# Graph-based Generative Face Anonymisation with Pose Preservation

Nicola Dall'Asen[1,2], Yiming Wang[3[0000−0002−5932−4371]], Hao Tang[4], Luca Zanella[3], and Elisa Ricci[2,3[0000−0002−0228−1147]]

[1] University of Pisa, Pisa, Italy
nicola.dallasen@phd.unipi.it
[2] University of Trento, Trento, Italy
[3] Fondazione Bruno Kessler, Trento, Italy
[4] ETH Zürich, Zürich, Switzerland

**Abstract.** We propose AnonyGAN, a GAN-based solution for face anonymisation which replaces the visual information corresponding to a source identity with a condition identity provided as any single image. With the goal to maintain the geometric attributes of the source face, i.e., the facial pose and expression, and to promote more natural face generation, we propose to exploit a Bipartite Graph to explicitly model the relations between the facial landmarks of the source identity and the ones of the condition identity through a deep model. We further propose a landmark attention model to relax the manual selection of facial landmarks, allowing the network to weight the landmarks for the best visual naturalness and pose preservation. Finally, to facilitate the appearance learning, we propose a hybrid training strategy to address the challenge caused by the lack of direct pixel-level supervision. We evaluate our method and its variants on two public datasets, CelebA and LFW, in terms of visual naturalness, facial pose preservation and of its impacts on face detection and re-identification. We prove that AnonyGAN significantly outperforms the state-of-the-art methods in terms of visual naturalness, face detection and pose preservation[5].

## 1 Introduction

In the era of deep learning, the availability of large scale data has undoubtedly brought technological advances. However, the very same fact has also fostered the growing concern regarding privacy issues. Visual privacy preservation is mostly achieved via video redaction methods by obfuscating the personally identifiable information (PII) of a data subject, whose face is often the most identity-informative part. Classic face anonymisation techniques, e.g., blurring [5] or pixelation [8], can effectively remove PII. However, this comes at a high cost of degrading other vision-related tasks, in particular for the action/emotion recognition where the poses play an essential role.

Thanks to recent advances in generative adversarial networks (GANs) [9], several anonymisation solutions have been proposed to generate *natural-looking*

[5] Code and pretrained model are available at https://github.com/Fodark/anonygan

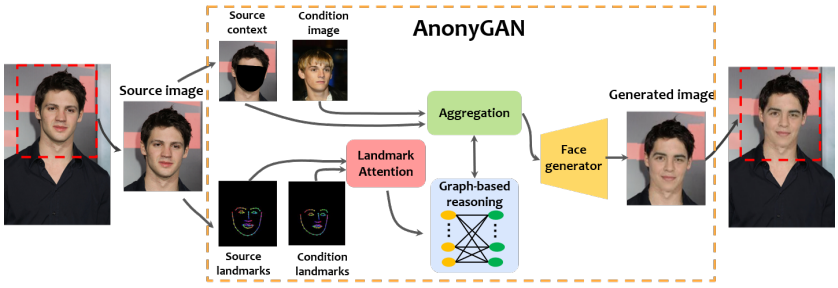


**Fig. 1. AnonyGAN** anonymises faces by generating visually similar faces to *any* condition image while preserving the facial pose of the source image, with a novel *landmark attention* model and a face generator with *graph-based landmark-landmark reasoning*.

faces that correspond to different identities [27,7,4,19], while preserving the original facial poses using the landmarks as the guidance [28,23,13]. Yet, it is challenging to produce realistic images in this context due to the lack of ground-truth images in real-world, i.e., images of different persons with the same facial pose. This also draws a fundamental distinction to the pose-guided image generation task [10,6,29], whose ground-truth images, i.e. images of the same person with varying poses, are available to provide direct pixel-level supervision. While for the pose preservation, the main challenge lies in the relation reasoning between the condition pose and the source pose, where graph-based modelling has demonstrated its strengths in such geometric reasoning [29]. Finally, the facial landmarks choice can impact the visual naturalness, pose preservation and face anonymisation, which is often heuristically handled with no optimality guaranteed [23,13].

