GANonymization: A GAN-based Face Anonymization Framework for Preserving Emotional Expressions

FABIO HELLMANN, University of Augsburg, Germany
SILVAN MERTES, University of Augsburg, Germany
MOHAMED BENOUIS, University of Augsburg, Germany
ALEXANDER HUSTINX, University of Bonn, Germany
TZUNG-CHIEN HSIEH, University of Bonn, Germany
CRISTINA CONATI, University of British Columbia, Canada
PETER KRAWITZ, University of Bonn, Germany
ELISABETH ANDRÉ, University of Augsburg, Germany

In recent years, the increasing availability of personal data has raised concerns regarding privacy and security. One of the critical processes to address these concerns is data anonymization, which aims to protect individual privacy and prevent the release of sensitive information. This research focuses on the importance of face anonymization. Therefore, we introduce GANonymization, a novel face anonymization framework with facial expression-preserving abilities. Our approach is based on a high-level representation of a face, which is synthesized into an anonymized version based on a generative adversarial network (GAN). The effectiveness of the approach was assessed by evaluating its performance in removing identifiable facial attributes to increase the anonymity of the given individual face. Additionally, the performance of preserving facial expressions was evaluated on several affect recognition datasets and outperformed the state-of-the-art methods in most categories. Finally, our approach was analyzed for its ability to remove various facial traits, such as jewelry, hair color, and multiple others. Here, it demonstrated reliable performance in removing these attributes. Our results suggest that GANonymization is a promising approach for anonymizing faces while preserving facial expressions.

CCS Concepts: • **Security and privacy → Privacy-preserving protocols**; **Pseudonymity, anonymity and untraceability**.

Additional Key Words and Phrases: face anonymization, emotion recognition, data privacy, emotion preserving, facial landmarks

1 INTRODUCTION

In the current machine learning landscape, models are getting more and more complex. This complexity places a significant demand on the availability of large, high-quality datasets, particularly when leveraging deep learning (DL) techniques. However, building such datasets is not always easy - besides the time-consuming process of

Authors' addresses: Fabio Hellmann, fabio.hellmann@informatik.uni-augsburg.de, University of Augsburg, Universitaetsstrasse 6a, Augsburg, Bavaria, Germany, 86159; Silvan Mertes, silvan.mertes@informatik.uni-augsburg.de, University of Augsburg, Universitaetsstrasse 6a, Augsburg, Bavaria, Germany, 86159; Mohamed Benouis, mohamed.benouis@informatik.uni-augsburg.de, University of Augsburg, Universitaetsstrasse 6a, Augsburg, Bavaria, Germany, 86159; Alexander Hustinx, ahustinx@uni-bonn.de, University of Bonn, Venusberg-Campus 1, Bonn, North Rhine-Westphalia, Germany, 53127; Tzung-Chien Hsieh, thsieh@uni-bonn.de, University of Bonn, Venusberg-Campus 1, Bonn, North Rhine-Westphalia, Germany, 53127; Cristina Conati, conati@cs.ubc.ca, University of British Columbia, 2366 Main Mall Vancouver, Vancouver, BC, Canada, V6T1Z4; Peter Krawitz, pkrawitz@uni-bonn.de, University of Bonn, Venusberg-Campus 1, Bonn, North Rhine-Westphalia, Germany, 53127; Elisabeth André, andre@informatik.uni-augsburg.de, University of Augsburg, Universitaetsstrasse 6a, Augsburg, Bavaria, Germany, 86159.

ACM 1551-6857/2024/1-ART
https://doi.org/10.1145/3641107

acquiring and annotating data, privacy is a serious obstacle here. While extensive datasets exist for non-sensitive data, the acquisition of data for sensitive use cases, especially those involving human data, is an intricate task when the subjects' privacy needs to be ensured. Particularly when it comes to scenarios involving the human face, it is generally a hard task to collect appropriate data, especially if datasets are to be made publicly available. On the other hand, developing DL models that use images of human faces offers promising opportunities. For instance, assessing affective states like emotions or stress might be beneficial to infer more serious conditions, such as chronic overload or depression, and react accordingly. However, not only training data for DL algorithms run the risk of violating humans' privacy - it is inference data too. When employing fully trained DL models in real-world scenarios, dealing with data that reveals a human's identity poses additional difficulties, as sovereignty over one's data is endangered. In general, it can be stated that different use cases require different degrees of anonymization to assure human privacy. On the other hand, different DL models require a different set of undiluted features in order to be able to model the problem at hand. In the case of facial affective state assessment, most of the context information is unimportant and should be eliminated to reduce the features for re-identification. Therefore, an approach is needed that offers the research community a pipeline to anonymize faces while only preserving affective state relevant information.

Further, face anonymization can be vital in promoting ethics and fairness in machine learning. Not anonymized data can lead to unfair AI decisions, as facial recognition models have been shown to exhibit bias against people of color and women [21]. However, current research on face anonymization algorithms often neglects the fact that mere anonymization does not necessarily remove those traits. For instance, a face image of a woman of color might still show a woman of color after applying state-of-the-art face anonymization techniques, although her exact identity might not be recognized anymore. For the task of emotion recognition, in particular, traits like skin color, gender, or hairstyle are not needed, which might introduce bias when being considered.

Additionally, the importance of face anonymization is evident in its ability to protect individual privacy, promote ethical considerations, and ensure compliance with legal requirements. By employing face anonymization techniques, researchers can prevent the misuse of personal information and enable the development of machine learning models that are more broadly applicable and ethical. Face anonymization conceals personal information such as identity, race, ethnicity, gender, or age, reducing the risk of re-identification. It is essential in sensitive datasets like medical records and criminal justice data, where anonymity is critical for individuals' privacy and safety. It is crucial in healthcare to ensure patient confidentiality when sharing medical images with researchers or medical professionals. In the criminal justice system, face anonymization can protect the identity of witnesses, victims, and suspects from potential harm. The protection of personal data by anonymization or pseudonymization is also enforced in the European Union by law with the General Data Protection Regulation (GDPR) [8]. Industries such as healthcare and finance are also subject to additional regulations and standards that require anonymization to protect sensitive data. For example, US law states that the Health Insurance Portability and Accountability Act (HIPAA) mandates anonymizing Protected Health Information (PHI) to ensure compliance with privacy and security regulations.

To address these shortcomings, this work presents a novel approach to face anonymization that addresses that problem specifically in the context of emotion recognition. Existing work predominantly tries to find a trade-off between anonymization and task performance by formalizing the problem as a min-max game in which the objective is to find a good compromise between both requirements [34, 45, 46]. However, features that are neither benefiting the task at hand nor taking away from identity obfuscation (i.e., not affecting either of the two objectives) are mostly ignored. As such, traits like skin color, gender, or age are still apparent in the anonymized versions of the images, conserving bias and inequity in the underlying data. Instead of engaging in the aforementioned min-max game, as done by previous approaches, we follow a different paradigm: we completely discard all information except a minimal feature representation that is needed for our chosen use case - emotion recognition - and subsequently re-synthesize arbitrary information for the lost features. By doing so,

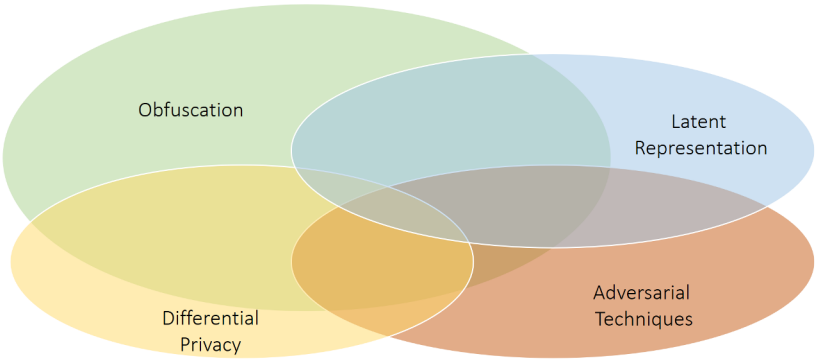


Fig. 1. Existing privacy preservation concepts in the context of face anonymization.

we obtain a complete face image with the same facial expression as the original face while, contrary to existing approaches, removing irrelevant traits for the use case of emotion recognition. After reviewing relevant literature [9, 22, 37, 43], we found that facial landmarks can be a good feature set for that task while not exposing too much unnecessary information. Therefore, as this work focuses on emotion recognition, we chose to extract facial landmarks as an intermediate representation. To disregard all unimportant information, we chose to extract facial landmarks as an intermediate representation. Subsequently, we use a Generative Adversarial Network (GAN) architecture, namely *pix2pix* [14], to re-synthesize a realistic face that incorporates exclusively the features included in the landmarks. By doing so, our approach - which we call *GANonymization* - has the advantage of not preserving any traits that were not present in the landmark representation. As such, features like hairstyle, skin color, or gender are diluted from the intermediate representation, which sets our approach apart from most existing methods.

We evaluate our approach in a three-fold manner:

(1) We validate if our anonymization method can anonymize faces sufficiently by using a standard measure in this research [41, 42].
(2) We validate if our anonymization method keeps important features to preserve emotional expressions by analyzing how the anonymization process affects the predictions of an auxiliary emotion classifier in both a training as well as an inference setting.
(3) We seek to explain the outcomes of the evaluation steps above by analyzing which facial traits are preserved or removed with our anonymization method. To do so, we study how the anonymization process affects the predictions of an auxiliary facial feature detection model.

We show that our approach significantly outperforms state-of-the-art methods in preserving most facial emotional expressions in an anonymized synthesized face.

2 RELATED WORK

In this section, we provide an overview of previous research on privacy preservation in the context of facial anonymization. The discussion is organized into four key concepts: Obfuscation, Adversarial Techniques, Differential Privacy, and Latent Representations. Note that those concepts are not distinct mechanisms, but different approaches can make use of several of those ideas, as depicted in Figure 1.

2.1 Obfuscation

Obfuscation techniques have been pivotal in anonymizing facial data by modifying or masking sensitive areas in images or videos. These techniques, including pixelation, blurring, and masking, aim to obscure facial features related to identity while retaining identity-independent characteristics [36].

For instance, Jourabloo et al. [15] presented an attribute-preserving face de-identification approach. While this approach achieved a commendably low face recognition rate, it succeeded in preserving critical facial attributes. The method employed an Active Appearance Model and the K-same algorithm to reconstruct new face images while averaging selected features. Wu et al. [47] introduced a face-blurring approach to obfuscate faces in the ImageNet dataset, and Raval et al. [39] employed an adversarial perturbation mechanism to protect visual information in camera feeds without substantially impairing application functionality.

