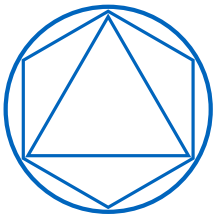


Technical University of Munich

Department of Mathematics

Face Shape Segmentation Mask Anonymization using a Generative Adversarial Network

Master's Thesis

von

Filippo Simonazzi

Supervisor:	Prof. Dr. Laura Leal-Taixé
Advisor:	Maxim Maximov
Submission Date:	07.07.2022

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Munich, 07.07.2022

German Abstract

Die folgende Masterarbeit basiert auf der früheren Arbeit von M. Maximov (Thesis Supervisor), I. Elezi und L. Leal-Taixé (Thesis Advisor) in *CIAGAN: Conditional Identity Anonymization Generative Adversarial Network* [1].

Das Ziel dieser Masterarbeit ist es, eine neue Pipeline für die Anonymisierung von Gesichtern zu definieren, dabei die gleiche visuelle Qualität wie die modernsten Modelle zu erreichen und gleichzeitig eine bessere Kontrolle über den Anonymisierungsprozess zu erhalten.

Der Hauptbeitrag dieser Arbeit ist ein Modell zur Formanonymisierung, das auf einem generativen adversen Netzwerk basiert, welches auf einer Gesichtssegmentierungsmaske und einem Gitternetz basiert. Die Verwendung von Segmentierungsmasken ermöglicht eine bessere Kontrolle über die einzelnen Gesichtskomponenten, einschließlich der Haare und des Halses, die andernfalls gnoriert werden. Die Pipeline erzeugt neue Bilder, die bestimmte Komponenten des ursprünglichen Gesichts enthalten, aber nicht als die ursprüngliche Identität identifizierbar sind.

English Abstract

The following master thesis is based on the previous work conducted by M. Maximov (Thesis Supervisor), I. Elezi and L. Leal-Taixé (Thesis Advisor) in *CIAGAN: Conditional Identity Anonymization Generative Adversarial Network* [1].

The goal of this master thesis is to define a new pipeline for face obfuscation, aiming to achieve the same visual quality as state-of-the-art models while gaining more control over the anonymization process.

The main contribution of the thesis is the shape anonymization model, based on a generative adversarial network conditioned on face segmentation mask and face mesh. The use of segmentation masks allows greater control over the individual face components, including hair and neck, otherwise ignored by face landmarks. The pipeline generates new images retaining specific components of the original face while not being identifiable as the original identity.

Contents

1 Introduction

Privacy Concerns and Face Anonymization
Computer vision technology is enabling automatic understanding of large-scale visual data, becoming a crucial component of many societal applications due to the exponential growth in the number of cameras. For instance, cities are adopting networked camera systems for policing and intelligent resource allocation, individuals are recording their lives using wearable devices, and service robots at homes and public places are becoming increasingly popular. Each of this application contains personal data and while we are all eager to take advantage of this quickly improving technology, be it video conferencing or surveillance, we should not be willing to do so by giving away our personal privacy. On the one hand, we want camera systems to recognize important events and assist our daily lives by understanding the videos they record, but on the other hand we want to ensure that they do not intrude on people's privacy.

In fact, data privacy is an increasing concern, and the European Union has already passed laws such as the General Data Protection Regulations (GDPR) to guarantee this privacy. GDPR requires consent from the individual for any use of their personal data. While being necessary to protect privacy, such regulation may represent an obstacle to the development of new technology that requires large amount of data containing identity information. Data anonymization represents a solution to these problems since most applications do not actually need to *identify* the subjects that appear on videos or pictures, they just need to be able to *detect* them.

The conventional hand-crafted face anonymization methods commonly try to obfuscate the original identity using masks, blurring or pixelation. Unfortunately, most of these methods lead to images where it is difficult, if not impossible, to detect and track the obfuscated faces. Consequently, robust machine learning models that hide the original identity by generating a new face that seamlessly blends into the existing scene are required to effectively achieve anonymization. In particular, M. Maximov, I. Elezi and L. Leal-Taixé in their work *CIAGAN: Conditional Identity Anonymization Generative Adversarial Network* [1] proposed five criteria that any anonymization network should satisfy:

- Anonymization: the produced output must hide the identity of the person in the original image.
- Control: the network should allow for control over the mapping between the real and the fake identity.
- New identities: the generated images must contain new identities not present in the training set.

- Realistic: output images must look realistic in order to be used by state-of-the-art detection and recognition systems.

- Temporal consistency and pose preservation: both properties must be ensured in video anonymization for tasks like people tracking or action recognition.

Aim of this work

The goal of the proposed work is to design a two-stage face anonymization pipeline consisting of a shape anonymization network and a face generator model, in order to achieve the same visual quality as state-of-the-art architectures while gaining more control over the anonymization process. The shape anonymization network introduces a novel layer of control as it generates segmentation masks in which specific facial components match the original ones, while the shape of the other components is modified. Additionally, the presented anonymization pipeline ensures control over the appearance of the obfuscated face as it incorporates the identity guidance network proposed in CIAGAN [1].

We decided to condition this shape anonymization network on segmentation masks because they allow to modify hair and neck, which are not included in state-of-the-art anonymization methods. such as CIAGAN [1] and *Natural and Effective Obfuscation by Head Inpainting* [2]. Additionally, we condition our shape anonymization network on face meshes to retain the structural information of the original face. Then we provide as input to the face generator model an inpainted version of the original background to blend the final output of the pipeline into the original scene.

2 Related Work

2.1 Conditional Generative Adversarial Networks

GAN Framework
Generative Adversarial Networks, proposed in 2014 by I. Goodfellow et al. [3], are a class of machine learning framework that combine two neural networks: a generator G that aims to generate realistic samples, and a discriminator D whose purpose is to differentiate between real and generated samples. These two networks are trained in an adversarial manner: the discriminator D is trained to maximize the probability of correctly assigning the proper label to both training and generated images while G is being trained to minimize the same probability. Given a training set, this technique learns to generate new data with the same statistics as the training set. Training GANs is known to be a very challenging task since it requires maintaining a balance between generator and discriminator and since the advent of this framework many ideas have been proposed [4], [5] [6], [7]. In this thesis we decided to train our models with the Least-Square loss function (LSGAN) [5] which is designed to help the generator better, as it penalizes also samples that are correctly classified but that at the same time lie far away from real samples, thus moving the fake samples toward the decision boundary. As a result, models trained with LSGAN loss tend to generate samples that are closer to real samples when compared to other losses, such as cross-entropy.
The objective function of the discriminator D trained with the LSGAN loss is defined as follows:

$$\begin{aligned}\min_D V_{LSGAN}(D) = &\frac{1}{2}\mathbb{E}x \sim p_{\text{data}}(x)[(D(x) - b^2)] + \\ &\frac{1}{2}\mathbb{E}z \sim p_{\text{z}}(z)[(D(G(z)) - a^2)]\end{aligned} \tag{2.1}$$

where a and b are the labels for real and generated data.
The loss of the generator G is defined as:

$$\min_G V_{LSGAN}(G) = \frac{1}{2}\mathbb{E}z \sim p_{\text{z}}(z)[(D(G(z)) - b^2)] \tag{2.2}$$

Conditional GANs
While in a classic GAN training setting the input to the generator is a vector of random noise whose goal is to provide diversity to the generated images, in the context of identity obfuscation it is necessary to modify this input to achieve the goals of the anonymization task. In particular, the generated faces should seamlessly blend with the original background. A more detailed discussion on this will be held in the next section, now we will briefly introduce the framework any state-of-the-art face anonymization method uses: Conditional Generative Adversarial Networks [8].
A conditional GAN framework is a type of GAN where additional information is passed to both the generator and discriminator networks. In the particular case of face anonymization, this information may be necessary to blend the generated faces with the background or to gain control over some specific features of the anonymized images.

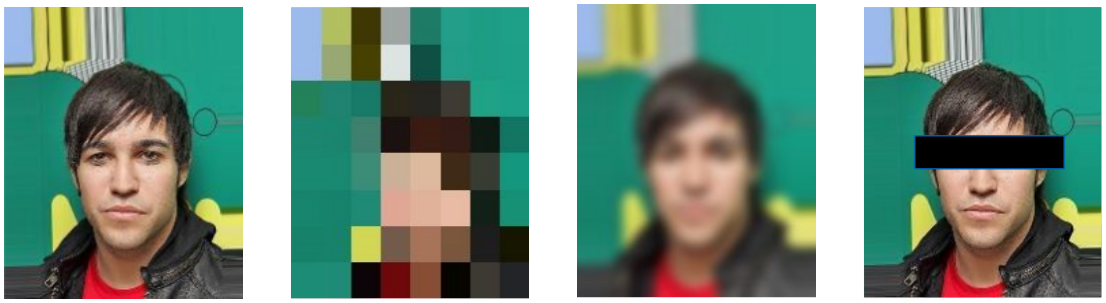
2.2 Face Generation

Generating realistic faces has been an extremely active research area since the rise in popularity of Generative Adversarial Networks [8] and recent methods [9], [10] can generate high-resolution faces achieving extremely realistic results. However, even though the quality of the generated faces is impressive, none of these models is suitable for the face anonymization task. By conditioning only on random noise, and thus having no information regarding the original face or background, those methods do not have any sort of control over the appearance or the pose of the generated face. As a result, blending the generated face with the original background and the other parts of the body is an almost impossible task, so the usability of such methods for the identity obfuscation task is limited if not impossible.

2.3 Face Anonymization

2.3.1 Heuristic Methods

Until recently, face anonymization has been achieved using heuristic methods such as pixelization, blurring or masking [11]. An example of these methods is presented in Figure 2.1.
Although they are easy to use and may result extremely effective to a human observer, they have two main drawbacks. They do not guarantee a successful anonymization against machine learning models and they can not be used in all those applications - such as social media or security cameras - where it is necessary to conceal the original identity, but where at the same time it is important that the generated face integrates with the image and appears realistic.
In fact, it is necessary for any anonymization method to preserve critical features for computer vision tasks, such as detection and tracking.