To address the above-mentioned challenges, we propose a novel graph-based GAN architecture, **AnonyGAN**, to perform landmark-guided face anonymisation (see Figure 1). Our network takes as input a *context* image and a condition image, as well as the facial landmarks extracted from the source and target images. The *context* image is the source image with the face (excluding the forehead) masked out, and it provides the necessary contextual information, e.g. the skin tone and the background, for naturally blending the generated faces to the source image. In order to improve the pose preservation, we propose to first disentangle the geometrical reasoning from the appearance, where the source landmarks and the condition landmarks are modelled as a bipartite graph with Graph Convolution Networks (GCNs). The appearance of the condition image is then aggregated to the pose reasoning module to generate more natural faces with the source facial pose preserved. Moreover, we introduce a novel landmark attention model to allow the network to automatically learn the importance of the facial landmarks, avoiding sub-optimal manual decisions. Finally, we propose a novel hybrid training strategy to address the training difficulty caused by the lack of ground-truth images. We form the source and condition pairs using both the same image and different images to facilitate the appearance learning. The former provides direct pixel-level supervision, while the latter applies a weak

context-level supervision by exploiting the high-level features extracted from the appearance discriminator. We validate **AnonyGAN** on two public datasets, i.e., CelebA [22] and LFW [12], and demonstrate that **AnonyGAN** can greatly improve the generated faces in terms of visual naturalness and pose preservation compared to baselines, i.e., blurring and pixelation, and the state-of-the-art method [23]. To summarise, our main contributions are listed below:

- We propose a novel GAN-based architecture, **AnonyGAN**, to perform landmark-guided face anonymisation, achieving the *highest visual quality* with the *best pose preservation* on two benchmark dataset.
- We propose to exploit a graph formulation on the landmarks of the source and condition face images to perform geometric reasoning using GCNs, and prove its effectiveness in improving the pose preservation.
- We propose a landmark attention model to automatically weight the facial landmarks, achieving a higher visual quality and perception performance.
- We propose a hybrid training strategy with a strong pixel-level and a weak context-level supervision to address the training without ground-truth images, achieving the best visual quality.

## 2 Related Work

We discuss recent face obfuscation techniques for anonymising visual data and briefly cover related works addressing pose-guided image generation, a related yet different task.

**Visual Anonymisation** often refers to irreversible obfuscation techniques for removing PII of the data subject in visual content, a.k.a. de-identification in some works [7]. Many works anonymise visual data by obfuscating the faces, the most privacy-concerning content, using classic techniques, such as blurring via filters [5] or pixelation by enlarging the pixels [8]. Such methods remove the most identifiable visual information, but greatly compromise other perception tasks, such as object detection and action recognition. Recently, with the progresses in GANs, face anonymisation techniques have advanced by generating realistic faces of a different identity, leaving intact most of the non-identifiable visual and geometrical attributes [13,7,27,28,23]. Sun *et al.* [27] proposed a two-step face inpainting technique for anonymisation by first generating the 68 facial landmarks, and then synthesising the faces guided by the landmarks. With a blurred face as condition, the generated faces have a rather high visual quality, yet resemble to the original face. DeepPrivacy [13] exploits the generator of StyleGAN [14] to generate a face of a fake identity, conditioned both on the context image with the face masked out and on 7 facial landmarks. The generated faces are limited in anonymisation and pose preservation. Conditional GANs are also proposed to explicitly control the identity on the generated faces [28,23]. Recently, CIAGAN [23] has introduced an identity discriminator to enforce the generated faces to be different from the source image, achieving a better anonymisation performance, while the visual naturalness and pose preservation are not yet satisfactory although the subset of landmarks are carefully chosen. Moreover, CIAGAN cannot

be easily applied to unknown condition identity, as the condition identities are encoded within the network during the training.

**AnonyGAN** takes any condition image as a reference for appearance, and improves the visual naturalness and facial pose preservation by first reasoning on landmark to landmark relations, and then generating the image aggregating the appearance features. Moreover, the choice of facial landmarks impacts both the image quality and the perception task with a certain trade-off. This aspect has not been properly addressed in the literature yet [23]. Instead, in this work we propose a landmark attention model to allow the network to learn the relative importance among the landmarks for face anonymisation.

