Explainability-Driven Incremental Image Anonymization

Rami Haffar, David Sánchez, Younas Khan, Josep Domingo-Ferrer

Universitat Rovira i Virgili, Department of Computer Engineering and Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Av. Països Catalans 26, 43007 Tarragona, Catalonia

E-mails: rami.haffar@urv.cat, david.sanchez@urv.cat, younas.khan@urv.cat, josep.domingo@urv.cat

Received 18 February 2025; received in revised form 23 May 2025; accepted 25 May 2025

Abstract. Privacy regulations require that images depicting humans be anonymized before they are publicly released or shared for secondary use. However, current image anonymization methods significantly degrade the analytical utility of protected images. This paper addresses the challenge of balancing privacy protection and utility preservation in image anonymization. We propose a general disclosure risk-aware anonymization framework that leverages explainability techniques to target identity-revealing features in images. Contrary to conventional methods, which uniformly perturb all image pixels, our proposal focuses on perturbing the pixels that contribute most to disclosure. Moreover, pixel perturbation is enforced incrementally and it is driven by the observed residual risk. Our framework is not tied to a specific pixel perturbation mechanism, and is versatile enough to support a wide variety of techniques, including blurring, pixelation, noise addition and pixel masking. Empirical results show that even with the simplest perturbation techniques, our approach significantly improves the privacy/utility trade-off compared to conventional and advanced state-of-the-art methods.

Keywords. Image Anonymization, Privacy Protection, Utility Preservation, Explainability, Machine Learning

1 Introduction

In today's ubiquitous world of high-resolution mobile devices, people easily share personal images on social networks. Instagram alone witnessed the posting of more than 1,100 images every second in 2022 [13]. In particular, images with human faces are prevalent because they are 38% more likely to attract engagement on social networks [1].

However, when an image depicts a human being, it poses privacy risks due to the identifying features of human faces [28]. Facial recognition (FR) methods leverage these features to create a distinctive identifier for re-identification, and this has privacy implications. Specifically, when an image is successfully linked to a real identity, the image content can be leveraged to infer sensitive information [42], such as personal preferences, habits, political affiliations, or health conditions. Moreover, aggregating extracted data from multiple re-identified images can lead to invasive profiling, which can be exploited for various purposes, including harassment and surveillance [29]. To forestall these privacy threats, image anonymization methods are used to remove or obfuscate the re-identificative features in the images [27].

On the other hand, face images are very useful for research, including emotion recognition or biometric research [4, 34, 14, 17, 41]. Most of these applications do not actually require re-identifiable images, but they need (privacy-protected) source images that retain enough *analytical utility* [16]. Preventing re-identification while retaining as much utility as possible is therefore the ultimate goal of image anonymization methods.

The most widely used image anonymization methods in real applications are blurring and pixelation [26, 45]. Differential privacy-based pixelation and blurring also add noise, which is bounded by a specific distribution [11]. A common feature of all these methods is that they apply a uniform distortion to all the pixels in the image. However, several studies have concluded that both humans and algorithms focus on specific facial features for re-identification, such as the eyes and forehead [6, 18, 44]. Hence, blindly and uniformly distorting all image pixels unnecessarily degrades the utility of the anonymized images. Even if some methods focus on masking specific facial features such as eyes or mouth [37], they still enforce a fixed masking criterion (e.g., all pixels in the targeted areas are blacked out) and, therefore, result in a suboptimal privacy/utility balance.

This paper presents a novel image anonymization framework whose balance between privacy protection and utility preservation is better than that of existing methods. Rather than using a fixed masking criterion, the proposed framework is disclosure risk-aware and it leverages explainability techniques to detect and target identity-revealing pixels within the images. As a result, and in contrast to conventional methods that uniformly perturb image pixels, our framework focuses on perturbing only those pixels that contribute the most to disclosure. Moreover, pixel perturbation is enforced incrementally and it is driven by the observed residual risk in order to add the minimal distortion required to attain the desired level of privacy protection. Thus, our anonymization framework might be dubbed "utility-first", in that it strives to preserve as much utility as possible. The empirical results we report show that our approach is able to encompass a wide variety of image perturbation techniques, and provides a significantly better privacy/utility balance than the state of the art even when using the simplest perturbation techniques.

The remainder of this paper is organized as follows. Section 2 discusses related work on image anonymization. Section 3 presents our image anonymization framework. Section 4 reports and discusses experimental results. Finally, Section 5 gathers the conclusions and depicts several lines of future research.

2 Related work

The earliest and most widely used methods for preventing face-based re-identification are pixelization and blurring [47]. Pixelization reduces the amount of information by segmenting the image into $n \times n$ blocks and setting all the pixels within each block to the same intensity resulting from averaging neighboring pixels. On the other hand, blurring is used to reduce the sharpness and clarity of an image by introducing Gaussian noise [35]. This noise is applied by using a Gaussian kernel of size k, which averages the pixel values with those of their neighbors. Since more weight is given to the pixels closer to the center of the kernel, the resulting images have a smoother appearance.

In [10] the author enforced pixelation under the umbrella of ϵ-differential privacy (DP). DP offers a formal privacy guarantee, which states that the presence or absence of any individual's data in a dataset should not be noticeable in the protected output up to an exponential factor denoted as ϵ [9]. DP is enforced by adding Laplacian noise to the data proportional to i) the ϵ value (the lower the ϵ value, the larger the noise and the higher the protection), and ii) the global sensitivity (GS) of the output to the presence or absence of any individual's data in the dataset to be protected. DP was originally defined to protect answers to statistical queries on a remote database; in that setting, aggregated queries are quite insensitive to changes in individual data points and relatively little noise may suffice for protection. However, DP poses significant challenges when applied to data releases [8], as is the case for the release of collections of images. For images, noise should be proportional to the maximum change that can occur to any individual's image; such a change may entail replacing all image pixels with max/min RGB values. Applying DP under this premise would produce nearly random images with no utility.

To avoid this issue, [10] employs the notion of m-neighboring images, which limits to m the maximum number of pixels that can be changed when calibrating noise. The rationale is that not all the pixels of an image are related to the depicted individual (e.g., background pixels); therefore, it is possible to confine all individual-related/disclosive pixels to a set of size m. The value of m can be customized to account for the individual's silhouette or specific facial features (e.g., the eyes) but it should be quite low w.r.t. the image resolution for the utility of the protected images to be reasonably preserved. For instance, the author used a value of $m = 16$ for face images with more than 10,000 pixels and, even in that case, the loss of visual quality was noticeable.