(a) Original Face (b) Pixelization (c) Blurring (d) Masking

Figure 2.1: Segmentation Mask and Face Mesh

2.3.2 Data-Driven Models

In order to overcome the issues with heuristic methods, deep learning models have been developed such as the ones proposed in [12], [13], [14], [15], [2], [1].

In particular, *Learning to Anonymize Faces for Privacy Preserving Action Detection* [14] and *DeepPrivacy: A Generative Adversarial Network for Face Anonymization* [13] are the first GAN-based methods achieving good obfuscation results. Unfortunately, they still have notable drawbacks, such as the fact that the generated images do not blend naturally with the background or the lack of diversity resulting from the fact that each identity is mapped onto a single fictitious identity.

In 2019, Gafni et al. proposed a novel method in *Live Face De-Identification in Video* [12] that shows good qualitative results and an unprecedented de-identification rate. Furthermore, their work proved effective in anonymizing complete video sequences, as it preserves temporal coherence. However, even this proposed method does not allow any control over the appearance or shape of the generated image and shows a lack of diversity in the output results when applied to anonymize the same input face several times.

The investigation of our thesis focused primarily on two networks whose design addresses the lack of diversity in the generated images. Both are based on a conditional GAN framework that passes face landmarks and background as input to the face generator.

The first network, proposed by Sun et al. [2] - consists of a two-stage head inpainting framework as presented in Figure 2.2. It is composed of two networks: a facial landmarks generator that hypothesizes realistic facial structures and a landmark guided face generation approach that naturally blends the obfuscated face into the given body pose and scene context.

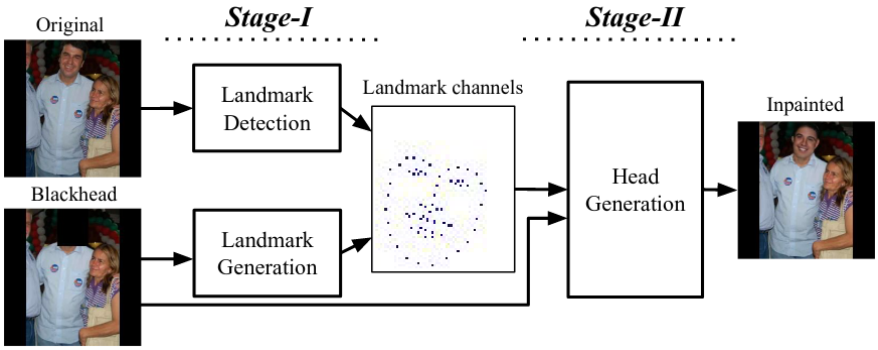


Figure 2.2: *Natural and Effective Obfuscation by Head Inpainting* Architecture

Figure 2.3 shows the network results for two different identities and multiple generated facial landmarks with different poses. The results achieved by this first framework are great and represent a huge improvement over previously proposed methods but at the same time they highlight two main issues that still need to be addressed:

- Generated poses may differ too much from the original one and consequently the inpainted head can not be naturally blended with the existing scene.
- The head generator is conditioned solely on the facial landmarks and background; as a result there is no control over the appearance of the generated identity.

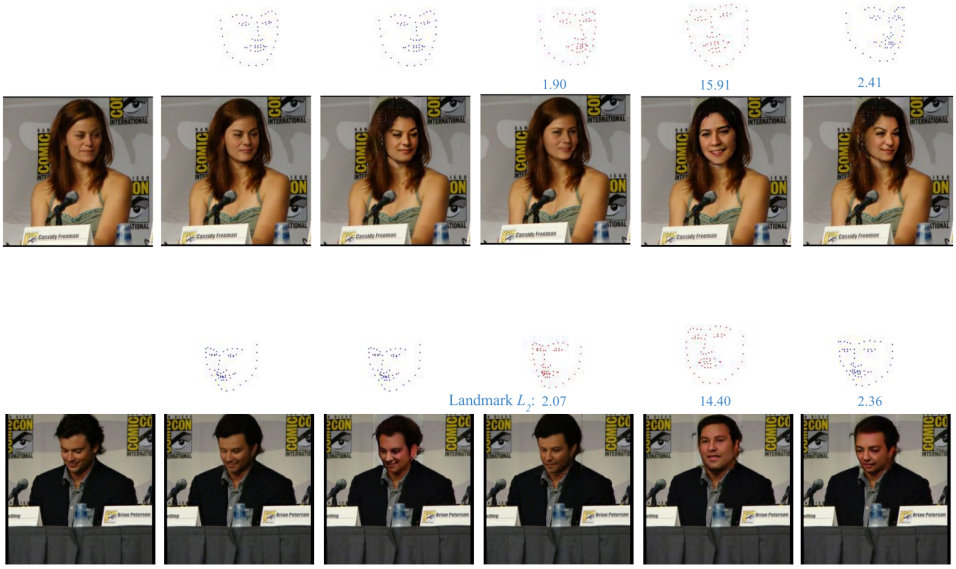


Figure 2.3: *Natural and Effective Obfuscation by Head Inpainting* Results

The second face anonymization network we focused on has been proposed in *CIA-GAN: Conditional Identity Anonymization Generative Adversarial Network* [1] by M. Maximov, I. Elezi and L. Leal-Taixé. It takes a step further addressing the lack of control over the appearance of the generated face with their novel framework. In fact, this conditional GAN framework is not only conditioned on facial landmarks and masked background, but it introduces the concept of identity guidance. An identity from the training set - the CelebA dataset [16] - is selected and encoded in a one-hot vector, which is then given as input to a transposed convolutional neural network. This network produces a parametrized version of the identity which is fed into the bottleneck of the generator. In addition to the usual discriminator, whose goal is to differentiate between real and generated samples, the framework also includes an identity discriminator that provides a guidance signal to the generator. This signal is used to guide the generator towards creating images whose representational features are similar to those of the chosen identity. A visual representation of this framework can be seen in Figure 2.4.

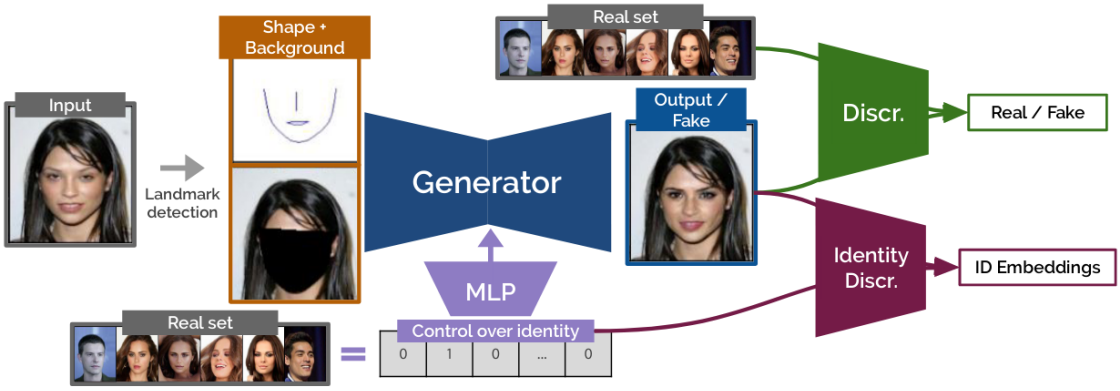


Figure 2.4: CIAGAN Architecture

Comparing the two architectures respectively proposed in *Natural and Effective Obfuscation by Head Inpainting* [2] and *CIAGAN: Conditional Identity Anonymization Generative Adversarial Network* [1] - Figure 2.3 and Figure 2.4 - it is evident that the conditional inputs of the face generator model differ significantly. In particular, the generator proposed in CIAGAN is conditioned on the detected facial landmarks corresponding to the nose, mouth and jaw line. Being provided less information regarding the facial structure of the face helps with the anonymization task, especially because there is no constraint on of the generated face eyes. At the same time, conditioning on the original facial landmarks does not allow any changes in shape or the pose - as can be seen in Figure 2.5 - and ultimately limits the flexibility of the entire network.

Figure 2.5: CIAGAN Results

In conclusion, we conducted a research investigation on the state-of-the-art face anonymization models, with a focus on the two architectures proposed in [2] and [1]. Our thesis

project aims to build a novel anonymization pipeline that addresses the limitations of the two networks discussed in this Section.

3 Methodology

3.1 Overview

In this chapter we will describe the final design of the anonymization pipeline and its development process. The pipeline consists of three distinct stages: preprocessing, shape anonymization and face generation, as shown in Figure 3.1.

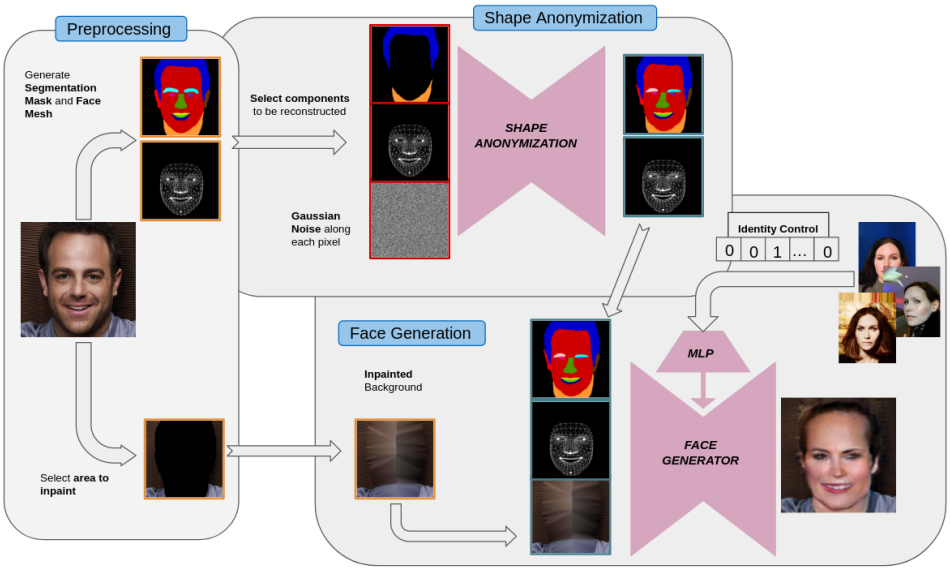


Figure 3.1: Pipeline Visualization

Each stage will be discussed in details in the following sections, here we will present the key components used in the development of the anonymization pipeline.

