Password-conditioned Anonymization and Deanonymization with Face Identity Transformers

Xiuye Gu[1,2][0000−0001−5568−564X], Weixin Luo[2,3][0000−0002−0754−6458], Michael S. Ryoo[4][0000−0002−5452−8332], and Yong Jae Lee[2][0000−0001−9863−1270]

[1]Stanford University [2]UC Davis [3]ShanghaiTech [4]Stony Brook University

Abstract. Cameras are prevalent in our daily lives, and enable many useful systems built upon computer vision technologies such as smart cameras and home robots for service applications. However, there is also an increasing societal concern as the captured images/videos may contain privacy-sensitive information (*e.g.*, face identity). We propose a novel *face identity transformer* which enables automated photo-realistic password-based anonymization and deanonymization of human faces appearing in visual data. Our face identity transformer is trained to (1) remove face identity information after anonymization, (2) recover the original face when given the correct password, and (3) return a wrong—but photo-realistic—face given a wrong password. With our carefully designed password scheme and multi-task learning objective, we achieve both anonymization and deanonymization using the same single network. Extensive experiments show that our method enables multimodal password conditioned anonymizations and deanonymizations, without sacrificing privacy compared to existing anonymization methods.

1 Introduction

As computer vision technology is becoming more integrated into our daily lives, addressing privacy and security questions is becoming more important than ever. For example, smart cameras and robots in homes are widely being used, but their recorded videos often contain sensitive information of their users. In the worst case, a hacker could intrude these devices and gain access to private information.

Recent anonymization techniques aim to alleviate such privacy concerns by redacting privacy-sensitive data like face identity information. Some methods [3,29] perform low-level image processing such as extreme downsampling, image masking, etc. A recent paper proposes to *learn* a face anonymizer that modifies the identity of a face while preserving activity relevant information [26]. However, none of these techniques consider the fact that the video/image owner (and his/her friends, family, law enforcement, etc.) may want to see the *original* identities and not the anonymized ones. For example, people may not want their real faces to be saved directly on home security cameras due to privacy concerns; however, remote family members may want to see the real faces from time to time. Or when crimes arise, to catch criminals, police need to see their real faces.

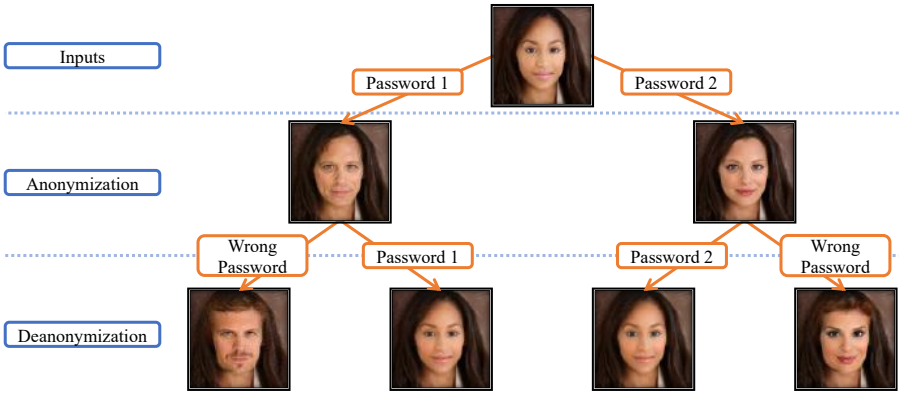


Fig. 1: Our system never stores users' faces on disk, and instead only stores the anonymized faces. When a user provides a correct recovery password, s/he will get the deanonymized face back. If a hacker invading their privacy inputs a wrong password, s/he will get a face whose identity is different from the original as well as the anonymized face. The photo-realism of the modified faces is meant to fool the hacker by providing no clues as to whether the real face was recovered.

This problem poses an interesting tradeoff between privacy and accessibility. On the one hand, we would like a system that can anonymize sensitive regions (face identity) so that even if a hacker were to gain access to such data, they would not be able to know who the person is (without additional identity revealing meta-data). On the other hand, the owner of the visual data inherently wants to see the original data, not the anonymized one.