It is important to note that the original formulation of DP requires that *any* data related to the individual to be protected be accounted for when calibrating noise. Assuming that only 0.16% of pixels in a face image are related to the depicted individual, i.e., 16 out of 10,000 pixels, as done in [10], hardly captures the guarantee DP aims to offer.

To further reduce the noise to be added, [10] first pixelates the input image, which reduces the sensitivity GS to changes in individual pixels. The value of GS is thus $255m/b^2$ for each $b \times b$ grid cell. The perturbation is applied to the pixelated image using the Laplacian noise distribution with mean 0 and scale of $255m/(b^2\epsilon)$. This method, called *DP-Pix*, is formalized as

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

where DP-Pix$_{(i,j)}$ is the pixel value in the DP-pixelated image and $P_{(i,j)}$ is the pixel value in the pixelated image.

In [11], the author extended the method above by adding a post-processing step based on Gaussian blur to produce smoother images. This method, called *DP-Blur*, is formalized as

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

Regardless of the questionable application of the DP principles, *DP-Pix* and *DP-Blur* produce more privacy-preserving anonymizations than the plain application of pixelation and blurring [10], but at the cost of a larger utility loss.

All the methods considered so far apply a uniform perturbation to all image pixels. In contrast, some image protection methods, such as eye and mouth masking, focus on specific facial features. These techniques superimpose black circles or boxes on the images to emulate sunglasses or face masks in order to obfuscate facial features that are crucial for re-identification [12, 20]. However, these methods also present several limitations. Firstly, research shows that masking these facial regions may not be entirely effective, and that their performance closely depends on the size of the drawn circles or boxes [37]. Also, some studies suggest the importance of masking other facial features, such as eyebrows [38].

Moreover, determining common parameters for detecting and masking eyes and mouths in a large and heterogeneous dataset can be complex, often requiring manual adjustment due to variations in facial structures, poses and feature sizes.

More recently, there has been a rise in machine learning (ML)-based methods for anonymizing images based on generative adversarial networks (GANs) [24, 33, 22]. GANs are distinguished for their ability to create synthetic images that conceal identifiable information [15]. GANs comprise two neural networks: i) a generator network responsible for generating synthetic images, and ii) a discriminator network tasked with discerning between real and synthetic images. Through adversarial training, these networks collaborate until they converge and generate synthetic images that satisfy privacy requirements.

StyleGAN-2 [23] is a GAN-based method for anonymizing images that employs latent vectors as replacements for original images to generate new ones. To enhance image quality, multiple latent vectors are provided to the model, and the anonymized image with the best quality is selected from the generated outputs. Perceptual Indistinguishability-Net (PI-Net) [5] employs an encoder-decoder architecture to produce perceptually indistinguishable obfuscated images. Like StyleGAN-2, PI-Net operates in the latent space; however, instead of replacing original images by latent vectors, it extracts the vector from the original image and uses it to generate a single anonymized version. Unfortunately, both methods lack proper privacy evaluation, as the data used in their experiments are not annotated for face identification. Furthermore, neither the code nor the trained models are publicly available, hindering their reproducibility.

Other state-of-the-art GAN-based methods, such as CIAGAN [33] and DeepPrivacy [22], leverage conditional generative adversarial networks (cGANs) for image anonymization. These methods preserve background information by incorporating it into the GAN generator while using an autoencoder to extract a latent vector that captures pose and facial features. CIAGAN introduces a one-hot identity vector into the generator's bottleneck to ensure that the generated face corresponds to an identity not present in the dataset. In contrast, DeepPrivacy does not impose such constraints. Based on this, DeepPrivacy2 [21] enhances the original DeepPrivacy approach by ensuring that the generated face does not appear in the dataset and extending anonymization to the entire body.

3 Explainability-Driven Incremental Image Anonymization

The conventional anonymization techniques discussed in Section 2, such as pixelation and blurring, operate in a one-shot and untargeted manner: they take an input image and apply a uniform perturbation across all pixels according to a predefined privacy parameter to produce an anonymized image in a single step. Facial masking also operates in a one-shot manner, but confines perturbation to the eyes or the mouth of the individual. Still, the perturbation applied to those facial features –which need to be defined beforehand– is uniform. This conventional image anonymization workflow is depicted in Fig. 1.

We argue that these conventional approaches incur an unnecessary loss of utility due to a fundamental intuition: not every pixel contributes the same to re-identification and, therefore, does not require the same level of perturbation (if any). This insight led us to develop a targeted anonymization framework that aims to perturb or obfuscate only the identity-revealing pixels, and up to a degree proportional to their contribution to re-identification. The goal of this disclosure risk-aware solution is to reduce the re-identification risk (RIR) of images while preserving as much analytical utility as possible. It is noteworthy that our framework neither proposes nor is tied to a specific image perturbation technique, but can

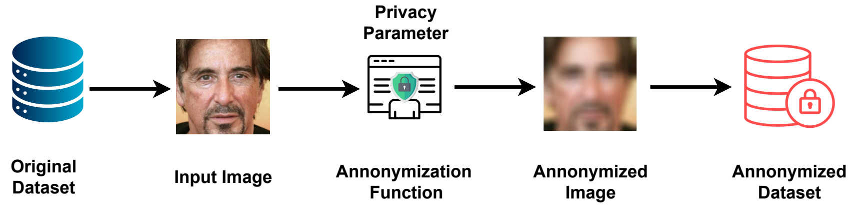


Figure 1: Workflow of conventional image anonymization techniques

enforce existing techniques –including pixelation, blurring, noise addition, and masking–, under this new image anonymization paradigm.

Disclosure awareness is enabled in our approach by using an ML FR model trained on the images to be anonymized.

