Adversarial Identity Injection for Semantic Face Image Synthesis

Giuseppe Tarollo[1], Tomaso Fontanini[1], Claudio Ferrari[1], Guido Borghi[2], Andrea Prati[1]

[1] Department of Engineering and Architecture, University of Parma, Parma, Italy

[2] Department of Computer Science and Engineering, University of Bologna, Cesena, Italy

{claudio.ferrari2, tomaso.fontanini, andrea.prati}@unipr.it, guido.borghi@unibo.it

Abstract

Nowadays, deep learning models have reached incredible performance in the task of image generation. Plenty of literature works address the task of face generation and editing, with human and automatic systems that struggle to distinguish what's real from generated. Whereas most systems reached excellent visual generation quality, they still face difficulties in preserving the identity of the starting input subject. Among all the explored techniques, Semantic Image Synthesis (SIS) methods, whose goal is to generate an image conditioned on a semantic segmentation mask, are the most promising, even though preserving the perceived identity of the input subject is not their main concern. Therefore, in this paper, we investigate the problem of identity preservation in face image generation and present an SIS architecture that exploits a cross-attention mechanism to merge identity, style, and semantic features to generate faces whose identities are as similar as possible to the input ones. Experimental results reveal that the proposed method is not only suitable for preserving the identity but is also effective in the face recognition adversarial attack, i.e. *hiding a second identity in the generated faces.*
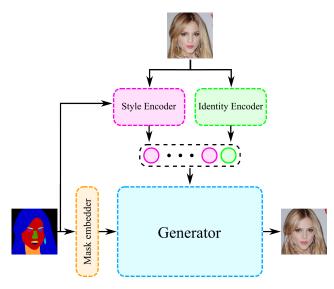
Figure 1. Overview of the proposed architecture. Starting from a face image, style and identity features are extracted through the encoders (Style Encoder ($\mathcal{E}_s$), Identity Encoder ($\mathcal{E}_{id}$) and Mask Embedder ($\mathcal{E}_m$)) and used by the Generator ($\mathcal{G}$), together with the semantic segmentation mask to generate the final image.

1. Introduction

In recent years, deep learning models have reached outstanding results in image generation, with several architectures standing out in the field of human face generation [17, 18, 28] and editing [1, 40, 44]. Indeed, through such powerful models, both humans and automated systems struggle to distinguish between real and generated face images [13, 29]. However, even the most realistic generative model has difficulty in preserving the perceived identity of the generated subject after reconstructing or manipulating a real face image of a specific individual [7, 24, 34]. Whereas this aspect is often neglected or partially discussed in related works, preserving the perceived identity is crucial to make synthetic data exploitable in biometrics applications.

Thus, in this paper, we investigate a possible solution to maximize the identity preservation property without sacrificing the generation quality, with a specific focus on face editing models. In this context, one of the most effective techniques to perform this task is through Semantic Image Synthesis (SIS) [43]. The goal of SIS is to generate realistic face images conditioning the model with a semantic mask, *i.e.* an image in which each pixel represents a semantic class *e.g.* hair, eyes, or mouth. Therefore, the semantic mask is a key element in defining the final shape of the edited face.

The clearest advantage of SIS methods is that the semantic mask can be used to learn explicit mappings between each semantic class and its style *i.e.* texture. In doing so, modern SIS methods can independently generate, control, or manipulate the style of local face regions [11, 12, 36, 46]. Whereas most SIS methods aim at learning such mapping to generate new images conditioned on the mask (*noise-*

based) [20, 25], some literature methods such as [12, 46] focus on extracting styles from a real (*reference*) image in order to map them in the corresponding semantic regions (*reference-based*). In this manner, it is specifically possible to perform editing of real images, for example by changing the hair color, makeup, and other different attributes in the generated samples. In this particular scenario, where the edited face belongs to a real individual, the ability to preserve the perceived identity is of utmost importance both if presented to a human observer or an automatic face recognition system [26].

Unfortunately, we observe that in semantic-based methods, especially for reference-based ones, the identity preservation of the edited face is not taken into account (*e.g.* [36, 46]). The large majority of available models neglect this aspect and grasp the input subject identity only to a limited extent, so lacking the ability to preserve it in the edited face as we show in our experimental evaluation (see Table 1). As a consequence, state-of-the-art Face Recognition (FR) systems [6, 30] would struggle to match the identity of the reconstructed face with that in the input.

Therefore, in this paper, we propose a solution to inject the identity information into a reference-based semantic image synthesis architecture. In particular, the proposed architecture builds upon that proposed in [12], and is composed of four different modules, as depicted in Figure 1: the Style Encoder and Identity Encoder models are responsible for extracting style and identity features, respectively, from the input face image. These features are then concatenated and fed as input to the Generator, responsible for the image generation. The Generator receives as input also the output of the Mask Embedder, which embeds the semantic information of the mask through a fully connected network; style, identity, and semantic information are finally merged through a cross-attention mechanism for face generation.

