

FIVA: Facial Image and Video Anonymization and Anonymization Defense

Felix Rosberg^{1,2} Eren Erdal Aksoy² Cristofer Englund² Fernando Alonso-Fernandez²

¹Berge Consulting, Gothenburg, Sweden ²Halmstad University, Halmstad, Sweden

felix.rosberg@berge.io, {eren.aksoy, cristofer.englund, fernando.alonso-fernandez}@hh.se



Figure 1: Identity consistent anonymization generated by *FIVA* (left). Anonymization, reconstruction attacks and defense against reconstruction attacks (right). Reconstruction attacks allow for facial recognition to succeed as they manage to undo the anonymization. This is preventable by applying a small amount of noise to the anonymized image pixels, causing the reconstruction attack model to collapse.

Abstract

In this paper, we present a new approach for facial anonymization in images and videos, abbreviated as FIVA. Our proposed method is able to maintain the same face anonymization consistently over frames with our suggested identity-tracking and guarantees a strong difference from the original face. FIVA allows for 0 true positives for a false acceptance rate of 0.001. Our work considers the important security issue of reconstruction attacks and investigates adversarial noise, uniform noise, and parameter noise to disrupt reconstruction attacks. In this regard, we apply different defense and protection methods against these privacy threats to demonstrate the scalability of FIVA. On top of this, we also show that reconstruction attack models can be used for detection of deep fakes. Last but not least, we provide experimental results showing how FIVA can even enable face swapping, which is purely trained on a single

target image.

1 . Introduction

Privacy holds significant importance in various aspects of society, and data collection and storage are no exceptions. The increasing demand and interest in data, coupled with the implementation of recent regulations such as the General Data Protection Regulation, have made data anonymization a necessity. Numerous challenges arise where identity information becomes irrelevant, while attribute information remains crucial. Anonymization techniques aim to obscure, remove, or replace identity information with arbitrary pseudo-identities while preserving essential attribute information. However, obscuring or removing identity information through direct manipulation of the data distribution often results in the loss of significant attributes. For instance, techniques like blurring faces or replacing them with black boxes eliminate vital details such as

eye gaze, pose, and expressions. In contrast, replacement-oriented methods focus on preserving key attributes while altering the identities of individuals.

In this study, we particularly concentrate on replacement-based approaches for anonymizing faces in both images and videos. We propose a novel anonymization method that utilizes target-oriented face-swapping models such as FaceDancer [27] and SimSwap [5] in conjunction with a model based on [9]. It is here worth noting that our method can be implemented by using any existing target-oriented face manipulation model (e.g., [5, 20, 27, 33]) that utilizes identity embeddings for facial manipulation. These models ensure strong consistency across video frames when the employed identity embedding is *stable*. To both enable and enhance this consistency, we introduce a simple and efficient method that not only tracks identities but also samples fake identities. Thus, our proposed method enables robust anonymization and consistency across frames.

Furthermore, our research addresses an important aspect of security, specifically the vulnerability to reconstruction attacks, which has been under-investigated in facial anonymization. In a reconstruction attack, an adversarial model attempts to translate the anonymized face back to the original identity. We hypothesize and provide compelling evidence that anonymization models leave traces in the images that can be exploited for successful reconstruction attacks. To mitigate and verify this threat, we investigate the effectiveness of various noise types, including adversarial noise, uniform noise, and parameter noise, while disrupting the reconstruction attack. Additionally, we present results demonstrating how a reconstruction attack model can be utilized for deep fake detection.

Lastly, our contributions highlight that maximizing the identity distance is detrimental to privacy. This is because, state-of-the-art facial recognition models constrain embeddings to a hyper-sphere, which allows us to easily find the original identity by negating one of the embeddings.

2 . Related Work

In this work, we focus on the direct manipulation of identity information within the data, such as identity masking and identity manipulation methods. Manipulating identity involves masking or altering identity information to preserve privacy. In the context of image and video faces, two common masking approaches are blurring and obscuring with black boxes. While these methods may ensure strong privacy, they directly eliminate valuable information such as eye gaze, potentially affecting the data distribution. Naively training models on this distorted data may lead to the model becoming dependent on the introduced distortions. Therefore, an emerging approach is to leverage advancements in generative models to replace the identity with realistic faces

[4, 9, 13, 19, 14, 21, 22, 26]. There exist various works, such as those by Ma *et al.* [22], Li *et al.* [19], Li and Han [21], and Ren *et al.* [26], that directly employ face modification techniques. However, the current evaluations differ significantly, and none of these works focus on realism in a spatiotemporal context. Gafni *et al.* [9] utilize a face modification autoencoder network with a focus on spatiotemporal consistency. Their approach operates consistently across frames, generating a learned mask for occlusion awareness. However, there is no quantitative evaluation of the temporal consistency, and the network is trained to push the identity away while preserving attribute information, similar to the training approach of the face-swapping method FaceDancer [27]. DeepPrivacy and DeepPrivacy2 [14, 13] employ a U-net-based model trained to inpaint a removed face, conditioned on pose information to preserve pose consistency. However, by completely removing the face, crucial information such as expression and eye gaze is lost. Temporal consistency is also disregarded, resulting in the generation of a new face for frames that differ slightly. Çiftçi *et al.* [4] utilize a face-swapping model, SimSwap [5], which we evaluate in this work using our proposed method. Therefore, our work complements theirs in this context.