Obfuscation techniques are indeed effective in achieving high degrees of anonymity, but they invariably degrade the naturalness and quality of face images, limiting their reusability for diverse facial applications [23]. In contrast, our approach takes a different path. Although it involves the removal of various facial traits, it excels in producing high-quality, naturalistic face images. We achieve this by re-synthesizing complete face images using a GAN-based architecture.

2.2 Adversarial Techniques

Many existing approaches to facial anonymization are based on training anonymization models using adversarial techniques. Generally, the term *adversarial* refers to the paradigm of two *contrary* objectives being maximized at the same time. For face anonymization, these objectives are the anonymization performance and the so-called *Utility*, i.e., the ability to preserve features that are relevant to solving a certain auxiliary task. This dual objective can create a min-max game, where improving one objective often results in the degradation of the other. As such, solving a min-max game with methods of DL inevitably results in finding a compromise between the two objectives.

For example, Nasr et al. [34] developed an adversarial regularization training method aimed at minimizing classification loss while maximizing defense against membership inference attacks. Wu et al. [46] utilized GANs to learn a degradation transformation that balances action recognition performance with video privacy. Wu et al. [45] introduced a face de-identification framework that generated de-identified faces with high feature attributes and minimized identity information by incorporating a contrastive verification loss and a structure similarity loss into the GAN training process.

Our approach differs from these methods in that we don't formulate the anonymization problem as a min-max game. Instead, we make use of adversarial learning techniques within our framework, particularly by employing a GAN-based architecture to re-synthesize full-face images from our latent representations. However, our method stands apart as we don't incorporate *privacy norms* into the GAN training but focus on feature reduction before GAN training. This unique approach enables us to remove traits that affect neither anonymization nor utility, setting our method apart from mere compromises between the two.

2.3 Differential Privacy

Differential privacy depends on the specific application's notion of neighboring databases, which is the core of privacy preservation. In deep learning, differential privacy involves the introduction of random noise into a training inference model, which is computed from the underlying stochastic gradient descent (SGD) training gradient. This noise is added to ensure a balanced distribution of the results, aligning both utility and privacy considerations [1]. Complementing differential privacy and the SGD helps balance accurate model predictions and privacy protection.

This method has found extensive application in face recognition technology, specifically in removing personal identity information to safeguard privacy. For hard biometric protections (i.e., identity), Chamikara et al. [2] added an amount of noise equally across different values of eigenvalues extracted from the image through the principal component analysis (PCA) method, protecting against the unauthorized disclosure of identity information. However, this approach only partially changes the reconstructed human face, but it is still vulnerable to adversary machine learning threats. To mitigate this issue, Wen et al. [44] proposed a framework based on differential privacy and GANs to hide the attribute that discloses the identity information while keeping only the attribute used for the downstream tasks. Kansal et al. [16] first mapped the original image into the embedding space garnered by an encoder deep learning model. During the training stage, they used an adversarial model-based differential mechanism to suppress the personally identifiable information. Then, they applied a decoder model to reconstruct the original image for the auxiliary task. Croft et al. [3] achieved successful image anonymization by incorporating differential privacy into a generative model's latent representation for soft biometrics such as gender and expression. Nevertheless, the accuracy of these methods is highly limited. Besides, the practical implementation of differential privacy in real-world scenarios presents a significant challenge. Determining precise privacy boundaries is critical, as adding noise to protect sensitive information may disrupt the entire data distribution, leading to unrecognizable output images [49].

In contrast, our approach does not introduce noise during training or generation. Instead, we focus on information reduction before training, retaining only a minimal latent representation, such as facial landmarks. While this approach may pose challenges in finding a suitable representation for domains other than emotion recognition, it distinctly sidesteps the pitfalls associated with noisy and/or unrecognizable data.

2.4 Latent Representations

Traditional GAN-based models often struggle to preserve complex facial attributes, such as emotion, pose, and background, due to image space's high dimensionality and complexity. This challenge often results in latent representations being softer in facial style change compared to image space manipulation. Latent representation, as an abstract and compressed representation inferred from data, captures essential features while discarding redundant information. This makes it easier for models to perform tasks like classification and generation.

Le et al. [24] introduced StyleID, a GAN that brings images into a latent representation, uncovers features with significant identity disentanglement, and changes these features in latent space or pixel space. However, StyleID may preserve facial traits that have the potential to introduce bias or unfairness, even if they don't correlate directly with identity. Other methods, such as Sun et al. [43], Hu et al. [11], and Maximov et al. [30] with CIAGAN, employed inpainting mechanisms in conjunction with GANs to anonymize faces based on facial landmarks. These approaches, while effective, retain context-relevant information outside of the face-segmented area, such as hair color, hairstyle, and gender. On the other hand, Hukkelås and Lindseth introduced DeepPrivacy2 [13], an enhanced guided GAN framework for anonymizing human figures and faces. The DeepPrivacy2 framework entails three detection components for each task: i) face detection with a Dual Shot Face Detector [25], ii) dense pose estimation with Continuous Surface Embeddings [35], and iii) Mask R-CNN [10] for instance segmentation. Additionally, three task-specific Surface-guided GANs [12] were trained to synthesize either human figures with conditions, human figures without conditions, or faces. However, the use of inpainting mechanisms in these approaches may inadvertently retain context-relevant information, potentially introducing bias or unfairness.

In contrast, our approach focuses on excluding context-relevant information by removing all context information except the facial structure with many facial landmarks. By concentrating on the elimination of contextual traits, we aim to reduce the potential for bias or unfairness in the dataset.

Overall, DeepPrivacy2 can be regarded as a state-of-the-art full-body anonymization method since it outperformed a variety of other methods in the past [13]. Furthermore, CIAGAN can be considered as another

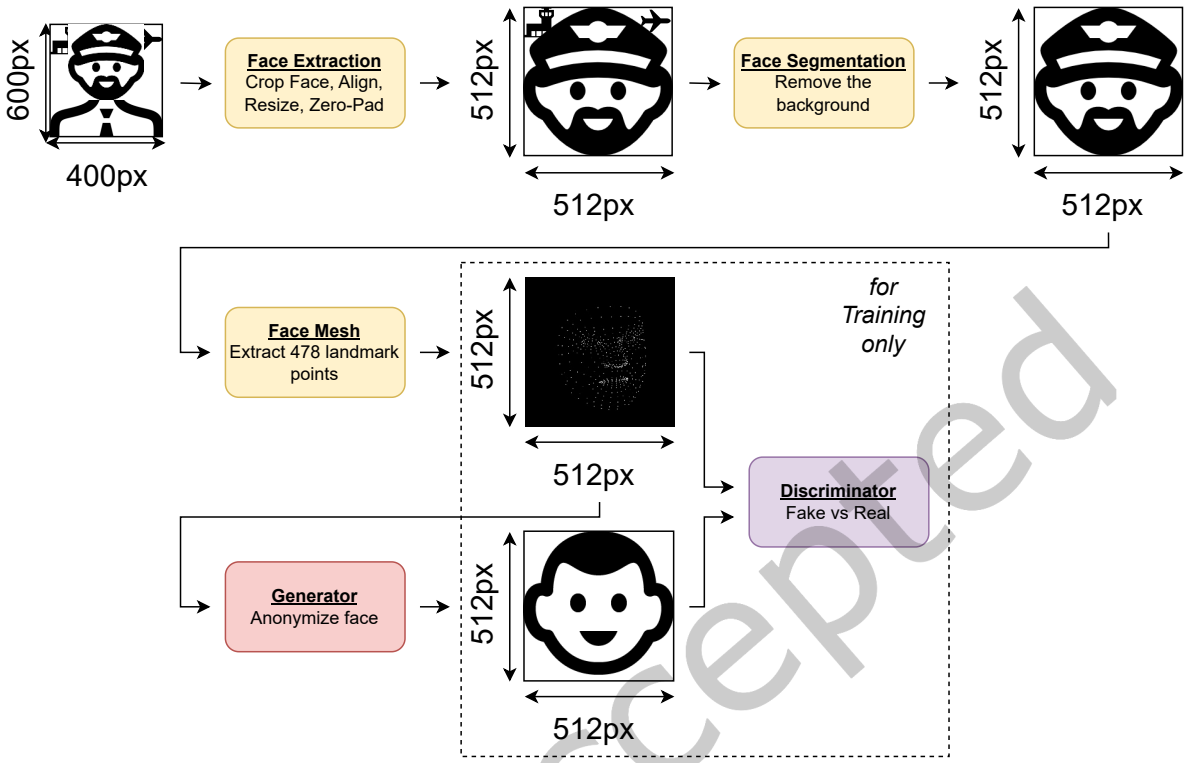


Fig. 2. Architecture of the GANonymization pipeline.

state-of-the-art face anonymization method, which is also based on landmarks [30]. While CIAGAN utilizes inpainting mechanisms to only anonymize the face area below the forehead, DeepPrivacy2 anonymizes the full facial area, including the forehead. Consequently, we used DeepPrivacy2 and CIAGAN as the baseline for all our performance evaluations.

3 METHOD

This section introduces the structure of our GANonymization framework (see Figure 2) and gives a detailed description of each component and the steps taken for training.[1] The complete GANonymization framework entails four components.

Training Scenario. In the first step, faces are detected, extracted, and brought into the right format afterward. The image's background is removed in the second step to eliminate distracting features. In the third step, facial landmarks are extracted from the face. In the last step, the GAN's generator synthesizes a new, anonymized face based on those landmarks. The discriminator evaluates the facial landmarks and the synthesized face to determine whether it is real or fake.

[1]Our framework's implementation can be found at https://github.com/hcmlab/GANonymization.

Inference Scenario. The inference requires fewer steps than the training scenario, as the first and second steps are unnecessary. Only the extraction of the facial landmarks is required to feed the generator to synthesize an anonymized face.