By exploiting the versatility of cross-attentions, we are able to condition the image generation with high-level information such as the identity, in addition to low-level style features, ultimately improving the identity similarity with respect to the input face. Nonetheless, another noticeable feature that arises with this design choice is the ability to change the identity embedding. In [12], the model can be used to swap specific style embeddings, so to perform "local" style transfer, even if not explicitly trained to do so. Thus, we expect our design to let us change the identity embedding, so conditioning the generation of a face belonging to a subject A with the identity of another subject B.

Therefore, when injecting the identity embedding of the input face image, our model helps in preserving its perceived identity during the whole generation process. As a result, the generated face presents an identity that is closer to the input one, which can be appreciated both visually and quantitatively when it is presented to a face recognition model. Conversely, we can also use our method to swap the identity between two different subjects. In other words, we can concatenate the style embeddings of a subject A with the identity embedding of another individual B. As a result, the generated face qualitatively looks like A, but it actually conceals the identity embedding of B. We wish to highlight here that, despite looking conceptually similar, this paradigm is different than common face-swapping methods [21]. In fact, face swapping approaches aim at changing the perceived identity of a subject in a way such that a human observer easily recognizes the *new* identity *i.e.* the one that was swapped. Differently, our solution "hides" the identity in a way that a human observer can hardly tell the difference, but at the same time, it will fool recognition algorithms. In both tests, the proposed method achieves competitive results, overcoming the large part of literature methods. To summarize, the main contributions of this paper are listed in the following:

- We explore the use of attention-based mechanism to merge identity, semantic and style information into the generation process of Semantic Image Synthesis (SIS);
- We test the proposed method in the task of identity preservation, and show our solution is promising in preserving the input identity during the generative process.
- We investigate the use of the proposed architecture in the task of adversarial attacks on face recognition, presenting an alternative pathway to achieve this goal.

2. Related Work

Semantic Image Synthesis. SIS models aim at generating images starting from a semantic mask. Several approaches were proposed to solve this task.

Firstly, SPADE [25] proposed a spatially adaptive denormalization module to modulate the activations with semantic information. Later, MaskGAN [20] proposed a method to manipulate human faces with semantic masks. MaskGAN was introduced simultaneously with SEAN [46] which allowed to extract semantic styles from a reference images and apply them to the generated samples. Along this line, multiple methods were developed, like CLADE [37] and INADE [36], which introduced the concept of instances in the semantic masks. Recently, Semantic-StyleGAN [32] allowed to control the generation of StyleGAN images [17] through semantic information. Finally, Semantic Diffusion [39] adapted a diffusion model adding SPADE normalization layers in order to control the generation with semantic masks. Methods for semantic image synthesis are domain agnostic, meaning that they can be applied to several different scenarios *e.g.* faces, outdoor scenes, objects. For this reason, domain-specific information such as the perceived identity as in the human face domain are neglected. Differently, the proposed model is specifically tailored for the human face domain.

Adversarial attack on face recognition. Deep learning architectures have been proven susceptible to adversarial attacks [35], which are slightly-perturbed versions of the original examples that eventually trick the networks into outputting a wrong prediction. The peculiar characteristic of adversarial examples is that they are hardly distinguishable from their "clean" counterpart for the human eye, because they have been specifically optimized to have the minimum possible perturbation. Adversarial attacks can be broadly divided into two categories: white-box or black-box. The former ones are the most difficult to counteract, as they assume full knowledge of the attacked system including model parameters, so that the attacker can exploit the model's gradient to craft the adversarial example. Black-box attacks are instead more difficult to craft since they only have access to the classifier prediction. Both can be also targeted or un-targeted. Targeted attacks aim at making the attacked model predict a *specific class*, while un-targeted methods only care for making the model predict a wrong class. Despite being applicable whenever a classification task is involved, attacks can also be designed for specific domains and tasks, such as face recognition. Attacks on face recognition can be divided in multiple categories. Firstly, gradient-based methods like [9, 14, 23] aim at adding perturbations in the pixels, but suffer from common denoising models. Next, patch-based methods focus on printing on the images adversarial hat [19] or glasses [31], but in this case the attack is easily spottable. Finally, stealthy-based methods inject the adversarial attack in the face attributes [15, 27, 42]. On the other side, our system treats identity information as an additional style and therefore the generated samples will not be changed by adding glasses or makeup making the attack almost invisible. Recently, a new type of attack was proposed by Li *et al.* [22] which utilizes an additional Attribute Recognition (AR) task to improve the attacking transferability.