To address this issue, we introduce a novel *face identity transformer* that can both *anonymize and deanonymize (recover)* the original image, while maintaining privacy. We design a discrete password space, in which the password conditions the identity change. Specifically, given an original face, our face identity transformer outputs different anonymized face images with different passwords (Fig. 1 Anonymization). Then, given an anonymized face, the original face is recovered only if the correct password is provided (Fig. 1 Deanonymization, 'Password 1/2'). We further increase security as follows: Given an anonymized face, if a wrong password is provided, then it changes to a new identity, which is still different from the original identity (Fig. 1 Deanonymization, 'Wrong Password'). Moreover, each wrong password maps to a unique identity. In this way, we provide security via ambiguity: even if a hacker guesses the correct password, it is extremely difficult to know that without having access to any other identity revealing meta-data, since each password—regardless of whether it is correct or not—always leads to a different realistic identity.

To enforce the face identity transformer to output different anonymized face identities with different passwords, we optimize a multi-task learning objective, which includes maximizing the feature-level dissimilarity between pairs of

anonymized faces that have different passwords and fooling a face classifier. To enforce it to recover the original face with the correct password, we train it to anonymize and then recover the correct identity only when given the correct password, and to produce a new identity otherwise. Lastly, we maximize the feature dissimilarity between an anonymized face and its deanonymized face with a wrong password so that the identity always changes. Moreover, considering the limited memory space on devices, we propose to use the same single transformer to serve both anonymization and deanonymization purposes.

We note that our approach is related to cryptosystems like RSA [27]. The key difference is that cryptosystems do not produce encryptions that are visually recognizable to human eyes. However, in various scenarios, users may want to understand what is happening in anonymized visual data. For example, people may share photos/videos over public social media with anonymized faces, but only their real-life friends have the passwords and can see their real faces to protect identity information. Moreover, with photorealistic anonymizations, one can *easily apply existing computer vision based recognition algorithms on the anonymized images* as we demonstrate in Sec. 5.5. In this way, it could work with e.g., smart cameras that use CV algorithms to analyze content but in a privacy-preserving way, unlike other schemes (*e.g.*, homomorphic encryption) that require developing new ad-hoc recognition methods specific to nonphotorealistic modifications, in which accuracy may suffer.

In our approach, only the anonymized data is saved to disk (*i.e.*, conceptually, the anonymization would happen at the hardware-level via an embedded chipset – the actual implementation of which is outside the scope of this work). The advantage of this concept is that the hacker could never have direct access to the original data. Finally, although there may be other identity-revealing information such as gait, clothing, background, etc., our work entirely focuses on improving privacy of face identity information, but would be complementary to systems that focus on those other aspects.

Our experiments on CASIA [34], LFW [11], and FFHQ [13] show that the proposed method enables multimodal face anonymization as well as recovery of original face images, without sacrificing privacy compared to existing advanced anonymization [26] and classical image processing techniques including masking, noising, and blurring, etc. *Please see* `https://youtu.be/FrYmf-CL4yk` *and Fig. 6 in the supp for image/video in the wild results.*

2 Related work

Privacy-preserving visual recognition. This is the problem of detecting humans, their actions, and objects without accessing user-sensitive information in images/videos. Some methods employ extreme low-resolution downsampling to hide sensitive details [32,6,29,28] but suffer from lower recognition performance in downstream tasks. More recent work propose a head inpainting obfuscation technique [31], a four-stage pipeline that first obfuscates facial attributes and then synthesizes faces [16], and a video anonymizer that performs pixel-level

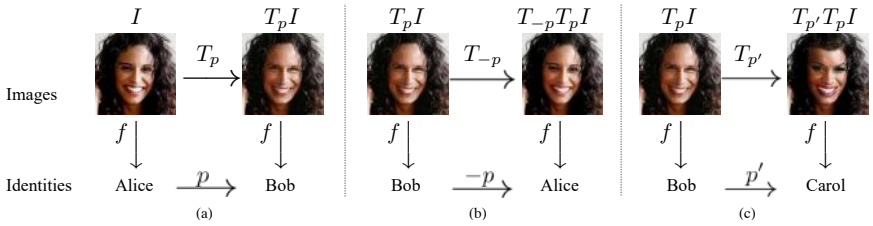


Fig. 2: Privacy-preserving properties that our face identity transformer T learns. (a) Anonymization stage. (b) Deanonymization stage with correct recovery password. (c) Deanonymization stage with incorrect recovery password.

modifications to remove people's identity while preserving motion and object information for activity detection [26]. Unlike our approach, none of these existing work employ a password scheme to condition the anonymization, and also do not perform deanonymization to recover the original face. Moreover, even if one could brute-forcely train a deanonymizer for these methods, there is no way to provide wrong recoveries upon wrong passwords, as our method does.