The complete face segmentation mask - called s - corresponding to the original face is generated in the preprocessing stage using a pre-trained network provided by the CelebAMask-HQ dataset [17]. This network is based on a U-Net [18] architecture and it outputs a segmentation mask with 19 classes. We decided to exclude those we have deemed not relevant for the anonymization task and we focused on ten facial components: neck, lips (upper and lower), mouth, nose, eyes and eyebrows (left and right) and hair. We refer to those classes as c_i, where $i \in \{1, \ldots, 10\}$.

In the following stage (shape anonymization) we construct two additional segmentation masks

$$s_I^+ := \sum_{i \in \mathcal{I}} c_i \quad \text{and} \quad s_I^- := \sum_{i \notin \mathcal{I}} c_i$$

where $\mathcal{I} := \{i \in \{1, \ldots, 10\} \mid c_i \text{ is selected to be reconstructed}\}$. These partial segmentation masks are key components of the shape anonymization network since they are used both as a conditional input for the GAN framework and in the loss computation. The shape anonymization model is also conditioned on face meshes and orientation landmarks, both discussed in detail in Section 3.3.

The third and final stage is designed to generate an anonymized face whose identity cannot be traced back to the original one. The face generator model is conditioned on the outputs of both previous stages: the anonymized mask and mesh provide information on facial shape and structure while the inpainted background is necessary to blend the generated face into the original scene.

3.2 Preprocessing

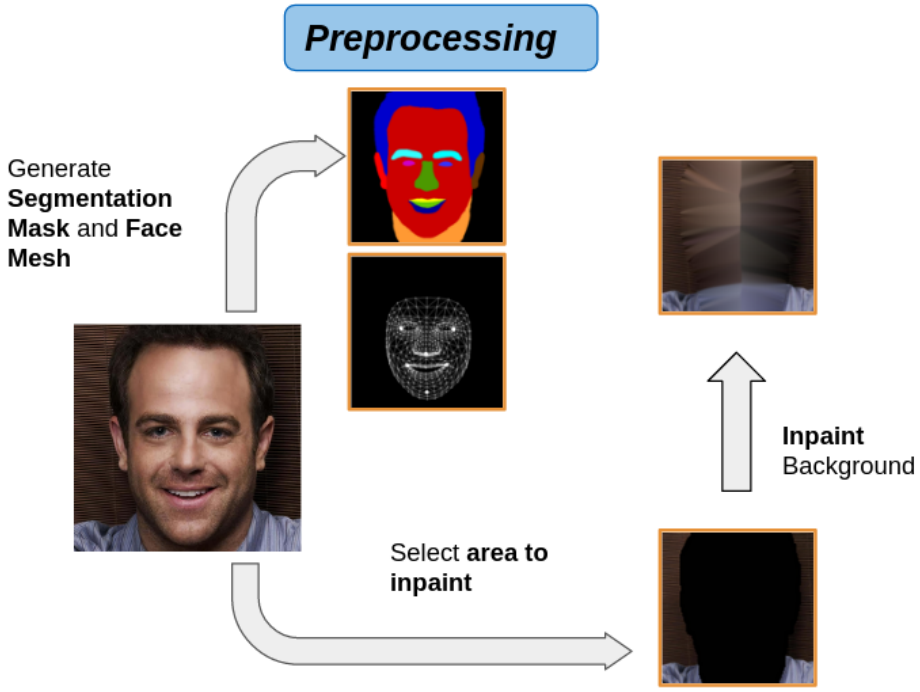


Figure 3.2: Preprocessing Stage Visualization

The first stage of the pipeline is the preprocessing step. It is designed to allow for any image to be anonymized, without any constraint on its size. The outputs of the preprocessing step are used as inputs in the second and third stages of the pipeline. In particular, we generate the segmentation mask, the face mesh and the orientation landmarks which are going to be fed into the shape anonymization model. Additionally, we remove the face from the existing scene and we inpaint the background, which will then be used for the blending step in the face generator model.

3.2.1 Segmentation Mask and Face Mesh Generation

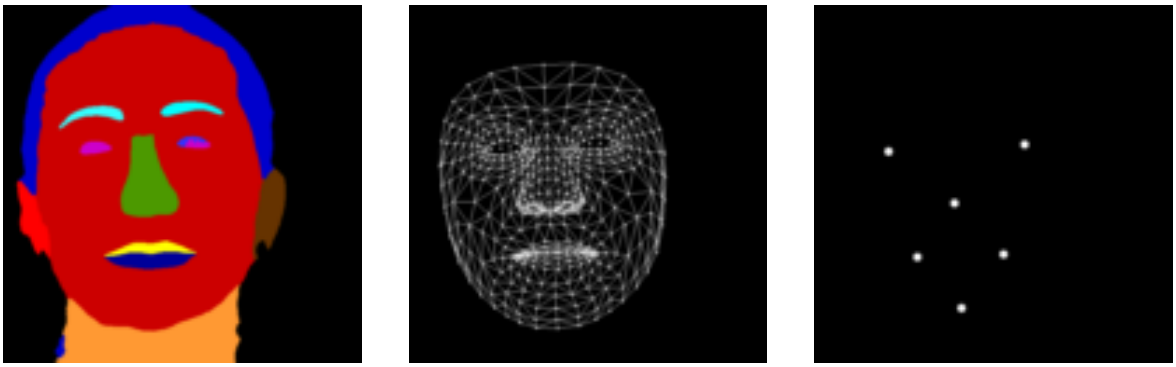
As a first step, the pipeline detects the face and determines an appropriate bounding box around it. Detection is performed using HOG [19], which returns the coordinates of the top-left and bottom-right corners of the box. Unfortunately, this detected bounding box does not include neither neck nor hair, hence we need to expand it - as in Figure 3.3 - because an image containing the full head is required to generate the segmentation mask. To obtain the best results from the pretrained model that generates the segmentation masks, the expanded bounding box must not include a sizable portion of the background. Therefore, to correctly expand the originally detected box, we determine where the subject's hair is located by generating a segmentation mask using the full image as input. This approach is not guaranteed to detect the proper hair location, but we

tested it on multiple images from CelebA, CelebAMask-HQ and FaceForensics++ and it worked as expected for most of them. When evaluating a single image, it is also possible to explicitly define the coordinates of the bounding box to obtain the best results. However, this is not feasible when evaluating thousands of images.

(a) Original detected area (b) Extended detected area

Figure 3.3: Face Detection

Once we obtain the extended bounding box, we crop the determined area and we use this portion of the original image to generate the segmentation mask, the face mesh and the orientation landmarks. Firstly, we resize the cropped image to 512x512, then we feed it into the pretrained face parsing algorithm that returns the segmentation mask, which is finally resized to a resolution of 128x128. Secondly, we use MediaPipe [20] - which provides 468 3-dimensional points - to generate the face mesh. Finally, we select the points corresponding to the corners of the eyes and of the mouth and the ones corresponding to the tip of the nose and chin. These selected points are used as orientation landmarks. Figure 3.4 shows the preprocessing step applied to the image cropped in Figure 3.3. It may be noted that the face parsing algorithm does not always generate a perfect segmentation mask, which may lead to worse performance in the later stages of the pipeline.



(a) Segm. Mask (b) Face Mesh (c) Orientation Landmarks

Figure 3.4: Segmentation Mask and Face Mesh

3.2.2 Background Inpainting

State-of-the-art anonymization models condition the generator on both facial landmarks and a masked version of the original image, which is used to blend the generated face with the existing scene. Conditioning on facial landmarks leads to anonymized faces where neither hair nor neck differ significantly from the original ones and as a result the masked background provides enough information for the blending phase.

In contrast, our work includes the segmentation mask of both hair and neck aiming to modify them as well as every other facial component. As a result, whenever the shape of hair or neck is significantly different compared to the original one, conditioning on the background where only the face is masked is not a suitable solution. Another problem arises when the generated segmentation mask covers a smaller area than the original one. In particular, the generated face would not entirely cover the masked background, resulting in empty regions that would not belong neither to the background nor the anonymized face.

We decided to solve these problems by training a face generator network that takes as input a fully inpainted version of the original background instead of the masked one. At first we thought of painting only the region exactly corresponding to the segmentation mask. However, we noticed that in most cases the generated mask did not match perfectly the original face. This leads to sub-optimal results because portions of the head, in particular the hair, are wrongly considered to be part of the background. Therefore, we decided to dilate the segmentation mask to ensure that the entire head is part of the region that will first be removed and then inpainted. This produces slightly worse results, as the region to be painted is larger, but it is a much more consistent option.

At first, we tried a deep learning based approach - *Generative Image Inpainting with Contextual Attention* proposed in [21] - but it did not perform as well as we expected. This is most likely due to the fact that the background information provided in our test images is extremely limited, as most of the image consists of the face, neck and hair. We therefore decided to use the most consistent option represented by the inpainting function provided by the opencv-python package [22].

The resulting backgrounds are not inpainted perfectly, but they provide enough information to blend the anonymized face with the original background in all those situations where conditioning the face generator directly on the masked background would not be sufficient.

3.3 Shape Anonymization

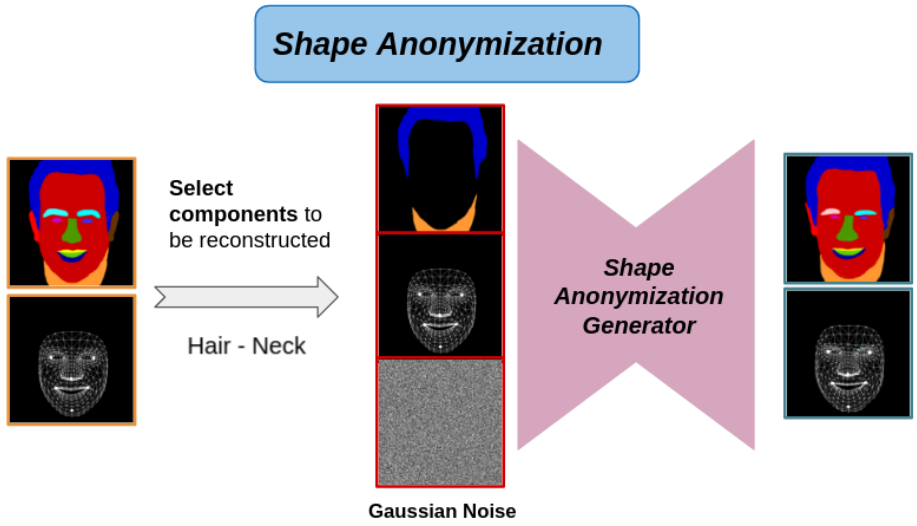


Figure 3.5: Shape Anonymization Visualization

In this section we will discuss the design development and the final architecture of the main contribution of the thesis: a novel network that modifies the shape of a face segmentation mask. The model is designed with the idea of changing the shape of the face in a controlled way, where the user can select which components should not be modified and which ones should instead be anonymized by changing their shape.

