image process-

- ing, *Acoustics, Speech and Signal Processing, IEEE Transactions*, 29(6):1153–1160, (1981)
- [12] 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)
 - [13] 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)
 - [14] 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)
 - [15] Z. Shen, W. Lai, T. Xu, J. Kautz and S.M. Yang, Deep semantic face deblurring *CVPR* pages 8260–8269, (2018)
 - [16] D. Garcia., srez: Adversarial super resolution, <http://github.com/david-gpu/srez>, (2016)
 - [17] 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)
 - [18] Y. Li, S. Liu, J. Yang, and M.-H. Yang., Generative face completion, *arXiv preprint arXiv:1704.05838*, (2007)
 - [19] 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)
 - [20] 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)
 - [21] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework, *Proceedings 2nd Int. Conf. on Knowledge Discovery and Data Mining*, pp. 82–88, (1996)
 - [22] 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)
 - [23] S. Majumdar Image Super Resolution, <https://github.com/titu1994/Image-Super-Resolution>, (2016)
 - [24] X. Yu and F. Porikli, Ultra-Resolving Face Images by Discriminative Generative Networks, Springer International Publishing, pages 318–333, (2016)
 - [25] Z. Shen Deep-Semantic-Face-Deblurring, <https://github.com/joanshen0508/Deep-Semantic-Face-Deblurring>, (2016)
 - [26] C.B. Jin, Semantic-image-inpainting, <https://github.com/ChengBinJin/semantic-image-inpainting>, (2018)
 - [27] M. Turk and A. Pentland, Face recognition using eigenfaces, *Computer Vision and Pattern Recognition, Proceedings CVPR '91., IEEE Computer Society Conference*, (1991)
 - [28] R. Gross, L. Sweeney, F. De la Torre, and S. Baker. Model-based face

- de-identification, *CVPRW*, (2006)
- [29] R. Hardie, K. Barnard, and E. Armstrong. Joint map registration and high-resolution image estimation using a sequence of undersampled images, *IEEE Transactions on Image Processing*, 6(12):1621–1633, (1997)
 - [30] 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)
 - [31] A. Punnapurath, 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)
 - [32] M.-R. Ra, R. Govindan, and A. Ortega. P3: Toward privacy-preserving photo sharing, *NSDI*, (2013)
 - [33] H.-W. Ng and S. Winkler. A data-driven approach to cleaning large face datasets In *IEEE International Conference on Image Processing (ICIP)*, (2014)
 - [34] Laboratories Cambridge. The database of faces, (1994)
 - [35] Y. LeCun, C. Cortes, and C. J. Burges. The mnist database of handwritten digits, (1998)
 - [36] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images, (2009)
 - [37] 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)
 - [38] 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)
 - [39] 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)
 - [40] W. Dong, L. Zhang, G. Shi, and X.Wu. Image deblurring and superresolution by adaptive sparse domain selection and adaptive regularization, *Image Processing, IEEE Transaction*, 20(7):1838–1857, (2011)
 - [41] 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)