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

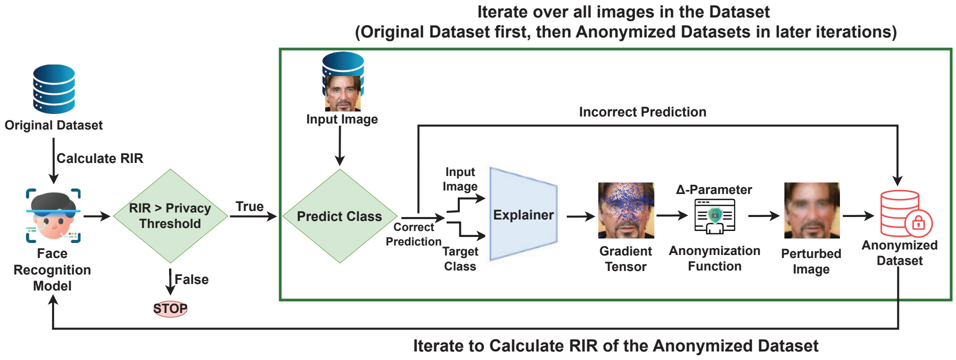


Figure 2: Workflow of EDI-Anon

3.1 EDI-Anon modules

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

3.1.1 FR model

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

On the one hand, the FR model is used to quantify the RIR of the (preliminary) anonymized dataset that results from each iteration. RIR is calculated as the accuracy achieved by the FR model on the input data, as follows:

$$RIR = \frac{Correct\_Predictions}{Total\_Predictions}. \tag{1}$$

On the other hand, the FR model is also the basis for detecting the pixels that contributed the most to successful re-identification, as described in Section 3.1.2.

3.1.2 Explainer

The *explainer* provides information about the impact of each individual pixel on the FR model's predictions, thereby allowing us to detect the pixels that contributed the most to re-identification, which we call *identity-disclosing pixels set* (IDPS). We opted to use the explainer introduced in [18], which leverages adversarial examples to alter the model predictions. This method provides a comprehensive weighted list that reflects the influence of each pixel on the model's prediction. In contrast, other explainers like LIME (Local Interpretable Model-agnostic Explanations) [36] employ linear models to interpret the model's predictions. However, they categorize the image pixels into only two labels, making them less suitable to guide the anonymization process.

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

These gradients represent the sensitivity of the loss to changes in individual pixels. Higher magnitude gradients indicate pixels that have a greater influence on the model's prediction, especially for misclassification (a sample of such pixels is depicted in Fig. 2 with a blue shade on the image to be protected). Therefore, by analyzing the gradients, we can detect which pixels, if changed, would increase the likelihood of the model misclassifying the input; or, in other words, which pixels, if obfuscated, will result in an incorrectly re-identified (that is, anonymized) image. These pixels, which consequently contribute the most to re-identification, constitute the IDPS, and will be subjected to perturbation.

3.1.3 Anonymization function

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

IDPS obtained in the previous step. In addition to being targeted, pixel perturbation is also applied *incrementally*: instead of perturbing pixels in a single shot, we perturb them incrementally over several times, by taking each time a *weak* value of the method's privacy parameter (e.g., the pixel block size for pixelation or the kernel size for Gaussian noise), that is, a value that achieves only a small privacy increment (and presumably a small utility decrement). We call the privacy parameter used in this manner the Δ-parameter. As we will describe in Section 3.2, by iteratively applying perturbations to each image in small increments, we can re-evaluate the RIR of each previously perturbed image through the FR model and apply (if needed) further perturbations on a freshly re-assessed IDPS. The motivation underlying this targeted incremental approach is to minimize the amount of perturbation needed to reach a desired level of protection, and thereby maximize utility preservation.

3.1.4 Privacy threshold

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

In our approach, the desired level of privacy is set as a *privacy threshold* that defines the maximum RIR allowed for the anonymized dataset. RIR is recomputed with Equation (1) at the end of each iteration by giving the perturbed image dataset resulting from that last iteration as input to the FR model. The resulting RIR value is used as the stopping criterion: the process halts when the observed RIR falls below the user-defined privacy threshold, as depicted in Fig. 2. Defining the desired privacy level in terms of RIR is also more intuitive and aligns better with real-world privacy requirements than employing the abstract and non-privacy-grounded parameters of conventional methods (such as pixel block or kernel sizes).

3.2 EDI-Anon workflow and algorithm

Next, we describe in detail the workflow of EDI-Anon in detail, which is also formalized in Algorithm 1.

First, the FR model is trained on the original image dataset (line 3 of Algorithm 1). Then, the anonymized dataset is initialized to the original data set, and the RIR of the anonymized dataset is assessed through the FR model (line 5). If it does not fulfill the user-defined privacy threshold (line 6), each input image in the anonymized dataset is individually assessed for re-identification using the FR model (line 8). Each correctly re-identified input image (line 9) then undergoes the following process in order to be anonymized:

1. Calculate its gradient tensor using the explainer to quantify the contribution of each pixel toward re-identification as detailed in Section 3.1.2 (line 10).

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

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

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

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

In addition to the input parameters that define the type of anonymization function to be used (i.e., pixelation, blurring, noise addition, or masking) and the desired privacy threshold for the RIR, the algorithm relies on two interrelated constants for the anonymization function: the IDPS percentage and the Δ-parameter. Both control the amount of distortion added to the images at each iteration: whereas the IDPS percentage defines the number of pixels to be distorted, the Δ-parameter controls the amount of distortion added to those pixels. Their values thus influence i) the speed of the process (i.e., the larger they are, the fewer iterations are required to reach the desired level of privacy) and ii) the granularity of the added perturbation (i.e., the smaller the parameter values, the more targeted the added perturbation is, because more re-assessments of the IDPS are performed). The nature of the Δ-parameter depends on the type of anonymization function. For instance, for pixelation the Δ-parameter is block size, for blurring it is kernel size, and for DP-based noise addition it is ϵ. Even though the values of those parameters are, in principle, fixed for a given anonymization function, advanced users might want to tune them in order to encompass specific utility/performance needs. The influence of such tuning will be evaluated in Section 4.4.1.

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

4 Empirical results

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

All experiments were carried out on an AMD Ryzen 5 3600 CPU (base speed 3.6 GHz), 32 GB of RAM, and an NVIDIA GeForce RTX 3060 GPU (12 GB VRAM).