3.1 Face Extraction

The first component in the pipeline is face extraction. The RetinaFace framework[2] [41] is utilized for this component, which is based on the RetinaFace algorithm [5]. RetinaFace has been tested against the WIDER [48] dataset to ensure maximum efficiency in detecting and aligning faces in various scenarios correctly. However, RetinaFace does not detect all faces every time, especially when factors like poor image quality, extreme angles, or heavy occlusions are in play. This component includes the following steps:

(1) *Face Crop.* The input image is analyzed to detect and extract all visible faces.
(2) *Face Align.* According to the literature, aligning the faces supports an increase in accuracy for face recognition models [38]. Therefore, the faces are aligned before the GAN receives them as input. By doing so, the GAN is prevented from focusing too much on the head orientation and instead takes only the face itself into account.
(3) *Image Resize.* The input size of the images for the GAN is set to 512×512 pixels. Therefore, the cropped and aligned faces are up-scaled to 512 pixels for the greatest axis, while maintaining the aspect ratio.
(4) *Zero Padding.* To achieve the final 512×512 pixels for the required input shape of the GAN, we apply zero padding to the sides [(right and left) or (top and bottom)] of the image to keep the face centered in the image.

3.2 Face Segmentation

The second component of the pipeline is face segmentation. Even though this step could be skipped, we observed that the pix2pix architecture we used for re-synthesis of the faces (see Section 3.4) yielded visually better results when not having to deal with variations in the background. Consequently, the original background is removed by applying face segmentation and setting all pixel intensities outside the face segments to 0. Therefore, a head segmentation model[3] based on a U-Net is utilized.

3.3 Facial Landmarks

After the pre-processing steps, we generate intermediate representations of the faces. Here, we aim for a representation that (i) does not contain information that could be used to identify the original face and (ii) holds all necessary information needed for facial expression analysis tasks. Existing literature on the topic [9, 22, 37, 43] indicates that facial landmarks fulfill both of these requirements in the context of emotion recognition. Note that although this work focuses on the context of emotion recognition exclusively, the concept could be transferred to other domains as well. Therefore, a suitable intermediate representation, which might not be facial landmarks, would have to be found for the specific task. For our experiments, we extract 478 3-dimensional coordinate facial landmarks utilizing the media-pipe face-mesh model [20] to receive an abstract representation of the facial shape. The resulting 3D landmarks are projected onto a 2D image with a black background where each landmark point is represented by a single white pixel. It is necessary to translate the 3D landmarks into a 2D image due to the image-to-image type of model used for the re-synthesis of the faces (as described in the following section 3.4).

[2]https://github.com/serengil/retinaface
[3]https://github.com/wiktorlazarski/head-segmentation

3.4 Re-Synthesis

To obtain an anonymized version of the input that still looks highly realistic, we aim for a re-synthesis of high-quality faces. Therefore, we use the *pix2pix* architecture, a GAN-based image-to-image model. The original purpose of *pix2pix* is to convert image data from a particular source domain to a target domain. Our specific goal in the re-synthesis stage is to transfer the landmark representations back to random, high-quality face images that expose the same facial landmark structure. The *pix2pix* architecture has been successfully applied to similar use cases in the past, e.g., for synthetic data augmentation in the context of defect detection [31, 32], where segmentation masks of material defects (which, on a technical level, are quite similar to visual landmark representations) were converted to realistic looking data. More recent GAN-based architectures like ProGAN [17], StyleGAN [18], or StyleGANv2 [19], that impress with their ability to generate hyper-realistic data, are specifically designed to create new data from scratch. To use those models for image-to-image conversion tasks, a projection of the input image has to be found in the GAN's latent space, which is highly inefficient and might not be possible at all for some data instances. As such, we chose to use *pix2pix*, as it is specifically tailored for end-to-end image-to-image translation. For the training of the *pix2pix* model, we used existing face images as the *target* domain, whereas for the *source* domain, we used landmark representations that we priory extracted from those images. In other terms, we trained the *pix2pix* network to learn the inverse transformation of a landmark extractor - we perform an image-to-image translation from an image representation of landmark features to realistic-looking face images. By using that approach, we are able to automatically create geometrically aligned source/target image pairs for training. Contrary to architectures such as CycleGAN [50] that work with non-parallel training data, *pix2pix* directly takes advantage of having mapped training pairs, which again supports our architecture choice.

We process the CelebA [26] dataset within our pipeline to extract and align the faces (section 3.1), remove the background of the faces by face segmentation (section 3.2), and extract a face-mesh of each face which represents the landmark/image pairs for training. CelebA was used because of its size (202,599 images) and because it contains only images of high quality - using low-quality images would limit the quality of GANonymization's output images unnecessarily. We used the same pipeline for the landmark extraction to anonymize the images. Additionally, training images were normalized to $mean = (0.5, 0.5, 0.5)$ and $std = (0.5, 0.5, 0.5)$. Our implementation was built upon Erik Linder-Norén's pix2pix implementation[4], which in turn strongly adheres to the original *pix2pix* publication[14]. We trained the model for 25 epochs with a batch size of 32. The Adam optimizer was used with a learning rate of 0.0002, β_1 decay of 0.5, and β_2 decay of 0.999. After training, our model could transfer landmark representations to face images that show the same facial expression expressed by the original face. In the case of an issue with face detection and, therefore, no available facial landmarks, an empty (black) image can be inferred with our model with the result of a synthesized average face, which is based on the faces seen by the model during the training process. Exemplary outputs of our pipeline are shown in Figures 3, 4, 5, 6, and 8.

4 EVALUATION

In the following sections, we describe how we validate our approach using three different evaluations. First, we evaluate the anonymization capability of the approach. Second, we evaluate the suitability of the approach for the task of emotion recognition, i.e., whether our approach preserves information that is relevant to facial emotion recognition. Finally, we go into detail about the facial features that get preserved or removed with our anonymization approach.

[4]https://github.com/eriklindernoren/PyTorch-GAN#pix2pix

4.1 Anonymization Performance

In this first part of the evaluation, the anonymization performance of our approach was assessed. Hereby, with the term anonymization performance, we refer to the capability of the method to alter input images in a way that they ideally cannot be re-identified. Therefore, we compared the synthesized images of our approach with the original images and versions synthesized by DeepPrivacy2, CIAGAN, and basic methods like pixelation and blurring.

4.1.1 Dataset. The dataset used for the comparison was the WIDER [48] dataset, which is commonly used for benchmarking face detection algorithms. Further, the authors of DeepPrivacy2 had already used it in their original publication. Therefore, by using it in our experiments too, we do not introduce a bias towards GANonymization by using a dataset that DeepPrivacy2 might not be suited for. It contains images of people in various sceneries whose faces vary in scale, pose, and occlusion. In each image, one or more faces are apparent. In total, WIDER embodies 32,203 images in 61 event settings. The many different head orientations, obfuscations, facial expressions, lighting conditions, and others enable an optimal evaluation setting to measure the overall performance in anonymizing these faces. After we applied our pre-processing pipeline with the face extraction (section 3.1) and face segmentation (section 3.2) components, the images were split into a training and validation set of 92,749 and 22,738 face images, respectively.

4.1.2 Setup. The performance measurement is based on the comparison of the original images and their synthesized counterparts. The synthesized images are produced by our method, DeepPrivacy2, and CIAGAN, respectively. Exemplary anonymized images for WIDER can be seen in Figure 3.

4.1.3 Metric. A widely used method to assess the anonymization degree of a face image is to compute the cosine distance between image encodings of the original and anonymized image versions. Here, a lower cosine distance equals higher similarity between the faces and is commonly considered as the anonymized face being *more recognizable* to the original face. Specialized frameworks for face recognition like DeepFace[5] make use of that paradigm and thus can be used as an evaluation tool for anonymization algorithms [41, 42]. As such, for the comparison of the anonymization performance of our approach versus the other methods, we use the DeepFace framework. As a backbone model for image encoding, we use the state-of-the-art face recognition model Facenet512 [7], which is also integrated into DeepFace. The cosine distance is defined as follows:

$$cdistance = 1 - \frac{I_o \cdot I_a}{\|I_o\| \|I_a\|} \tag{1}$$

where I_o and I_a are the Facenet512 feature embedding space representations of the original and anonymized images, respectively. When the cosine distance exceeds 0.3, it indicates that the feature embedding space has diverged significantly from the original space, making re-identification impractical. We computed the cosine distance of the image pairs for each method with the original image.

4.1.4 Results. Our approach achieved a mean cosine distance of 0.7145, while DeepPrivacy2 and CIAGAN reached a greater cosine distance of 0.8119 and 0.9280, respectively (see Table 1). The pixelation with a kernel sized 8×8 achieved 0.8791, while the bigger kernel sized 16×16 achieved 0.6651. Blurring with a kernel sized 9×9 and 17×17 stayed below the threshold necessary for no re-identification with a cosine distance of 0.0102 and 0.0725, respectively.

4.1.5 Discussion. Our evaluation measures the mean cosine distance between the Facenet512-face-based image encodings of original and anonymized face images. Accordingly, the distance between two encodings marks the non-similar features and how complex the reconstruction of one encoding towards another encoding is, which is

[5] https://github.com/serengil/deepface

Fig. 3. Sample of synthesized faces based on the WIDER dataset.

Method	Cosine Distance
Original	0.0000
Ours	**0.7145**
DeepPrivacy2[13]	**0.8119**
CIAGAN[30]	**0.9280**
Pixel 8x8	**0.8791**
Pixel 16x16	**0.6651**
Blur 9x9	0.0102
Blur 17x17	0.0725

Table 1. The mean cosine distances between the original images and the anonymized versions obtained through GANonymization, DeepPrivacy2 (DP2), CIAGAN (CIA), pixelation with a kernel sized 8x8 and 16x16, and blurring with a kernel sized 9x9 and 17x17. The methods with a cosine distance in bold exceed the threshold of 0.3.

conventionally interpreted as *degree of anonymization*. Comparing the results of our approach with the others,

we found that DeepPrivacy2, CIAGAN, and pixelation achieved a mean cosine distance above the threshold of 0.3, which indicates that the feature embedding space diverged significantly from the original image.

While pixelation changes only the underlying image resolution to obfuscate the face, the quality of the image suffers accordingly and the face could still be re-identified - at least for the kernel sized 16×16. Blurring does not modify the image resolution but reduces the overall image quality nonetheless.

On the other hand, CIAGAN synthesized a new face inside of the facial landmark segment of the original image. The result of the synthesized face inside the original image by CIAGAN lacks in quality. However, a face with its emotional expression can still be determined. The low quality and high number of artifacts can be a reason for the high cosine distance to the original image.

DeepPrivacy2, on the other side, synthesized a face that does not necessarily preserve the orientation of the face or the facial expression. In some cases, it can be observed that the outputted face does not have much similarity to a face due to the extreme dysmorphism of facial areas. Accordingly, the dysmorphism can be a result of the increased cosine distance to the original images compared to our approach.