Face-swapping techniques have emerged as a promising approach for facial anonymization in recent years, serving as the foundation for the direct manipulation of identity information. Specifically, a target-oriented face-swapping model is employed to facilitate the anonymization process. face-swapping refers to the task of transferring a source face or identity onto a target face. Target-oriented methods, in particular, prove suitable for anonymization purposes as they typically employ identity embeddings from an identity encoder to directly manipulate the identity within the image [5, 20, 27, 33]. Working with these identity embeddings is efficient and straightforward. Source-oriented approaches tend to struggle with lighting, textures and demands an actual fake image, which is costly [2, 24, 25]. We demonstrate an approach that can be attached to an existing target-oriented face-swapping model to allow for both image and video anonymization.

3 . Method

3.1. Network Architecture

FIVA comprises a target-oriented encoder-decoder generator, a pre-trained ArcFace model [8], and the Identity Tracking Module (ITM) as illustrated in Figure 9. We here note that any target-oriented face manipulation model can be employed as the generator, for which ArcFace [8] provides identity conditioning information. ITM is for tracking and sampling fake identities. For clarity and separation from other target-oriented models investigated, we will from here on denote the newly introduced model in this

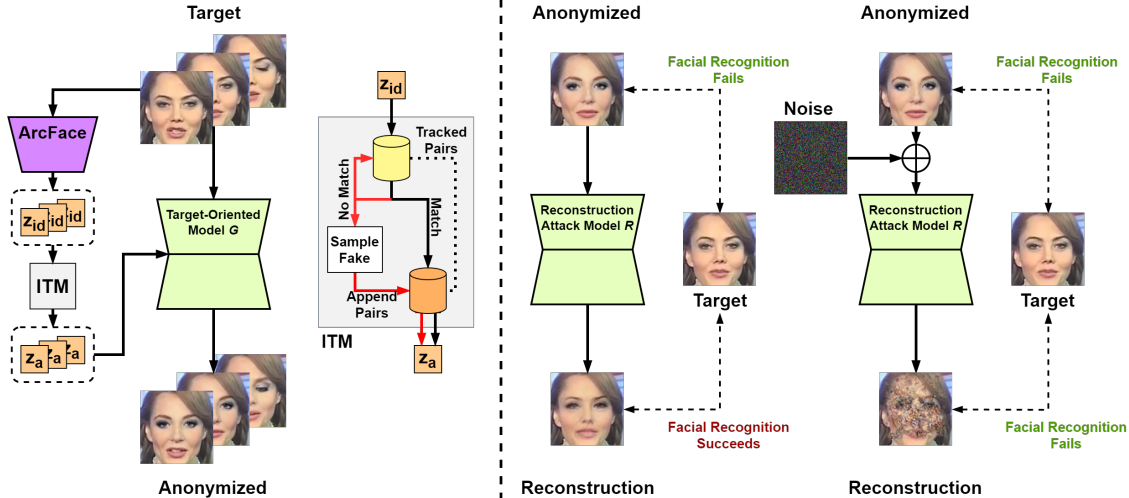


Figure 2: Overview of our anonymizing pipeline: the Identity Tracking Module (ITM) (left) and the implication of our reconstruction attacks (right). Here, z_{id} and z_a are the identity vector and the sampled fake identity vector, respectively. ITM checks if an identity exists and returns a corresponding fake identity (*Match*). If not (*No Match*), as indicated by the red arrows, we generate and/or save a fake identity and store the pair of vectors. We determine a match with the cosine distance and a manually chosen threshold (See Algorithm 1). The generator G , in this case, could be any target-orient facial manipulation model [27, 5, 33, 34, 20, 19, 9]. The reconstruction attack model R shown on the right is learned in a black-box setting (No access to G). The model R learns to reverse the transformation performed by G , thus, allowing for successful face recognition. By applying a small amount of noise to the anonymized image pixels, the reconstruction attacks can be defended, thus, yielding failed face recognition.

work as *FIVA*.

Generator (G): Our generator model G is based on the architecture presented in [9]. We apply several important modifications to G , which are motivated by recent advancements in the literature and the need to address certain missing details in [9]. First, the skip connection directly incorporates all feature maps from the encoder. Secondly, the identity information is broadcasted and concatenated with the bottleneck layer, preserving the spatial dimensions. The fully connected layers in the bottleneck are directly applied to the feature maps. Each encoder block utilizes a depth-wise convolution with a stride of 2 to downsample the feature maps, while the decoder block employs the pixel shuffle operation [31] for upsampling. Similar to [9], our model generates an anonymized image, along with an automatically learned mask, which is designed to blend seamlessly with the target image (See supplementary material for a detailed overview).