4.1 Evaluation dataset

Since we are interested in measuring both privacy protection and utility preservation of image anonymization, we chose the Multi-Task Face (MTF) dataset [19] because it is the only available dataset that provides comprehensive multi-label annotations necessary for this dual evaluation. The MTF dataset consists of a manually curated collection of 5,246

Algorithm 1 Explainability-Driven Incremental Anonymization

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

high-resolution images (1024 × 1024) depicting the faces of 240 individuals. Unlike other datasets, such as CelebA [31], which contains low-resolution images and suffers from significant labeling inaccuracies due to automatic annotation [30], MTF offers accurate and manually verified annotations for identity, age, gender, and race. These annotations are essential for evaluating both the success of anonymization in preventing face recognition (FR) and utility preservation across key classification tasks. Furthermore, MTF follows legally and ethically compliant curation practices, ensuring its long-term availability and reproducibility. Table 1 reports the distribution of annotation labels in all four tasks.

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

Table 1: Number of individuals for the various tasks and labels in the MTF dataset

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

4.2 Evaluation metrics

To evaluate privacy, we employed the RIR metric defined in Equation (1), which is calculated by using the FR model trained on the original FR data from the MTF dataset. Our evaluation considers a worst-case scenario in which the adversary is assumed to have complete access to all images of the individuals. The privacy figures we report are thus on the conservative side, and better privacy will be attained in more plausible attack settings, in which adversaries have access to smaller sets of images.

The FR model is a ConvNeXt convolutional neural network [32], which is recommended by the authors of the dataset. The base model is pre-trained on ImageNet [7]. The last layer is adapted to the number of classes (i.e., number of individuals) in the dataset. To train the FR model, we used cross-entropy loss as the optimization criterion [46] and Adam [25] as the optimizer for updating the model parameters. The model was trained for 100 epochs. Recall that RIR measures the accuracy of the predictions without making a distinction between false positives and false negatives, and thereby captures the overall correctness of the FR task.

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

To assess utility preservation, we focused on the other three tasks supported by the MTF dataset: age, gender and race classification. Following the methodology outlined in the MTF paper [19], we trained three corresponding ML models on the original (non-anonymized) images of the training data set by using the ConvNeXt pre-trained Base model. The hyperparameters for those three models were the same as used in the MTF paper, with a batch size of 32 images.

Utility preservation was measured as the performance of the three models on the anonymized images. The models' performance was quantified via the F_1 score, which is the harmonic mean of precision and recall [43], and is calculated as

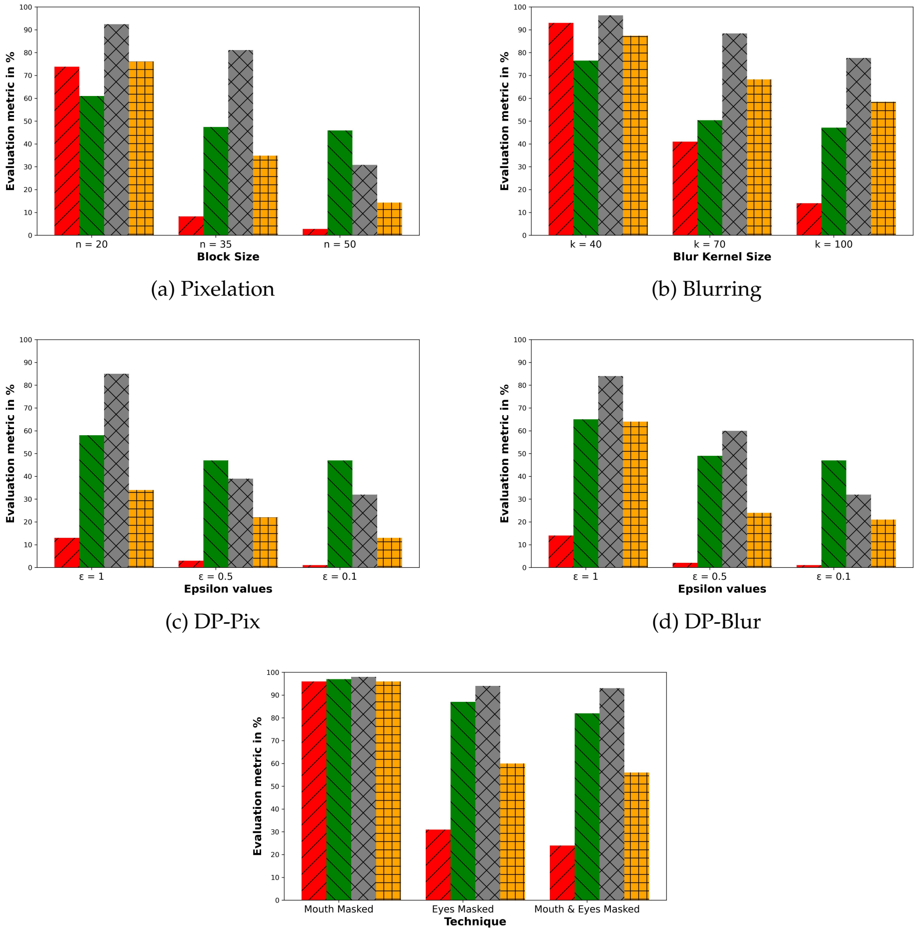
$$F_1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall},$$

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

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

4.3 Conventional methods

For pixelation and blurring, the privacy parameters used were n (block size) and k (blur kernel size), respectively. The proposed values for these parameters in [45] and [40] –which did not evaluate empirical re-identification– were $n \in [4, 10]$ and $k = 25$. However, according to our privacy evaluation, these resulted in negligible reductions in RIR. Therefore, we decided to use larger parameter values to produce observable differences, as shown in Figs.



(a) Pixelation

(b) Blurring

(c) DP-Pix

(d) DP-Blur

(e) Mouth and eye masking

Figure 3: Evaluation results of conventional anonymization methods for different privacy parameters. Within each group, from left to right, the bars correspond to RIR, F_1-Age, F_1-Gender, and F_1-Race.

3a and 3b. For DP-Pix and DP-Blur, [11] employed $\epsilon \in \{0.1, 0.5, 1\}$ and $m = 16$ neighboring pixels. Furthermore, for DP-Blur, a kernel size (k) of 99 was recommended. We used the same parameter values in our tests, even though $m = 16$ only accounts for 0.0015% pixels in our images. Considering that, in face images, a substantial proportion of pixels contribute to re-identification, it is clear that such a small set of pixels would not result in any meaningful DP guarantee. When no *ex ante* privacy guarantees hold, the attained level of privacy can only be evaluated empirically [2], as we do in this section. As a matter of fact, setting larger –and safer– m (ideally, the whole image, $m = 1024 \times 1024$) produced random pixels for any value ϵ, even for extremely large ones.