Therefore, we can claim that our approach has a great quality in synthesized faces and solid anonymization performance, indicated by surpassing the common threshold of 0.3 for the mean cosine distance, despite DeepPrivacy2, CIAGAN, and pixelation 8×8 achieving slightly better results here.

4.2 Preserved Emotional Expressions

After showing our approach's anonymization capabilities in section 4.1, we need to ensure that this performance does not come at the expense of the primary task that the data will be used for, in our case, affect recognition. Thus, in this section, we examine whether our method can anonymize faces while maintaining their original emotional expressions. For this evaluation, we use three different datasets which are commonly used in the research field of affect recognition, namely *AffectNet* [33], *CK+* [29], and *FACES* [6].

4.2.1 Datasets. We used three different datasets to cover a wide variety of different settings.

The first dataset we've chosen is the *AffectNet* dataset. We chose it because it contains in-the-wild data, resulting in emotions being expressed in a quite natural way. It contains around 0.4 million images manually labeled according to eight emotional expressions: *Neutral*, *Happy*, *Angry*, *Sad*, *Fear*, *Surprise*, *Disgust*, and *Contempt*. The faces in this dataset have a great variety of individuals, head orientations, lighting conditions, and ethnicities. The dataset was pre-processed with face extraction (section 3.1) and face segmentation (section 3.2). In the process, images in which no face was detected were discarded. Accordingly, the training and validation splits contained 210,174 and 2,874 images, respectively.

The second dataset, namely *CK+*, contains 593 video sequences with 123 subjects aged 18 to 50 years and of various genders and heritage. Each video sequence shows the transition from a neutral facial expression to a non-neutral one, recorded at 30 frames per second. We chose the dataset because, due to the emotional transitions, single image frames also cover facial expressions where the emotions are shown quite subtly. Overall, 327 of those videos are labeled with one of seven emotional expressions: *Anger*, *Contempt*, *Disgust*, *Fear*, *Happiness*, *Sadness*, and *Surprise*. Again, we applied our pre-processing pipeline with face extraction and face segmentation on the dataset and received a training and validation set of 259 and 68 images, respectively.

Lastly, the *FACES* dataset with a total of 2,052 images with different age groups and gender embodies six emotional expressions: *Neutral*, *Sad*, *Disgust*, *Fear*, *Anger*, and *Happy*. We used that dataset as it contains only images of acted emotions, making it a good counterpart for the other two datasets. By including it, we also cover emotional expressions that are shown in a rather exaggerated way. The images in this dataset have high quality. Further, the dataset contains only frontal shots of the faces with optimal lighting conditions. As was done for the previous datasets, we also applied pre-processing with face extraction and face segmentation, resulting in 1,827 images in the training split and 214 images in the validation split.

4.2.2 Setup. We created anonymized versions of the three datasets, resulting in 12 datasets in total: the three original ones, those anonymized with GANonymization, and those anonymized with DeepPrivacy2 and CIAGAN. Exemplary anonymized images for AffectNet, CK+, and FACES can be seen in Figure 4, 5, and 6, respectively. Note that although the CK+ dataset consists of greyscale images, the anonymized versions of our approach are colored - this is a nice byproduct of our approach since we only use the landmarks as an intermediate representation, whereas the re-synthesis is still based on the GAN that was trained on CelebA. We splitted the evaluation of the emotional expression preserving capabilities into two sub-evaluations. First, we assessed how the emotional expression gets preserved during an *inference* setting, thus, how a model trained on original data behaves when fed with anonymized data. Second, we evaluated how the model influences the training process of a model trained on anonymized data.

Inference Scenario Evaluation. To measure how well GANonymization can preserve emotional expressions, we first trained an emotion classifier separately for the three original datasets. Subsequently, we applied the trained models to the original and the anonymized datasets and studied the prediction changes caused by the anonymization methods. Here, big changes in prediction probability can be interpreted as poor preservation of features contributing to emotional expressions. We decided to go for three separate dataset-specific models instead of one across-dataset model, as our evaluation methodology relies on the classifiers accurately modeling the relation between data and emotion for the specific datasets. As the datasets differ substantially, we argue that an across-dataset model, although having the potential to gain a greater overall generalizability, would under-perform on the single datasets due to dataset-specific details that would get lost (e.g., the CK+ dataset is greyscale, FACES are frontal-only, etc.).

As classifier architecture, we chose the base version of the ConvNeXt, which is considered one of the state-of-the-art DL architectures for computer vision tasks [27]. Furthermore, the model was pre-trained on the ImageNet [4] dataset. The classification model's last linear layer's amount of output nodes was changed to match the number of classes, which differed for each dataset. We used the cross-entropy loss for training. Class weights were calculated on the train split of each dataset individually. The AdamW [28] optimizer was used with a learning rate of 0.0003 and a weight decay of 0.001. Additionally, the learning rate was reduced when the validation loss reached a plateau for three consecutive epochs. The images were pre-processed by normalizing with $mean = (0.485, 0.456, 0.406)$ and $std = (0.229, 0.224, 0.225)$ for both, training and testing. Hereby, the mean and standard values for normalization were based on the pre-trained model's dataset (ImageNet). During the training phase, images were randomly flipped horizontally with a probability of 50% for data augmentation. The classification models converged on the validation split within 3, 12, and 9 epochs for AffectNet, CK+, and FACES, respectively. For comparing the anonymization approaches, namely ours, DeepPrivacy2, and CIAGAN, we used the trained emotion classifiers to make predictions on the original images as well as for the anonymized versions. By doing so, we can assess to which degree the anonymization process preserves features that hold information on emotional expressions.

Training Scenario Evaluation. In this sub-evaluation, we assess how the performance of an emotion recognition model's performance degraded when trained on the anonymized versions. To do so, we used the same classifiers that were trained in the *Inference Scenario* but additionally trained the same architecture once with the data anonymized by GANonymization and once anonymized by DeepPrivacy2 and CIAGAN. Thus, we use 12 different models for this experiment, each trained on one of the datasets mentioned above. Subsequently, we compare the performance of the models on the original datasets' validation splits.

4.2.3 Metric.

Inference Scenario Evaluation. We measure the ability of each anonymization approach to preserve the original emotional expressions by looking at how the prediction probabilities for the emotion classifiers change when

Fig. 4. Sample of synthesized faces based on the AffectNet dataset.

Fig. 5. Sample of synthesized faces based on the CK+ dataset.

applied to the original datasets vs. each of the anonymized datasets. I.e., for each image, we measure how the class probability of a certain emotion predicted from the original image differs from the class probability of that same emotion in the anonymized version of the image. Subsequently, we average the resulting probability differences of the images in the validation sets for each emotion. Here, a higher mean difference indicates that the anonymization process obfuscated more features defining the respective emotion. In comparison, a lower difference implies that the anonymization process preserved more emotion-related features.

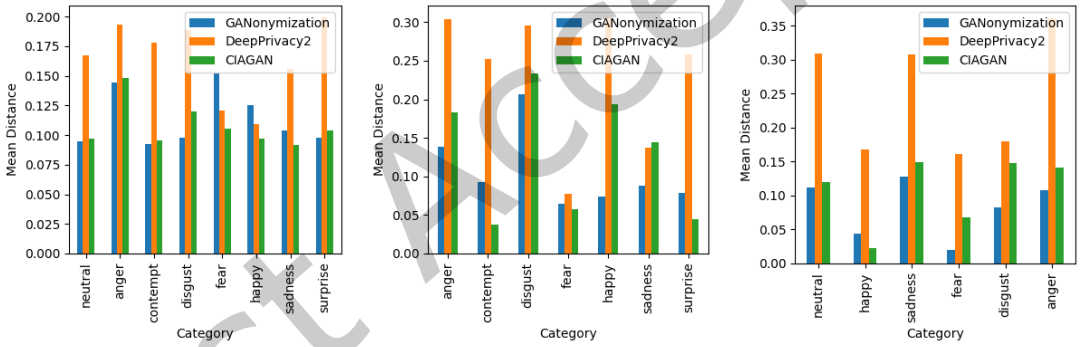
Training Scenario Evaluation. Here, we compare the F1 score of the different models on the respective validation splits. F1 score was chosen as, especially in the CK+ data, a relatively high class imbalance is apparent.

4.2.4 *Results.*

Inference Scenario Evaluation. The results are depicted in Figure 7 and Table 2. As can be seen, GANonymization outperformed DeepPrivacy2 in all emotions except *Fear* and *Happy* in the AffectNet dataset. Compared to CIAGAN,

Fig. 6. Sample of synthesized faces based on the FACES dataset.

our approach outperformed in most emotions except *Fear*, *Happy*, and *Sadness* in the AffectNet dataset, also *Contempt*, *Fear*, and *Surprise* in the CK+ dataset, and only *Happy* in the FACES dataset.



(a) AffectNet (b) CK+ (c) FACES

Fig. 7. The mean distance of the class probability prediction for each emotion for our method, DeepPrivacy2, and CIAGAN on each dataset (lower is better).

To assess if these differences are statistically significant, we conducted statistical hypothesis tests for each emotion as well as each dataset. As a Shapiro-Wilk test revealed that the data was not normally distributed for any of the datasets, Wilcoxon tests were used for the post-hoc analysis. Subsequently, we did a dataset-wise p-value correction using Bonferroni's method. We report the resulting statistics in Table 3. As can be seen, we found significant differences for all emotions in the AffectNet dataset for DeepPrivacy2 and CIAGAN except for *Neutral* and *Anger*. In CK+, we found significant differences for all classes except *Sadness* and *Surprise* for DeepPrivacy2 and *Disgust* for CIAGAN. In the FACES dataset, we found significant differences for all classes except *Happy* and *Fear* for DeepPrivacy2 and *Happy* and *Anger* for CIAGAN.

(a) AffectNet

	Ours	DP2	CIA
Neutral	**0.09**	0.17	0.10
Anger	**0.14**	0.19	0.15
Contempt	**0.09**	0.18	0.10
Disgust	**0.10**	0.19	0.12
Fear	0.15	0.12	**0.11**
Happy	0.13	0.11	**0.10**
Sadness	0.10	0.16	**0.09**
Surprise	**0.10**	0.20	0.10

(b) CK+

	Ours	DP2	CIA
Anger	**0.14**	0.30	0.18
Contempt	0.09	0.25	**0.04**
Disgust	**0.21**	0.30	0.23
Fear	0.06	0.08	**0.06**
Happy	**0.07**	0.31	0.19
Sadness	**0.09**	0.14	0.14
Surprise	0.08	0.26	**0.05**

(c) FACES

	Ours	DP2	CIA
Neutral	**0.11**	0.31	0.12
Anger	**0.11**	0.36	0.14
Disgust	**0.08**	0.18	0.15
Fear	**0.02**	0.16	0.07
Happy	0.04	0.17	**0.02**
Sadness	**0.13**	0.31	0.15

Table 2. The mean class probability distances between the original images and the anonymized versions obtained through GANonymization, DeepPrivacy2 (DP2), and CIAGAN (CIA).