We decided to start with a simple network and modify it accordingly to achieve the goals we had in mind. We built our conditional generative adversarial network on top of the existing CIAGAN architecture, described in Section 2.3.2. Since the shape anonymization network is not responsible to change the identity of the image, we removed the identity discriminator and we decided to inject Gaussian noise instead of encoding a source identity. Additionally, we changed the conditional input: we removed the masked background and we decided to condition the network on segmentation masks instead of facial landmarks. The initial idea was to use the GAN framework to generate a modified version of the segmentation mask with the aid of the L1 reconstruction loss. In particular, the generator objective function was defined as follows:

$$\mathcal{L}(G) = \lambda_1 \mathcal{L}_{LSGAN}(G) + \lambda_2 \mathcal{L}_{REC}(G) \tag{3.1}$$

where

$$\mathcal{L}_{LSGAN}(G) = \frac{1}{2}\mathbb{E}z \sim p_z(z)[(D(G(s,z)) - 1)^2] \tag{3.2}$$

$$\mathcal{L}_{REC}(G) = |G(s,z) - s| \tag{3.3}$$

The notation $G(s,z)$ indicates the output of the generator conditioned on the complete segmentation mask s and the Gaussian noise z. The discriminator LSGAN loss function is defined by Equation 2.1.

After training the network, we noticed that the generated outputs where extremely dependent on the two hyper-parameters λ_1 and λ_2 corresponding to GAN and reconstruction loss respectively. In particular, when λ_2 was too high we would simply reconstruct the original segmentation mask; whereas when we increased λ_1 the training process would collapse before obtaining any result due to the vanishing gradient problem. Finally, we came to the conclusion that, regardless of the choice of hyper-parameters, we would never have achieved the desired results without modifying the architecture or the loss function.

Since our initial idea was not successful, we decided to modify the conditional input of the generator. In particular, our design choice of using segmentation masks instead of facial landmarks allowed us to select specific facial components and condition the generator on those parts of the face only. At first, we decided to condition the model on the hair and nose. Conditioning on the hair is useful in the face generation phase of the pipeline because it facilitates blending the anonymized face with the background, while conditioning on the nose provides guidance for maintaining the orientation of the face. The segmentation mask consisting only of hair and nose is referred to as s_I^+. Moreover, we changed the reconstruction term of the objective function in the following way:

$$\mathcal{L}_{REC}(G) = \gamma_1 |G(s_I^+, z)_{hair} - s_{hair}| + \gamma_2 |G(s_I^+, z)_{nose} - s_{nose}| \tag{3.4}$$

where, instead of computing the L1 reconstruction loss on the full segmentation mask, we compute it only on the facial components that were used as a conditional input for the generator. With these design changes, we wanted to give the generator more guidance in producing the segmentation masks, so that the hair and nose matched the original ones, while all other facial components were randomly generated by the GAN framework. Unfortunately, these changes did not lead to any real improvement, and we could still not generate realistic masks with the aforementioned characteristics.

We re-evaluated every design choice we made to understand why our proposed network did not produce the expected results and, finally, we hypothesized that the underlying reason of our failed attempts could be identified in the lack of face structural information provided by conditioning only on segmentation masks.

To test our hypothesis, we decided to condition the network on segmentation masks as well as face meshes. We used MediaPipe Face Mesh [20] to estimate 468 3D face landmarks and connect them to generate the corresponding face mesh for each image in our dataset. After this change, the output of the generator consisted in the anonymized segmentation mask concatenated with the corresponding anonymized face mesh. Additionally, we introduced an new term - called negative reconstruction loss - in the objective function. The aim of this additional term is to provide positive feedback to the generator when the shape of the generated facial components differ significantly from the original, with the exception of hair and nose. As a result of this change, the objective function for the generator was defined as follows:

$$\mathcal{L}(G) = \lambda_1 \mathcal{L}_{LSGAN}(G) + \lambda_2 \mathcal{L}^+_{REC}(G, s_I^+) - \lambda_3 \mathcal{L}^-_{REC}(G, s_I^-) \tag{3.5}$$

In particular, $\mathcal{L}^+_{REC}$ is the reconstruction loss on neck and hair defined in Equation 3.4

and the negative reconstruction loss is defined as:

$$\mathcal{L}^{-}_{REC}(G, s^{-}_{I}) = \sum_{c_i \notin \{neck, hair\}} |G(s^{+}_{I}, z)_i - c_i| \tag{3.6}$$

where $G(s^{+}_{I}, z)_i$ represents the generated segmentation mask of the i^{th} facial component. For example, following the definition provided in Section 3.1, $G(s^{+}_{I}, z)_1$ corresponds to the generated segmentation mask of the nose.

Finally, with the introduction of face meshes and negative reconstruction loss, we were able to achieve the first milestone of our project: generate segmentation masks in which hair and nose are the same as the original ones, while every other component's shape is modified.

However, what we achieved was still far from the goal we set at the beginning of the thesis. There were still three main problems we had to overcome:

- Lack of diversity. We noticed that the model suffered from mode collapse; injecting Gaussian noise in the bottleneck of the generator had no effect on the output.
- Lack of flexibility. Our model could generate an anonymized segmentation mask while keeping hair and nose untouched, but it could not do the same with any other facial component.
- Maintaining face orientation without conditioning on the nose. Although keeping the same orientation is not strictly necessary to effectively anonymize a face, it is very useful during the blending phase to achieve more realistic results.

To tackle the lack of diversity we decided to add a regularization term in the loss function. At first, we implemented the one proposed by D. Yang et al. in [23]. The proposed method is to encourage diversity by introducing a regularization term that approaches its minimum when the generator collapses into a single mode. Unfortunately, this approach failed to produce more diverse results, so we decided to implement a different regularization term. In particular, we added the regularization term proposed by Q. Mao et al. in [24]. It maximizes the ratio of the distance between generated images with respect to the distance between their corresponding input latent code. Again, this approach did not affect the diversity of the generated images.

Finally, we implemented a third regularization term, defined by R. Liu et al. in *DivCo: Diverse Conditional Image Synthesis via Contrastive Generative Adversarial Network* [25]. It is defined as follows:

$$\mathcal{L}_{DIV}(G) = \mathbb{E}_{y\sim Y,\, z\sim N(0,1)} \left[-log \frac{exp(\langle f, f^{+} \rangle / \tau)}{exp(\langle f, f^{+} \rangle / \tau) + \sum_{i=1}^{N} exp(\langle f, f^{-}_{i} \rangle / \tau)} \right] \tag{3.7}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between two normalized feature vectors to measure their similarity, τ is a hyper-parameter for scaling the similarity and f, f^{+} and f^{-} are the generated images' feature representations. To better describe the contribution of this

regularization term we need to define the concept of positive and negative noise vectors. Given a random vector $z \sim N(0,1)$, we denote a positive noise z^+ as a vector randomly sampled within the small hyper-sphere with radius R centered at z. In contrast, we define a negative noise z^- as any vector outside the sphere within the latent space. On one hand, we can generate z^+ by sampling a random vector $u \sim U[-R, R]$ and adding it to the original vector z, obtaining $z^+ = z + u$. On the other hand, to generate a negative noise vector, we can directly sample $z^- \sim N(0,1)$. We then check if z^- lies in the hyper-sphere around z and, if that is the case, we sample it again until we obtain a vector outside the sphere. Once we obtained these three noise vectors, we can define $\hat{x} = G(z, y)$, where y is the conditional input of the generator, and we can generate a positive sample as well as a negative one. These samples are respectively defined as $\hat{x}^+ = G(z^+, y)$ and $\hat{x}^- = G(z^-, y)$. Finally, we can compute the generated images' feature representations $f = E(\hat{x})$, $f^+ = E(\hat{x}^+)$ and $f^- = E(\hat{x}^-)$, where E is a feature extraction model. In particular, we implemented E as a pretrained ResNet152.

The regularization term $\mathcal{L}_{DIV}(G)$ is proposed to regularize generated images' feature representations. That is, it aims to bring f and f^+ closer while spreading away f and f^-. Figure 3.6 shows a comprehensive diagram describing the implementation of this regularization term.

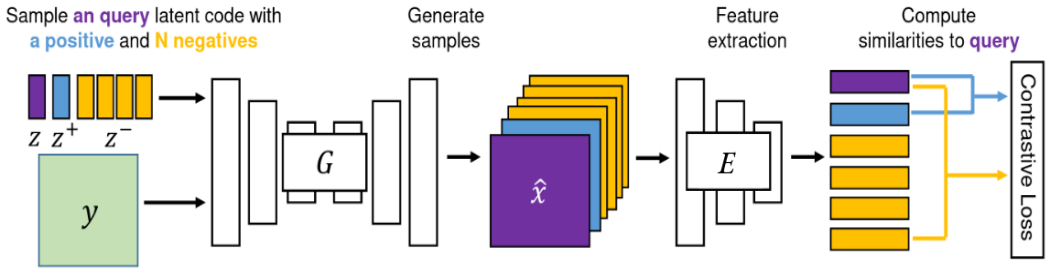


Figure 3.6: Diversity Loss $\mathcal{L}_{DIV}(G)$ Diagram

With the introduction of $\mathcal{L}_{DIV}(G)$ defined in Equation 3.7, the objective function of the generator is defined as:

$$\mathcal{L}(G) = \lambda_1 \mathcal{L}_{LSGAN}(G) + \lambda_2 \mathcal{L}^+_{REC}(G, s^+_I) - \lambda_3 \mathcal{L}^-_{REC}(G, s^-_I) + \lambda_4 \mathcal{L}_{DIV}(G)$$

Once we solved the mode collapse problem, we could focus on the orientation task. Until now, we decided to always condition the generator on the nose because reconstructing it allowed us to maintain the original orientation. However, we wanted our model to be more flexible by allowing to change the shape of the nose as well as the one of any other facial component. We decided to approach the problem by introducing a third conditional input: orientation landmarks. We selected six landmarks, among the ones generated with MediaPipe, that we believe are responsible for determining the orientation of the face. In particular, these landmarks correspond to the locations of the

chin, the mouth's and eyes' corners and the tip of the nose. We highlighted them by expanding their radius and we concatenated one extra channel to the existing conditional input, which now consisted of 3 channels for the segmentation mask, 3 channels for the face mesh and 1 channel for the orientation landmarks. Our idea for maintaining the original orientation was to force the position of the generated landmarks to be close to that of the original ones, so we decided to compute the L1 reconstruction loss between them and called it orientation loss ($\mathcal{L}_{OR}$).

