This ACCV 2022 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ACCV 2022
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/accv

Decoupling identity and visual quality for image and video anonymization

Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé

Technical University of Munich, Arcisstraße 21, 80333 Munich, Germany

Abstract. The widespread usage of computer vision applications in the public domain has opened the delicate question of image data privacy. In recent years, computer vision researchers have proposed technological solutions to anonymize image and video data so that computer vision systems can still be used without compromising data privacy. While promising, these methods come with a range of limitations, including low diversity of outputs, low-resolution generation quality, the appearance of artifacts when handling extreme poses, and non-smooth temporal consistency. In this work, we propose a novel network based on generative adversarial networks (GANs) for face anonymization in images and videos. The key insight of our approach is to decouple the problems of image generation and image blending. This allows us to reach significant improvements in image quality, diversity, and temporal consistency while making possible to train the network in different tasks and datasets. Furthermore, we show that our framework is able to anonymize faces containing extreme poses, a long-standing problem in the field.

Keywords: Anonymization · Image Synthesis.

1 Introduction

The increase of cameras in the real world offers the possibility of widespread usage for computer vision tools, with applications ranging from autonomous robots and cars to automatic monitoring of public spaces. The question of personal privacy is becoming more prominent, especially since people are often the subject of observation by these cameras. The European Union has passed laws on data protection such as the General Data Protection Regulations (GDPR) [6]. The research community has also accepted responsibility, from taking offline one of the most popular re-identification datasets [11] to making mandatory for one of the leading machine learning conferences to consider the ethical issues of scientific publications. From a technical point of view, as researchers, we can also contribute to the solution by proposing novel computer vision tools.

For many vision tasks such as person detection, person tracking, or action recognition, we do not need to *identify* the people in the videos, we only need to *detect* them [24]. Recent works proposed to use computer vision tools to remove identity information from people's faces in videos [7,24,30,36], while still trying to preserve the accuracy of the computer vision algorithms for the final task,

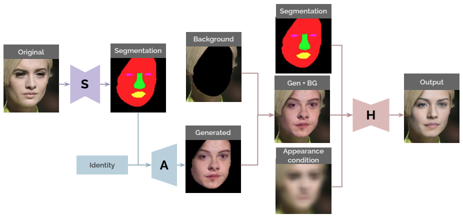

Fig. 1. The inference pipeline of our method. We use as input segmentation masks, given by a segmentation network (S). We use AnonymizationNet (A) to generate an anonymized version of the original image, and then use HarmonizationNet (H) to blend the generated image with the background.

e.g., face detection. These methods anonymize identities by replacing or altering the input faces, achieving overall good de-identification results. While a valuable first step, these methods come with several weaknesses.

First, the visual quality of the generated images is lacking [24,30]. This is often due to the lack of high-quality datasets suitable for training anonymization methods. In other cases, there is a trade-off between image quality and anonymization results, leading to photorealistic images that are still identifiable by humans [7]. Second, output diversity is low, i.e., for the same input identity, only a few types of anonymizations are produced. If every input identity is mapped to one anonymized version, then it is straightforward to de-anonymize all faces and establish correspondences to the original identity.

In this work, we propose to overcome these issues by decoupling the anonymization task from the image generation task. We argue that it is not efficient for a single network to focus on both diversity as well as producing realistically looking results. Our pipeline is therefore separated into two networks. The job of the first network is to generate anonymized and diverse versions of the original face, while preserving its pose, without focusing on integrating it with the background information. The second network receives the anonymized face and performs image blending, i.e., changing the appearance of the generated image according to external conditioning, such as illumination or background appearance. Overall, the second network is responsible for generating high-quality realistically looking images. We show how to train the second network on a proxy task not related to image anonymization. Interestingly, this allows us to separately train the second network on datasets that do not contain information about face identities. This is the key towards generating high-quality outputs, given high-resolution face datasets do not contain identity information. It is therefore challening for single-stage anonymization methods [7,24,29] to make use of them, and consequently, they are limited to low-resolution datasets such as CelebA [23].

In the experimental section, we show that our decoupled formulation achieves state-of-the-art results on output diversity and quality. We further compare the performance of state-of-the-art methods under extreme poses, an underestimated problem in the field [24], and evaluate the temporal consistency of the output when moving towards video anonymization. Finally, we show that our method can be intuitively extended to other domains, e.g., full-body anonymization.

Our **contribution** in this work is three-fold:

- We propose a novel two-step framework that decouples image generation from image blending. We show how to train the second step on a proxy task unrelated to image anonymization, which allows us to leverage high-resolution datasets designed for different tasks. This allows us to generate overall more diverse and high-quality outputs.
- We analyze the drawbacks of current anonymization methods with a comprehensive study on output diversity, quality, anonymization and detection rates, temporal consistency and performance on extreme poses.
- We present state-of-the-art results on six public datasets.