3. Identity-conditioned Image Synthesis

The proposed system builds upon the very recent SIS model proposed by Fontanini *et al.* [12] named CA^2SIS. We chose this specific architecture as, differently from the vast majority of SIS models that employ SPADE layers *e.g.* [25, 46], it uses spatial transformer blocks to condition the image generation with style features extracted from a RGB reference.

The versatility of the cross-attention layers included in the spatial transformer blocks allowed us to design an alternative solution to inject identity information into the generator. This could not be done with standard SIS models based on SPADE as they require an explicit spatial mapping of the style features. While this mapping is straightforward when style features represent a well-defined class *e.g.* hair, eyes, that becomes challenging if using features related to

high-level concepts such as identity. In fact, there is no clear prior on such information; in other words, which face parts influence the the identity perception the most? To what extent? Whereas some literature works do provide some hints in this regard [10, 33], what contributes to recognizing an individual is actually a combination of facial features. This makes SPADE-like layers difficult to use. Cross-attentions instead provide a nice alternative since the spatial mapping is implicitly learned by the attention mechanism. This allows the model to learn how to optimally map the identity information into the generated face image without requiring prior intervention (Fig. 4).

3.1. Architecture

The objective of the original architecture is that of generating a photo-realistic image given a semantic segmentation mask and a reference image. It is composed by three modules: a Cross-Attention Generator $\mathcal{G}$, a Multi-Resolution Style Encoder $\mathcal{E}_s$, and a Mask Embedder $\mathcal{E}_m$. Necessary details are provided in the paragraphs below so to make the paper self-contained, but we refer the reader to [12] for a detailed description.

Mask Embedder. Let a semantic mask be a C-channel image $\mathcal{M} \in \mathbb{N}^{C \times H \times W}$, where each channel $\mathcal{M}_j$ is a binary image encoding the pixel-wise spatial location of a specific class *e.g.* eyes, lips, hair. The module is an MLP that receives $\mathcal{M}$ and outputs C embeddings of size 256, one for each semantic class. These are reshaped to form 16×16 feature maps, and then stacked to form a mask descriptor $m_x = \mathcal{E}_m(\mathcal{M}) \in \mathbb{R}^{16 \times 16 \times C}$. The descriptor m_x will be the input to the generator $\mathcal{G}$.

Style Encoder. The Style Encoder $\mathcal{E}_s$ extracts style features from the input RGB reference images x. Specifically, it is equipped with Grouped Convolutions, Group Normalization layers and skip connections, and is designed to extracts a style code $s_c \in \mathbb{R}^{256}$ for each semantic class c by exploiting the mask $\mathcal{M}$. The style codes are concatenated to form a combined style code of size $256 \times C$, *i.e.* $s_x = \mathcal{E}_s(x) \in \mathbb{R}^{1280}$.

Generator. Finally, the generator $\mathcal{G}$ receives the mask descriptor as input and the style codes as condition, ultimately outputting a realistic image having the shape defined by the semantic mask, and the styles of the reference image, that is $\hat{x} = \mathcal{G}(m_x, s_x)$. More in detail, style codes are injected in the cross-attention (CA) layers of the Generator that are defined as follows:

$$CA(Q, K, V) = \mathcal{S}\left(\frac{QK^T}{\sqrt{d}}\right)V \qquad (1)$$

where $Q = W_Q^{(i)} \cdot \phi^{(i)}$ is obtained from the projection of the flattened features $\phi^{(i)}$ of previous convolutional layers,

while $K = W_K^{(i)} \cdot \mathcal{E}_s(x_i)$ and $V = W_V^{(i)} \cdot \mathcal{E}_s(x_i)$ are computed from the style codes. The Generator is also paired with a Discriminator $\mathcal{D}$ to exploit the adversarial loss.

3.2. Identity Module

Despite the original model itself can already preserve the perceived identity of the input face x way better than other approaches, it does not allow for manipulating or changing it explicitly, although it is possible to swap style codes of a different face image y for generating diverse images *i.e.* $\hat{x} = \mathcal{G}(m_x, s_y)$. Given that the goal of this work is to explore an alternative pathway to preserve the identity of a subject A or to conceal the identity of an individual A into a face image of another subject B without making the change being perceivable, we augmented this architecture by adding a pre-trained face recognition model $\mathcal{E}_{id}$. This module is used to extract an identity embedding from the input face x *i.e.* $id_x = \mathcal{E}_{id}(x)$, which is then used as additional style code for the generator (see Fig. 1). The idea is that in doing so, we can both increase the capability of the generator to preserve the original identity, while also being able to swap the identity code so to condition the generation with the identity embedding of a different individual. Ultimately, this leads the generated image to qualitatively appear as the original subject, at the same time concealing the identity information of a different subject.