Security/cryptography research on privacy-preserving recognition is also related *e.g.*, [8,9]. The key difference is that these methods encrypt data in a secure but visually-uninterpretable way, whereas our goal is to anonymize the data in a way that is still interpretable to humans and existing computer vision techniques can still be applied. Differential privacy [1,35] is also related but its focus is on protecting privacy in the training data whereas ours is on anonymizing visual data during the inference stage.

Face image manipulation and conditional GANs. Our work builds upon advances in pixel-level synthesis and editing of realistic human faces [15,22,30,13,23,2] and conditional GANs [20,24,12,36,41,5,7,25], but we differ significantly in our goal, which is to completely change the identity of a face (and also recover the original) for privacy-preserving visual recognition.

3 Desiderata

Our face identity transformer T takes as input a face image $I \in \Phi$ and a user-defined password $p \in P$, where Φ and P denote the face image domain and password domain. We use the notation T_pI to denote the transformed image with input image I and password p. Before diving into the details, we first outline desired properties of a privacy-preserving face identity transformer.

Minimal memory consumption. Considering the limited memory space on most camera systems, a *single* face identity transformer that can both anonymize and deanonymize faces is desirable.

Photo-realism. We would like the transformer to maintain photo-realism for any transformed face image:

$$T_p I \in \Phi, \quad \forall p \in P, \forall I \in \Phi. \tag{1}$$

Photo-realism has three benefits: 1) a human who views the transformed images will still be able to interpret them; 2) one can easily apply existing computer vision algorithms on the transformed images; and 3) it's possible to confuse a hacker, since photo-realism can no longer be used as a cue to differentiate the original face from an anonymized one.

Compatibility with background. The background $B(\cdot)$ of the transformed face should be the same as the original:

$$B(T_p I) = B(I), \quad \forall p \in P, \forall I \in \Phi. \tag{2}$$

This will ensure that there are no artifacts between the face region and the rest of the image (*i.e.*, it will not be obvious that the image has been altered).

Anonymization with passwords. Let $f : \Phi \rightarrow \Gamma$ denote the function mapping face images to people's identities. We would like to condition anonymization via a password p:

$$f(T_p I) \neq f(I), \quad \forall p \in P, \forall I \in \Phi. \tag{3}$$

Deanonymization with inverse passwords. We should recover the original identity only when the correct password is provided. To achieve our goal of minimal memory consumption, we can model the additive inverse of the password used for anonymization as the correct password for deanonymization. In this way, we can use the same transformer for deanonymization, *i.e.* we model $T_{-p} = T_p^{-1}$:

$$f(T_{-p} T_p I) = f(T_p^{-1} T_p I) = f(I), \ \forall p \in P, \forall I \in \Phi. \tag{4}$$

Wrong deanonymization with wrong inverse passwords. We would like the transformer to change the anonymized identity into a *different* identity that is different from both the original as well as the anonymized image when given a wrong inverse password:

$$f(T_{p'} T_p I) \neq f(I), \quad \forall p, p' \in P, p' \neq -p, \forall I \in \Phi, \tag{5}$$

$$f(T_{p'} T_p I) \neq f(T_p I), \ \forall p, p' \in P, p' \neq -p, \forall I \in \Phi. \tag{6}$$

In this way, whether the password is correct or not, the identity is always changed so as to confuse the hacker.

Diversity. The image I should be transformed to different identities with different passwords, to increase security in both anonymization and deanonymization. Otherwise, if multiple passwords produce the same identity, a hacker could realize that the photo is anonymized or his attempts have failed in deanonymization:

$$f(T_{p_1} I) \neq f(T_{p_2} I), \ \forall p_1, p_2 \in P, p_1 \neq p_2, \forall I \in \Phi. \tag{7}$$

Fig. 2 summarizes our desiderata for anonymization and deanonymization.