**Pose-Guided Image Generation** is a related task to visual anonymisation where the main challenge is due to the pose deformation between the source and the target image. Modelling the relations between the source pose and the target pose is the key to solve this challenging task. Existing methods such as [2,26,30,1,32,3,20,21] are based on the stacking of several convolutional layers, which can only leverage the relations between the source pose and the target pose locally. For instance, Zhu *et al.* [32] proposed a Pose-Attentional Transfer Block, in which the source and target poses are simply concatenated and then fed into an encoder to capture their dependencies. Different from existing methods for modelling the relations between the source and target poses in a localised manner, BiGraphGAN [29] reasons and models the crossing long-range relations between different features of the source pose and target pose in a bipartite graph, which are then used for the person image generation process. In general, for pose-guided image generation, the ground-truth images of the same person with various poses are available. In contrast, there is no set of images of different identities with the exact same pose, thus making the face anonymisation a more challenging problem due to the lack of direct pixel-level supervision.

## 3 Proposed Method

The proposed **AnonyGAN**, as illustrated in Figure 2, aims to anonymise a source image of any identity $I_s$ by replacing the face with any condition identity provided as a condition image $I_c$ while preserving the facial pose of source image $Lm_s$. Both the source image $I_s$ and the condition image $I_c$ are preprocessed to extract their facial landmarks $Lm_s$ and $Lm_c$, respectively, for the graph-based landmark-landmark reasoning. Both facial poses are defined by 68 landmarks and encoded as 68-channel images as in [24] with a channel per landmark. In order to generate the face with a consistent context of the source image, in particular, for the skin tone, we also input to the network the context of the source image $I_{sc}$, i.e., the background image with the face masked out by the contour defined by the facial landmarks.

The main components of our network are: a *Landmark Attention Model* that takes as input $Lm_s$ and $Lm_c$ defined by the full 68 landmarks, and learns to weight the landmarks to optimally trade-off between the face naturalness and re-identification performance; a *Landmark2Landmark Face Generator* that follows a similar architecture of the generator of BiGraphGAN [29]. We use GCNs

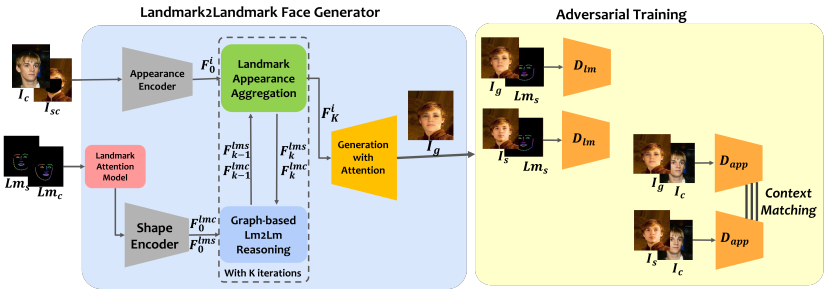


**Fig. 2.** Network architecture of **AnonyGAN**. The condition image $I_c$ and source context image $I_{sc}$ are encoded with an appearance encoder, while the source and condition landmarks $Lm_s$ and $Lm_c$ are encoded by a shape encoder. The shape codes $F_0^{lmc}$ and $F_0^{lms}$ are passed to the graph-based landmark reasoning model, and iteratively aggregated with the appearance code for K iterations. The final appearance code $F_K^i$ is used for the face generation. The adversarial training is performed with two discriminators: The Appearance Discriminator $D_{app}$ optimises for naturally blending the facial attributes of the target into the source image, and the Landmark Discriminator $D_{lm}$ optimises for preserving the source pose.

to learn the spatial relations between $Lm_s$ and $Lm_c$, aggregated with the visual features extracted from the condition image $I_c$. With several iterations of the landmark appearance aggregation, the final image $I_g$ is generated with an attention model. For *adversarial training*, we adopt a *Landmark Discriminator* $\mathbf{D}_{lm}$ and an *Appearance Discriminator* $\mathbf{D}_{app}$. $\mathbf{D}_{lm}$ is designed to preserve the facial pose of the source image, while $\mathbf{D}_{app}$ is designed to preserve the appearance of the condition face with the skin tone matched to the source face. We will provide more details on the *Landmark Attention Model*, *Landmark2Landmark Face Generator* and *Adversarial Training* in the following sections.