3.3. Identity Preservation Loss

In order to inject identity information into the Generator, we treat the identity embedding as an additional style code. In particular, we employed a pretrained face recognition model to extract an identity embedding from a reference image x of some subject i. The embedding is then concatenated to the style codes extracted from $\mathcal{E}_s$, and the new identity-style representation is injected in the cross-attention layers of $\mathcal{G}$. Formally, we indicate this addition as $\hat{x} = \mathcal{G}(m_{x_i}, s_{x_i}, id_{x_i})$ During training, an *identity preservation loss* $\mathcal{L}_{id}$ is employed in order to force the model to utilize this additional information. $\mathcal{L}_{id}$ is as follows:

$$\mathcal{L}_{id} = 1 - cos\left(\mathcal{E}_{id}\left(\mathcal{G}\left(m_{x_i}, s_{x_i}, id_{x_i}\right)\right), \mathcal{E}_{id}(x_i)\right) \quad (2)$$

where $\mathcal{G}(m_{x_i}, s_{x_i}, id_{x_i})$ is the generator output starting from reference mask m_{x_i}, style codes s_{x_i} and identity embedding id_{x_i}. The term cos is the cosine similarity function. This loss forces the generated samples to match the identity embeddings that are injected in the model during training. At inference time, the style codes and identity can be extracted from two difference images x_i and x_j resulting in a sample having the same appearance as x_i but that will be recognized as x_j by a FR network.

3.4. Training Objective

In addition to the identity preservation loss, during training, we employ a set of losses as in [12]. More in detail, we implemented an adversarial loss $\mathcal{L}_{adv}$, a feature matching loss $\mathcal{L}_{FM}$ [38] and a perceptual loss $\mathcal{L}_{prc}$ [16]. The full training objective becomes:

$$\mathcal{L}_{tot} = \mathcal{L}_{adv} + \lambda_{FM}\mathcal{L}_{FM} + \lambda_{prc}\mathcal{L}_{prc} + \lambda_{id}\mathcal{L}_{id} \quad (3)$$

where $\lambda_{FM}, \lambda_{prc}$ and λ_{id} are the weights for feature matching, perceptual and identity preservation loss, respectively. More in detail, they are set during training as follows: $\lambda_{FM} = 10, \lambda_{prc} = 10$ and $\lambda_{id} = 10$.

4. Experiments

The proposed approach is thoroughly validated through a set of experiments aimed at verifying the ability of our model to *(i)* improve the identity preservation when it is applied to the task of reconstructing a face image; *(ii)* hide the identity of another individual while maintaining the face visually unchanged as in an impersonation attack. We employ three different FR models during evaluation: IR152 [6], MobileFace [3] and FaceNet [30]. The pre-trained weights for these models are the same as those used in [42].

In order to verify the robustness of our solution to different FR networks, we train two different models: one employs a pre-trained FaceNet model (referred to as Ours-FaceNet) using the implementation and weights from the repository[1]. The other uses a pre-trained ArcFace model [6] (Ours-ArcFace) taken from the InsightFace repository[2]. Note that the pre-trained weights of these models differ from those used for evaluation.

We carry out the experimental validation on the CelebMask-HQ dataset [20], which comprises 30k face images, of which 28k for training, and 2k for testing. Each image is paired with its own semantic segmentation mask, comprising 19 semantic classes.

4.1. Identity Preservation

Injecting the identity embedding in our system has an immediate positive effect: the perceived identity is better preserved during the image generation. This is incidentally a critical issue in generative models for human face generation which often lack this property. To prove that, in Table 1, we report the average cosine similarity obtained between original and reconstructed faces for different SIS methods on several FR models, which is computed as:

$$C = \frac{1}{N}\sum_{i=1}^{N} \cos\left(\mathcal{E}_{id}\left(x_i\right), \mathcal{E}_{id}\left(\hat{x}_i\right)\right) \quad (4)$$

---

[1] https://github.com/timesler/facenet-pytorch
[2] https://github.com/nizhib/pytorch-insightface

Method	IR152 ↑	MobileFace ↑	FaceNet ↑	FID ↓
SEAN [46]	0.51	0.73	0.74	18.7
V-INADE [36]	0.23	0.49	0.43	18.3
Semantic StyleGAN [32]	0.38	0.66	0.62	26.8
CA^2SIS [12]	0.52	0.75	0.74	**15.8**
Ours - FaceNet	**0.64**	<u>0.80</u>	*0.90*	18.1
Ours - ArcFace	<u>0.62</u>	**0.83**	<u>0.81</u>	<u>16.5</u>