For masking, we simulated the effect of wearing a face mask and/or sunglasses. We used Haar-cascade [39] for mouth and eye detection via the OpenCV library [3], and applied enough large round and square masks to hide the detected facial features.

As shown in Fig. 3e, the least effective technique w.r.t. RIR was mouth masking, due to the mouth region lacks identity-revealing features. Although eye masking proved more effective at concealing the individual's identity, it still yielded an RIR larger than most other methods, while combining mouth and eyes masking produced a minor decrease in RIR over just masking the eyes. However, this minor decrease in RIR came at the cost of a small reduction in utility for the classification of age and gender and a more noticeable reduction for the classification of races. The latter can be explained due to the importance of the shape of the eyes to recognize Asian faces.

Pixelation and blurring, which apply a uniform perturbation across all pixels, achieved stronger privacy (i.e., lower RIR) for $n > 20$ and $k > 40$. However, both methods significantly hampered classification by age and race, as shown in Figs. 3a and 3b. The accuracy of the age classification model was particularly affected by the misclassification of the *Old* class, because pixelation and blurring can hardly preserve facial wrinkles that are a distinctive feature of elderly individuals. On the other hand, the performance of the race classification model was hampered by the loss of finer details, including the shapes of the eyes, lips, and noses, which are key for identifying the *Asians*. In particular, blurring performed better to preserve the *black* class, due to the unchanged facial contours and skin tones.

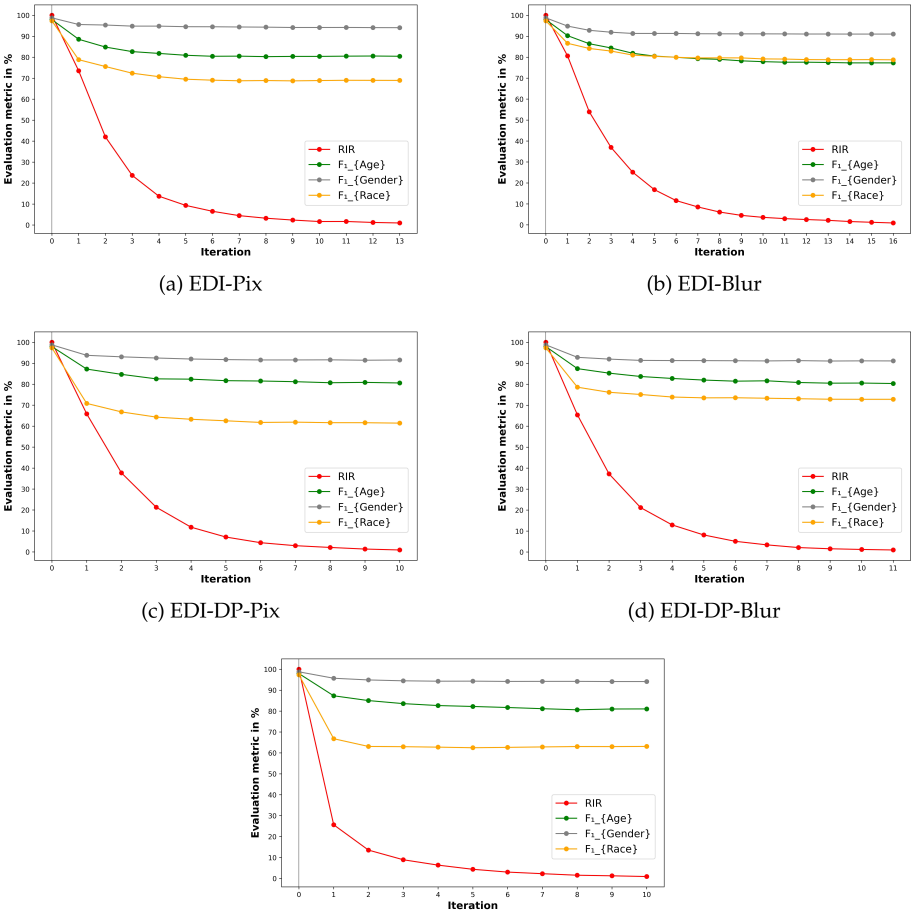
Although DP-based methods lack formal privacy guarantees for face images due to the small value of m employed, both DP-Pix and DP-Blur achieved the strongest empirical privacy, as reported in Figs. 3c and 3d. Lower ϵ values resulted in more effective anonymization for both methods. DP-Pix exhibited the lowest utility preservation, whereas DP-Blur offered a slightly better privacy/utility trade-off. The relatively better performance of DP-Blur can be attributed to the application of Gaussian blur, which improved image quality by smoothing the images and reducing the effect of the added (colored) noise. This was especially useful for race classification. Nevertheless, the utility reached for both methods with $\epsilon < 1$ approached the random guess for all tasks.

4.4 EDI-Anon

As explained in Section 3.1.3, EDI-Anon is enforced via an anonymization function that can implement a variety of conventional methods. In this section, we evaluate the application of the conventional methods considered in the previous section within the EDI-Anon framework. We refer to them by preceding their names by "EDI-". To accommodate the iterative nature of EDI-Anon, we took weak values of the privacy parameters of the different methods (i.e., what we refer to as Δ-*parameter*, as explained in Section 3.1.3). Specifically, for EDI-Pix and EDI-Blur, their respective Δ-parameters were set to $n = 20$ and $k = 40$ (i.e., the weakest values employed in the previous experiments), whereas for EDI-DP-Pix and EDI-DP-Blur we used a notably higher $\epsilon = 100$ while keeping m and k unchanged from their conventional counterparts with one shot. As discussed in Section 4.3, these loose parameters do not provide *ex ante* privacy guarantees for EDI-DP, but neither did the one-shot counterparts.