Finally, in order to design a more flexible network, we decided to change the generator architecture to take as input all the segmentation masks corresponding to the ten facial components and the set I describing which ones are to be used as a conditional input. This design allowed to train on randomly selected sets I, which proved to be a successful strategy to obtain a robust and flexible model.

In conclusion, the shape anonymization network is designed as a GAN framework conditioned on:

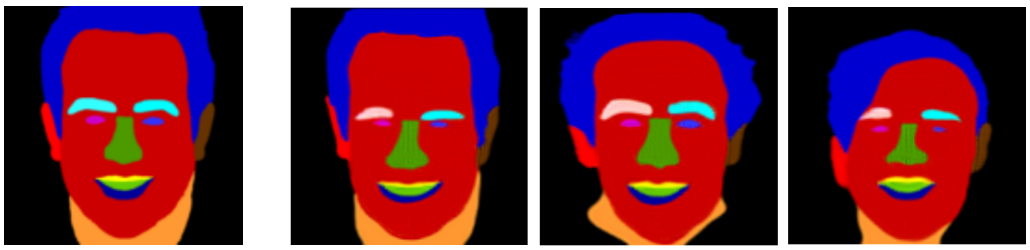
- The segmentation mask s_I^+ consisting of all the facial components we aim to reconstruct. The network is flexible enough to take as input any combination of the 10 valid facial components.
- The face mesh generated with MediaPipe.
- The six orientation landmarks.

The final objective function for the generator is

$$\begin{aligned}\mathcal{L}(G) =& \lambda_1 \mathcal{L}_{LSGAN}(G) + \lambda_2 \mathcal{L}^+_{REC}(G, s_I^+) - \\ & \lambda_3 \mathcal{L}^-_{REC}(G, s_I^-) + \lambda_4 \mathcal{L}_{DIV}(G) + \lambda_5 \mathcal{L}_{OR}\end{aligned}$$

where $\mathcal{L}_{LSGAN}$ is the GAN loss responsible to generate realistic output, $\mathcal{L}^+_{REC}$ and $\mathcal{L}^-_{REC}$ are L1 reconstruction loss that allow the network to generate masks reconstructing specific components while modifying all the other ones, $\mathcal{L}_{OR}$ helps to maintain the same face orientation and finally $\mathcal{L}_{DIV}$ is a regularization term that prevents mode collapse.

The final architecture of the shape anonymization network can be seen in Figure 3.5. Figure 3.7 compares the outputs of the shape anonymization network when conditioned on different facial components of the same original segmentation masks.



(a) Original Mask (b) Hair, Neck (c) Eyes Region (d) Mouth Region

Figure 3.7: Different components comparison

3.4 Face Generation

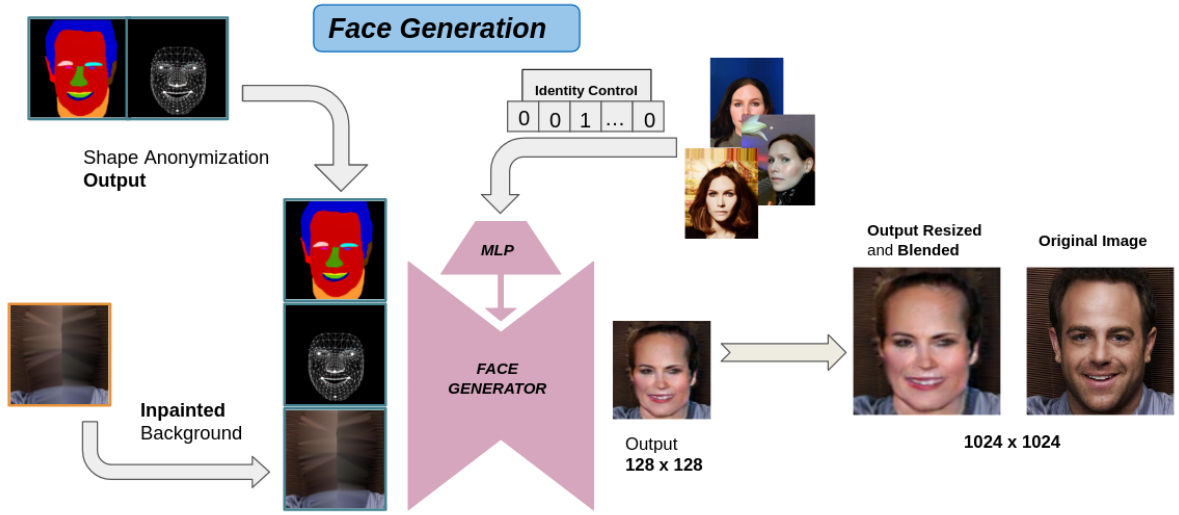


Figure 3.8: Face Generation Visualization

The third and final stage of the pipeline consists of the face generation network. The face generator network is a GAN framework conditioned on the anonymized outputs obtained generated in the shape anonymization phase and the inpainted background.

As with the shape anonymization model, we based our design on CIAGAN architecture which was previously discussed in Section 2.3.2. We implemented the identity guidance discriminator as well as the transposed convolutional network to encode the source identity. The generated image has a 128x128 resolution and its appearance is determined by the segmentation mask and face mesh generated in the shape anonymization stage, as well as the selected source identity. Finally, this image is resized accordingly to perfectly fit the bounding box defined in the preprocessing phase and it is pasted into the original image, producing the final output of the anonymization pipeline.

This implementation of the face generator, together with the shape anonymization model, ensure a complete control over each stage of the anonymization process. More details about the training process and the loss function are discussed in Section 4.2

4 Evaluation

4.1 Datasets

To train and evaluate our models we used the following datasets: CelebA, CelebaAMask-HQ and FaceForensics++.

CelebA [16] consists of 202.599 face images of 10.177 unique identities. We use the aligned version where each image is centered on a point in between person's eyes, and then padded and resized to have 178 $\times$ 218 resolution, while maintaining original face proportions. Each identity consists up to 35 photos, but we only considered those with at least 30 images for a total of 1415 identities. The training set consists of the first 1.200 identities, while the remaining 215 constitute the test set. First, we use the CelebA dataset to pretrain the identity discriminator and the face generator network - described in Section 3.4 - and the baseline model that will be discussed in Section 4.3. Then, in the evaluation phase, we used this dataset to compute all image-related metrics.

The CelebAMask-HQ dataset [17] has 30.000 high-resolution (1024x1024) face images selected from the CelebA dataset. Each image has a segmentation mask of 19 facial attributes, each of size 512x512. This dataset was used for the training phase of the shape anonymization model, as described in Section 3.3. We used the first 25.000 to define our training set.

FaceForensics++ [26] is a dataset of 1000 original video sequences, sourced from 977 youtube videos which contain a trackable face without occlusions. We evaluated all the video-related metrics, discussed in Section 4.3, on 100 videos from the FaceForensics++ dataset.

4.2 Training

In the following section we will discuss the training process of the two networks of the anonymization pipeline, namely the shape anonymization model and the face generator. All networks were implemented and trained using PyTorch [27] and training was performed using images with a resolution of 128x128 pixels. We used a learning rate of 1e-4, a batch size of 8 and the Adam optimizer [28] without weight decay. Training times were approximately 24 hours for both the shape anonymization model and the face generator.

Shape Anonymization Network
The training set of the shape anonymization model consists of segmentation masks provided by the CelebAMask-HQ dataset in addition to face meshes and orientation landmarks obtained using MediaPipe, as explained in Section 3.3. Since most of the faces in the dataset are centered, and consequently so are the corresponding segmentation masks, we introduced randomness in the data loading process, with the intention of making the model more robust against non-centered input faces. This simple but effective data augmentation is performed as follows: first the segmentation mask is padded, then it is randomly cropped - with strong limitations to ensure consistency - and finally it is translated up to 20 pixels both vertically and horizontally.

The feature extractor network - used in the diversity loss computation described in Equation 3.7 - is a pre-trained ResNet152 [29].

Figure 4.1 shows a visual overview of the loss function used to train the shape anonymization generator. We set the hyper-parameters corresponding to each loss component as follows: $\mathcal{L}_{LSGAN}$: 1.0, $\mathcal{L}^{+}_{REC}$: 1.0, $\mathcal{L}^{-}_{REC}$: 3.0, $\mathcal{L}_{OR}$: 0.1, $\mathcal{L}_{DIV}$: 0.5.

Furthermore, for each facial component, we introduced a different weight for the computation of both the positive and negative reconstruction losses, to take into account the different areas covered by each component. These weights are defined as: Neck 0.1, Lips 8.0, Mouth 3.0, Nose 3.0, Eyes 5.0, Eyebrows 10.0, Hair 0.1.