Table 1. Cosine similarity metric and FID comparison between original x and reconstructed $\hat{x}$ faces when conditioning with its own identity embedding id_x. Injecting the identity information increases the similarity to a great extent for various FR models (see Sect. 4.1). **Bold**=best result, <u>underlined</u>=second best. *Italic* font indicates the validation FR architecture is the same as that used to train our model, but the pre-trained weights differ.

where $\hat{x}_i = \mathcal{G}(m_{x_i}, s_{x_i}, id_{x_i})$ is the reconstructed face. A higher cosine similarity score implies an automatic face recognition system will likely verify the two faces as belonging to the same individual. Maintaining higher similarity scores is crucial since face verification systems rely on fixed thresholds to reject or accept face image pairs; the larger the score for genuine pairs, the less the number of false negatives that is returned from the system.

As expected, state-of-the-art SIS methods struggle to maintain the identity in the generated results. This is especially true for StyleGAN-based methods that, in order to reconstruct and manipulate a given input, rely on GAN-inversion techniques [45] to find an embedding in the StyleGAN latent space that is as similar as possible to the real image. On the other side, our method exhibits superior performance in identity preservation both w.r.t. to the baseline model CA^2SIS [12] and the state of the art.

In Table 1 we also report the Frechet Inception Distance (FID), which measures the similarity between real and generated data distributions. This is usually employed for validating the quality of the fake samples generated by a model. This is intended to verify that, by conditioning the generation with an identity embedding, we do not compromise the realism of the generated faces. Overall, this conditioning actually negatively impacts on the FID score to a little extent. Although we cannot prove this formally, we believe this is likely due to the way in which FID is calculated: the additional information carried in the identity embedding that is mapped into the pixel space by the generator might shift the fake image distribution, hindering such score.

Nevertheless, even if FID is widely employed to compare different image generation methods, it has also drawn a several criticism [4] and should therefore be always paired with a qualitative evaluation of the results. For this reason, in Fig. 2 we report a qualitative example of the effect that the identity injection elicits to our model. At first glance, the results produced by the model with and without identity preservation seem almost identical. Things change if
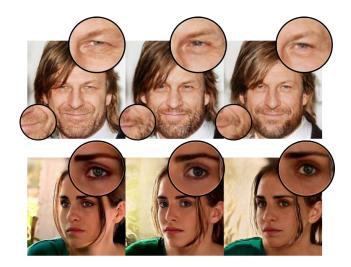

Source   CA^{2}SIS   **Ours**

Figure 2. Qualitative comparison between the original CA^2SIS model [12] and the proposed architecture with cross attention-based identity injection (see Sect. 4.1). Identity-related details such as the color of the eyes, the eyebrows, and mouth shape or subtle details such as the teeth are better preserved when conditioning with identity embedding. Better seen on screen.

some details are zoomed in and better highlighted. In particular, the eyes and mouth region are the most affected by the identity injection, and more closely resemble the original image. On the other side, the identity information looks almost completely hidden in the generated samples, opening the way to inconspicuous identity swapping as it will be presented in the next section.

4.2. Adversarial Attack on Face Recognition

In this section, we show that the proposed architectural change leads to the possibility of generating a face image of some subject i (*attacker*) conditioned with the identity embedding of another individual j (*target*) *i.e.* $\hat{x}_{i \to j} = \mathcal{G}\left(m_{x_i}, s_{x_i}, id_{x_j}\right)$. When doing so, this will be almost completely hidden in the output image, but FR models will recognize the image as belonging to identity j. This is very similar to an adversarial attack as in [15], where the goal is to exploit semantic clues to attack state-of-the-art FR systems. At the same time, there are some key differences: a) our model is not explicitly trained to perform adversarial attacks, but simply for reconstructing an image given its semantic mask, styles and identity; b) it does not require additional training for each attack, but simply swaps the identities of target and attacker at inference time; and, finally, c) the proposed model does not hide the identity into a specific attribute, like eyeglasses as in [15], but treats it as an additional style that is applied globally to the attacker face.

Given this task similarity, we evaluate the performance using standard metrics in the field and adopt the Attack Success Rate (ASR) as metric, which is computed as:

$$\frac{1}{N}\sum_{i=1}^{N} cos\left(\mathcal{E}_{id}\left(\mathcal{G}\left(m_{x_i}, s_{x_i}, id_{x_j}\right)\right), \mathcal{E}_{id}\left(x_i\right)\right) > \tau \quad (5)$$

where m_{x_i}, s_{x_i} are the mask and style codes associated to the attacker i, $id_{x_j} = \mathcal{E}_{id}(x_j)$ is the identity embedding associated to the target identity j. Finally, τ value is taken from [15] and is set considering a False Acceptance Rate (FAR) of 0.01 w.r.t. the attacked FR system. Briefly, this metric quantifies how many times a face recognition systems accepts the input face image as belonging to the target identity and not the attacker.