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

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



(a) EDI-Pix

(b) EDI-Blur

(c) EDI-DP-Pix

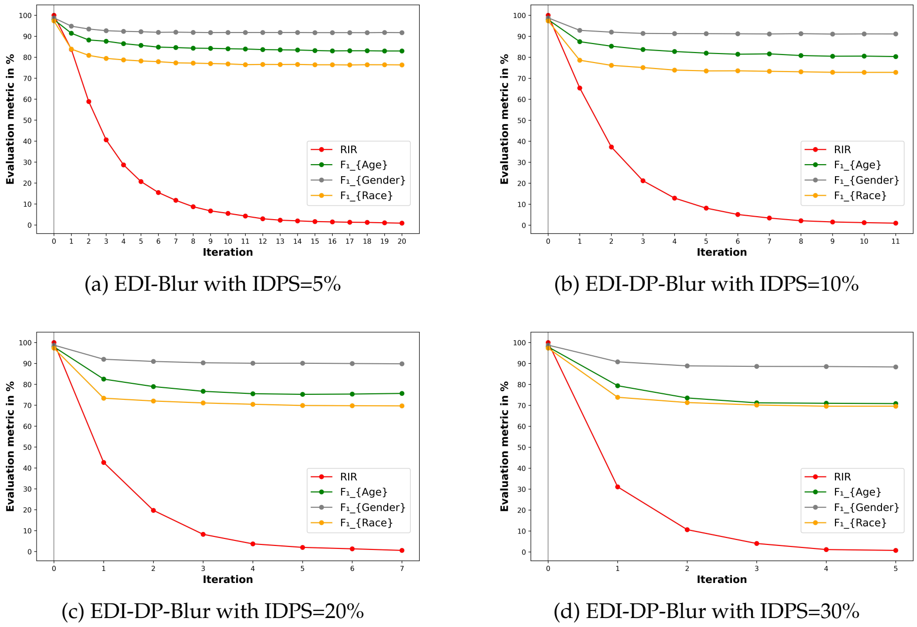
(d) EDI-DP-Blur

(e) EDI-Masking

Figure 4: Evolution of RIR and utilities for EDI-Anon across iterations with different anonymization functions until reaching a 1% privacy threshold

The privacy threshold for the EDI-Anon methods was set to the lowest RIR achieved by any conventional method, that is, the 1% achieved by DP-Blur. In Fig. 4 we report the evolution of the RIR and the utility metrics across iterations until that privacy threshold is reached.

We can see that, regardless of the underlying anonymization function, EDI-Anon was able to achieve the target privacy threshold with a similar number of iterations. Despite their differences, all EDI-Anon methods significantly lowered the RIR within the initial iterations, and most iterations were devoted to lower the RIR from 10% to 1%. The sharp decline in RIR was accompanied by a milder decrease in the utility metrics. Differences across methods manifest in some utilities being slightly better preserved than others. In general, all EDI-Anon methods affected all three utilities proportionally to their number of classes; that is, the more classes, the less utility because of having to deal with a more difficult classification problem. Age classification was similarly preserved across all meth-



(a) EDI-Blur with IDPS=5%

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

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

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

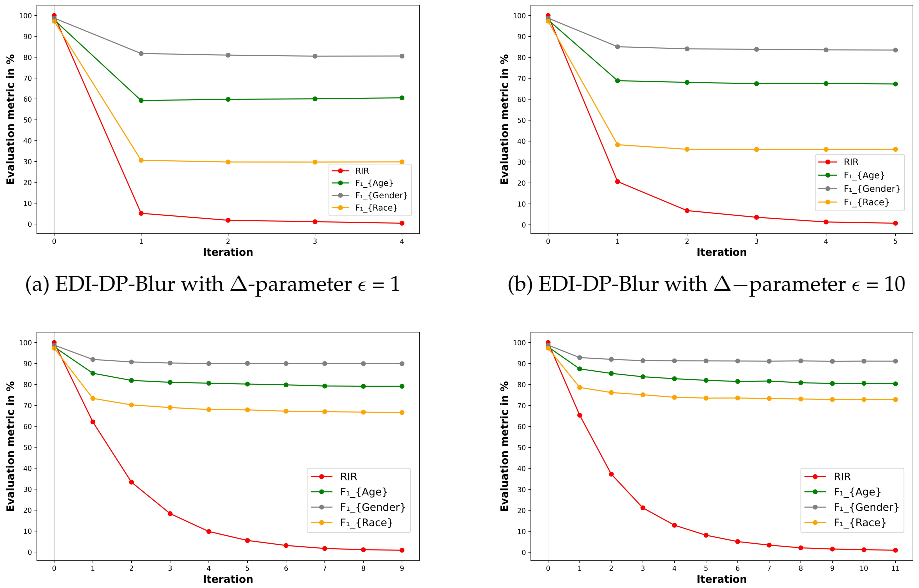
Figure 5: Evolution of RIR and utilities for EDI-DP-Blur across iterations with different IDPS Percentages until reaching a 1% privacy threshold

ods, indicating that age-related facial features were minimally affected by the EDI-Anon process. On the other hand, EDI-DP and EDI-Masking hampered race classification the most due to the loss of race-related facial features (e.g., those depicting eye shapes, facial contours, and skin tones).

4.4.1 On EDI-Anon constants

As explained in Section 3.2, in addition to the parameters that define the type of anonymization function to be used by EDI-Anon and the desired privacy threshold, two other values also influence the behavior of the algorithm: the IDPS percentage and the Δ-parameter of the chosen anonymization function. In the following, we analyze the impact of these values on EDI-DP-Blur, which provided a good balance between RIR reduction and utility preservation.

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



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

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

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

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

Figure 6: Evolution of RIR and utilities for EDI-DP-Blur across iterations with different Δ-parameter values until reaching a 1% privacy threshold

impact on the three utilities. This illustrates that the lower the IDPS percentage, the more accurate and targeted the added distortion is; yet, one should take into account the added runtime resulting from the additional iterations (see the runtime analysis in Section 4.5.3).

Second, we examined the influence of the Δ-parameter on EDI-DP-Blur by setting $\epsilon \in (1, 10, 50, 100)$ while keeping all other parameters unchanged. The findings are presented in Fig. 6. The results show a similar tendency as when modifying the IDPS percentage: the stricter the Δ-parameter (i.e., the lower ϵ is), the fewer iterations are needed to attain the privacy threshold but, at the same time, the more affected the utilities are. In fact, the strictest Δ-parameter produced a sharp decline of all three utilities, which barely changed in subsequent iterations. This indicates that the affected pixels were almost randomized due to the large perturbation added by DP in just one iteration. Thus, it is important to choose a relatively weak Δ-parameter, which will allow keeping the added perturbation per iteration small enough for the affected pixels to still retain (some of) their utility.