### 3.1 Landmark Attention Model

The landmark attention model is designed to learn the optimal weighting strategy on the landmarks to achieve jointly the best visual naturalness and pose preservation. We first concatenate the 68-channel landmark maps of both the source facial pose and the condition facial pose, and feed it to an Efficient Channel Attention module [31] formulated as:

$$\omega = \sigma(Conv1D_J(GAP(Concat(Lm_s, Lm_c)))), \tag{1}$$

where $\sigma(\cdot)$ is the Sigmoid function, $Conv1D_J(\cdot)$ is 1-D convolution with kernel size $J$, $GAP(\cdot)$ represents the operation of the channel-wise Global Average Pooling, and $Concat(\cdot)$ is the concatenation operation.

### 3.2 Landmark2Landmark Face Generator

The landmark2landmark face generator follows the architecture of [29] that iteratively reasons the relations between the source landmarks $Lm_s$ and the condition

landmarks $Lm_c$ following a bipartite graph formulation, and aggregates with the appearance feature of the condition image $I_c$ and the context of the source image $I_{sc}$. The final aggregated feature is used for the landmark-guided face generation with the condition identity.

The condition image $I_c$ and the source context image $I_{sc}$ are first concatenated and fed to an appearance encoder to generate the initial appearance code $F_0^i$, while the $Lm_s$ and $Lm_c$ after the Landmark Attention Model are passed to a shape encoder to obtain the shape codes $F_0^{lms}$ and $F_0^{lmc}$, respectively. The shape codes are then fed to a graph-based landmark-to-landmark reasoning model in a bipartite graph via GCNs to update the shape codes, which are then fed to the Landmark Appearance Aggregation model to synchronise the updates in both appearance and shape codes. Such operation of landmark-to-landmark reasoning and appearance aggregation is performed *iteratively* to form a thorough reasoning from low to high level.

At the $K$-th iteration of the graph-based landmark-landmark reasoning and appearance aggregation, the final appearance code $F_K^i$ is passed to both an image decoder to generate the intermediate result $\tilde{I}_g$, and an attention decoder to produce the attention mask $A_i$, a one-channel attention mask with the pixel value between 0 to 1. The final generated image is obtained by $I_g = I_c \otimes A_i + \tilde{I}_g \otimes (1 - A_i)$, where $\otimes$ denotes element-wise product.

### 3.3 Adversarial Training

Two discriminators are designed for the adversarial training. Specifically, the *Landmark Discriminator* $\mathbf{D}_{lm}$ is fed with the landmark-image pairs of the source $\{Lm_s, I_s\}$ and the generated $\{Lm_s, I_g\}$ to encourage the generation of similar facial pose of the source image. The *Appearance Discriminator* $\mathbf{D}_{app}$ takes the source-condition image pair $\{I_s, I_c\}$ and the generated-condition pair $\{I_g, I_c\}$ to guide the face generation with the facial attributes of the condition face within the same context of the source image by coupling $I_s$ and $I_g$ with $I_c$.

In order to facilitate the appearance learning, we train the Appearance Discriminator in a hybrid manner with the source-condition image pairs composed by either the same image or a couple of images with different poses and contexts. When the source and condition images are the same, i.e. $I_s = I_c$, we apply a L1 loss on $I_g$ and $I_s$ to provide the strong pixel-level supervision. When the source and condition images are different, we apply a weak supervision for *Context Matching* by enforcing similar high-level features of $I_g$ and $I_s$ that correspond to skin tone and background.