In Fig. 3 several qualitative results of the adversarial attack are presented: an *attacker* picture is injected with the identity of a *target* face. Indeed, the difference between the attacker face generated with the correct identity and the one generated with the target identity is negligible and a human eye would almost certainly fail when asked to recognize which one was forged with the incorrect identity. The last column of Fig. 3, where heatmaps of the pixel-wise $L1$ difference between the two different generated samples are shown, further highlighting this characteristic, exhibiting values close to zero almost everywhere. Still, at a closer look, some minor changes in the swapped samples can be appreciated. In particular, the eyebrows and eyes shape is slightly altered (see third and fourth rows) as well as the nose appearance (see first row). In addition, sometimes also the eye color is affected (see last row). This is in line with the considerations made in [10] that the identity information is concentrated in the eyes and eyebrows areas of the face.

In Fig. 4, we show the cross-attention maps estimated by

| Attacker | Target | Reconstr. | Id. Swap | Diff. |

Figure 3. Results of our architecture obtained as described in Sect. 4.2. The first column is the attacker, the second column is the target, the third column is the reconstruction result of the attacker using the correct identity embedding, fourth column is the reconstruction result when injecting the identity of the target in the attacker. Finally, the last column represents the pixel difference between the two reconstruction results, highlighting that the identity information is effectively concealed in the manipulated face.

the model when injecting identity information, in order to visualize in which face areas the identity gets mapped by the attention layer. This figure suggests our intuition was correct, that is a high-level complex information such as the identity influences several different regions of the face, making state-of-the-art SIS methods based on SPADE unsuitable for this task.

Finally, in Table 2 an extensive comparison with state-of-the-art methods for adversarial attack on human face recognition is presented. All the numbers are taken from [15], and our results are calculated using the same settings and pre-trained FR models. Our method is able to improve the ASR score of the state-of-the-art, by a great margin in some cases. When the attacked FR architecture is the same as that used to train our model, the improvement is clearly larger. We note this is a fair setting since the architecture is the same, but the pre-trained weights are not. Thus, it resembles a *transferable adversarial attack* (or gray-box) scenario, that is when the attacker is aware of the victim's
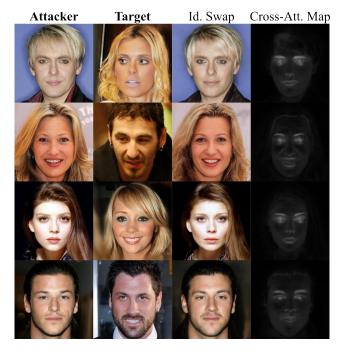
Figure 4. Cross-Attention layer visualization when swapping the identity of a target face to the attacker. The areas that are most affected by the identity injection are the eyes, eyebrows, nose, and mouth. This result suggests the perceived identity is complex information that is carried by several different facial traits.

Method	IR152 ↑	MobileFace ↑	FaceNet ↑
FGSM [14]	2.70	5.10	1.90
PGD [23]	26.00	29.90	3.50
MI-FGSM [8]	26.80	21.70	4.60
C&W [2]	27.30	28.20	3.30
Adv-Hat [19]	2.50	8.40	4.70
Adv-Glasses [31]	4.50	5.60	9.10
Gen-AP [41]	19.50	24.40	15.80
Adv-Face [5]	31.40	36.40	21.60
Adv-Makeup [42]	10.80	14.60	10.50
Semantic-Adv [27]	10.30	19.40	9.00
Adv-Attribute [15]	44.30	_50.20_	31.80
Ours - FaceNet	_48.30_	40.60	***77.60***
Ours - ArcFace	***67.80***	**98.10**	_72.20_

Table 2. ASR comparisons against adversarial attacks methods targeting different models on CelebA-HQ, as detailed in Sect. 4.2. **Bold**=best result, underlined=second best. *Italic* font indicates the attacked FR architecture is the same as that used to train our model, but the pre-trained weights differ.

architecture but does not have access to its internal weights. In the other cases (the attacked FR model is different from that of our architecture), our solution still obtains largely higher ASR score in all the cases except one. Other than demonstrating that our solution is highly effective, it suggests that a significant overlap across recognition models to the relevant facial features useful for recognition occurs. Compared to the state-of-the-art, our results are quite impressive considering that we do not specifically train our system to perform adversarial attacks (nor it requires fine-tuning). Differently, all the compared methods perform a specific optimization.