4.5 Comparisons

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

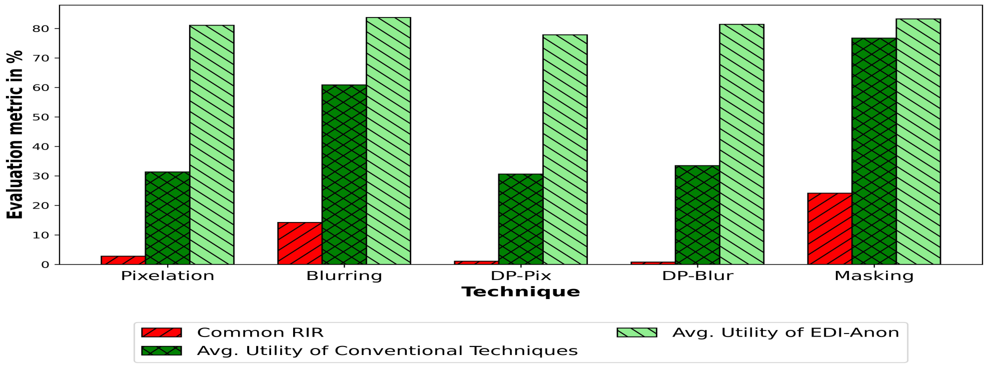


Figure 7: Average utility comparison between EDI-Anon and its conventional counterparts at equal RIRs

4.5.1 Comparison with conventional methods

Whereas EDI-Anon can be set to reach a target RIR/privacy threshold, the RIR attained by the conventional one-shot methods depends on the values of their parameters and is hard to fix *ex ante*. To enable a fair comparison, in the following we report EDI-Anon utilities for the same RIR achieved by its conventional counterpart with the privacy parameter that resulted in the lowest RIR (see Fig. 3). To facilitate the comparison, Fig. 7 reports the average of the three utilities considered for the different methods.

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

4.5.2 Comparison with DeepPrivacy2

As discussed in Section 2, GANs are employed in state-of-the-art image anonymization methods due to their ability to create synthetic images that conceal identifiable information while retaining relevant facial features. In this section, we compare EDI-Anon with DeepPrivacy2 [21], a state-of-the-art cGAN-based approach.

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

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

Applying DeepPrivacy2 to the MTF dataset, we obtained an anonymized dataset with a RIR = 16.27%. To allow a fair comparison with EDI-Anon, in Table 2 we report the average utility obtained by DeepPrivacy2 and all versions of EDI-Anon when setting the privacy threshold of the latter to 16%.

Table 2: Average utility comparison between DeepPrivacy2 and EDI-Anon methods for equal RIR $\approx 16\%$

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

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

4.5.3 Runtime comparison

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

For each method, we picked the instantiation that resulted in the best privacy (i.e., lowest RIR). Even if the level of privacy attained does not affect the runtime of conventional methods applied in one shot, it does affect the runtime of EDI-Anon methods. Thus, for the latter, we report the time of their longest run. The results are shown in Table 3. To put EDI-Anon figures in context, we also specify the number of iterations needed to reach the privacy threshold.

Table 3: Average runtime per image for conventional methods, their EDI-Anon counterparts and DeepPrivacy2

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

As expected, the simplest methods, such as pixelation and blurring, resulted in the lowest runtimes. DP-based methods were costlier, due to the additional effort required to compute noise distributions. In comparison, their EDI-Anon counterparts took 3-17 times longer runtimes, a magnitude that is proportional to the number of iterations required to reach the desired privacy threshold. In contrast, DeepPrivacy2 was, by far, the slowest in generating anonymized images, which can be attributed to the complexity of the GAN model used.

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

In spite of its longer runtime, EDI-Anon has an intrinsic advantage over conventional one-shot methods. Whereas the latter require re-running the whole process from scratch to create anonymizations with different privacy parameters, using a strict privacy threshold with EDI-Anon immediately yields not only the results corresponding to that threshold but any of the (weaker) anonymizations resulting from intermediate iterations. This is quite convenient for practitioners who aim to optimize the privacy/utility trade-off *ex post*, a task that is usually implemented as a trial and error process for conventional methods.

5 Conclusions and future work

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

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

As future work, we plan to extend EDI-Anon to other image obfuscation techniques such as motion blur. We also plan to evaluate the robustness of our framework against image-reconstruction attacks.

Acknowledgements

This research was funded by the European Commission (project H2020-871042 "SoBigData++"), the Government of Catalonia (ICREA Acadèmia Prizes to J.Domingo-Ferrer and to D. Sánchez), MCIN/AEI/ 10.13039/501100011033 and "ERDF A way of making Europe" under grants PID2021-123637NB-I00 "CURLING" and PRE2019-089210, and INCIBE and European Union NextGenerationEU/PRTR (project "HERMES" and INCIBE-URV Cybersecurity Chair). Y.K. acknowledges financial support from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Grant Agreement No. 945413 and from the Universitat Rovira i Virgili. This paper reflects only the authors' view and the European Research Executive Agency is not responsible for any use that may be made of the information it contains.

References

[1] S. Bakhshi, D. A. Shamma, and E. Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 965–974, 2014.

[2] A. Blanco-Justicia, D. Sánchez, J. Domingo-Ferrer, and K. Muralidhar. A critical review on the use (and misuse) of differential privacy in machine learning. *ACM Computing Surveys*, 55(8):160:1–160–16, 2023.

[3] G. Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000.

[4] F. Z. Canal, T. R. Müller, J. C. Matias, G. G. Scotton, A. R. de Sa Junior, E. Pozzebon, and A. C. Sobieranski. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582:593–617, 2022.

[5] J.-W. Chen, L.-J. Chen, C.-M. Yu, and C.-S. Lu. Perceptual indistinguishability-net (pi-net): Facial image obfuscation with manipulable semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6478–6487, 2021.