2 Related Work

GANs for Face Generation and Translation. The advent of Generative Adversarial Networks [9,28] brought a lot of research interest in the field of face generation [16,17,22]. Recent methods [17,18] are able to generate realistically looking high-resolution face images and provide high diversity. However, these methods have no mechanism for conditioning their output on the original face, making the blending of the generated face with the other parts of the body challenging. Additionally, they cannot keep the stature and the direction of the face, making their usability in face anonymization limited. More successful have been the methods based on temporal consistency [39]. While the method provides a simple image and video translation method, the generated faces are similar to the original identities, making it not usable for anonymization.

Face Anonymization. Traditionally, face anonymization has been achieved by heuristic methods, such as pixelization, blurring, masking or segmentation [33]. The pioneering work of [10] used for the first time model-based learning for face de-identification. Recently, deep learning models have been used for the same problem [7,14,24,29,35,36]. In particular, [14,29] are one of the first methods to use GANs that reach promising de-identification results. However, the generated outputs are not naturally blended with the rest of the image, and [29] maps every identity to a unique fake identity, not allowing the generation of diverse images. The work of [35] uses a different GAN scheme to further improve the results, but the generated images still remain unnatural looking and provide no explicit way of controlling the generated appearance. Some of these problems were remedied in [36], providing a method for generating faces that are more de-identifiable and natural-looking. However, the method is based on a parametric face model with an additional alignment procedure that does not offer a direct way of extending it to other domains such as full bodies and works only on images.

Current state-of-the-art models [7,24] mitigate some of the issues mentioned above. The work of [7] generates high-quality looking images, temporally consistent videos, and reaches high de-identification results. However, as argued in [24], the generated images can be recognized by humans and the method does not provide a way to generate diverse images. On the other hand, [24] reaches state-of-the-art de-identification results, and the generated images are not easily identifiable by humans. At the same time, the images do not look natural and video results are not as temporally consistent as in [7]. Finally, both methods can be only trained on datasets that provide multiple images for each identity, thus making it harder to use many high-quality datasets.

In this work, by decoupling the anonymization from the blending process, our method allows higher control over the de-identification process and can be used with datasets that provide only one image per identity.

3 Methodology

Most image translation methods typically take input representations, e.g., semantic segmentation, landmarks, or a background image, and use a network to encode the input information into a low-dimensional latent space. A decoder then translates the information into a new image, which, in the case of anonymization, is a face with a new identity. The decoder is optimized for two tasks: (i) image anonymization: generating image parts that form a new identity; (ii) image blending: changing the appearance of the generated image according to external conditioning, such as illumination or background appearance. Considering that these networks are trained with a single adversarial loss, often one of the tasks is neglected in favor of the other. Neglecting image anonymization results in low diversity of the outputs, i.e., the network only generates a few types of anonymizations, thus compromising privacy. Neglecting image blending results in artifacts and unrealistically looking faces, which often leads to computer vision algorithms not being able to perform face detection. Our proposed decoupling architecture allows us to create diverse outputs and high-quality outputs.

Overview of our method. In this section, we describe in depth the methodology of our proposed framework. Our model includes: (i) a pre-trained segmentation network, (ii) AnonymizationNet, a network that anonymizes the face and is based on a GAN and an identity network (see Fig. 3.1), and (iii) HarmonizationNet, a network that blends the produced face with the rest of the image (see Fig. 3.2). We first explain each of the networks' tasks and analyze why decoupling is the key element that allows us to generate high-quality diverse outputs. We then discuss which elements of our pipeline allow for identity control during anonymization, pose preservation, and temporal consistency.

3.1 Decoupling anonymization and visual quality

Inspired by rendering pipelines, where geometry creation and the shading process are divided, we advocate for the idea that the image generation should be separated from image blending, thus, we use two networks.

AnonymizationNet. The goal of AnonymizationNet (see Fig. 3.1) is to define the facial geometry and the characteristics of the face. We train a conditional