Discriminator: Our discriminator, follows the design principles of discriminators used in FaceDancer [27], HiFiFace [33], and StarGAN-v2 [6]. We employ the non-saturated GAN loss [15, 16, 17] for training.

ArcFace: In order to encourage the generator to deviate from the target image in terms of identity, we condition the bottleneck layer with identity vectors extracted from ArcFace [8], which utilizes a ResNet100 backbone [11].

Reconstruction Attack Model (R): Our reconstruction attack model, denoted as R , is a U-Net model. We choose to use similar encoder and decoder blocks to the ones described in [9]. The main difference is that the attack model is not conditioned on any identity information and utilizes several skip connections. See supplementary materials for details.

3.2. Identity Tracking Module (ITM)

When performing live anonymization in the wild, we need to keep track of the identities present in the video and the conditional information used to anonymize. To solve this, we simply use an Identity Tracking Module (ITM) that checks if the detected face has been observed before (Figure 9). If it has not, we sample a fake identity based on Equation 8 as described in section 3.3. If it has been observed, we skip the sampling and return the matching identity’s corresponding fake identity. Let z_{id} be the embedded identity vector, \mathcal{T}_{id} is the stored previous identity vectors, \mathcal{F}_{id} defines the dictionary containing the stored fake identity vectors z_a , t denotes the threshold, *key_pointer* represents the next key for adding new fake identities to \mathcal{F}_{id} , and V_i is the sampling function. The process for tracking and generating new fake identities is then defined as shown in Algorithm 1.

Algorithm 1: Identity tracking algorithm for using consistent fake identities and generating new fake identities.

```

1 ITM  $z_{id}$ ;
   Input : Extracted identity vector  $z_{id}$ 
   Output: Fake identity  $z_a$ 
2  $D = 1 - \text{cosine\_similarity}(z_{id}, \mathcal{T}_{id})$ 
3  $idx = \text{argmin}(D)$ 
4  $d = D[idx]$ 
5 if  $d < t$  then
6   | return  $\mathcal{F}_{id}[idx]$ ;
7 else
8   |  $\mathcal{T}_{id}.\text{append}(z_{id})$ 
9   |  $z_a = V_i(z_{id})$ 
10  |  $\mathcal{F}_{id}[\text{key\_pointer}] = z_a$ 
11  |  $\text{key\_pointer} = \text{key\_pointer} + 1$ 
12  | return  $z_a$ ;
13 end

```

3.3. Sample Fake Identities

For both our generator and target-oriented face-swapping models, we need to sample fake vectors, on which the models are conditioned. Considering those facial recognition models such as ArcFace [8] and CosFace [32] are trained to strengthen the cosine similarity between the same identity images, which is constraining the embeddings to a unit hypersphere, we can prove that the absolute opposite of a vector z_{id} is simply put $-z_{id}$, see Equation 2 and 3. This means that we can guarantee a strong anonymization (difference in identity) that facial recognition can not match correctly. However, this is a double-edged sword in which facial recognition guarantees a hit by searching for the most similar identity to $-z_{id}$. This means we need to sample identity information that is far away from *both* z_{id} and $-z_{id}$. This is done by preparing an anchor search space \mathcal{S}_a . In this regard, we extract average embeddings of identities in the VGGFace2 [3] train set. These are spherically interpolated with a shifted version of itself, thus creating an anchor search space of identity embeddings that all are a mix of two people (A visualization in a 3-dimensional coordinate system can be viewed in the supplementary material). This step can be repeated to a desired size of the anchor search space. We then search for an appropriate embedding as follows:

$$z_a = \mathcal{S}_a[\text{argmin}(|\cos(z_{id}, \mathcal{S}_a)| + m)], \quad (1)$$

where z_a is a fake identity and m is a margin to constrain the search result close to a specific distance. For instance, a margin m of 0 would result in finding an embedding of a cosine similarity of 0.

Cosine similarity $\cos(\theta)$ is defined as follows

$$\cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}, \quad (2)$$

where A and $B \in \mathcal{R}^D$, and the formula is constrained between -1 and 1 . Assuming we measure the cosine similarity of A with itself, we obtain

$$\cos(\theta) = \frac{A \cdot A}{\|A\| \|A\|} = 1, \quad (3)$$

which is equal to 1, meaning exactly the same. Therefore, if we compare A and $-A$ we obtain $\cos(\theta) = -1$, meaning exactly opposite.

3.4. Loss Functions

The loss function has two components: one is for the face manipulation model and the other is for the reconstruction attack, described below.