[6] V. Dalai and K. Kadambari. Face identification using visible eye region via vanilla cnn and siamese networks. In *2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, pages 2597–2600. IEEE, 2022.

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Proceedings of the ieee conference on computer vision and pattern recognition, 2009.

[8] J. Domingo-Ferrer, D. Sánchez, and A. Blanco-Justicia. The limits of differential privacy (and its misuse in data release and machine learning). *Communications of the ACM*, 64(7):33–35, 2021.

[9] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

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

[11] L. Fan. Differential privacy for image publication. In *Theory and Practice of Differential Privacy (TPDP) Workshop*, volume 1, page 6, 2019.

[12] I. H. Fraser, G. L. Craig, and D. M. Parker. Reaction time measures of feature saliency in schematic faces. *Perception*, 19(5):661–673, 1990.

[13] A. Gajic. Instagram marketing statistics. 99firms.com, 2023.

[14] S. Gao, I. W.-H. Tsang, and L.-T. Chia. Kernel sparse representation for image classification and face recognition. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 1–14. Springer, 2010.

[15] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[16] R. Gross, E. Airoldi, B. Malin, and L. Sweeney. Integrating utility into face de-identification. In *Privacy Enhancing Technologies: 5th International Workshop, PET 2005, Cavtat, Croatia, May 30-June 1, 2005, Revised Selected Papers 5*, pages 227–242. Springer, 2006.

[17] S. K. Gupta and N. Nain. Single attribute and multi attribute facial gender and age estimation. *Multimedia Tools and Applications*, 82(1):1289–1311, 2023.

[18] R. Haffar, N. M. Jebreel, J. Domingo-Ferrer, and D. Sánchez. Explaining image misclassification in deep learning via adversarial examples. In *International Conference on Modeling Decisions for Artificial Intelligence*, pages 323–334. Springer, 2021.

[19] R. Haffar, D. Sánchez, and J. Domingo-Ferrer. Multi-task faces (MTF) data set: A legally and ethically compliant collection of face images for various classification tasks. *CoRR*, abs/2311.11882, 2023.

[20] N. D. Haig. Exploring recognition with interchanged facial features. *Perception*, 15(3):235–247, 1986.

[21] H. Hukkelås and F. Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1329–1338, 2023.

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

[23] S. M. S. M. Khorzooghi and S. Nilizadeh. Examining stylegan as a utility-preserving face de-identification method. *Proceedings on Privacy Enhancing Technologies*, 2023.

[24] H. Kim, Z. Pang, L. Zhao, X. Su, and J. S. Lee. Semantic-aware deidentification generative adversarial networks for identity anonymization. *Multimedia Tools and Applications*, 82(10):15535–15551, 2023.

[25] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[26] S. R. Klomp, M. Van Rijn, R. G. Wijnhoven, C. G. Snoek, and P. H. De With. Safe fakes: Evaluating face anonymizers for face detectors. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8. IEEE, 2021.

[27] Z. Kuang, H. Liu, J. Yu, A. Tian, L. Wang, J. Fan, and N. Babaguchi. Effective de-identification generative adversarial network for face anonymization. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3182–3191, 2021.

[28] H.-P. Lee, Y.-J. Yang, T. S. Von Davier, J. Forlizzi, and S. Das. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2024.

[29] B. Leong. Facial recognition and the future of privacy: I always feel like… somebody's watching me. *Bulletin of the Atomic Scientists*, 75(3):109–115, 2019.

[30] B. Lingenfelter, S. R. Davis, and E. M. Hand. A quantitative analysis of labeling issues in the celeba dataset. In *International Symposium on Visual Computing*, pages 129–141. Springer, 2022.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[32] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11976–11986, 2022.

[33] M. Maximov, I. Elezi, and L. Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020.

[34] W. Mellouk and W. Handouzi. Facial emotion recognition using deep learning: review and insights. *Procedia Computer Science*, 175:689–694, 2020.

[35] C. Neustaedter, S. Greenberg, and M. Boyle. Blur filtration fails to preserve privacy for home-based video conferencing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 13(1):1–36, 2006.

[36] M. T. Ribeiro, S. Singh, and C. Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[37] E. A. Roberts, C. Troiano, and J. H. Spiegel. Standardization of guidelines for patient photograph deidentification. *Annals of plastic surgery*, 76(6):611–614, 2016.

[38] J. Sadr, I. Jarudi, and P. Sinha. The role of eyebrows in face recognition. *Perception*, 32(3):285–293, 2003.

[39] M. C. Santana, O. Déniz-Suárez, L. Antón-Canalís, and J. Lorenzo-Navarro. Face and facial feature detection evaluation - performance evaluation of public domain haar detectors for face and facial feature detection. In A. Ranchordas and H. Araújo, editors, *VISAPP 2008: Proceedings of the Third International Conference on Computer Vision Theory and Applications, Funchal, Madeira, Portugal, January 22-25, 2008 - Volume 2*, pages 167–172, Setúbal, Portugal, 2008. INSTICC - Institute for Systems and Technologies of Information, Control and Communication.

[40] C. J. Schuler, M. Hirsch, S. Harmeling, and B. Schölkopf. Learning to deblur. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1439–1451, 2015.

[41] X. Shi, X. Chai, J. Xie, and T. Sun. Mc-gcn: a multi-scale contrastive graph convolutional network for unconstrained face recognition with image sets. *IEEE Transactions on Image Processing*, 31:3046–3055, 2022.

[42] Z. Sun and Z. Liu. Ensuring privacy in face recognition: a survey on data generation, inference and storage. *Discover Applied Sciences*, 7(5):441, 2025.

[43] C. J. van Rijsbergen. *Information Retrieval*. Butterworth, London, 1979.

[44] B. Xu, L. He, J. Liang, and Z. Sun. Learning feature recovery transformer for occluded person re-identification. *IEEE Transactions on Image Processing*, 31:4651–4662, 2022.

[45] L. Yuan, L. Liu, X. Pu, Z. Li, H. Li, and X. Gao. Pro-face: A generic framework for privacy-preserving recognizable obfuscation of face images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1661–1669, 2022.

[46] Z. Zhang and M. R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8792–8802, 2018.

[47] Q. A. Zhao and J. T. Stasko. Evaluating image filtering based techniques in media space applications. In *Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 11–18, 1998.