Training Scenario Evaluation. The full evaluation results for the training scenario can be found in Table 4 in the appendix, whereas the confusion matrices for the single models are shown in Figure 10 in the appendix. For the AffectNet dataset, the classifier trained on the original data achieved an overall F1 score of 0.58. In contrast, the classifier trained on the data anonymized with GANonymization achieved an overall F1 score of 0.37. DeepPrivacy2 led to a worse performance, reaching only an F1 score of 0.30. CIAGAN could acquire a slightly increased F1 score of 0.38 than our method.

The other datasets continue the trend for DeepPrivacy2 but worsen the performance for CIAGAN: CK+ (Original data: 0.99, GANonymization data: 0.69, DeepPrivacy2 data: 0.46, CIAGAN data: 0.62) and FACES (Original data: 0.97, GANonymization data: 0.81, DeepPrivacy2 data: 0.67, CIAGAN data: 0.75).

4.2.5 *Discussion.*

Inference Scenario Evaluation. The overall results indicate the superior performance of our approach in preserving facial expressions.

It outperformed the mean distance of DeepPrivacy2 for all emotions except *Fear* and *Happy* in AffectNet. However, we did not find statistical evidence (see Table 3) for the performance differences for *all* of those classes in CK+ and FACES (which might be because those two datasets include a substantially lower amount of images than AffectNet). This could be because many predictions from *Fear* and *Happy* of the synthesized images of our approach were mixed classified (see Figure 10 in the appendix). For example, the emotions *Happy* and *Surprise* were mainly predicted as *Fear* by our classification model.

Compared to the synthesized images by CIAGAN, the cosine distances are closer to our method. CIAGAN preserved *Fear*, *Happy*, and *Sadness* significantly better in the AffectNet dataset (see Table 3). Additionally, the emotions *Contempt*, *Fear*, and *Surprise* also performed significantly better in CK+ judging by the mean distance. In the FACES dataset, CIAGAN outperformed our method only for the emotion *Happy*. However, the results from the significance test for the FACES dataset in Table 3 show that it does not have any statistical significance. An explanation for the small gap between the cosine distances from our method and CIAGAN could be a similar approach to the facial landmarks. The facial landmarks preserve the facial expression mostly accurately. However, in increasing the number of facial landmark points with our approach, it becomes clear that the affective state preservation can be enhanced.

Training Scenario Evaluation. Here, we could observe that data obtained through the anonymization methods led to substantially worse F1 scores for the trained classifiers than the original data. However, GANonymization

(a) AffectNet

	Ours vs. DeepPrivacy2			Ours vs. CIAGAN			
	p	Z	r	p	Z	r	N
neutral	<0.001***	-26.149391	-0.499194	0.119	-26.183915	-0.492635	2744
anger	<0.001***	-10.844502	-0.207023	1.000	-10.331517	-0.194381	2744
contempt	<0.001***	-26.046682	-0.497233	0.013*	-26.187686	-0.492706	2744
disgust	<0.001***	-25.431412	-0.485488	<0.001***	-25.441451	-0.478666	2744
fear	<0.001***	-15.905089	-0.303630	<0.001***	-15.850979	-0.298227	2744
happy	<0.001***	-12.989478	-0.247970	<0.001***	-12.821720	-0.241233	2744
sadness	<0.001***	-4.724173	-0.090185	<0.001***	-4.195525	-0.078936	2744
surprise	<0.001***	-25.921903	-0.494851	0.034*	-26.142871	-0.491863	2744

(b) CK+

	Ours vs. DeepPrivacy2			Ours vs. CIAGAN			
	p	Z	r	p	Z	r	N
anger	<0.001***	-3.941178	-0.477938	<0.001***	-3.415688	-0.414213	68
contempt	<0.001***	-4.326130	-0.524620	<0.001***	-6.495306	-0.787672	68
disgust	<0.001***	-3.635660	-0.440889	0.158	-1.411492	-0.171169	68
fear	0.002**	-3.067397	-0.371977	0.001**	-3.201825	-0.388278	68
happy	<0.001***	-3.415688	-0.414213	<0.001***	-3.440129	-0.417177	68
sadness	0.146	-1.454264	-0.176355	<0.001***	-4.741634	-0.575008	68
surprise	0.051	-1.949203	-0.236376	<0.001***	-4.069495	-0.493499	68

(c) FACES

	Ours vs. DeepPrivacy2			Ours vs. CIAGAN			
	p	Z	r	p	Z	r	N
neutral	<0.001***	-6.869167	-0.469567	<0.001***	-4.080480	-0.278936	214
happy	0.965	-0.043556	-0.002977	0.102	-1.633626	-0.111672	214
sadness	<0.001***	-4.305428	-0.294313	0.018*	-2.366910	-0.161799	214
fear	0.287	-1.064641	-0.072777	<0.001***	-8.291628	-0.566804	214
disgust	0.014*	-2.459535	-0.168130	<0.001***	-8.192387	-0.560020	214
anger	<0.001***	-6.771028	-0.462858	0.052	-1.945685	-0.133004	214

Table 3. The statistics for the cosine distance of GANonymization the DeepPrivacy2, and CIAGAN method to the original based on dataset a), b), and c). If a p-value is less than 0.05, it is flagged with one star (*). If a p-value is less than 0.01, it is flagged with 2 stars (**). If a p-value is less than 0.001, it is flagged with three stars (***)

still performed better for each dataset, except for a very slightly worsening performance in the AffectNet dataset compared to CIAGAN.

4.3 Analysis of Facial Feature Anonymization

To better understand which features are being preserved and which are discarded by GANonymization, we performed an analysis using a pre-trained model for facial feature classification on the CelebA [26] dataset. By analyzing how the predictions of that model change when applied to original versus anonymized images, we aim to infer insights about which facial features our model removes.

4.3.1 Dataset. We've chosen the CelebA dataset due to its vast amount of 202,599 face images with 10,177 identities and 40 binary features representing about facial attributes per subject. For example, those attributes entail eyeglasses, hairstyle, hair color, facial shape, and beard. Thus, this dataset is well suited for analyzing which attributes might change with our anonymization method. However, it should be noted that the dataset contains primarily images of young celebrities - as those might visually not be a representative sample of the entirety of people, it might influence the analysis. We applied our pre-processing pipeline with face extraction and face segmentation on the dataset and received a training and validation set of 166,223 and 20,259 images, respectively.

4.3.2 Setup. Similar to section 4.1 and section 4.2, the analysis of which traits of the original face images are removed through our anonymization pipeline is based on utilizing an auxiliary classifier to compare original versus anonymized images. We trained the same model architecture described in Section 4.2.2, but this time to classify facial features rather than emotions. The only changes made to the architecture were matching the output layer to fit the number of features incorporated in the CelebA dataset, switching to a binary-cross-entropy loss, and changing the output activation function to Sigmoid, as in this case, we dealt with a multi-label task (i.e., multiple traits can be present at once). Here, each feature can be interpreted as a facial trait that is apparent in the model's face input. Exemplary anonymized images for CelebA can be seen in Figure 8. The performance of the classification model can be looked up in the appendix in Figure 11 and Table 5.

Fig. 8. Sample of synthesized faces based on the CelebA dataset.

4.3.3 Metric. To examine which of those traits get removed, for each trait we take the subset of images in the original dataset where the classifier predicted that trait, i.e., the classifier assigned it a probability of > 0.5. Subsequently, we assess the portion of anonymized versions of those images where the classifier did not predict the respective trait.

4.3.4 Results. The results are depicted in Figure 9. Here, we ordered the features according to the percentage of cases where they have been removed. The actual percentages are in the appendix in Table 6.

4.3.5 Discussion. As can be seen in the results, some traits were removed in 100% of the cases, whilst others were preserved in almost all images. Traits that refer to head or facial hair, e.g., *Bald*, *Gray Hair*, *Mustache*, or *Goatee* are removed quite frequently. This is not surprising since the only information our re-synthesis model can rely on is the landmark representation of the input face. Changing the head or facial hair style or color does not necessarily obscure the face to a degree of unrecognizable and can, therefore, be regarded as unimportant in terms of anonymization. Also, wearing specific accessories like neckties, hats, or necklaces is not encoded