We employ several losses to drive the network learning. We train $D_{app}$ with the adversarial loss $L_{app}$:

$$L_{app} = \min_G \max_{D_{app}} \mathbb{E}[log(\mathbf{D}_{app}(I_s, I_c))] + \mathbb{E}[1 - log(\mathbf{D}_{app}(I_g, I_c))]. \quad (2)$$

The landmark discriminator $D_{lm}$ is trained with $L_{lm}$, driving the generator towards the correct pose:

$$L_{lm} = \min_G \max_{D_{lm}} \mathbb{E}[log(\mathbf{D}_{lm}(I_s, Lm_s))] + \mathbb{E}[1 - log(\mathbf{D}_{lm}(I_g, Lm_s))]. \quad (3)$$

Moreover, when the source and condition pair is of different images, the *Context Matching* supervision is realised through the weak Feature Matching (wFM) loss [4] defined as:

$$L_{wFM}(D_{app}) = \sum_{i=m}^{M} \frac{1}{N_i} ||D_{app}^{(i)}(I_g, I_c) - D_{app}^{(i)}(I_s, I_c)||_1, \tag{4}$$

where $D_{app}^{(i)}$ denotes the feature map produced by the $i$-th layer of the discriminator $D_{app}$; $N_i$ is the number of elements in the feature map produced by the $i$-th layer; $m$ is the first layer from which the weak feature matching loss computation starts; and $M$ is the total number of layers of the discriminator $D_{app}$.

When the condition and source pair is composed of the same image, we provide the generator with the direct pixel-level supervision using the image reconstruction loss $L_{Recon} = ||I_g - I_s||_1$.

The *final loss* function can be expressed as:

$$L = \lambda_{app} L_{app} + \lambda_{lm} L_{lm} + \lambda_{wFM} L_{wFM} + \lambda_{Recon} L_{Recon}, \tag{5}$$

where the weighting parameters are empirically set.

## 4 Experiments

We evaluate our proposed method **AnonyGAN** and its variants in comparison with state-of-the-art methods using two public face datasets. We validate the effectiveness of our design choices by evaluating the generated faces in terms of the visual naturalness, pose preservation, face detection and anonymisation.

**Datasets.** Our model is evaluated on two benchmark datasets created for face-related computer vision tasks, i.e., CelebA [22] and Labeled Faces in the Wild (LFW) [12,18]. **CelebA** is a large-scale dataset of 202,599 images with different poses and backgrounds of 10,177 celebrities. **LFW** is composed of 13,233 face images collected from the web covering 5,749 identities with 1,680 people having two or more images. We use CelebA for both training and testing, while LFW is used for testing only. We follow the same train/test split of CelebA as in [23], where in total 1,563 identities with more than 30 images per identity are selected. The training set is composed of 1,200 identities, where 24,000 pairs are formed for the training. We use the images of the remaining 363 identities for testing, where each source image is paired with a condition image that is randomly selected from the images of the next identity as in [23]. For the test set of LFW, we follow the protocol defined in [7], where we form 6,000 pairs of images organised in 10 different folds with each folder containing half of the pairs corresponding to the same person, and the other half corresponding to different persons. We use Dlib [16] to preprocess the 68 facial landmarks, which are then used to define the mask area of each face.

**Evaluation Metrics.** We evaluate the performance of our model in terms of visual quality, pose preservation, face detection, and face re-identification. The **visual quality** of generated faces is measured by the Fréchet inception distance

(FID) [11], which calculates the distance between real and synthetic images in a feature space given by a specific layer of Inception Net. The lower the FID score is, the higher the quality of generated images[6]. The **pose preservation** is measured by the $L1$ distance between the detected 68 landmarks and the ground truth landmarks, normalised by the inter-ocular distance [15]. The model should generate faces that minimally impact the visual detection task. We evaluate the **face detection** performance on the generated images using two face detection algorithms, i.e., Dlib [17] and FaceNet [25]. A higher detection rate is more desired. Meanwhile, the generated faces should maximally prohibit the **face re-identification** for the best anonymisation. We report the rate of the correct matches of the generated faces and source faces using FaceNet without fine-tuning. In particular, for the test of LFW, we compute for each fold the re-identification rate, and report the mean and the standard deviation among all folds. A lower re-identification rate indicates a better face anonymisation.

**Implementation Details.** We set the learning rate for the generator during training to $2 * 10^{-4}$ and the learning rate for discriminators to $2 * 10^{-6}$. We set the weights in final loss in Eq. (5), i.e. $\lambda_{app}$, $\lambda_{lm}$, $\lambda_{CM}$ and $\lambda_{Recon}$ as 1, 1, 1 and 5, respectively. We perform hybrid training using the condition and source image pairs with the same images and different images to facilitate the appearance learning. During training, we we set 75% of pairs with $I_c = I_s$ for each batch.