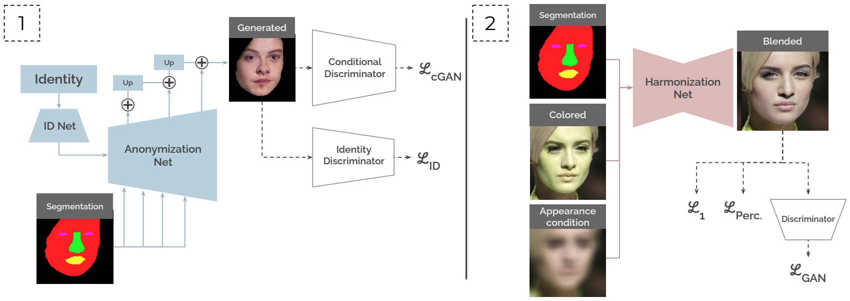

Fig. 2. The training pipeline of our model. Our model consists of two networks trained separately. 1) AnonymizationNet takes a segmentation mask and an identity condition to generate a face without any appearance conditioning. It uses losses from a conditional discriminator, and from an identity discriminator to guide generation to the desired identity. 2) HarmonizationNet is an encoder-decoder network where the encoder embeds the image information into a low dimensional space. It takes a triplet of segmentation mask, a colored face image, and an appearance condition. The appearance condition is given as a heavily blurred version of the original face, while the segmentation mask is given to indicate the image region that needs to be blended. The network learns to blend the input face to the given background by using information from the appearance condition.

GAN [15] to generate an anonymized version of the input face, without considering how the face fits with the other part of the image, e.g., the background. We achieve that by conditioning the generation process only on a semantic map of the original image. By doing so, we ensure that the pose of the face is preserved and prevents identity leakage. Similar to [24], the network collaborates with an identity network that guides the generator towards generating a face with different identifying characteristics.

HarmonizationNet. The task of HarmonizationNet (see Fig. 3.2) is to blend the generated face in order to naturally fit with the background and overall illumination. We use a conditional GAN to generate a realistic-looking version of the image produced by AnonymizationNet. Importantly, we do not train the network together with AnonymizationNet. Instead, we train on a *proxy task*, which greatly simplifies the training procedure and allows us to achieve high-quality and diverse image generation. During inference, the network takes the anonymized version of a face and blends it with the rest of the image, (Fig. 1).

Advantages of decoupling. In the experimental section, we validate that the presented decoupling achieves a higher image quality and output diversity, and has a higher degree of control over the blending process. A clear advantage of decoupling anonymization from visual quality is that it allows HarmonizationNet to be trained on a different dataset. AnonymizationNet needs to be trained on a dataset that contains multiple images per identity, with its goal being to provide anonymized versions of the images. Single-stage methods [7,24,29] are therefore limited to such datasets, e.g., CelebA, to train their anonymization pipeline. In contrast, our method can be additionally trained in high-resolution datasets, such as CelebA-MaskHQ, and consequently we are able to generate high-resolution images (see Fig 3).

3.2 Proxy training

As mentioned before, we propose to train the networks separately, as opposed to training our entire model in an end-to-end fashion. By training the networks separately, we ensure that the second network never sees the original image, which could be a cause for identity leakage. Furthermore, we can train our second network in parallel on datasets that provide images without identity information [23]. Last, but not least, we simplify the training and as shown in the ablation studies, the model trained with the proxy task reaches higher results compared to training the networks jointly.

We propose to train HarmonizationNet on a *proxy task*, which we design to be a relaxed version of the blending task. More concretely, we use a colored foreground image as an input, see Fig. 3.2, during training. The task of HarmonizationNet is to reconstruct the original image color. In other words, we train the network to change the appearance of the foreground in the input image to match the appearance of the overall image. The model takes as input: (i) a semantic segmentation map of the foreground, i.e., the face, (ii) the altered colored image, and (iii) a blurred version of the original image. The motivation for the third input is to provide some guidance to the blending network instead of allowing it to blindly reconstruct the image from the semantic input. As we show in Table 1 of experiments, the heavily blurred image removes identity information.

Our intuition is that the proxy task is teaching HarmonizationNet two functions: to disassociate the general facial (shape) details from the rest of the input appearance in the encoder and to inpaint the missing textures on top of those details in the decoder. During the inference, encoder activates on any high-level shape details despite the domain gap and embeds them in the bottleneck. In our evaluation and supplementary, we show a robust generalization of HarmonizationNet to the output of AnonymizationNet regardless of datasets used.

3.3 Identity guidance

The goal of AnonymizationNet is to generate a new anonymized image, given the segmentation map of the original face as input. Note, the semantic map allows us to preserve the pose of the face without allowing identity leakage. In order to control the anonymization output, we make use of a *control identity*.

For any given image, we randomly choose a control identity, parameterized by a one-hot vector. This information is fed into the generator of AnonymizationNet, with its goal being to embed identifying features of the control identity to the original semantic mask. This process is achieved with the use of an identity discriminator that provides a guiding signal to AnonymizationNet so that the generated image has similar characteristics to the control identity. The identity discriminator is a siamese neural network pre-trained on the real images using Proxy-NCA loss [25] and finetuned using the contrastive loss [1]. During finetuning, the network learns to bring together the identity representation of the fake images and the real images. We note that the identity discriminator is trained with AnonymizationNet in a collaborative (not adversarial) manner. As a result, the generator mixes the semantic segmentation map of the original identity, with the identifying features of the control identity, thereby creating a new non-identifiable identity.