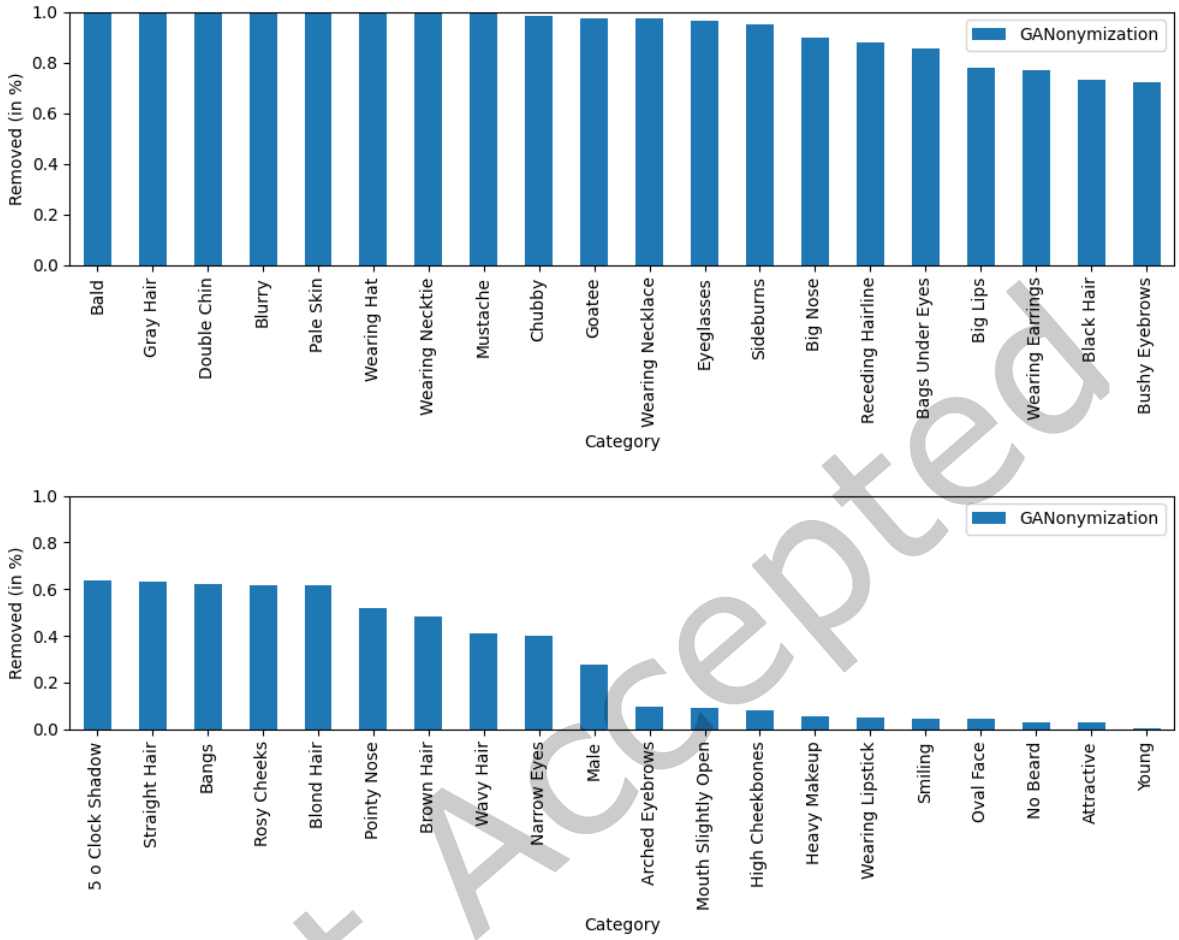


Fig. 9. The removed categories in % over the total number of available samples for each category in the CelebA dataset.

in landmark representations, resulting in them getting reliably removed. Again, changing the accessories can also help obscure the face but does not help anonymization. The *Smiling* feature, which is highly correlated to emotional expressions, gets preserved quite well, which again supports our claim of being able to preserve such expressions. The emotional expression falls under the category of facial shape and, therefore, should be preserved to keep the original shape of the face intact. On the other hand, a surprising observation is that *Heavy Makeup* and *Wearing Lipstick* predominantly are getting preserved. The training data we used for our GAN model is a possible explanation. For that, the CelebA dataset, containing exclusively celebrity face images, was used, too. In the world of celebrities, it is common practice for women to dress up and apply makeup for their appearance at public events. As the GAN model aims to resemble the data distribution imposed by the training data, these traits are also apparent in the anonymized versions. The same goes for features like *No Beard* or *Young* - the vast majority of subjects in the CelebA dataset are relatively young and do not wear a beard [40]. This bias, regarding *Heavy Makeup*, *Wearing Lipstick*, *Young*, and *No Beard*, introduced by the CelebA dataset indicates a limitation of

this approach and the necessity to include more images with contradicting features to balance the training data. Besides that, an interesting observation is that the *Chubby* trait was removed in the vast majority of cases where it was apparent. Intuitively, the facial landmark representation should have covered that trait, but apparently, it wasn't. The same goes for *Big Nose* and *Big Lips* - which were also removed frequently. Removing those traits advocates our approach since they are typical examples of features that could introduce unfairness and bias into datasets.

The feature *Male* got removed in 27.62% of the cases. It has to be noted that there is no *Female* feature in the dataset, and as such, the absence of the *Male* trait is mainly interpreted as the face being female. Therefore, it is good that the *Male* trait did not get removed in 100% of the cases - which would mean that the anonymized versions are *always* female. Here, a medium removal rate indicates that the gender sometimes changes and sometimes does not, indicating that it indeed gets diluted by GANonymization.

Finally, the feature *Blurry* was removed in over 99% of the cases. Although this trait doesn't refer to the face itself but to the image quality, it is a good indicator that the results of GANonymization are of high quality - even if the original images are not.

5 CONCLUSION

This research aimed to evaluate the anonymization performance using our method - GANonymization. Our method is a generative adversarial network (GAN) based image-to-image translation approach to anonymize faces and preserve their original facial expression. Facial landmarks serve as image input into a *pix2pix* architecture to re-synthesize high-quality, anonymized versions of the input face image.

First, we measured the efficiency of our approach in removing identifiable facial attributes to increase the anonymity of the given individual face. Our method proved its anonymization performance in the chosen metric on the WIDER dataset.

Second, we evaluated the performance regarding preserving emotional facial expressions on the AffectNet, CK+, and FACES datasets. Our approach significantly outperformed DeepPrivacy2 in most categories. However, DeepPrivacy2 significantly outperformed our approach in the emotion *Fear* and *Happy* from the AffectNet dataset. Compared to CIAGAN we could show a significant improvement in most of the preserved emotional facial expressions for *Neutral*, *Anger*, *Contempt* (in AffectNet), *Disgust*, *Fear* (in FACES), *Happy* (in CK+), *Sadness* (in CK+ and FACES), and *Surprise* (in AffectNet). Furthermore, a noticeable quality difference in the image could be seen between the different methods. Here, our method showed the highest quality in the synthesized faces.

Last, analyzing facial traits removed by our approach showed that some traits were eliminated in almost 100% of the cases while others were preserved. Especially jewelry, clothing, and hair, e.g., *Bald*, *Gray Hair*, *Mustache*, or *Goatee* are removed quite reliably.

In future efforts, training the GAN with a wider variety of facial expressions and traits might be supportive in increasing the overall performance in preserving the facial expressions, especially in adding more diversity to the generated faces. Additionally, evaluating each facial feature's contribution to the anonymization performance might be valuable - it might help to investigate further how anonymization models can mitigate bias and fairness issues. Finally, how GANonymization can be transferred to other, not emotion-related, contexts has to be studied. Therefore, suitable intermediate representations for the specific tasks have to be investigated. The medical domain is one field where adapting our approach might have a major impact. Here, anonymizing patient photos could reduce the sparsity of available data, which is crucial for that field. Doing so might enhance the data basis researchers can work with without being restricted by data privacy regulations.

ACKNOWLEDGMENTS

This paper was partially funded by the DFG through the Leibniz Award of Elisabeth André (AN 559/10-1).

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 308–318.

[2] Mahawaga Arachchige Pathum Chamikara, Peter Bertok, Ibrahim Khalil, Dongxi Liu, and Seyit Camtepe. 2020. Privacy preserving face recognition utilizing differential privacy. *Computers & Security* 97 (2020), 101951.

[3] William L Croft, Jörg-Rüdiger Sack, and Wei Shi. 2021. Obfuscation of images via differential privacy: from facial images to general images. *Peer-to-Peer Networking and Applications* 14 (2021), 1705–1733.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[6] Natalie C Ebner, Michaela Riediger, and Ulman Lindenberger. 2010. FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. *Behavior research methods* 42 (2010), 351–362.

[7] Andrian Firmansyah, Tien Fabrianti Kusumasari, and Ekky Novriza Alam. 2023. Comparison of Face Recognition Accuracy of ArcFace, Facenet and Facenet512 Models on Deepface Framework. In *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*. IEEE, 535–539.

[8] Nils Gruschka, Vasileios Mavroeidis, Kamer Vishi, and Meiko Jensen. 2018. Privacy issues and data protection in big data: a case study analysis under GDPR. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 5027–5033.

[9] Shivam Gupta. 2018. Facial emotion recognition in real-time and static images. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. 553–560. https://doi.org/10.1109/ICISC.2018.8398861

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. *CoRR* abs/1703.06870 (2017). arXiv:1703.06870 http://arxiv.org/abs/1703.06870

[11] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. 2022. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15014–15023.

[12] Håkon Hukkelås, Morten Smebye, Rudolf Mester, and Frank Lindseth. 2022. Realistic Full-Body Anonymization with Surface-Guided GANs. *CoRR* abs/2201.02193 (2022). arXiv:2201.02193 https://arxiv.org/abs/2201.02193

[13] Håkon Hukkelås and Frank Lindseth. 2022. DeepPrivacy2: Towards Realistic Full-Body Anonymization. https://doi.org/10.48550/ARXIV.2211.09454

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2016. Image-to-Image Translation with Conditional Adversarial Networks. *CoRR* abs/1611.07004 (2016). arXiv:1611.07004 http://arxiv.org/abs/1611.07004

[15] Amin Jourabloo, Xi Yin, and Xiaoming Liu. 2015. Attribute preserved face de-identification. In *2015 International conference on biometrics (ICB)*. IEEE, 278–285.

[16] Kajal Kansal, Yongkang Wong, and Mohan Kankanhalli. 2024. Privacy-Enhancing Person Re-Identification Framework-A Dual-Stage Approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 8543–8552.

[17] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=Hk99zCeAb

[18] Tero Karras, Samuli Laine, and Timo Aila. 2021. A Style-Based Generator Architecture for Generative Adversarial Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 43, 12 (2021), 4217–4228. https://doi.org/10.1109/TPAMI.2020.2970919

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 8107–8116. https://doi.org/10.1109/CVPR42600.2020.00813

[20] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. 2019. Real-time facial surface geometry from monocular video on mobile GPUs. *arXiv preprint arXiv:1907.06724* (2019).

[21] Brendan F Klare, Mark J Burge, Joshua C Klontz, Richard W Vorder Bruegge, and Anil K Jain. 2012. Face recognition performance: Role of demographic information. *IEEE Transactions on information forensics and security* 7, 6 (2012), 1789–1801.

[22] Byoung Chul Ko. 2018. A brief review of facial emotion recognition based on visual information. *sensors* 18, 2 (2018), 401.

[23] Zhenzhong Kuang, Huigui Liu, Jun Yu, Aikui Tian, Lei Wang, Jianping Fan, and Noboru Babaguchi. 2021. Effective de-identification generative adversarial network for face anonymization. In *Proceedings of the 29th ACM International Conference on Multimedia*. 3182–3191.

[24] Minh-Ha Le and Niklas Carlsson. 2022. StyleID: Identity Disentanglement for Anonymizing Faces. *arXiv preprint arXiv:2212.13791* (2022).

[25] Jian Li, Yabiao Wang, Changan Wang, Ying Tai, Jianjun Qian, Jian Yang, Chengjie Wang, Ji-Lin Li, and Feiyue Huang. 2018. DSFD: Dual Shot Face Detector. *CoRR* abs/1810.10220 (2018). arXiv:1810.10220 http://arxiv.org/abs/1810.10220

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

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

[28] Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *CoRR* abs/1711.05101 (2017). arXiv:1711.05101 http://arxiv.org/abs/1711.05101

[29] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. 2010. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 94–101. https://doi.org/10.1109/CVPRW.2010.5543262

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

[31] Silvan Mertes, Andreas Margraf, Steffen Geinitz, and Elisabeth André. 2020. Alternative data augmentation for industrial monitoring using adversarial learning. In *International Conference on Deep Learning Theory and Applications*. Springer, 1–23.