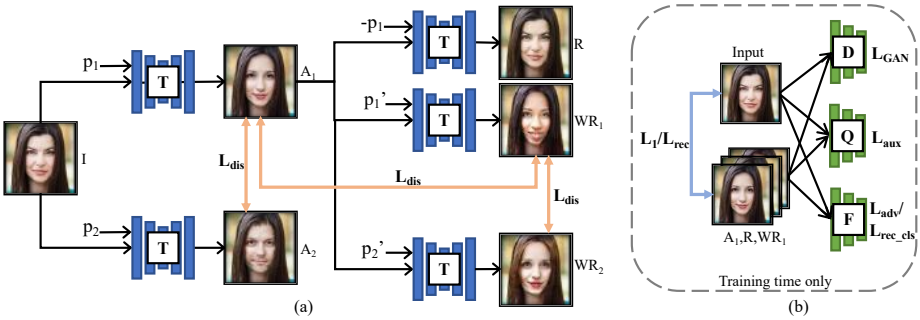


Fig. 3: (a) Face identity transformer network architecture. (b) Objectives we apply to synthesized images during training (not included in (a) for clarity). I: Input image, $A_{1,2}$: Anonymized faces, R: Recovered face, $WR_{1,2}$: Wrongly Recovered faces. $\mathcal{L}_{feat}$ is the sum of the three $\mathcal{L}_{dis}$'s.

4 Approach: *Face Identity Transformer*

Our face identity transformer T is a conditional GAN trained with a multi-task learning objective. It is conditioned on both the input image I and an input password p. Importantly, the function of p is different from the usual random noise vector z in GANs: z simply tries to model the distribution of the input data, while p in our case makes the transformer hold the desired privacy-preserving properties (Eq. 3-7). We next explain our password scheme, multimodal identity generation, and multi-task learning objective.

4.1 Password scheme

We use an N-bit string $p \in \{0,1\}^N$ as our password format, which consists of 2^N unique passwords. Given image $I \in \mathbb{R}^{H\times W\times 3}$, we form the input to the transformer as a depthwise concatenation $(I,p) \in \mathbb{R}^{H\times W\times(3+N)}$, where p is replicated in every pixel location. To make the transformer condition its identity change on the input password, we design an auxiliary network $Q(I, T_pI) = \hat{p}$. It learns to predict the embedded password from the input and transformed image pair, and thus maximizes the mutual information between the injected password and the identity change in the image domain, similar to InfoGAN [7]. We use cross entropy loss for the classifier Q, and denote it as $\mathcal{L}_{aux}(T,Q)$. See supp Sec. 1 for the detailed formula.

4.2 Multimodal identity change

Conditional GANs with random noise do not produce highly stochastic outputs [19,12]. To overcome this, BicycleGAN [40] uses an explicitly-encoded multimodality strategy similar to our auxiliary network Q. However, even with Q,

we only observe multimodality on colors and textures as in [40], but not on high-level face identity.

Thus, to induce diverse high-level identity changes, we propose an explicit feature dissimilarity loss. Specifically, we use a face recognition model F to extract deep embeddings of the faces, and minimize their cosine similarity when they are associated with different passwords:

$$\mathcal{L}_{dis}(M_1, M_2) = \max\Big(0, \cos\big(F_{embed}(M_1), F_{embed}(M_2)\big)\Big), \tag{8}$$

where cos is cosine similarity, and M_1 and M_2 are two transformed face images with two different passwords. We do not penalize pairs whose cosine similarity is less than 0; *i.e.*, it is enough for the faces to be different up to a certain point.

We apply the dissimilarly loss between (1) two anonymized faces with different passwords, (2) two incorrectly deanonymized faces given different wrong passwords, and (3) the anonymized face and wrongly recovered face:

$$\begin{aligned}\mathcal{L}_{feat}(T) = \mathbb{E}_{(I,p_1\neq p_2)}&\mathcal{L}_{dis}(T_{p_1}I, T_{p_2}I)\\ &+ \mathbb{E}_{(I,p'_1\neq p'_2,p'_1\neq -p,p'_2\neq -p)}\mathcal{L}_{dis}(T_{p'_1}T_pI, T_{p'_2}T_pI)\\ &+ \mathbb{E}_{(I,p'\neq -p)}\mathcal{L}_{dis}(T_pI, T_{p'}T_pI).\end{aligned} \tag{9}$$

This loss can be easily satisfied when the model outputs extremely different content that do not necessarily look like a face, and thus can adversely affect other desideratum (*e.g.*, photo-realism) of a privacy-preserving face identity transformer. We next introduce a multi-task learning objective to restrict the outputs to lie on the face manifold, as a form of regularization.