Is attribute preservation desirable? We argue that preserving the attributes of the original identity is not desirable when it comes to face anonymization, and makes the pipeline less robust to identity attacks. For example, knowing the gender reduces the search space by half. Preserving other attributes, e.g., age, skin color, or specific attributes for eyes, nose, forehead, lowers the search space significantly and makes it easier to *guess* the identity. In our work, preserving attributes, e.g., gender, would be as simple as giving as control identity an identity that contains the same attribute. However, considering that some face verification methods [20,37] rely on facial attributes, we decide to not force any attribute preservation. The only exception is skin tone, considering that we generate only the facial region, and therefore need to match the face skin to the neck and the other exposed body skin areas.

3.4 Pose preservation and temporal consistency

Previous works rely on facial landmarks [24], faces [29] or statistical 3D shade models [36]. While a landmark representation is simple and easy-to-use, it fails on extreme poses and cannot properly represent certain body parts such as hair [24]. Statistical 3D shade models are quite robust, but add additional computation complexity and are domain-specific. Due to privacy reasons, we avoid working directly on faces. In our work, we use face segmentation as input representation. Using segmentation allows us to outline the area we want to modify, and we are able to estimate specific occlusions. Furthermore, we can use the same framework on other domains, e.g., for full-body anonymization, with few changes on the expected input type.

Additionally, to improve temporal consistency for video sequences, we transform HarmonizationNet into a frame recurrent network by concatenating the output of the previous frame to the input of the current frame and replacing a spatial discriminator with a temporal one. The temporal discriminator [4] takes three consecutive frames as an input and judges both temporal smoothness and visual quality. This simple change to HarmonizationNet leads to less color jittering in the final video.

3.5 Architectures and training

AnonymizationNet. We use spatially-adaptive denormalization (SPADE) [27] residual blocks as building blocks of the network. We give the same segmentation map as input to every SPADE block, and each block produces an RGB image. The control identity, represented as one hot-vector, is given as input to a transposed convolutional neural network. The network then produces a parametrized version of the identity and gives it as input to the generator. We also use a simplified design of upsampling and summing RGB outputs to avoid progressive GAN training, leading to a more stable and efficient training in higher resolutions. We sum all RGB outputs in order to get the final result. These two changes lead to a more robust training, while achieving higher quality compared to the regular encoder-decoder architecture used in image translation [24].

HarmonizationNet. We base the network's architecture on a U-Net composed of residual blocks [31]. We give the detailed architecture of both networks in the supplementary.

Loss functions. We use LSGAN loss function to train the networks. The loss of AnonymizationNet (A) generator is defined as:

$$\min_{G_A} V(G_A) = \frac{1}{2}\mathbb{E}_{i\sim p_{data}(i)}[(D_A(G_A(i)) - b)^2] + L_{id} \tag{1}$$

where L_{id} is the loss of the identity discriminator as explained in section 3.3, b is the label for the real data, i is the input to the generator G_A, and D_A is the discriminator.

The loss of HarmonizationNet (H) generator is defined as:

$$\begin{aligned}\min_{G_H} V(G_H) = \frac{1}{2}\mathbb{E}_{i\sim p_{data}(i)}[(D_H(G_H(i)) - b)^2] + \\ VGG_P(I, I') + L1(I, I')\end{aligned} \tag{2}$$

where G_H and D_H represent the generator and the discriminator of HarmonizationNet, VGG_P represents the perceptual loss of VGG network [8], I and I' represent the original and the generated image.

The loss function for the discriminators A and H is given below:

$$\begin{aligned}\min_{D} V(D) = \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x) - b)^2] + \\ \frac{1}{2}\mathbb{E}_{i\sim p_{data}(i)}[(D(G(i)) - a)^2]\end{aligned} \tag{3}$$

where a is the label for the fake data, x is the real data and D is valid for both discriminator D_A and D_H.

4 Experiments

In this section, we compare our method with several classic and learning-based methods commonly used for identity anonymization. We analyze the drawbacks of current anonymization methods in terms of diversity, image quality, anonymization and detection rates, temporal consistency, performance on extreme poses, and show state-of-the-art qualitative and quantitative results. We also present a set of comprehensive ablation studies to demonstrate the effect of our design choices. We detail the implementation details in the supplementary material.

Datasets. We perform experiments on 6 public datasets: two face datasets with annotated identities: CelebA [23] and Labeled Faces in the Wild (LFW) [13], a high-quality face dataset without identity information: CelebA-MaskHQ [23], a video dataset: FaceForensics++ [32], a dataset on extreme poses: AFLW2000 [19], and a dataset with annotated body segmentations: MOTS [38].