Face Manipulation Loss: A combination of different losses is employed to train the generator model, such as IFSR \mathcal{L}_{ifsr} introduced in [27], a cosine distance loss \mathcal{L}_a , L1 pixel-wise reconstruction loss, and L1 mask loss in [9]. IFSR calculates the cosine distance between entire feature maps within an identity encoder to regularize the model to retain attribute information such as expression, pose, lighting and make-up. We modify IFSR slightly, instead of using a feature map for each residual block, we use a feature map for each resolution scale, and completely omit the proposed margins by setting them to 0. For identity manipulation, we choose to train it counterfactual, similar to [9]. We use the cosine distance as:

$$\mathcal{L}_a = 2 + \cos(I(X_t), I(X_a)) + \cos(I(X_t), I(X_{\hat{a}})), \quad (4)$$

where $\cos(\cdot)$ denotes the cosine similarity, X_t is the target image, X_a is the anonymized face, $X_{\hat{a}}$ is the anonymized face blended with the target face and $I(\cdot)$ is the pretrained ArcFace [8].

Reconstruction Attack Loss: To perform a reconstruction attack against a target-oriented face-swapping and/or anonymization model, we train a U-Net architecture to reconstruct the original image. The reconstruction attack model was trained using an L1 pixel-wise reconstruction loss, an identity loss, the same version of the IFSR loss \mathcal{L}_{ifsr} from [27], and an adversarial loss. The reconstruction loss is as follows:

$$\mathcal{L}_r = \|X_t - R(X_c)\|, \quad (5)$$

where X_c is the anonymized/swapped target image, X_t defines the unaltered target image of X_c , and $R(\cdot)$ represents the reconstruction attack model. The identity loss is used to further enforce the original identity information:

$$\mathcal{L}_i = 1 - \cos(I(X_t), I(R(X_c))), \quad (6)$$

where $\cos(\cdot)$ denotes the cosine similarity. Once the reconstruction attack model is trained, we evaluate its capabilities in conjunction with the reconstruction attack vulnerability of the face-swapping model. In this case, we (1) sample a fake identity vector (see Section 3.3) for the attacked face-swapping model, (2) perform anonymization, (3) reconstruct the original identity with the reconstruction attack model, and finally (4) try to retrieve the original identity in the dataset using a separate identity embedding model, CosFace [32].

3.5. Evaluation

Temporal Consistency: Ideally, when performing anonymization on video, should the new identity be consistent over time. To gauge the identity consistency temporally, we first extract the face from N frames from M videos. Secondly, we extract the identity vector for each frame using CosFace [32]. Finally, we calculate the pair-wise cosine distance between all frames and measure the mean standard deviation of these distances:

$$\mathcal{M}_{tc}^\sigma = \frac{1}{M} \sum \sqrt{\frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N (D_{i,j} - \mu)^2}, \quad (7)$$

where D is the pair-wise cosine distance matrix and μ is the mean distance of D . We also report the mean of the mean distances in D , denoted as \mathcal{M}_{tc}^μ . This lets us represent the mean variation of the anonymized face when the identity is the same as the input for the anonymization algorithm.

Anonymization: We follow previous works [9, 23, 19, 28] for evaluating how well our method manages to hide the identity of the target in contrast to other approaches. In most of the face-swapping algorithms [5, 20, 27, 33, 34] and reconstruction attack approaches, the identity performance is evaluated using the percentage of successful retrieval of the original source identity. We here rather report the percentage of successful retrieval of the target identity, where the goal is to negate the recapturing of the identity. Another difference is that successful retrieval of the target identity is not necessarily successful in a practical setting. More specifically, we follow common practices from recognition tasks and reject the retrieved identity as a success if and only if the closest identity is the target identity with a distance that is still larger than a practical threshold. In this case, the threshold is 0.63 for a false acceptance rate of 0.001.

Adversarial Defense: The theory for why reconstruction works is because it learns a function in the image that the anonymization model “*applies*”. Therefore, we investigate potential approaches that could disrupt that function.

For defending against reconstruction attacks we investigate adversarial attacks, more specifically, the fast gradient sign attack [10], against the reconstruction attack model, standard uniform noise, and Gaussian noise applied to the parameters of the model.

3.6. Reconstruction Attacks as Deep Fake Detectors

Since the reconstruction attack model is trained to take a manipulated face image as input and perform an alteration in the image itself to restore the original identity, we investigate the cases when the input is not manipulated by a target-oriented face-swapping model. Since our hypothesis states that the reconstruction attack model learns a function in the image that the anonymization model or face-swapping model leaves behind, we expect that images that have not been manipulated cannot be changed by the reconstruction attack model. This claim is supported by the experimental results shown in Section 4. Because of this behavior, we can use the reconstruction attack model as a deep fake detector. This can be easily done by measuring the cosine distance between identity embeddings of the input image and output image, similar to Eq. 6. A high distance value would indicate a deep fake, while a low distance does not.

4. Results

Implementation Details: *FIVA* is trained on the dataset VGGFace2 [3]. All faces are aligned with five-point landmarks extracted with RetinaFace [7]. The alignment is performed to match the input into ArcFace [8]. We used the Adam [18] optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.99$, a learning rate of 0.0001, and exponential learning rate decay of 0.97 every 100K steps. The target (X_t) face is distorted during training with random rotation of 10 degrees and small random zooms are applied before being fed to *FIVA*. The input X_t of ArcFace remains undistorted. Image resolution is 256×256 .