[32] Silvan Mertes, Andreas Margraf, Christoph Kommer, Steffen Geinitz, and Elisabeth André. 2020. Data Augmentation for Semantic Segmentation in the Context of Carbon Fiber Defect Detection using Adversarial Learning. In *Proceedings of the 1st International Conference on Deep Learning Theory and Applications, DeLTA 2020, Lieusaint, Paris, France, July 8-10, 2020*, Ana L. N. Fred and Kurosh Madani (Eds.). ScitePress, 59–67. https://doi.org/10.5220/0009823500590067

[33] Ali Mollahosseini, Behzad Hassani, and Mohammad H. Mahoor. 2017. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *CoRR* abs/1708.03985 (2017). arXiv:1708.03985 http://arxiv.org/abs/1708.03985

[34] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2018. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*. 634–646.

[35] Natalia Neverova, David Novotný, Vasil Khalidov, Marc Szafraniec, Patrick Labatut, and Andrea Vedaldi. 2020. Continuous Surface Embeddings. *CoRR* abs/2011.12438 (2020). arXiv:2011.12438 https://arxiv.org/abs/2011.12438

[36] Elaine M Newton, Latanya Sweeney, and Bradley Malin. 2005. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering* 17, 2 (2005), 232–243.

[37] Binh T. Nguyen, Minh H. Trinh, Tan V. Phan, and Hien D. Nguyen. 2017. An efficient real-time emotion detection using camera and facial landmarks. In *2017 Seventh International Conference on Information Science and Technology (ICIST)*. 251–255. https://doi.org/10.1109/ICIST.2017.7926765

[38] O Parkhi, A Vedaldi, and A Zisserman. 2015. Deep face recognition. *BMVC 2015 - Proceedings of the British Machine Vision Conference 2015*, 1–12.

[39] Nisarg Raval, Ashwin Machanavajjhala, and Landon P Cox. 2017. Protecting visual secrets using adversarial nets. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 1329–1332.

[40] Ethan M Rudd, Manuel Günther, and Terrance E Boult. 2016. Moon: A mixed objective optimization network for the recognition of facial attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 19–35.

[41] Sefik Ilkin Serengil and Alper Ozpinar. 2020. LightFace: A Hybrid Deep Face Recognition Framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 23–27. https://doi.org/10.1109/ASYU50717.2020.9259802

[42] Sefik Ilkin Serengil and Alper Ozpinar. 2021. HyperExtended LightFace: A Facial Attribute Analysis Framework. In *2021 International Conference on Engineering and Emerging Technologies (ICEET)*. IEEE, 1–4. https://doi.org/10.1109/ICEET53442.2021.9659697

[43] Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2017. Natural and Effective Obfuscation by Head Inpainting. *CoRR* abs/1711.09001 (2017). arXiv:1711.09001 http://arxiv.org/abs/1711.09001

[44] Yunqian Wen, Bo Liu, Ming Ding, Rong Xie, and Li Song. 2022. Identitydp: Differential private identification protection for face images. *Neurocomputing* 501 (2022), 197–211.

[45] Yifan Wu, Fan Yang, Yong Xu, and Haibin Ling. 2019. Privacy-protective-GAN for privacy preserving face de-identification. *Journal of Computer Science and Technology* 34 (2019), 47–60.

[46] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. 2018. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European conference on computer vision (ECCV)*. 606–624.

[47] Kaiyu Yang, Jacqueline H Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. 2022. A study of face obfuscation in imagenet. In *International Conference on Machine Learning*. PMLR, 25313–25330.

[48] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. 2016. WIDER FACE: A Face Detection Benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[49] Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar. 2020. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE journal of biomedical and health informatics* 24, 8 (2020), 2378–2388.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer

Society, 2242–2251. https://doi.org/10.1109/ICCV.2017.244

A APPENDIX

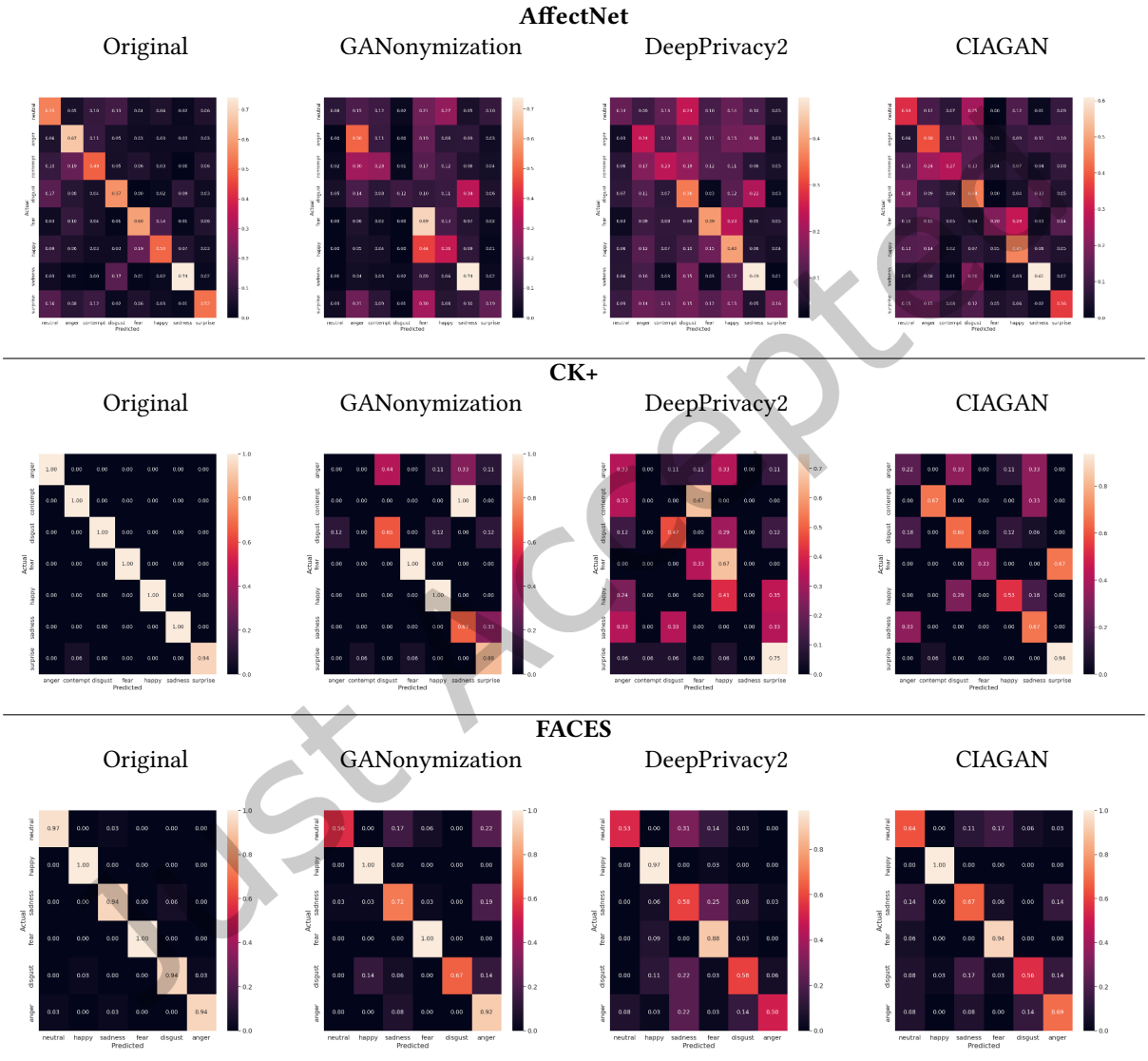


Fig. 10. For each specified dataset a multi-class classification model was trained. Accordingly, the confusion matrices depict the classification model's performance on the validation sets. The column "Original" model was trained on the original images from the training split. The column "GANonymization" and "DeepPrivacy2" contains the models trained on the synthesized images of the training split, respectively.

Table 4. For each specified dataset a multi-class classification model was trained. Accordingly, the classification reports show the classification model's performance on the validation sets. The column "Original" model was trained on the original images from the training split. The column "GANonymization" and "DeepPrivacy2" contains the models trained on the synthesized images of the training split, respectively. (P) Precision; (R) Recall; (F1) F1-Score; (N) Support