Baselines. We follow previous works [24] to use simple anonymization baselines such as pixelization or blurring and compare them with our method. We also compare with state-of-the-art learning-based methods [7,24].

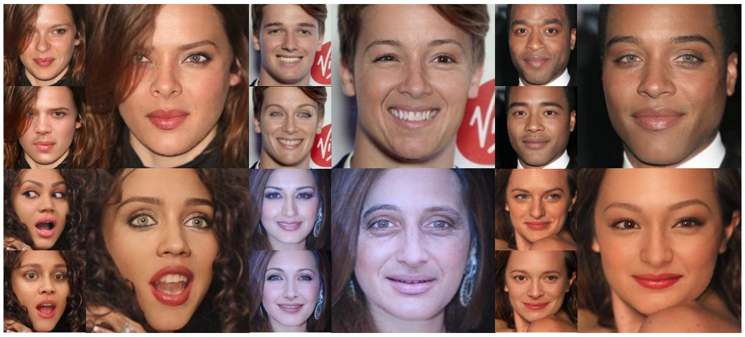

Fig. 3. A set of triplet images generated by our framework. In every triplet, the left-upper image is the original image in 128x128 resolution, the left-lower image is the anonymized version of it in 128x128 resolution, and the large image in the right is its anonymized version in 512x512 resolution. We present cases of different gender, skin-color, pose, and illumination.

4.1 Ablation study

To validate our two-step training concept, we do an ablation study measuring the importance of each component of our model.

Setup. We investigate how our proposed decoupled pipeline affects the quality and diversity of the generated images. We use FID [12] as a quality metric and Re-ID as a diversity metric. In Table 1, we present the results of three different configurations of our method: (1) a regular model without decoupling, (2) a decoupled model where the pipeline is jointly trained instead of proxy training, and (3) our model with decoupling of the task and proxy training.

We perform this ablation on two datasets: CelebA and CelebA-MaskHQ [23]. There is a domain gap between the datasets, as CelebA-MaskHQ consists of higher quality and sharper images compared to CelebA. Most importantly, CelebA-MaskHQ does not contain identity information by itself, hence we combine it with CelebA to be able to train AnonymizationNet. For the CelebA-MaskHQ evaluation, we train only HarmonizationNet on the CelebA-MaskHQ for decoupled configurations (networks (2) and (3)) and do simultaneous training on both datasets for the single model configuration (network (1)).

Models	CelebA			CelebA + HQ		
	FID (↓)	Re-ID (↑)	Detection (↑)	FID (↓)	Re-ID (↑)	Detection (↑)
1) w/o decoupling	**3.05**	33.4	.993	16.25	17.3	.903
2) w/o proxy training	3.15	88.6	.992	16.75	**170.5**	.917
3) Two-step framework	4.17	**96.1**	**.999**	**10.49**	124.1	**.994**

Table 1. Ablation study of our model. We measure the generation quality and diversity for different versions of our model. Lower (↓) results for FID imply a higher generation quality. Higher (↑) results for diversity imply a higher diversity in generation. Higher (↑) results for detection imply a higher detection rate.

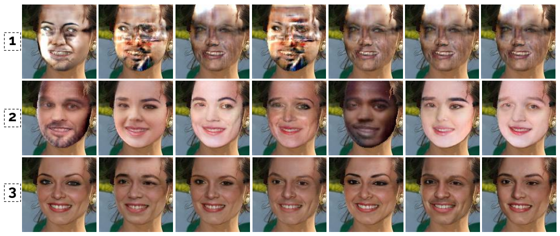

Fig. 4. Qualitative results of diversity on CelebA-MaskHQ dataset. *Top row:* The model trained without decoupling (1). *Middle row:* The model trained without the proxy task (2) produces anonymized faces, but they are not well-blended with the rest of the image. *Bottom row:* In contrast, our model (3) produces realistic anonymized faces that are diverse and blended with the other parts of the image.

Results. As shown in Table 1, our decoupled model reaches best results in Re-ID and detection rate on CelebA. More interestingly, we significantly outperform networks (1) and (2) when working on CelebA-MaskHQ, where there is a domain gap. As we show in the qualitative example in Fig. 4, simultaneously training with two different datasets leads to artifacts in (1) and a lack of blending in (2), which explains the high diversity and decline in quality in Table 1. The separately trained model (3) achieves the most balanced results in terms of quality and diversity across both datasets. It maintains a high diversity while properly blending the output faces. It is also easier to train and can be parallelized due to the separate nature of the training. In supplementary, we provide more qualitative analysis and a discussion of each model.