4.3 Multi-task learning objective

We describe our multi-task objective that further aids identity change, identity recovery, and photo-realism.

Face classification adversarial loss. We apply the face classification adversarial loss from [26], which helps change the input face's identity. We apply it on both the transformed face T_pI as well as the reconstructed face with wrong recovery password $T_{p'}T_pI$:

$$\begin{aligned}\mathcal{L}_{adv}(T, F) = &- \mathbb{E}_I\mathcal{L}_{CE}\big(F(I), y_I\big) - \mathbb{E}_{(I,p)}\mathcal{L}_{CE}\big(F(T_pI), y_I\big)\\ &- \mathbb{E}_{(I,p'\neq -p)}\mathcal{L}_{CE}\big(F(T_{p'}T_pI), y_I\big),\end{aligned} \tag{10}$$

where F is the face classifier, y_I is face identity label, and $\mathcal{L}_{CE}$ denotes cross entropy loss.

Similar to the dissimilarity loss ($\mathcal{L}_{dis}$), this loss pushes the transformed face to have a different identity. The key difference is that this loss requires face identity labels so cannot be used to push $T_{p_1}I$ and $T_{p_2}I$ to have different identities, but has the advantage of utilizing supervised learning so that it can change the identity more directly.

Reconstruction losses. We use L_1 reconstruction loss for deanonymization:

$$\mathcal{L}_{rec}(T) = \|T_{-p}T_pI - I\|_1. \tag{11}$$

With the L_1 loss alone, we find the reconstruction to be often blurry. Hence, we also introduce a face classification loss $\mathcal{L}_{rec\_cls}$ on the reconstructed face to enforce the transformer to recover the high-frequency identity information:

$$\mathcal{L}_{rec\_cls}(T, F) = \mathbb{E}_{(I,p)}\mathcal{L}_{CE}\big(F(T_{-p}T_pI), y_I\big). \tag{12}$$

This loss enforces the reconstructed face $T_{-p}T_pI$ to be predicted as having the same identity as I by face classifier F.

Background preservation loss. For any transformed face, we try to preserve its original background. To this end, we apply another L_1 loss (with lower weight):

$$\mathcal{L}_1(T) = \|T_pI - I\|_1 + \|T_{p'}T_pI - I\|_1. \tag{13}$$

Although employing a face segmentation algorithm is an option, we find that applying $\mathcal{L}_1$ on the whole image works well to preserve the background.

Photo-realism loss. We use a photo-realism adversarial loss $\mathcal{L}_{GAN}$ [10] on generated images to help model the distribution of real faces. Specifically, we use PatchGAN [12] to restrict the discriminator D's attention to the structure in local image patches. To stabilize training, we use LSGAN [18]:

$$\max_D \mathcal{L}_{GAN}(D) = -\frac{1}{2}\mathbb{E}_I[\big(D(I) - 1\big)^2] - \frac{1}{2}\mathbb{E}_{(I,p)}[D(T_pI)^2] \tag{14}$$

$$\min_T \mathcal{L}_{GAN}(T) = \mathbb{E}_{(I,p)}[\big(D(T_pI) - 1\big)^2] \tag{15}$$

4.4 Full objective

Overall, our full objective is:

$$\begin{aligned}\mathcal{L} = {} & \lambda_{aux}\mathcal{L}_{aux}(T, Q) + \lambda_{feat}\mathcal{L}_{feat}(T) \\ & + \lambda_{adv}\mathcal{L}_{adv}(T, F) + \lambda_{rec\_cls}\mathcal{L}_{rec\_cls}(T, F) \\ & + \lambda_{rec}\mathcal{L}_{rec}(T) + \lambda_{L_1}\mathcal{L}_1(T) + \mathcal{L}_{GAN}(T, D).\end{aligned} \tag{16}$$

We optimize the following minimax problem to obtain our face identity transformer:

$$T^* = \arg\min_{T,Q}\max_{D,F}\mathcal{L} \tag{17}$$

Training. Fig. 3 shows our network for training. For each input I, we randomly sample two different passwords for anonymization and two incorrect passwords for wrong recoveries, and then impose $\mathcal{L}_{dis}$ on the generated pairs and enforce $\mathcal{L}_{dis}$ between the anonymization and wrong reconstruction. We observe that during training, the auxiliary networks and backprop can consume a lot of GPU memory, which limits batch size. We propose a strategy based on symmetry:

except for the feature dissimilarity loss, we apply all other losses only to the first anonymization and first wrong recovery, which empirically works well.