**Method Comparisons.** We compare **AnonyGAN** against a set of baselines and state-of-the-art GAN-based methods that are closely related to the face obfuscation task: 1) **Blurring** applies a $101 \times 101$ Gaussian kernel to blur the face region. 2) **Pixelation** uses a $10 \times 10$ pixelation mask on the face region. 3) **CIAGAN** [23] is a state-of-the-art face anonymisation method based on conditional GAN, which accepts as input the context image and 29 facial landmarks of the source image and a conditioning ID, and generates a face matching both the context and the conditioning ID. We use the inference model provided by the authors for the evaluation.

Moreover, in order to demonstrate the validity of our design choices and, in particular, the importance of our Landmark Attention (LA) and Context Matching (CM), we ablate several variants of **AnonyGAN**: 1) **AnonyGAN**-$(\mathbf{CM}, \mathbf{LA})^-$ (68 lm) is trained without Landmark Attention and Context Matching, with the full 68 facial landmarks. 2) **AnonyGAN**-$(\mathbf{CM}, \mathbf{LA})^-$ (29 lm) is trained without Landmark Attention (LA) and Context Matching (CM), but with 29 facial landmarks that are manually selected as in CIAGAN [23], in order to show the impact of the landmark choice. 3) **AnonyGAN**-$(\mathbf{CM})^-$ (68 lm) is trained with Landmark Attention but without the Context Matching, with 68 facial landmarks, to prove the capability of the Landmark Attention module on automatically weighting the landmarks. Finally, **AnonyGAN**(68 lm) is trained with both Landmark Attention and Context Matching, with all 68 landmarks, to justify the capability of Context Matching for improving visual fidelity.

**Result Discussion.** Table 1 presents the results of all methods evaluated on the test set of CelebA. Classic face obfuscation techniques, i.e. blurring and pix-

[6] FID Implementation is taken from: `https://github.com/mseitzer/pytorch-fid`

| | $FID\downarrow$ | Face detection ↑ | | Face re-identification ↓ | | Pose↓ |
|---|---|---|---|---|---|---|
| | | dlib | Facenet | Casia | VGG | |
| Blurring | 95.13 | 4% | 4% | **0.07**% | **0.02**% | - |
| Pixelation | 59.82 | 1% | 28% | 0.28% | 0.12% | - |
| CIAGAN [23] (recomp.) | 37.94 | 96% | **100**% | 1.61% | 0.51% | 1.44 |
| **AnonyGAN**-$(\mathbf{CM}, \mathbf{LA})^-$ (68 lm) | 43.99 | **100**% | **100**% | 2.63% | 0.58% | **0.16** |
| **AnonyGAN**-$(\mathbf{CM}, \mathbf{LA})^-$ (29 lm) | 30.24 | **100**% | **100**% | 2.84% | 0.66% | **0.16** |
| **AnonyGAN-CM**$^-$ (68 lm) | 26.12 | **100**% | **100**% | 2.70% | 0.91% | **0.16** |
| **AnonyGAN**(68 lm) | **22.53** | **100%** | **100%** | 3.52% | 1.60% | **0.16** |

**Table 1.** Performance of the proposed **AnonyGAN** and its variants evaluated on the test set of CelebA in comparison to the baselines and state-of-the-art GAN-based method.

elation, reduce greatly the visual quality, thus hampering face detection. On the other hand, they achieve the lowest re-identification rate, thus the best anonymisation performance. The pose preservation metric is not available as the landmarks are not detectable from the anonymised faces. The visual quality of the images generated by CIAGAN [23] is in general inferior to our **AnonyGAN** models as indicated by the FID metric. The unnaturalness of the generated faces leads to a lower detection rate with the standard face detector Dlib, but also confuses the face re-identification module, leading to a better anonymisation performance. Moreover, our **AnonyGAN** models significantly improve the pose preservation performance thanks to the graph-based geometric reasoning among the source and condition landmarks.