4.2 Comparison to SOTA

Models	Detection (↑)		Identif. (↓)	Diversity (↑)	
	Dlib	SSH	PNCA	LPIPS	Re-ID
Original	100	100	70.7	-	-
Pixelization 8 by 8	0.0	0.0	0.4	-	-
Pixelization 16 by 16	0.0	0.0	**0.3**	-	-
Blur 9 by 9	90.6	38.6	16.9	-	-
Blur 17 by 17	68.4	0.3	1.9	-	-
CIAGAN	97.8	97.4	1.3	0.032	64.5
Ours	**98.9**	**99.9**	2.2	**0.036**	**96.1**

Table 2. Results of existing detection, recognition and diversity pre-trained methods. Lower (↓) results imply a better anonymization. Upper (↑) results imply a better detection and diversity. Diversity metric is not applicable to classic methods since they can only produce a single output.

Anonymization Anonymization on CelebA. Following [24], we present in Table 2 the detection and identification results of our method compared to the other methods on the CelebA dataset. For detection, we use the classical HOG detector [5] and the deep learning-based SSH [26]. Pixelization methods, while having a higher de-identification rate, generate faces that cannot be detected by either detector, making the anonymized images unusable for computer vision applications. The low-blurring method has both a significantly lower detection rate and a lower de-identification rate. The high-blurring has a slightly better de-identification rate, but it comes at the cost of a very low detection rate, e.g., the SSH detector cannot detect virtually any of the faces. Our main competitor, CIAGAN [24], has a better de-identification rate, 1.3 for them compared to 2.2 for us, but it comes at the cost of generating less realistic images, many of which cannot be detected by the detectors. For example, HOG detector misses 2.2% of the faces generated by CIAGAN compared to 1.1% of the faces generated by our method. Even more extremely, deep-learning based SSH misses 2.6% of the faces generated by CIAGAN, but it misses less than 0.1% of the faces from our method. We qualitatively compare with [24] in Fig. 5.2. It can be seen that the images generated by our method are more realistic. An interesting case is the first image that contains an extreme pose. Our method is able to anonymize it in a realistic manner, while [24] generates an undetectable face.

Anonymization on LWF dataset. We do a similar experiment on the LWF dataset. We follow the standard protocol, where the dataset is divided into 10 different splits, each containing 600 pairs. A pair is defined as positive if both elements share the same identity, otherwise as negative. In every split, the first 300 pairs are positive, and the remaining 300 pairs are negative. Following [7], we anonymize the second image of every pair. We use FaceNet [34] identification model, pre-trained on two public datasets: VGGFace2 [2] and CASIA-Webface [23]. The main evaluation metric is the true acceptance rate, i.e., the ratio of true

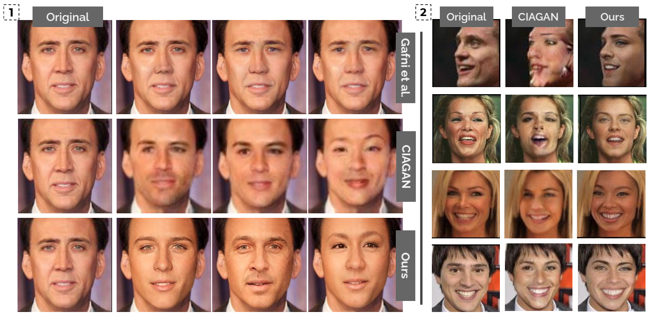

Fig. 5. Left: Qualitative comparisons with [7] and [24]. Right: Qualitative comparisons with [24], where the first two faces are extreme poses.

positives for a maximum 0.001 ratio of false positives. We present our results in Table 3. As shown, we reach state-of-the-art results with the network trained on VGGFace2, outperforming the other two methods [7,24]. When evaluated with the network trained on CASIA-Webface, we outperform [7] but do not reach as good de-identification rate as [24]. However, we also check the detection rate of the methods; after all, it is easy to reach a very good de-identification rate if the generated faces look extremely unnatural and are not detectable. As we can see, 98.4% of the faces generated by our method can be detected by SSH, compared to 95.4% of the faces generated by [24], showing that our method has a 65% lower error in detection rate. We conclude that our method has the best trade-off between the de-identification and the detection rates. We qualitatively compare with [7,24] in Fig. 5.1. We show that the visual quality of the generated images is high for our method and [7], while the images generated by [24] are less realistic. At the same time, we observe that the images generated by our method and [24] are both anonymized and diverse, while the images generated by [7] are easily identifiable. We argue that our method combines the best of both worlds, generating images that are non-identifiable, diverse, and realistic. Unfortunately, the code for [7] has not been released, hence, we cannot compute the detection rate for their method.