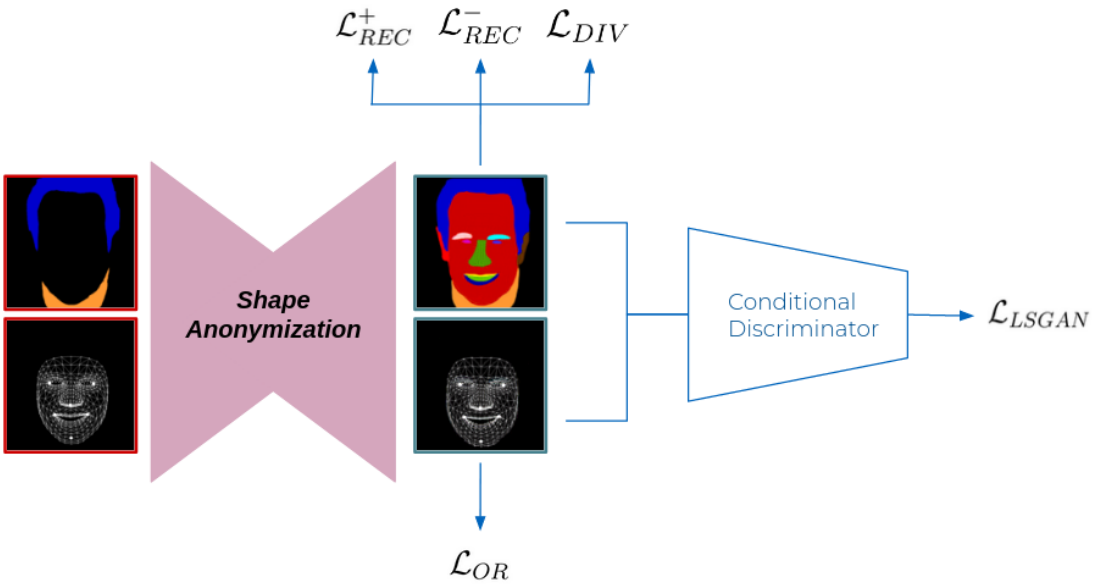


Figure 4.1: Loss Computation Shape Anonymization

Face Generator Network

Initially, we again used the data augmentation process described above for the shape anonymization network also to train the face generator model. Once we trained the model, we observed that the generated faces, at the output resolution 128x128, were both realistic and naturally blended with the background. Unfortunately, when we evaluated the results of the full pipeline we noticed that some output images - resized to the original resolution of the corresponding test image - presented visual artifacts, such as groups of randomly coloured pixels in the hair region. The explanation for this unintended effect lies in the fact that each image of the training set was scaled down to the training resolution (128x128) from the initial resolution of the CelebA dataset (178×218), while other images were resized to 128x128 from other resolutions. For example faces from CelebAMask-HQ have an original size of 1024x1024. The face generator model was not robust against images with a different original height-to-width ratio and this represented a problem because we wanted our pipeline to be able to handle images with arbitrary resolution.

Therefore, we implemented an additional step in the data augmentation process. Given an image from the CelebA dataset, first we resize it to the training resolution of 128x128, then we randomly crop it and finally we resize it back to 128x128 to effectively change the original height-to-width ratio. We have manually introduced some limitations to the function that performs random cropping to ensure that the cropped area covers at least three quarters of the original image, so that a substantial part of the face is always included. Once the height/width ratio of the faces has been modified, we complete the data augmentation process with the steps already described for the shape anonymization model.

(a) Original Face (b) Height/Width Modified

Figure 4.2: Data Augmentation for Face Generation

As discussed in Section 3.4, the face generator framework includes an identity discriminator that provides the identity guidance signal $\mathcal{L}_{SIAM}(G)$ to the generator. This identity discriminator network is designed as a siamese neural network, which was pretrained on the CelebA dataset using the Proxy-NCA loss [30]. The identity discriminator is finetuned during the training of the full face generator framework. In particular, it is trained jointly with the generator in a collaborative manner.

The loss function of the generator G is defined as:

$$\mathcal{L}(G) = \gamma_1 \mathcal{L}_{LSGAN}(G) + \gamma_2 \mathcal{L}_{SIAM}(G) + \gamma_3 \mathcal{L}_{REC}(G) + \gamma_4 \mathcal{L}_{PERC}(G)$$

where $\mathcal{L}_{LSGAN}(G)$ is the GAN loss we introduced in Section 2.1. $\mathcal{L}_{REC}(G)$ and $\mathcal{L}_{PERC}(G)$ are respectively the L1 reconstruction loss and the perceptual loss. Both these losses are useful in generating more high quality images, as argued by Johnson et al. [31].

Figure 4.3 shows a visual overview of the loss function used to train the face generator model.

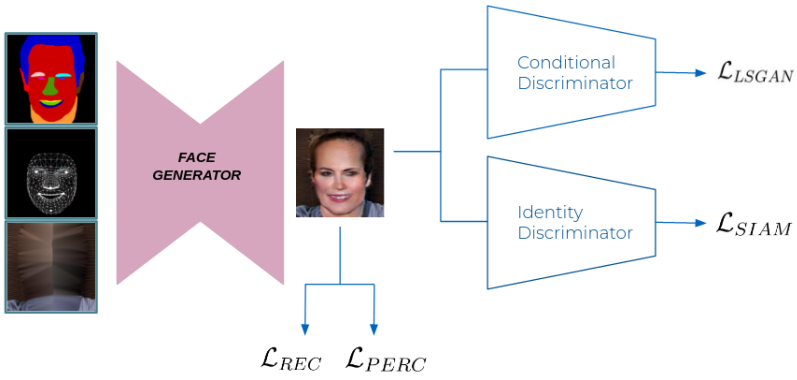


Figure 4.3: Loss Computation Face Generation

The hyper-parameters corresponding to each component of the loss function are defined as follows: $\mathcal{L}_{LSGAN}$: 1.0, $\mathcal{L}_{SIAM}$: 0.1, $\mathcal{L}_{REC}$: 0.1, $\mathcal{L}_{PERC}$: 1.5.

4.3 Metrics and Baseline

In this section, we compare our proposed pipeline with several heuristic models commonly used for identity anonymization. Additionally, we trained a baseline network using the CIAGAN architecture.

We evaluate all models in face detection and re-identification metrics. Detection is performed using both HOG [19] and SSH detectors [32]. For re-identification we use a pretrained FaceNet model [33] based on Inception-Resnet backbone [34]. We use the Recall@1 evaluation metric for re-identification, which measures the ratio of samples whose nearest neighbor belongs to the same identity. This metric takes values from 0 to 100 with 0 showing perfect de-identification rate and 100 representing perfect identification rate.

Additionally, we evaluate the visual quality of the generated images using the Fréchet Inception Distance (FID) [35]. This metric compares the statistics of generated samples to those of real samples, so the lower the FID, the more similar real and generated samples are. We evaluate the diversity of the samples using the Learned Perceptual Image Patch Similarity (LPIPS) metric [36], that is used to judge the perceptual similarity between two images and it has been shown to match human perception well. An higher LPIPS score means that the two images are perceptually dissimilar.

Finally, we evaluate the performance of the pipeline applied to video frames using the metrics proposed in *Learning Temporal Coherence via Self-Supervision for GAN-based Video Generation* [37]: PSNR, LPIPS, tOF, tLP. PSNR is a spatial metric that computes the difference between two corresponding frames based on mean squared error, hence higher PSNR indicates a better pixel-wise accuracy. As previously mentioned, lower LPIPS represents closer semantic similarity. Finally, toF measures the pixel-wise difference of motion estimated from sequences and tLP measures perceptual changes over time via deep feature map. Higher tOF and tLP scores correspond to a worst performance of the model.

4.4 Results

4.4.1 Quantitative Results

We present detection and identification results in the following Table.

Models	**Detection (↑)**		**Identification (↓)**
Evaluation on Neck, Hair	Dlib	SSH	FaceNet
Original	100	100	95.44
Pixelization 8 by 8	0	0	**0.43**
Blur 9 by 9	93.58	33.87	73.34
Blur 17 by 17	82.05	0.33	24.67
CIAGAN Baseline	99.73	54.38	2.94
Thesis Model	**99.80**	**65.43**	1.20

Table 4.1: Detection - Identification

We compare the results of the thesis pipeline with those obtained with the baseline network and two heuristic methods, namely pixelization and blurring. The first row of Table 4.1 is computed by evaluating detection and identification metrics directly on the test set and it shows that a face is detected in every image and more than 95% of them are correctly identified. It can be seen that the best identification score is achieved by the Pixelization 8 by 8 method. However, none of the face anonymized with this heuristic method can be detected, making pixelization an impractical option for any real-world application where face tracking is required. The thesis model achieves the second lowest identification rate, while at the same time it has the highest detection rate for both HOG and SSH detectors, making it the best model among those considered. All the results described in this table are achieved by conditioning the shape anonymization network on neck and hair.

Table 4.2 shows how detection and identification scores change when the pipeline is conditioned on different facial components. Detection rates do not change significantly, while it can be seen that by limiting the information given to the shape anonymization model - i.e. providing only the eyes region or no information at all - the identification score drops below 1%. We also compare how selecting different facial components affects the visual quality (FID) and the diversity (LPIPS) of the anonymized faces. On one hand, the lowest FID scores are achieved when both hair and neck are reconstructed, mainly because having hair and neck matching the original ones is extremely helpful in

the blending phase. That is, when the shape anonymization model does not change the shape of these components, the face generator does not have to rely on the inpainted part of the background to blend the output face with the original scene and as a result the final output is more realistic. On the other hand, it can be seen that when the FID score is low, so is the LPIPS score and viceversa when one is high, so is the other. The reason for this correlation lies in the fact that the shape anonymization network can generate a variety of hair styles, whereas the diversity of other facial components is limited by the fact that the anonymized segmentation mask must be realistic. Therefore, when the pipeline is conditioned on the hair, less diversity is expected than when the shape anonymization network has the freedom to change the hair structure.

Wanted Parts	**Detection (↑)** Dlib	SSH	**Identification(↓)** FaceNet	**FID (↓)**	**LPIPS(↑)**
Hair, Neck	**99.80**	65.43	1.20	**67.11**	0.025
Mouth, Lips	99.61	72.10	1.08	94.77	0.141
Eyes, Eyebrows	99.85	**72.15**	0.82	83.94	0.214
Everything	99.78	69.89	1.27	68.85	0.021
Nothing	99.72	67.14	**0.73**	88.44	**0.226**
CIAGAN Baseline	99.73	54.38	2.85	71.95	0.107

Table 4.2: Comparison of different conditional inputs

Finally, Table 4.3 shows the difference between evaluating a video sequence using the thesis pipeline and the baseline model. The evaluation is performed on the frames of 100 test videos selected from the FaceForensics++ dataset. For each video, the gaussian noise and the source identity are sampled only once and they are used in the anonymization process of every single frame to maintain temporal consistency.