4.3. Style Transfer Effect on Face Recognition

The SIS architecture, which was adapted to include identity information during generation, extracts a set of styles s_{x_i} via $\mathcal{E}_s$ from each semantic region of an RGB image and maps them to the corresponding semantic class in the input mask (see Sect. 3). This allows the style transfer between a source x_i and a target x_j image by mixing their corresponding style codes s_{x_i} and s_{x_j} *e.g.* hair color.

In this section, we explore how style transfer affects the attacks on FR. More in detail, the objective is to verify if swapping some styles can boost the ASR. To prove this, firstly, we performed style transfer on a set of styles combined with identity swap. In particular, *Skin*, *Eyes*, *Eye-*

brows, *Mouth*, *Hair* and *Full S. Swap* (*i.e.* all the styles are swapped) were chosen. Then, the results for each different style transfer experiment were evaluated using two different metrics: ASR and LPIPS. More in detail, the LPIPS metric was calculated between the images reconstructed with the correct identity and the ones generated with style transfer and identity swap. By doing so, we can measure how much the style transfer alters the overall appearance of the generated samples. Ideally, the best results are those in which the system is able to fool the FR models while maintaining the attack almost invisible to the human eye; in other words,
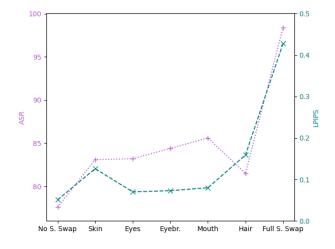

Figure 5. Graph showing different Attack Success Rate (ASR) and LPIPS metric values when swapping different styles along the identity. The style swapping procedure is described in Sect. 4.3.
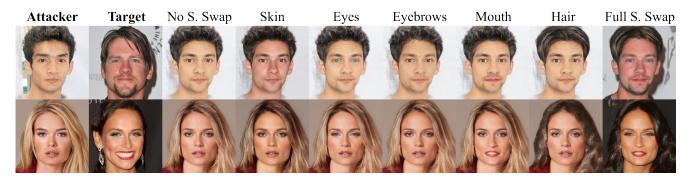
Figure 6. Results of different types of style swaps as described in Sect. 4.3.

we want the ASR to be as high as possible, while the LPIPS remains as low as possible

Quantitative results of this experiment can be seen in Fig. 5, while qualitative results can be seen in Fig. 6. In both figures, the "*No S. Swap*" value represents the baseline, *i.e.* the identity swapping is performed without any style transfer. As expected in this case, the LPIPS value is very low enforcing our claim of inconspicuous attacks. On the other side, when performing style transfer, ASR results increase for every different transferred part. This proves that the capability of pairing identity and style swaps can strengthen the attacks against FR systems giving our model an additional edge w.r.t. current state-of-the-art.

Interestingly, the parts that increased the ASR results the most are also the ones with the lower LPIPS. Indeed, they are also the ones that contain the most identity information. More in detail, the single semantic classes that obtained the highest ASR results were *Eyes*, *Eyebrows* and *Mouth*. This is in line with the results presented in Fig. 2. These findings prove that is possible to combine style and identity swaps maintaining the attack almost invisible to the human eye, which is of paramount importance for this kind of system. Finally, when swapping all the styles together ("*Full S. Swap*" in the figure), the ASR reaches 98.4% but, at the same time, LPIPS value is the highest.

5. Conclusions and Ethical Concerns

In this paper, we proposed a novel Semantic Image Synthesis (SIS) method for image manipulation that also employs identity information during the generation process. The identity injection into the model is based on the identity embedding, extracted from a pre-trained FR system, as an additional style that is concatenated to the other styles obtained by a style encoder from a reference RGB image. Then, a cross-attention mechanism is used in the generator. We observe the proposed identity injection procedure has two main effects: firstly, when the same identity of the input subject is used, it greatly improves the identity preservation during generation. Secondly, if the identity is swapped (*i.e.*

the injected identity is different with respect to the input one), the model is able to perform an adversarial attack to FR systems hindering their results. Extensive experiments on these two contributions were performed proving the effectiveness of the proposed architecture.

As a future work, we plan to further develop the identity injection mechanism, so as to have a strategy for making the identity of the second subject in the final generated image visible or not. This is important both for controlling the extent of the identity injected into the system and for the development and study of biometric systems based on facial identification, such as face swapping and morphing.

Lastly, we are aware that these types of systems could be used in malicious or criminal ways. On the other side, we strongly believe that studying and publicly sharing results in this field can increase awareness of the use of these systems in the academic community (and beyond), stimulate the development of new countermeasures, and lead to the creation of new datasets for training future systems.