4.1. Quantitative Results

Exhaustive experiments are conducted to demonstrate the effectiveness of our proposed *FIVA* model together with target-oriented face-swapping models. Table 1 contains quantitative results on the dataset FaceForensic++ [29]. This table not only compares *FIVA* with previous works, but also acts as an ablative benchmark for highlighting the effectiveness of our proposed anchor sampling and ITM modules. The metrics evaluated are identity retrieval (ID), negated identity retrieval (\neg ID), reconstruction attack identity retrieval (RA), and temporal consistency (See Section 3.5). As shown in Table 1, both *FIVA*, SimSwap [5], and FaceDancer [27] achieve the lowest ID scores and thus allow for strong anonymity of faces.

Table 1: Quantitative experiments on FaceForensics++ [29]. Evaluated with identity retrieval (ID), negated identity retrieval (\neg ID, searching for a match with $-z_{id}$), and reconstruction attack (RA) identity retrieval. Temporal identity consistency \mathcal{M}_{tc} calculated using 10 frames per video. The divide in the table separates inpainting-based methods from target-oriented ones. The \times indicates that RA is not applicable to the corresponding method. +Sampling means we used the anchor sampling method to assign anonymized identities (See Equation 8), while +ITM indicates both the anchor sampling and tracking (See Algorithm 1 and Figure 9). The \downarrow indicates lower is better.

Method	ID \downarrow	\neg ID \downarrow	RA \downarrow	$\mathcal{M}_{tc}^\mu \downarrow$	$\mathcal{M}_{tc}^\sigma \downarrow$
Real Data	-	-	-	0.150	0.074
CIAGAN [23]	0.035	0.000	\times	0.521	0.220
CIAGAN [23] + ITM	0.030	0.000	\times	0.300	0.151
DeepPrivacy [14]	0.004	0.000	\times	0.359	0.184
CFA-Net [22]	0.012	N/A	N/A	N/A	N/A
SimSwap [5] + Sampling	0.002	0.000	0.994	0.607	0.345
SimSwap [5] + ITM	0.002	0.000	0.994	0.084	0.051
FaceDancer [27] + Sampling	0.000	0.000	0.999	0.556	0.314
FaceDancer [27] + ITM	0.000	0.000	0.999	0.186	0.141
FIVA	0.000	0.966	0.998	0.227	0.101
FIVA + Sampling	0.000	0.000	0.996	0.550	0.310
FIVA + ITM	0.000	0.000	0.996	0.075	0.041

Our proposed method of sampling fake identities shows that we can avoid identity leakage by searching for $-z_{id}$. Recalling that *FIVA* was trained counterfactually and can, in theory, anonymize without any sampling, Table 1 shows a successful identity retrieval rate ($-z_{id}$) of 96.6% (\neg ID) even if anchor sampling is *not* employed. Note that SimSwap and FaceDancer are primarily designed as face-swapping approaches and both require an additional sampling process. Our proposed ITM approach allows for strong temporal consistency in all the target-oriented methods, as shown in Table 1 where the \mathcal{M}_{tc} scores drop once ITM is added to SimSwap [5], FaceDancer [27], and *FIVA*.

Next, we compare the performance of *FIVA*, SimSwap [5], and FaceDancer [27] as anonymizer (utilizing ITM for tracking and anchor sampling) with previous work of the LFW benchmark [12] in Table 2. The used facial recognition model differs here, however, we claim that the comparison is still valid as we use a more powerful model, CosFace [32], for identity retrieval. Table 2 further demonstrates the effectiveness of both *FIVA* and ITM in providing robust anonymization.

Finally, we claim that *FIVA* and other target-orient approaches leave a trace in the image that allows for an adversarial network to learn to reconstruct the original identity. This is highlighted by the lower scores in column 3 in Table 1. To strengthen this claim and demonstrate poten-

Table 2: Quantitative identity retrieval experiments on LFW [12]. CFA-Net [22] and Gafni et al. [9] demonstrate the true positive rate for a false acceptance rate of 0.001 using FaceNet [30] as the facial recognition model. We evaluate the remaining methods with CosFace and a threshold of 0.63 (Cosine *distance*), for a false acceptance rate of 0.001. The \downarrow indicates lower is better.

Method	ID \downarrow
Gafni et al. [9]	0.035
CIAGAN [23]	0.034
CFA-Net [22]	0.012
DeepPrivacy [14]	0.002
FaceDancer [27] + ITM	0.002
SimSwap [5] + ITM	0.001
FIVA + ITM	0.000

tial approaches, we show in Table 3 that a disrupting noise can collapse the output of the reconstruction attack model. Furthermore, in Figure 3 we investigate how many pixels are needed to disrupt the reconstruction attack model. As shown in Figure 3, for *FIVA*, 47% noised pixels yielded the best disruption of the reconstruction attack model, but as little as 10% causes a severe collapse of the reconstruction attack model. Figure 4 provides qualitative results of the reconstruction attack, different defenses, and anonymization using *FIVA*.