De-ID method	VGGFace2 (↓)	CASIA (↓)	Detec. (↑)
Original	0.986 ±0.01	0.965 ±0.02	100
Gafni [7]	0.038 ±0.02	0.035 ±0.01	-
CIAGAN [24]	0.034 ± 0.02	**0.019 ± 0.01**	95.4
Ours	**0.032 ± 0.02**	0.032 ± 0.01	**98.4**

Table 3. Comparisons with SOTA in LWF dataset. Lower (↓) identification rates imply better anonymization. Higher (↑) detection rate implies better generation quality.

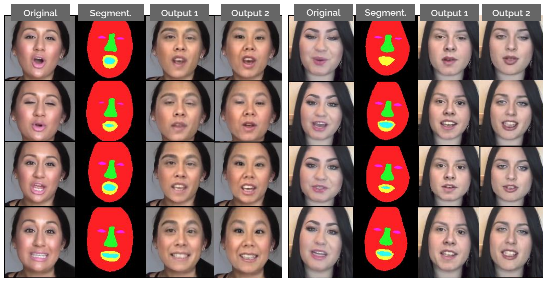

Fig. 6. Results of our model on FaceForensics++ dataset. For each sequence we show input faces, segmentation masks, and two different anonymizations results.

Diversity We test the diversity of our metric by generating 500 anonymized versions of 100 randomly chosen input images. For each image, we randomly sample 500 pairs, where the number of possible pairs is $\frac{500\times499}{500}$, and measure the LPIPS score [40] between all pairs. LPIPS score measures the similarity between two images, the higher it is, the more different the two images are. As we show in Table 2, we reach a 12.5% relative higher score than CIAGAN. This means that the images our method generates look more diverse compared to the ones generated by CIAGAN. Furthermore, for every generated image, we compute the nearest neighbor in the training set (Re-ID). Intuitively, considering that for every generated image, we use 500 identities, a method that generates perfectly diverse images would result in every generated image having a different nearest identity in the training set. In this upper bound case, the number of unique identities for every image would be 500. As we show in Table 2, CIAGAN shows an average of 64.5 identities, while our method shows an average of 96.1 identities, for a 49% relative improvement, showing again that our method generates more diverse images.

Temporal consistency We show a quantitative evaluation of the temporal consistency in Fig. 6. We use FaceForensics++ [32] dataset and we measure tLP, as defined in [4], in addition to visual quality. tLP measures the similarity between all consecutive frames in a video. Intuitively, the better the temporal consistency is, the more similar two consecutive frames are, hence, the lower the tLP metric is. Our method reaches 0.023 tLP, better than CIAGAN which reaches 0.047 score. Additionally, our method reaches a significantly better FID score (14.7 for our method compared to 62.7 for CIAGAN), indicating its higher visual quality. Furthermore, if we finetune HarmonizationNet on FaceForensics++, the temporal consistency improves to 0.016 tLP and FID to 8.3.

Extreme poses We check the detection rate in the challenging AFLW dataset [19], which contains extreme poses. We run a landmark detector to generate the landmarks which are needed for CIAGAN [24]. We remove every face that the

Models	time sec(↓)		
	x128	x256	x512
CIAGAN + SR method	35.89	38.82	37.98
Ours	**10.49**	**8.41**	**11.60**

Table 4. Results on the different resolutions of HQ. Lower (↓) results imply better.

detector cannot find in order to have the same number of generated faces as CIAGAN which needs landmarks. We use the same detector on the anonymized versions of the faces generated by our method and CIAGAN. The detector detects 90.99% of the faces generated by our method, but only 72.58% of the faces generated by CIAGAN, showing that our method is more robust.

Different domain. We train our method on MOTS dataset [38]. We use whole-body segmentation masks and estimated body joints, using OpenPose [3], as an input to AnonymizationNet. We provide qualitative results and additional details in the supplementary.

Super-resolution. In order to show the potential of our two-step framework, we re-train HarmonizationNet to output a higher resolution image compared to the original image. The first step remains the same with AnonymizationNet generating images of size 128 × 128. We train HarmonizationNet separately on CelebA-MaskHQ dataset. It takes an input with dimensions 128 × 128 and outputs images of resolution 256 × 256 and 512 × 512. We compare with CIAGAN, which is trained with 128 × 128 resolution. Due to its nature, we cannot re-train it on a dataset that has no identity information, as is the case for the CelebA-MaskHQ dataset. Therefore, we upscale its output to higher resolution using off-the-shelf super-resolution method [21]. We evaluate the quality of output on different resolution levels using the FID metric. As shown in Table 4, our method achieves significantly better results in all cases. We show qualitative results in Fig. 3 and in the supplementary.