Models Evaluation on Neck, Hair	**PSNR (↑)**	**LPIPS (↓)**	**tOF (↓)**	**tLP (↓)**
Thesis Model	21.66	0.133	1.02	5.97
CIAGAN Baseline	**25.99**	**0.077**	**0.22**	**1.65**

Table 4.3: Video evaluation with TecoGAN metrics

It can be seen that the thesis pipeline performs worse than the baseline network in every evaluated metric. This results can be explained as the shape anonymization model may fail in creating a temporal consistent mapping - between the original segmentation masks and the generated ones - when the pose of the original face is not perfectly centered and aligned. In addition, most videos have a heterogeneous background and the portion of the background detected around the subjects' heads can change between the initial and final frames. Although this is not an issue for the baseline model - which is conditioned directly the masked background - it represents a further challenge for the temporal consistency of our proposed pipeline. In fact, inpainting the detected portion of the background results in the loss of some detail, making the next frame slightly different from the previous one not only because of differences in the generated face but also because the background itself changes.

4.4.2 Qualitative Results

In this section, we present the qualitative results of the complete anonymization pipeline, highlighting both the effect of conditioning the shape anonymization model on different facial components and the effect of selecting different source identities as guidance for the face generation phase. The latter is shown in Figure 4.4: the column on the left contains the source images - whose identity we aim to anonymize - and the top row represents the identities selected as guidance.

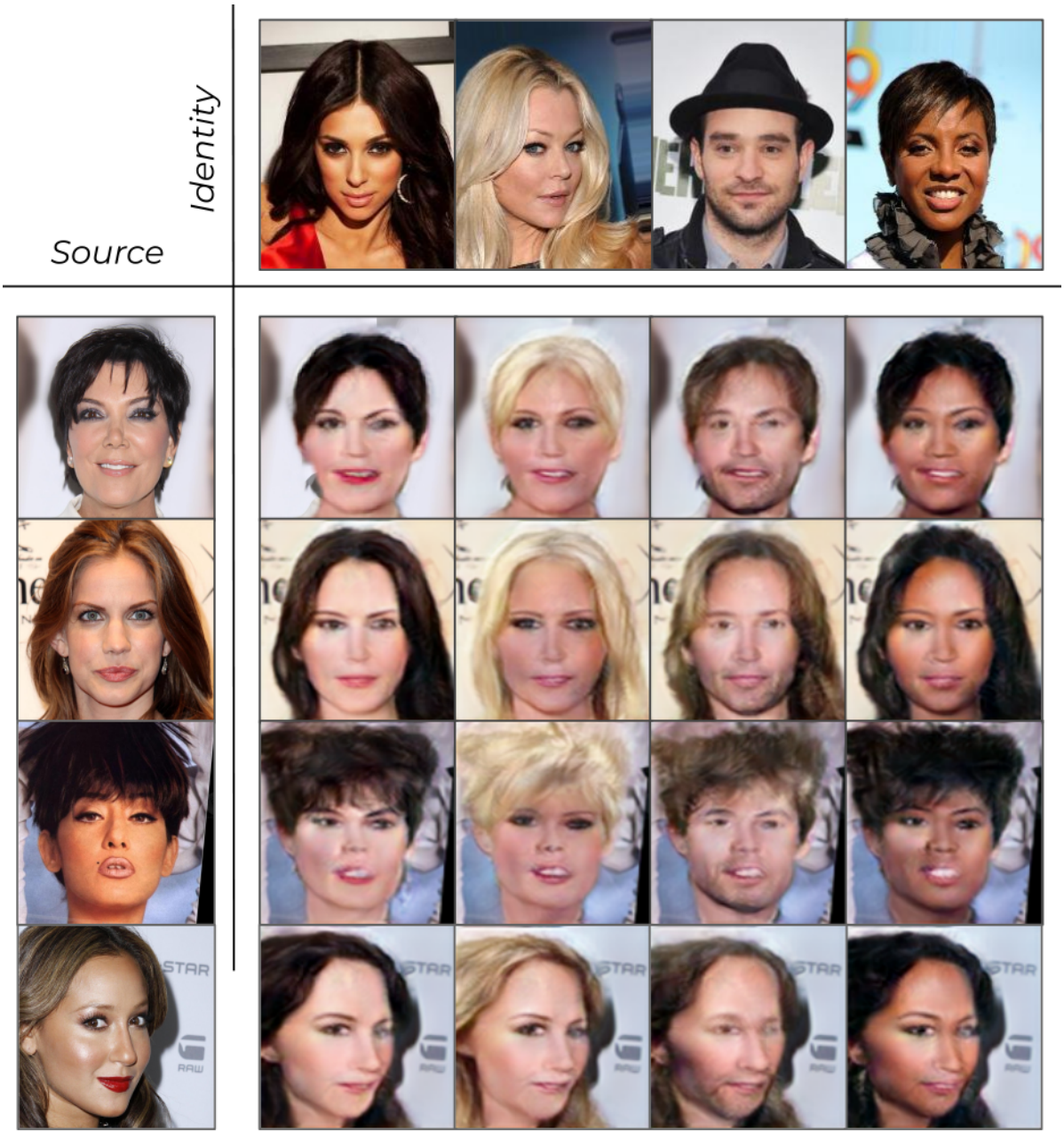


Figure 4.4: Qualitative Results: Effect of Identity Guidance

Figure 4.5 shows the anonymized segmentation masks conditioned on different inputs as well as the final output of the pipeline. Each image is generated using the same source identity.

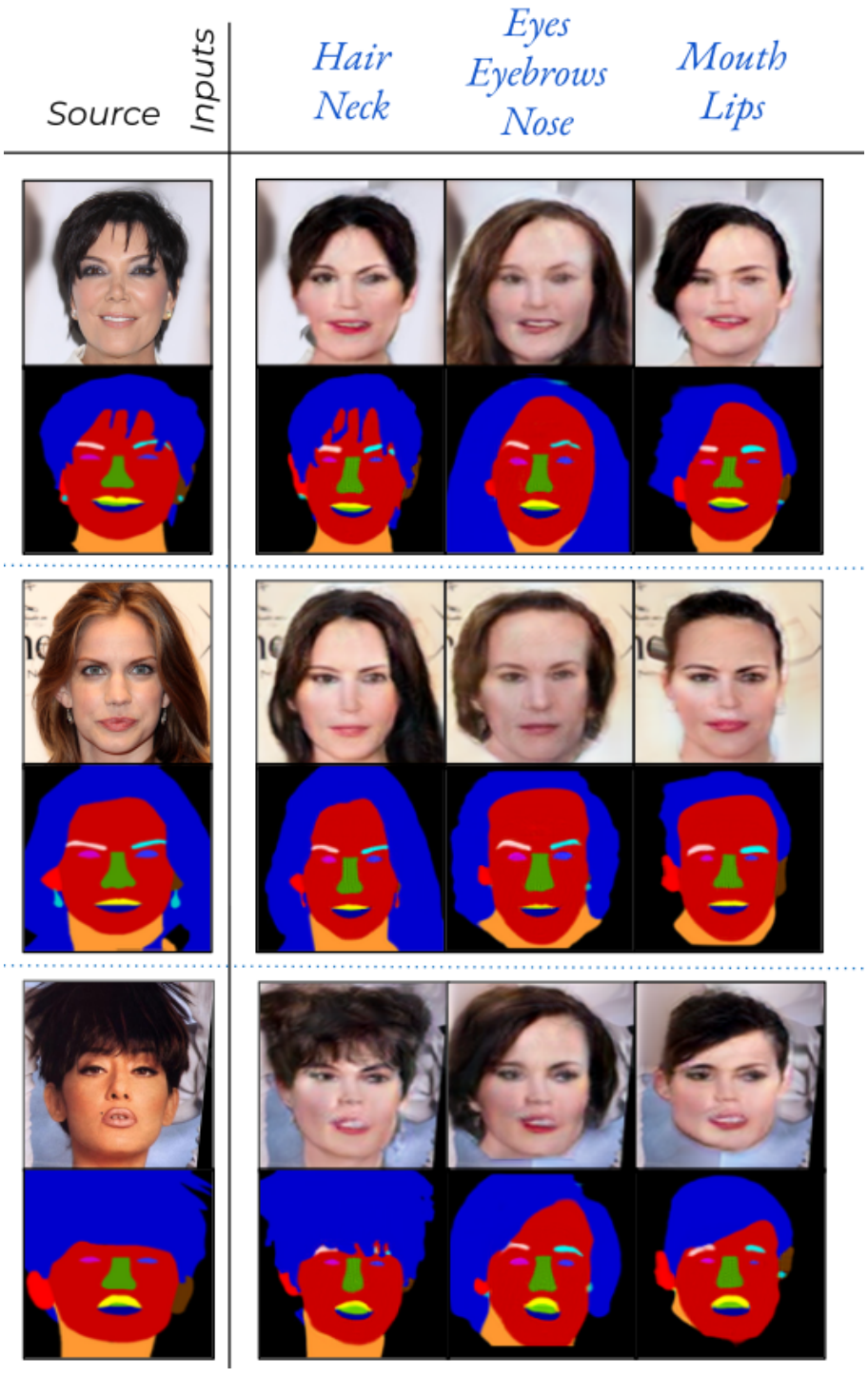


Figure 4.5: Qualitative Results: Effect of Conditioning on different Facial Components

4.4.3 Ablation Study

In this section we perform an ablation study of our proposed shape anonymization model to evaluate the value of our design choices. In the following tables we present the results obtained by evaluating the metrics discussed in Section 4.3 on different versions of the models using the following notation: *Full Segmentation Mask* indicates a model which has been trained conditioning every training image on the complete segmentation mask instead of a randomly selected one. *No Face Mesh* and *No Orientation Landmarks* represent networks where, during training, the generator does not receive any information regarding face meshes and a orientation landmarks, respectively. Additionally, we trained and evaluated multiple networks in which a specific component of the generator's objective function was removed. In particular, we studied the effects of both the positive and negative reconstruction losses and diversity loss.