4.2. Reconstruction Attack as Deepfake Detection

The success of reconstructing the original identity from target-oriented face-swapping or anonymization approaches raises the question of what happens if one inputs an image

Table 3: Defense against reconstruction attack in *FIVA*, evaluated on FaceForensics++ [29]. Adversarial Defense in the form of a fast sign gradient method. Noise Defense just adds regular uniform noise to the image. Parameter Noise means adding a small Gaussian noise to the parameters. We report the fraction of successful retrievals of the original identity after applying the reconstruction attack. ϵ highlights how much the noise was scaled. The \downarrow indicates lower is better. Black-box means it does not need access to the reconstruction attack model.

Method	ϵ	ID \downarrow	Black-box
Parameter Noise	0.10	0.442	yes
Adversarial Defense	0.15	0.002	no
Noise Defense	0.15	0.004	yes

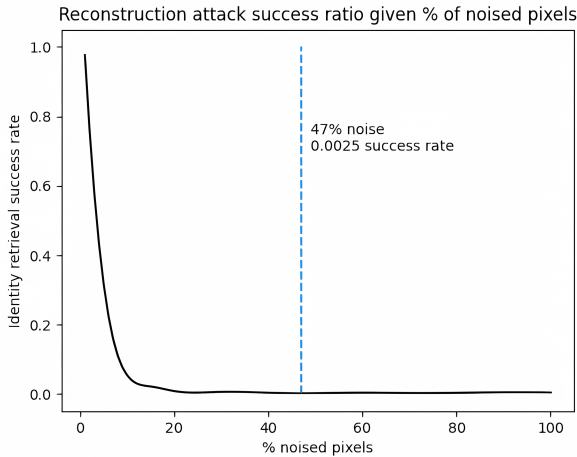


Figure 3: Identity retrieval success rate of the reconstruction attack, depending on the number of noisy pixels. This graph demonstrates the identity retrieval success rate for *FIVA* while using the uniform noise. The blue dashed line points out the % which prevents reconstruction attacks best (0.0025).

that is not manipulated. We notice that the reconstruction attack model does not really change anything at all with these images. As demonstrated in Figure 5, a potential approach for deep fake detection is simply measuring the cosine distance of the CosFace embeddings between the input image X and the output image X' . We note that a threshold of 0.6 would reach a false positive of almost 0. One drawback is that the reconstruction attack model is as of now not model agnostic.

4.3. Face-swapping using *FIVA*

Since *FIVA* is trained to drive the identity away using the cosine distance (Eq. 4), we hypothesize that *FIVA* can also perform face-swapping even if it is not the main goal.

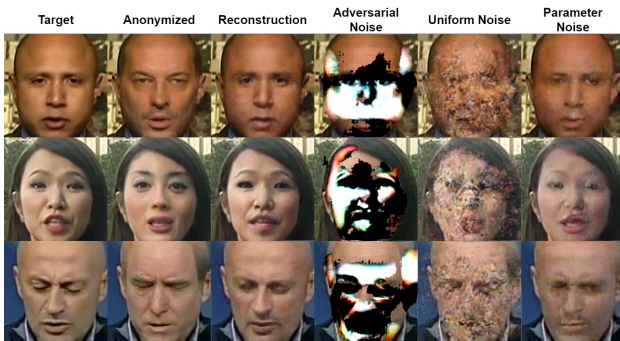


Figure 4: Qualitative results of reconstruction attack, different defenses and anonymization using *FIVA*.

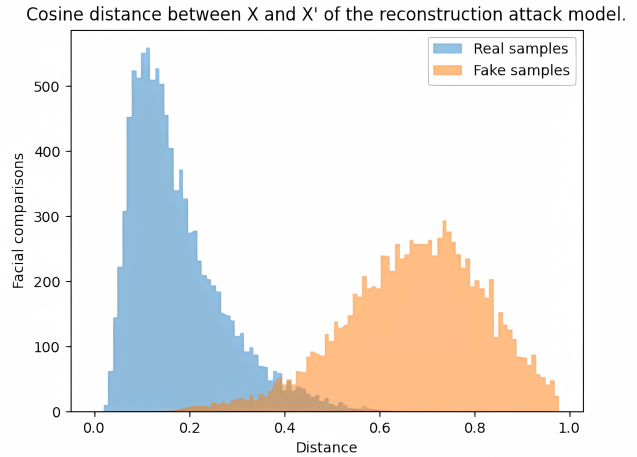


Figure 5: Cosine distance distributions between the input (X) and output (X') identity vectors of the reconstruction attack for genuine images (Real samples) and manipulated images (Fake samples).