		Original			GANonymization			DeepPrivacy2			CIAGAN			
Dataset		P	R	F1	P	R	F1	P	R	F1	P	R	F1	N
AffectNet	Neutral	0.49	0.41	0.45	0.43	0.08	0.14	0.26	0.14	0.18	0.30	0.34	0.32	360
	Anger	0.58	0.59	0.58	0.34	0.50	0.41	0.22	0.24	0.23	0.28	0.38	0.32	346
	Contempt	0.48	0.67	0.56	0.37	0.28	0.32	0.28	0.23	0.25	0.41	0.27	0.33	354
	Disgust	0.58	0.54	0.56	0.62	0.12	0.20	0.25	0.36	0.29	0.32	0.44	0.38	357
	Fear	0.65	0.62	0.63	0.32	0.69	0.44	0.38	0.39	0.38	0.54	0.20	0.29	357
	Happy	0.57	0.51	0.53	0.30	0.36	0.33	0.28	0.40	0.33	0.39	0.45	0.42	362
	Sadness	0.68	0.77	0.72	0.48	0.74	0.58	0.41	0.49	0.45	0.60	0.61	0.60	352
	Surprise	0.66	0.55	0.60	0.39	0.19	0.25	0.36	0.16	0.22	0.39	0.36	0.37	337
	accuracy			0.58			0.37			0.30			0.38	2825
	macro avg	0.59	0.58	0.58	0.41	0.37	0.33	0.31	0.30	0.29	0.40	0.38	0.38	2825
	weighted avg	0.58	0.58	0.58	0.41	0.37	0.33	0.30	0.30	0.29	0.40	0.38	0.38	2825
CK+	Anger	1.00	1.00	1.00	0.00	0.00	0.00	0.25	0.33	0.29	0.33	0.22	0.27	9
	Contempt	0.75	1.00	0.86	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.67	0.67	3
	Disgust	1.00	1.00	1.00	0.73	0.65	0.69	0.73	0.47	0.57	0.58	0.65	0.61	17
	Fear	1.00	1.00	1.00	0.75	1.00	0.86	0.25	0.33	0.29	1.00	0.33	0.50	3
	Happy	1.00	1.00	1.00	0.85	1.00	0.92	0.39	0.41	0.40	0.75	0.53	0.62	17
	Sadness	1.00	1.00	1.00	0.25	0.67	0.36	0.00	0.00	0.00	0.20	0.67	0.31	3
	Surprise	1.00	0.94	0.97	0.78	0.88	0.82	0.55	0.75	0.63	0.88	0.94	0.91	16
	accuracy			0.99			0.69			0.46			0.62	68
	macro avg	0.96	0.99	0.97	0.48	0.60	0.52	0.31	0.33	0.31	0.63	0.57	0.55	68
	weighted avg	0.99	0.99	0.99	0.62	0.69	0.65	0.45	0.46	0.44	0.67	0.62	0.62	68
FACES	neutral	0.97	0.97	0.97	0.95	0.56	0.70	0.86	0.53	0.66	0.64	0.64	0.64	36
	happy	0.97	1.00	0.99	0.86	1.00	0.92	0.78	0.97	0.86	0.97	1.00	0.99	36
	sadness	0.94	0.94	0.94	0.70	0.72	0.71	0.44	0.58	0.50	0.65	0.67	0.66	36
	fear	1.00	1.00	1.00	0.92	1.00	0.96	0.64	0.88	0.74	0.78	0.94	0.85	34
	disgust	0.94	0.94	0.94	1.00	0.67	0.80	0.68	0.58	0.63	0.74	0.56	0.63	36
	anger	0.97	0.94	0.96	0.62	0.92	0.74	0.86	0.50	0.63	0.69	0.69	0.69	36
	accuracy			0.97			0.81			0.67			0.75	214
	macro avg	0.97	0.97	0.97	0.84	0.81	0.81	0.71	0.67	0.67	0.75	0.75	0.74	214
	weighted avg	0.97	0.97	0.97	0.84	0.81	0.80	0.71	0.67	0.67	0.75	0.75	0.74	214

Fig. 11. A multi-label classification model was trained on the CelebA dataset. Accordingly, the confusion matrices depict the classification model's performance on the validation sets. The column "Original" model was trained on the original images from the training split. The column "GANonymization" and "DeepPrivacy2" contains the models trained on the synthesized images of the training split, respectively.

Table 5. A multi-label classification model was trained on the CelebA dataset. Accordingly, the classification reports show the classification model's performance on the validation sets for each label. The column "Original" model was trained on the original images from the training split. The column "GANonymization" and "DeepPrivacy2" contains the models trained on the synthesized images of the training split, respectively. (P) Precision; (R) Recall; (F1) F1-Score; (N) Support

	Original			GANonymization			DeepPrivacy2			CIAGAN			
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	N
5 o Clock Shadow	0.72	0.79	0.75	0.00	0.00	0.00	0.71	0.67	0.69	0.66	0.20	0.30	2345
Arched Eyebrows	0.74	0.69	0.72	0.75	0.40	0.52	0.63	0.49	0.55	0.65	0.59	0.62	5134
Attractive	0.79	0.86	0.83	0.87	0.47	0.61	0.77	0.81	0.79	0.78	0.79	0.79	10332
Bags Under Eyes	0.67	0.52	0.59	0.60	0.07	0.12	0.54	0.47	0.50	0.62	0.29	0.40	4120
Bald	0.74	0.48	0.58	0.00	0.00	0.00	0.73	0.46	0.56	0.74	0.23	0.35	410
Bangs	0.84	0.86	0.85	0.83	0.05	0.10	0.81	0.86	0.84	0.84	0.85	0.85	2913
Big Lips	0.62	0.22	0.33	0.37	0.43	0.40	0.54	0.21	0.31	0.59	0.18	0.28	3044
Big Nose	0.79	0.44	0.56	0.61	0.49	0.54	0.63	0.52	0.57	0.65	0.48	0.56	4940
Black Hair	0.78	0.75	0.76	0.65	0.13	0.21	0.76	0.72	0.74	0.68	0.80	0.74	4143
Blond Hair	0.82	0.85	0.84	0.88	0.05	0.09	0.77	0.87	0.82	0.77	0.86	0.82	3054
Blurry	0.72	0.45	0.55	0.77	0.01	0.02	0.62	0.38	0.47	0.65	0.35	0.45	929
Brown Hair	0.68	0.64	0.66	1.00	0.00	0.00	0.70	0.56	0.62	0.74	0.42	0.53	4792
Bushy Eyebrows	0.79	0.67	0.73	0.93	0.00	0.01	0.58	0.45	0.51	0.73	0.43	0.54	2830
Chubby	0.68	0.48	0.57	0.62	0.04	0.07	0.50	0.50	0.50	0.63	0.29	0.40	1215
Double Chin	0.70	0.50	0.59	0.57	0.01	0.02	0.51	0.50	0.50	0.69	0.29	0.40	975
Eyeglasses	0.97	0.96	0.97	0.64	0.23	0.34	0.90	0.86	0.88	0.84	0.45	0.58	1380
Goatee	0.81	0.69	0.75	0.60	0.01	0.01	0.79	0.61	0.69	0.67	0.17	0.28	1460
Gray Hair	0.81	0.70	0.75	1.00	0.00	0.00	0.77	0.68	0.72	0.82	0.57	0.67	966
Heavy Makeup	0.88	0.92	0.90	0.87	0.71	0.78	0.80	0.91	0.85	0.80	0.88	0.84	7751
High Cheekbones	0.92	0.80	0.86	0.78	0.87	0.82	0.81	0.78	0.79	0.75	0.87	0.81	8926
Male	0.97	0.98	0.98	0.91	0.85	0.88	0.94	0.95	0.94	0.94	0.93	0.93	8443
Mouth Slightly Open	0.94	0.94	0.94	0.73	0.96	0.83	0.84	0.69	0.76	0.86	0.92	0.89	9569
Mustache	0.72	0.49	0.59	0.00	0.00	0.00	0.52	0.34	0.41	0.42	0.04	0.08	1002
Narrow Eyes	0.51	0.67	0.58	0.64	0.23	0.33	0.40	0.03	0.06	0.33	0.00	0.00	1491
No Beard	0.97	0.98	0.98	0.87	0.97	0.92	0.95	0.98	0.96	0.91	0.95	0.93	16326
Oval Face	0.67	0.29	0.40	0.87	0.00	0.01	0.58	0.31	0.41	0.51	0.39	0.44	5564
Pale Skin	0.58	0.66	0.62	0.88	0.06	0.11	0.71	0.32	0.44	0.63	0.38	0.48	856
Pointy Nose	0.65	0.45	0.53	0.74	0.02	0.04	0.55	0.34	0.42	0.61	0.24	0.35	5658
Receding Hairline	0.64	0.43	0.52	0.60	0.01	0.02	0.64	0.41	0.50	0.54	0.43	0.48	1429
Rosy Cheeks	0.77	0.40	0.52	1.00	0.00	0.00	0.51	0.58	0.54	0.54	0.48	0.50	1358
Sideburns	0.84	0.75	0.79	0.00	0.00	0.00	0.84	0.65	0.73	0.75	0.20	0.32	1366
Smiling	0.95	0.90	0.92	0.83	0.94	0.88	0.86	0.80	0.83	0.83	0.91	0.87	9601
Straight Hair	0.60	0.41	0.49	0.00	0.00	0.00	0.55	0.40	0.47	0.52	0.26	0.34	4082
Wavy Hair	0.68	0.64	0.66	0.57	0.22	0.32	0.66	0.62	0.64	0.67	0.56	0.61	5492
Wearing Earrings	0.77	0.59	0.67	0.80	0.01	0.02	0.72	0.58	0.64	0.76	0.46	0.57	3789
Wearing Hat	0.87	0.89	0.88	0.87	0.14	0.25	0.84	0.88	0.86	0.89	0.82	0.86	939
Wearing Lipstick	0.88	0.96	0.92	0.87	0.82	0.85	0.83	0.95	0.89	0.83	0.94	0.89	8860
Wearing Necklace	0.51	0.15	0.23	0.00	0.00	0.00	0.48	0.06	0.10	0.38	0.01	0.02	2396
Wearing Necktie	0.60	0.29	0.39	0.00	0.00	0.00	0.54	0.29	0.38	0.57	0.09	0.16	1442
Young	0.87	0.97	0.92	0.78	0.98	0.87	0.86	0.96	0.90	0.86	0.95	0.90	14821
micro avg	0.84	0.76	0.80	0.79	0.51	0.62	0.78	0.71	0.74	0.79	0.68	0.73	176143
macro avg	0.76	0.65	0.69	0.63	0.25	0.28	0.69	0.59	0.62	0.69	0.50	0.55	176143
weighted avg	0.82	0.76	0.78	0.73	0.51	0.52	0.76	0.71	0.72	0.75	0.68	0.69	176143
samples avg	0.83	0.75	0.78	0.79	0.51	0.60	0.78	0.70	0.72	0.78	0.67	0.70	176143

	GANonymization
Bald	1.000000
Gray Hair	1.000000
Double Chin	0.998494
Blurry	0.996370
Pale Skin	0.996337
Wearing Hat	0.993348
Wearing Necktie	0.992764
Mustache	0.992661
Chubby	0.984813
Goatee	0.973311
Wearing Necklace	0.972358
Eyeglasses	0.966012
Sideburns	0.949251
Big Nose	0.899965
Receding Hairline	0.877510
Bags Under Eyes	0.852971
Big Lips	0.780942
Wearing Earrings	0.768467
Black Hair	0.729177
Bushy Eyebrows	0.721409
5 o Clock Shadow	0.636142
Straight Hair	0.630562
Bangs	0.620606
Rosy Cheeks	0.615530
Blond Hair	0.615213
Pointy Nose	0.516256
Brown Hair	0.480853
Wavy Hair	0.410118
Narrow Eyes	0.400334
Male	0.276199
Arched Eyebrows	0.097596
Mouth Slightly Open	0.089010
High Cheekbones	0.083279
Heavy Makeup	0.054131
Wearing Lipstick	0.048915
Smiling	0.046791
Oval Face	0.044784
No Beard	0.031004
Attractive	0.028488
Young	0.001595

Table 6. The table shows the percentage of removed traits over the total number of available samples for each trait in the validation set of the CelebA dataset.