Table 4.4 and Table 4.5 report detection and identification rates, as well as FID and LPIPS scores, for each of the ablation study model. Each result reported in these two tables refers to images generated while conditioning only on neck and hair. As it can be seen by looking at the detection column of Table 4.4, both the structural information provided by the face meshes and the positive reconstruction loss are necessary for the pipeline to generate realistic images. Consequently, the best scores achieved by the *No Face Mesh* model for both re-identification and diversity (LPIPS) do not provide valuable information.

Models Evaluation on Neck, Hair	**Detection (↑)** Dlib	**Identification (↓)**
Thesis Model	99.80	1.20
Full Segmentation Mask	99.18	1.20
No Face Mesh	1.09	**0.22**
No Orientation Landmarks	99.59	0.46
No Reconstruction Loss	0.41	0.33
No Inverse Reconstruction Loss	99.61	1.15
No Diversity Loss	**99.92**	0.29

Table 4.4: Ablation Study: Detection - Identification

Table 4.5 highlights that the lowest FID score is achieved by the original model, whose architecture has been described in Section 3.3. In the same table we can observe that removing either the orientation landmark or the negative reconstruction loss does not lead to significant decrease in the visual quality of the generated images. Additionally, the identification rate obtained with the model trained without the negative L1 loss is very similar to the one of the final model, suggesting that the contribution of this component of the generator's objective function is quite limited.

Models Evaluation on Neck, Hair	**FID (↓)**	**LPIPS (↑)**
Thesis Model	**67.11**	0.025
Full Segmentation Mask	71.70	0.049
No Face Mesh	142.72	**0.229**
No Orientation Landmarks	71.93	0.103
No Reconstruction Loss	274.77	0.225
No Inverse Reconstruction Loss	72.77	0.065
No Diversity Loss	85.55	0.045

Table 4.5: Ablation Study: FID - LPIPS

Table 4.6 illustrates what effect conditioning on different facial components has on the final output, assessing both visual quality and diversity for each of the ablation study networks. The most important information we can extrapolate confirms what we already discussed in Section 4.4.1: the pipeline works best when the shape anonymization model is conditioned on neck and hair. In fact, when the model has no information regarding these components - that is all the columns except for the one called *Everything* - the lowest FID score is always achieved by the ablation network *Full Segmentation Mask*. Moreover, the diversity score of this network does not change depending on the conditional input. These two observations prove that the *Full Segmentation Mask* network - trained on the complete segmentation mask instead of a randomly selected one - has mostly learned to reconstruct the original segmentation mask instead of anonymizing it and, as a result, it is less influenced by the facial components it is conditioned on. In contrast, the other networks depend more on the information provided by the segmentation mask they are given as input and the visual quality of their results is significantly worse when hair and neck are not included.

Models	**Mouth, Lips**		**Eyes, Eyebrows**	
	FID (↓)	LPIPS (↑)	FID (↓)	LPIPS (↑)
Thesis Model	94.77	0.141	83.94	0.214
Full Segmentation Mask	**72.11**	0.050	**72.35**	0.051
No Face Mesh	204.33	**0.356**	266.22	**0.362**
No Orientation Landmarks	89.75	0.173	86.66	0.192
No Reconstruction Loss	277.76	0.277	274.21	0.278
No Inverse Reconstruction Loss	95.09	0.095	85.13	0.144
No Diversity Loss	106.39	0.035	91.46	0.048

Models	**Everything**		**Nothing**	
	FID (↓)	LPIPS (↑)	FID (↓)	LPIPS (↑)
Thesis Model	**66.85**	0.021	88.44	0.226
Full Segmentation Mask	74.60	0.051	**75.00**	0.051
No Face Mesh	132.93	0.134	253.08	**0.345**
No Orientation Landmarks	73.02	0.074	85.53	0.196
No Reconstruction Loss	274.98	**0.213**	274.34	0.282
No Inverse Reconstruction Loss	71.75	0.036	85.36	0.147
No Diversity Loss	100.42	0.018	96.58	0.064

Table 4.6: Ablation Study: Comparison of different conditional inputs

5 Conclusions

This master thesis aimed to develop a complete anonymization pipeline that provides control over both the shape of specific components of the generated face and the appearance of the new identity. This has been achieved by proposing an architecture consisting of three stages: preprocessing, shape anonymization and face generation.

The main contribution of the thesis is represented by the shape anonymization network: a GAN framework conditioned on specific facial components that generates a segmentation mask in which the selected components match the ones of the original face while all the remaining ones are modified. We decided to condition the shape anonymization network on segmentation masks instead of facial landmarks, because they allow both the modification of hair and neck shape and more control over individual components. However, this approach has a downside: segmentation masks do not contain information on the structure of the face. Consequently, we had to include face meshes as well as segmentation masks as the conditional input for the shape anonymization network generator. The face generator model is responsible for generating realistic faces and to blend them with the original scene. Furthermore, it guarantees control over the appearance of the generated identity appearance because it includes the identity guidance method proposed in CIAGAN [1].

We evaluated the performance of the pipeline using face detection and re-identification as key anonymization metrics; FID and LPIPS are calculated as metrics assessing the visual quality and diversity of the generated images, respectively (see Section 4.4.1). This quantitative evaluation shows that our proposed pipeline achieves high identification rates while maintaining the lowest re-identification rate amongst the evaluated methods. Qualitative results are shown in Section 4.4.2. Even though we achieved satisfying results in every image-related metric that we computed - proving the effectiveness of our pipeline -, the baseline network performed better than our proposed method for all video-related metrics. In fact, the quality of our generated images worsens both when the subject is not aligned and when the background is not homogeneous. When the former condition applies, the segmentation mask generated in the preprocessing phase may not correspond exactly to the shape and pose of the original face, and this has a negative effect both on the shape anonymization network and on the background inpainting step; and it is precisely the inpainting step that represents the weak point of the entire pipeline. Although it is necessary, as explained in Section 3.2, it is extremely challenging because the information provided by the background may not be sufficient for the task. This has a negative effect especially when anonymizing a video sequence, as the loss of detail due to background inpainting results in the loss of temporal consistency between frames.

As future work, a new method to better blend the anonymized faces with the original background would drastically improve the quality of the generated images.

List of Figures

List of Tables

Bibliography

[1] M. Maximov, I. Elezi, and L. Leal-Taixé, "CIAGAN: conditional identity anonymization generative adversarial networks," *CoRR*, vol. abs/2005.09544, 2020.

[2] Q. Sun, L. Ma, S. J. Oh, L. V. Gool, B. Schiele, and M. Fritz, "Natural and effective obfuscation by head inpainting," *CoRR*, vol. abs/1711.09001, 2017.

[3] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

[4] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015.

[5] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, and Z. Wang, "Multi-class generative adversarial networks with the L2 loss function," *CoRR*, vol. abs/1611.04076, 2016.

[6] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved training of wasserstein gans," 2017.

[7] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *CoRR*, vol. abs/1606.03657, 2016.

[8] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *CoRR*, vol. abs/1611.07004, 2016.

[9] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *CoRR*, vol. abs/1812.04948, 2018.

[10] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," *CoRR*, vol. abs/1912.04958, 2019.

[11] M. S. Ryoo, K. Kim, and H. J. Yang, "Extreme low resolution activity recognition with multi-siamese embedding learning," *CoRR*, vol. abs/1708.00999, 2017.

[12] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," *CoRR*, vol. abs/1911.08348, 2019.

[13] H. Hukkelås, R. Mester, and F. Lindseth, "Deepprivacy: A generative adversarial network for face anonymization," *CoRR*, vol. abs/1909.04538, 2019.

[14] Z. Ren, Y. J. Lee, and M. S. Ryoo, "Learning to anonymize faces for privacy preserving action detection," *CoRR*, vol. abs/1803.11556, 2018.

[15] Q. Sun, A. Tewari, W. Xu, M. Fritz, C. Theobalt, and B. Schiele, "A hybrid model for identity obfuscation by face replacement," *CoRR*, vol. abs/1804.04779, 2018.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[17] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, 2005.

[20] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C. Chang, M. G. Yong, J. Lee, W. Chang, W. Hua, M. Georg, and M. Grundmann, "Mediapipe: A framework for building perception pipelines," *CoRR*, vol. abs/1906.08172, 2019.

[21] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *CoRR*, vol. abs/1801.07892, 2018.

[22] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[23] D. Yang, S. Hong, Y. Jang, T. Zhao, and H. Lee, "Diversity-sensitive conditional generative adversarial networks," *CoRR*, vol. abs/1901.09024, 2019.

[24] Q. Mao, H. Lee, H. Tseng, S. Ma, and M. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," *CoRR*, vol. abs/1903.05628, 2019.

[25] R. Liu, Y. Ge, C. L. Choi, X. Wang, and H. Li, "Divco: Diverse conditional image synthesis via contrastive generative adversarial network," *CoRR*, vol. abs/2103.07893, 2021.

[26] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to detect manipulated facial images," in *International Conference on Computer Vision (ICCV)*, 2019.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch:

An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, pp. 8024–8035, Curran Associates, Inc., 2019.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[30] Y. Movshovitz-Attias, A. Toshev, T. K. Leung, S. Ioffe, and S. Singh, "No fuss distance metric learning using proxies," *CoRR*, vol. abs/1703.07464, 2017.

[31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *CoRR*, vol. abs/1603.08155, 2016.

[32] dechunwang, "dechunwang/ssh-pytorch: First release," Nov. 2018.

[33] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," *CoRR*, vol. abs/1503.03832, 2015.

[34] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, inception-resnet and the impact of residual connections on learning," *CoRR*, vol. abs/1602.07261, 2016.

[35] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017.

[36] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *CoRR*, vol. abs/1801.03924, 2018.

[37] M. Chu, Y. Xie, J. Mayer, L. Leal-Taixe, and N. Thuerey, "Learning Temporal Coherence via Self-Supervision for GAN-based Video Generation (TecoGAN)," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, 2020.