Table 4 shows quantitative comparisons with the state-of-the-art face-swapping methods on the dataset FaceForensics++ [29]. We follow the same evaluation protocols in [5, 20, 27, 33] and obtain that *FIVA* reaches state-of-the-art performance for the identity transfer (ID). This concludes that target-oriented face-swapping methods can actually be trained in a counterfactual way, which eliminates the need for sampling pairs of faces during training. However, we note that *FIVA* naturally tends to keep attributes such as gender, ethnicity, and face shape as shown in Figure 6. Consequently, the obtained swapped faces are not perceptually convincing for humans but rather are for facial recognition models. This feature is useful for the task of anonymization, and further addresses ethical questions in regard to deep fakes in the context of anonymization. We elaborate

Table 4: Quantitative face-swapping comparisons on FaceForensics++ [29]. The \downarrow indicates lower is better, while \uparrow indicates higher is better.

Method	ID \uparrow	Pose \downarrow
FaceSwap [1]	54.19	2.51
FaceShifter [20]	97.38	2.96
MegaFS [35]	90.83	2.64
FaceController [34]	98.27	2.65
HifiFace [33]	98.48	2.63
SimSwap [5]	92.83	1.53
FaceDancer [27]	98.84	2.04
FIVA (Ours)	99.25	2.16



Figure 6: Face swapping results using *FIVA*.

more on this matter in the supplementary materials.

4.4. Qualitative Results

For qualitative evaluation, we compare the output of *FIVA* with previous works, as shown in Figures 7 and 8. In-depth comparisons are performed with the already available models such as Gafni et al. [9], CIAGAN [23], CFA-NET [22], and DeepPrivacy [14]. Images in Figures 7 and 8 clearly show that CFA-NET struggles with maintaining the color and eye-gaze, whereas CIAGAN has issues with the resolution of the image and returns rather low-quality outputs. DeepPrivacy often produces artifacts while struggling with eye-gaze and facial expression. Gafni et al. [9] together with CFA-NET and *FIVA*, is the only approach that demonstrates successful results on video. In this particular frame, the change is arguably small. CFA-NET does use similar identity control as target-oriented face-swapping and *FIVA*, which means that they need to manually assign identity embedding for each face in a video, restricting their use for in-the-wild anonymization. This problem is solved by our contribution ITM, which can be directly applied to their work. Gafni et al. [9] train their model in a counterfactual fashion, and produce stable videos without the need to track identities. This is true for *FIVA* as well, however, as pointed out in section 3.3, a maximized cosine distance allows facial recognition to find the identity by searching for $-z_{id}$.



Figure 7: Qualitative comparison between Gafni et al. [9], CIAGAN [23], CFA-NET [22], DeepPrivacy [14], and *FIVA*.

We visualize and discuss this further in the supplementary material. For video results, we refer to the supplementary material.



Figure 8: Qualitative temporal comparison between CIAGAN [23], DeepPrivacy [14] and *FIVA*. Note we used ITM for tracking the identity in CIAGAN. Video results demonstrating this can be found in the supplementary material.

5. Conclusion

In this work, we introduce a new facial anonymization framework *FIVA*, which together with our proposed identity sampling and identity tracking, reaches state-of-art performance for facial anonymization for both video and images. We also show that target-oriented models are very easy to attack and, thus, introduce adversarial models that can completely undo the masked identity in the frame. To the best of our knowledge, this potential security issue has so far not been addressed. We furthermore demonstrate that regardless of what the attack model looks for in the image, it can be disrupted by noise. We expect this to become a cat-and-mouse game where attack models can learn to ignore the noise, thus making anonymization more challenging. The attack model can also be used as deep fake detector, since the model does not change anything in the image when the input has not been manipulated. Last but not least, *FIVA* is also capable of face-swapping, reaching state-of-the-art performance for identity transfer, thus demonstrating its excellent control over identity information. Interestingly it does so, while keeping the changes to a minimum.

References

- [1] Faceswap. Accessed 2022-02-18.
- [2] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. Creating a photoreal digital actor: The digital emily project. In *2009 Conference for Visual Media Production*, pages 176–187, 2009.