References

[1] Guido Borghi, Annalisa Franco, Gabriele Graffieti, and Davide Maltoni. Automated artifact retouching in morphed images with attention maps. *IEEE Access*, 9:136561–136579, 2021. 1

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 7

[3] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobile-facenets: Efficient cnns for accurate real-time face verifica-

tion on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer, 2018. 4

[4] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079, 2020. 5

[5] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 7

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2, 4

[7] Nicolò Di Domenico, Guido Borghi, Annalisa Franco, and Davide Maltoni. Face restoration for morphed images retouching. In *Proceedings of the 12th International Workshop On Biometrics And Forensics (IWBF)*, 2024. 1

[8] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 7

[9] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019. 3

[10] Claudio Ferrari, Matteo Serpentoni, Stefano Berretti, and Alberto Del Bimbo. What makes you, you? analyzing recognition by swapping face parts. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 945–951. IEEE, 2022. 3, 6

[11] Tomaso Fontanini, Claudio Ferrari, Massimo Bertozzi, and Andrea Prati. Automatic generation of semantic parts for face image synthesis. In *International Conference on Image Analysis and Processing*, pages 209–221. Springer, 2023. 1

[12] Tomaso Fontanini, Claudio Ferrari, Giuseppe Lisanti, Massimo Bertozzi, and Andrea Prati. Semantic image synthesis via class-adaptive cross-attention. *arXiv preprint arXiv:2308.16071*, 2023. 1, 2, 3, 4, 5

[13] Anjith George, Zohreh Mostaani, David Geissenbuhler, Olegs Nikisins, André Anjos, and Sébastien Marcel. Biometric face presentation attack detection with multi-channel convolutional neural network. *IEEE transactions on information forensics and security*, 15:42–55, 2019. 1

[14] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3, 7

[15] Shuai Jia, Bangjie Yin, Taiping Yao, Shouhong Ding, Chunhua Shen, Xiaokang Yang, and Chao Ma. Adv-attribute: Inconspicuous and transferable adversarial attack on face recognition. *Advances in Neural Information Processing Systems*, 35:34136–34147, 2022. 3, 6, 7

[16] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 4

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 1, 2

[18] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[19] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826. IEEE, 2021. 3, 7

[20] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 2, 4

[21] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2

[22] Zexin Li, Bangjie Yin, Taiping Yao, Junfeng Guo, Shouhong Ding, Simin Chen, and Cong Liu. Sibling-attack: Rethinking transferable adversarial attacks against face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24626–24637, 2023. 3

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3, 7

[24] Masahiro Mori, Karl F MacDorman, and Norri Kageki. The uncanny valley [from the field]. *IEEE Robotics & automation magazine*, 19(2):98–100, 2012. 1

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2, 3

[26] Stefano Pini, Guido Borghi, Roberto Vezzani, Davide Maltoni, and Rita Cucchiara. A systematic comparison of depth map representations for face recognition. *Sensors*, 21(3): 944, 2021. 2

[27] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 19–37. Springer, 2020. 3, 7

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of*

the *IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1

[29] Ulrich Scherhag, Christian Rathgeb, Johannes Merkle, Ralph Breithaupt, and Christoph Busch. Face recognition systems under morphing attacks: A survey. *IEEE Access*, 7: 23012–23026, 2019. 1

[30] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 2, 4

[31] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pages 1528–1540, 2016. 3, 7

[32] Yichun Shi, Xiao Yang, Yangyue Wan, and Xiaohui Shen. Semanticstylegan: Learning compositional generative priors for controllable image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11254–11264, 2022. 2, 5

[33] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006. 3

[34] Sanjana Sinha, Sandika Biswas, and Brojeshwar Bhowmick. Identity-preserving realistic talking face generation. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE, 2020. 1

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 3

[36] Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Bin Liu, Gang Hua, and Nenghai Yu. Diverse semantic image synthesis via probability distribution modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7962–7971, 2021. 1, 2, 5

[37] Zhentao Tan, Dongdong Chen, Qi Chu, Menglei Chai, Jing Liao, Mingming He, Lu Yuan, Gang Hua, and Nenghai Yu. Efficient semantic image synthesis via class-adaptive normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4852–4866, 2021. 2

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 4

[39] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models. *arXiv preprint arXiv:2207.00050*, 2022. 2

[40] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021. 1

[41] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11845–11854, 2021. 7

[42] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Advmakeup: A new imperceptible and transferable attack on face recognition. *arXiv preprint arXiv:2105.03162*, 2021. 3, 4, 7

[43] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Jiahui Zhang, Shijian Lu, Lingjie Liu, Adam Kortylewski, Christian Theobalt, and Eric Xing. Multimodal image synthesis and editing: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[44] Shangchen Zhou, Kelvin C.K. Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, 2022. 1

[45] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *European conference on computer vision*, pages 592–608. Springer, 2020. 5

[46] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 1, 2, 3, 5