Interestingly, by choosing a subset of the landmarks (29 out of 68 landmarks) carefully as in CIAGAN, we observe that **AnonyGAN**-$(\mathbf{CM}, \mathbf{LA})^-$ (29 lm) improves the visual quality without impacting the pose preservation, compared to **AnonyGAN**-$(\mathbf{CM}, \mathbf{LA})^-$ (68 lm). Moreover, the result of **AnonyGAN-CM**$^-$ (68 lm) in comparison to **AnonyGAN**-$(\mathbf{CM}, \mathbf{LA})^-$ (29 lm) shows that the introduction of Landmark Attention enables the network to learn the importance of landmarks automatically, achieving a better visual quality without impacting the pose preservation. Finally, **AnonyGAN** with both Landmark Attention and Context Matching achieves the best results in terms of face visual quality, face detection and pose preservation, however with a slight compromise in the anonymisation performance compared to the variants and CIAGAN. These above-mentioned observations are also in line with the results evaluated on the LFW dataset as reported in Table 2.

Figure 3 shows the faces generated by **AnonyGAN** and the compared method CIAGAN [23], with the condition (the 1st column) and source (the 2nd column) images. For the condition identities, we use those that CIAGAN has been trained with for a fair comparison. We can observe that **AnonyGAN** generates more natural looking images with the facial attributes of the condition face better transferred to the context of the source image. The pose of the source face is better preserved, confirming the effectiveness of the proposed Landmark Attention module that allows for automatic weighting on the full 68 landmarks.

| | $FID \downarrow$ | Face detection ↑ | | Face re-identification ↓ | | Pose↓ |
|---|---|---|---|---|---|---|
| | | dlib | Facenet | Casia | VGG | |
| Blurring | 85.03 | 7% | 17% | **0.03% ± 0.07** | **0.02% ± 0.05** | - |
| Pixelation | 47.97 | 6% | 22% | **0.03% ± 0.07** | **0.02% ± 0.05** | - |
| CIAGAN [23] (recomp.) | 18.37 | 99% | **100%** | 1.28% ± 0.32 | 0.08% ± 0.12 | 0.45 |
| **AnonyGAN-(CM, LA)**$^-$ (68 lm) | 44.41 | 99% | **100%** | 3.93% ± 0.67 | 0.38% ± 0.19 | 0.11 |
| **AnonyGAN-(CM, LA)**$^-$ (29 lm) | 19.42 | 97% | **100%** | 4.18% ± 0.45 | 0.35% ± 0.31 | 0.11 |
| **AnonyGAN-(CM)**$^-$ (68 lm) | 14.66 | **100%** | **100%** | 3.35% ± 0.78 | 0.32% ± 0.27 | 0.06 |
| **AnonyGAN**(68 lm) | **8.93** | **100%** | **100%** | 3.55% ± 0.94 | 0.48% ± 0.29 | **0.05** |

**Table 2.** Performance of the proposed **AnonyGAN** and its variants evaluated on the test set of LFW in comparison to the baselines and state-of-the-art GAN-based method.

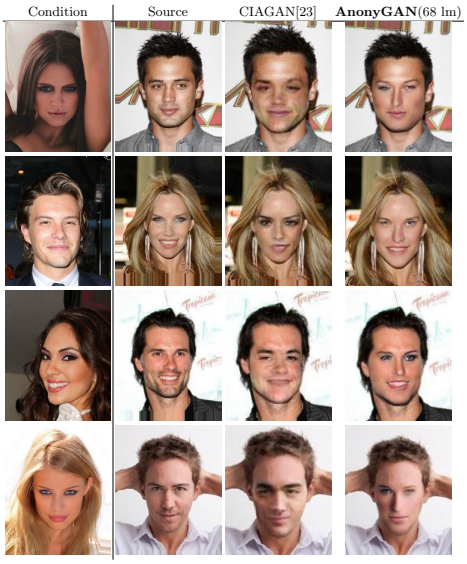


**Fig. 3.** Qualitative results of SOTA and proposed **AnonyGAN** and variants with images from CelebA.

## 5 Conclusions

In this paper, we proposed **AnonyGAN**, a GAN-based solution for face anonymisation, generating faces with the appearance of any condition image and the facial pose of the source image. Our approach leverages landmark-to-landmark geometric reasoning via GCNs to model the relations between the condition and source facial landmarks. We also introduced a landmark attention model to automatically learn the importance of the facial landmarks. We compared our approach with the state-of-the-art approaches both quantitatively and qualitatively, demonstrating a better performance both in term of visual quality and pose preservation. As future work, we will explore alternatives to further improve the face anonymisation performance and adapt the model to operate on real-world video streams.