5 Conclusions

The exponential increase in the deployment of cameras in public spaces and its subsequent use in computer vision applications has raised the difficult question of how to deal with data privacy. In this work, we proposed a framework to anonymize faces and bodies based on conditional generative adversarial networks. The key contribution of our approach is to separate two important concepts in face anonymization: generation and blending. We showed the benefits of our decoupled formulation, reaching state-of-the-art results in quality, diversity, and temporal consistency. We also showed the benefits of the training procedure, which can leverage datasets that do not contain identity annotations. Finally, we show how our method can be easily adapted to other tasks like full-body anonymization and can also be used to produce high-resolution images.

Acknowledgements. This research was partially funded by the Humboldt Foundation through the Sofja Kovalevskaja Award.

References

1. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a" siamese" time delay neural network. In: Advances in Neural Information Processing Systems (NIPS). pp. 737–744 (1994)
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: International Conference on Automatic Face & Gesture Recognition. pp. 67–74 (2018)
3. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
4. Chu, M., Xie, Y., Mayer, J., Leal-Taixé, L., Thuerey, N.: Learning temporal coherence via self-supervision for gan-based video generation. ACM Trans. Graph. **39**(4), 75 (2020)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 886–893 (2005)
6. 2018 reform of eu data protection rules. `https://gdpr-info.eu` (2018)
7. Gafni, O., Wolf, L., Taigman, Y.: Live face de-identification in video. In: International Conference on Computer Vision (ICCV) (2019)
8. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Conference on Computer Vision and Pattern Recognition, (CVPR). pp. 2414–2423 (2016)
9. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS). pp. 2672–2680 (2014)
10. Gross, R., Sweeney, L., la Torre, F.D., Baker, S.: Model-based face de-identification. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). p. 161. IEEE Computer Society (2006)
11. Harvey, Adam. LaPlace, J.: Megapixels: Origins, ethics, and privacy implications of publicly available face recognition image datasets (2019)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems NIPS, Long Beach, CA, USA. pp. 6626–6637 (2017)
13. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
14. Hukkelås, H., Mester, R., Lindseth, F.: Deepprivacy: A generative adversarial network for face anonymization. CoRR **abs/1909.04538** (2019)
15. Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5967–5976 (2017)
16. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: International Conference on Learning Representations (ICLR) (2018)
17. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. CoRR **abs/1812.04948** (2018)
18. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8107–8116 (2020)

19. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: International Conference on Computer Vision Workshops (ICCVW). pp. 2144–2151 (2011)
20. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: International Conference on Computer Vision (ICCV). pp. 365–372 (2009)
21. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 105–114 (2017)
22. Liu, M., Tuzel, O.: Coupled generative adversarial networks. In: Advances in Neural Information Processing Systems (NIPS). pp. 469–477 (2016)
23. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: International Conference on Computer Vision (ICCV) (2015)
24. Maximov, M., Elezi, I., Leal-Taixé, L.: CIAGAN: conditional identity anonymization generative adversarial networks. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5446–5455 (2020)
25. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: International Conference on Computer Vision (ICCV). pp. 360–368 (2017)
26. Najibi, M., Samangouei, P., Chellappa, R., Davis, L.S.: SSH: single stage headless face detector. In: International Conference on Computer Vision (ICCV). pp. 4885–4894 (2017)
27. Park, T., Liu, M., Wang, T., Zhu, J.: Semantic image synthesis with spatially-adaptive normalization. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2337–2346 (2019)
28. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: International Conference on Learning Representations (ICLR) (2016)
29. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: European Conference on Computer Vision (ECCV). pp. 639–655 (2018)
30. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: European Conference on Computer Vision (ECCV) (2018)
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention - MICCAI. pp. 234–241 (2015)
32. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: International Conference on Computer Vision (ICCV). pp. 1–11 (2019)
33. Ryoo, M.S., Kim, K., Yang, H.J.: Extreme low resolution activity recognition with multi-siamese embedding learning. In: Conference on Artificial Intelligence (AAAI). pp. 7315–7322 (2018)
34. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 815–823 (2015)
35. Sun, Q., Ma, L., Oh, S.J., Gool, L.V., Schiele, B., Fritz, M.: Natural and effective obfuscation by head inpainting. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5050–5059 (2018)

36. Sun, Q., Tewari, A., Xu, W., Fritz, M., Theobalt, C., Schiele, B.: A hybrid model for identity obfuscation by face replacement. In: European Conference on Computer Vision (ECCV). pp. 570–586 (2018)
37. Taherkhani, F., Nasrabadi, N.M., Dawson, J.M.: A deep face identification network enhanced by facial attributes prediction. In: Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 553–560 (2018)
38. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: MOTS: multi-object tracking and segmentation. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7942–7951 (2019)
39. Wang, T., Liu, M., Zhu, J., Yakovenko, N., Tao, A., Kautz, J., Catanzaro, B.: Video-to-video synthesis. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 1152–1164 (2018)
40. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)