- [3] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.
- [4] Umur A. Çiftçi, Gokturk Yuksek, and İlke Demir. My face my choice: Privacy enhancing deepfakes for social media anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1369–1379, January 2023.
- [5] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. *SimSwap: An Efficient Framework For High Fidelity Face Swapping*, page 2003–2011. Association for Computing Machinery, New York, NY, USA, 2020.
- [6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020.
- [7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotzia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020.
- [8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019.
- [9] Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9378–9387, 2019.
- [10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [12] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.
- [13] Håkon Hukkelås and Frank Lindseth. Deepprivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1329–1338, 2023.
- [14] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *Advances in Visual Computing*, pages 565–578. Springer International Publishing, 2019.
- [15] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021.
- [16] 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.
- [17] 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.
- [18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [19] Jingzhi Li, Lutong Han, Hua Zhang, Xiaoguang Han, Jingguo Ge, and Xiaochun Cao. Learning disentangled representations for identity preserving surveillance face camouflage. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9748–9755. IEEE, 2021.
- [20] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020.
- [21] Tao Li and Lei Lin. Anonymousnet: Natural face de-identification with measurable privacy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.
- [22] Tianxiang Ma, Dongze Li, Wei Wang, and Jing Dong. Cfa-net: Controllable face anonymization network with identity representation manipulation. *arXiv preprint arXiv:2105.11137*, 2021.
- [23] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Cigan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020.
- [24] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7183–7192, 2019.
- [25] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 98–105, 2018.
- [26] Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pages 620–636, 2018.
- [27] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: Pose- and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3454–3463, January 2023.
- [28] Felix Rosberg, Cristofer Englund, Martin Torstensson, and Boris Durán. Towards privacy aware data collection in traffic. In *FAST-zero - International Symposium on Future Active Safety Technology toward zero-traffic-accident*, 2021.
- [29] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In

Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 1–11, 2019.

- [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.
- [31] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016.
- [32] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.
- [33] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hiface: 3d shape and semantic prior guided high fidelity face swapping. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1136–1142. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.
- [34] Zhiliang Xu, Xiyu Yu, Zhibin Hong, Zhen Zhu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Facecontroller: Controllable attribute editing for face in the wild. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(4):3083–3091, May 2021.
- [35] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4834–4844, June 2021.

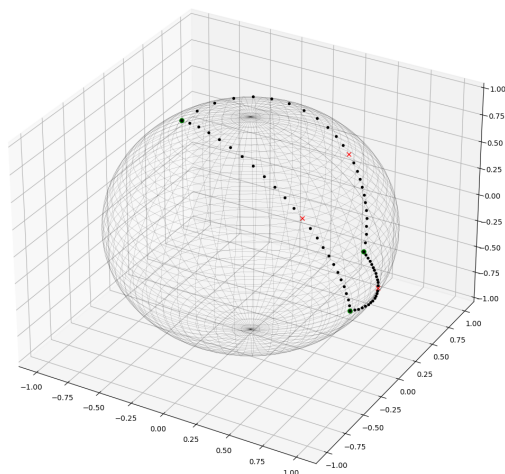


Figure 11: Illustration of how we create the anchor vectors that we sample fake identities from. The green points illustrates an identity extracted by ArcFace, the black dots shows the spherical interpolation path on the unit sphere and the red crosses represent the resulting vectors used in the anchor set which becomes a mix of actual identities.

$$z_a = \mathcal{S}_a[\operatorname{argmin}(|\cos(z_{id}, \mathcal{S}_a)| + m)], \quad (8)$$

we can control a desired approximate cosine distance from the target identity by changing the margin m . In Figure 12 we illustrate anchor matches as the margin changes, where the green line represent the match that would occur for the margin m of 0.7. Because *FIVA* was trained counter-factually to drive the identity away we want to find an anchor close. When using target-oriented face swapping methods, which are trained to drive the identity towards the source, you have to sample identities far away.

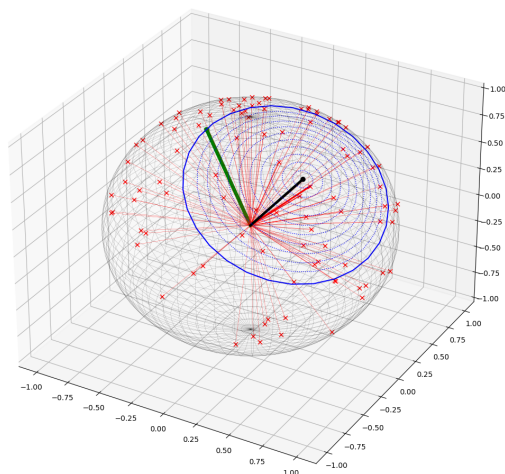


Figure 12: Illustration of matching a desired anchor. The red lines illustrates matches to a desired anchor based on desired approximate distance from the target vector (black line). The green line illustrates the match that would occur for when sampling for *FIVA*. Blue circle illustrates the desired distance.

B . Gender and Ethnicity Preservation

Because *FIVA* is trained counter-factually to drive the identity away it only uses target faces, compared to face swapping which generally uses pairs of target and source faces. It makes sense that the generator learns to preserve gender and ethnicity to be able to generate convincing samples for the discriminator. We find interesting evidence for this when looking into face swapping using *FIVA*. *FIVA* manage to reach state-of-the-art for identity transfer for face-swapping. However, we notice qualitatively that the gender and ethnicity it preserved, even if you would use a male target and female source or Asian target with a Caucasian source. Shown in Figure 10, we can see that *FIVA* tends to preserve gender and ethnicity for face swaps compared to FaceDancer. Even if *FIVA* performs better for identity transfer quantitatively, it does not qualitative. However this behaviour is useful for other tasks, such as anonymization, allowing us to ignore the need to sample identities based of gender.