## References


1. AlBahar, B., Huang, J.B.: Guided image-to-image translation with bi-directional feature transformation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9016–9025 (2019)
2. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8340–8348 (2018)
3. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody dance now. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5933–5942 (2019)
4. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap. Proceedings of the ACM International Conference on Multimedia (Oct 2020)
5. Du, L., Zhang, W., Fu, H., Ren, W., Zhang, X.: An efficient privacy protection scheme for data security in video surveillance. Journal of Visual Communication and Image Representation **59**, 347–362 (2019)
6. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8857–8866 (2018)
7. Gafni, O., Wolf, L., Taigman, Y.: Live face de-identification in video. In: Proceedings in IEEE/CVF International Conference on Computer Vision. pp. 9377–9386 (2019)
8. Gerstner, T., DeCarlo, D., Alexa, M., Finkelstein, A., Gingold, Y., Nealen, A.: Pixelated image abstraction with integrated user constraints. Computers & Graphics **37**(5), 333–347 (2013)
9. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Proceedings of Conference on Neural Information Processing Systems (2014)
10. Grigorev, A., Sevastopolsky, A., Vakhitov, A., Lempitsky, V.: Coordinate-based texture inpainting for pose-guided human image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12135–12144 (2019)
11. 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. arXiv preprint arXiv:1706.08500 (2017)
12. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In: Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie, Marseille, France (Oct 2008)
13. Hukkelås, H., Mester, R., Lindseth, F.: Deepprivacy: A generative adversarial network for face anonymization. In: Advances in Visual Computing. pp. 565–578. Springer International Publishing, Cham (2019)
14. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings in IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4396–4405 (2019)
15. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 1867–1874 (2014)
16. King, D.E.: Dlib-ml: A machine learning toolkit. Journal of Machine Learning Research **10**, 1755–1758 (2009)

17. King, D.E.: Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research **10**, 1755–1758 (2009)
18. Learned-Miller, G.B.H.E.: Labeled faces in the wild: Updates and new reporting procedures. Tech. Rep. UM-CS-2014-003, University of Massachusetts, Amherst (May 2014)
19. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
20. Liang, D., Wang, R., Tian, X., Zou, C.: Pcgan: Partition-controlled human image generation. In: AAAI (2019)
21. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: ICCV (2019)
22. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of International Conference on Computer Vision (ICCV) (December 2015)
23. Maximov, M., Elezi, I., Leal-Taixé, L.: Ciagan: Conditional identity anonymization generative adversarial networks. In: Proceedings in IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5446–5455 (2020)
24. Sagonas, C., Antonakos, E., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: database and results. Image and Vision Computing **47**, 3–18 (2016), 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge
25. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
26. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable gans for pose-based human image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)
27. Sun, Q., Ma, L., Oh, S.J., Van Gool, L., Schiele, B., Fritz, M.: Natural and effective obfuscation by head inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5050–5059 (2018)
28. Sun, Q., Tewari, A., Xu, W., Fritz, M., Theobalt, C., Schiele, B.: A hybrid model for identity obfuscation by face replacement. In: Proceedings of the European Conference on Computer Vision. pp. 553–569 (2018)
29. Tang, H., Bai, S., Torr, P.H., Sebe, N.: Bipartite graph reasoning gans for person image generation. In: Proceedings of the British Machine Vision Conference (2020)
30. Tang, H., Xu, D., Liu, G., Wang, W., Sebe, N., Yan, Y.: Cycle in cycle generative adversarial networks for keypoint-guided image generation. In: Proceedings of ACM Multimedia (2019)
31. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
32. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)