We adopt a two-stage training strategy for the minimax problem [10]. In the discriminator's stage, we fix the parameters of T, Q, and update D, F; in the generator's stage, we fix D, F, and update T, Q.

Inference. During testing, the transformer T takes as input a user-defined password and a face image, anonymizes the face, and saves it to disk. When the user/hacker wants to see the original image, the transformer takes the recovery password and the anonymized image, and either outputs the identity-recovered image or a hacker-fooling image depending on password correctness. Throughout the whole process, the original images and passwords are never saved on disk for privacy reasons.

5 Experiments

In this section, we demonstrate that our face identity transformer achieves password conditioned anonymization and deanonymization with photo-realism and multimodality. We also conduct ablation studies to analyze each module/loss.

Implementation details. Our identity transformer T is built upon the network from [39]. We use size 128x128 for both inputs and outputs. We subtract 0.5 from p before inputting it to the transformer to make the password channels have zero mean. We set N=16. We use the pretrained SphereFace [17] as our face recognition network F for both deep embedding extraction in the feature dissimilarity loss and face classification adversarial training. For each stage, we use two PatchGAN discriminators [12] that have identical structure but operate at different image scales to improve photo-realism. The coarser discriminator is shared among all stages, while three separate finer discriminators are used for anonymization, reconstruction, and wrong recovery. To improve stability, we use a buffer of 500 generated images when updating D. We set λ_{aux}=1, λ_{feat}=2, λ_{adv}=2, λ_{rec_cls}=1, λ_{L_1}=10 and λ_{rec}=100, based on qualitative observations.

Datasets. 1) CASIA [34] has 454,590 face images belonging to 10,574 identities. We split the dataset into training/validation/testing subsets made up of 80%/10%/10% identities. We use the validation set to select our model. All reported results are on the test set. 2) LFW [11] has 13,233 face images belonging to 5,749 identities. As our network is never trained on LFW, we evaluate on the entire LFW to test generalization ability. 3) FFHQ [13] is a high-quality face dataset for benchmarking GANs. It is not a face recognition dataset, so we use it to only test generalization. We directly test our model on its validation set at 128x128 resolution, which contains 10,000 images.

Evaluation metrics. ***Face verification accuracy:*** We measure our transformer's identity changing ability with a standard binary face verification test, which

Method	Anonymize?	Deanonymize?	Password-conditioned?
Ren *et al.* [26]	✔	✔	✘
Super-pixel	✔	✔	✘
Edge	✔	✔	✘
Blur	✔	✘	✘
Noise	✔	✔	✘
Masked	✔	✘	✘
Ours	✔	✔	✔

Table 1: Privacy-preserving ability comparison. Our method is the only one that supports password-conditioned face (de)anonymization without sacrificing privacy.

scores whether a pair of images have the same identity or not. Since different face recognition models may have different biases, we use two popular pretrained face recognition models: SphereFace [17] and VGGFace2 [4].

Face recovery quality: We measure face recovery quality using **LPIPS distance** [38], which measures perceptual similarity between two images based on deep features, and **DSSIM** [33], which is a commonly-used low-level perceptual metric. We also use pixel-level L_1 and L_2 distance.

AMT perceptual studies: We use Amazon Mechanical Turk (AMT) to test how well our method 1) changes and recovers identities, 2) achieves photo-realism, and 3) attains multimodal anonymizations, as judged by human raters.

Runtime: On a single Titan V, averaged over CASIA testset, runtime is 0.0266 sec/batch with 12 images per batch. Though we use multiple auxiliary networks to help achieve our desiderata, they are all discarded during inference time.

5.1 Anonymization and deanonymization

To our knowledge, *no prior work achieves password-conditioned anonymization and deanonymization on visual data like ours*, see Table 1. Hence, we cannot directly compare with any existing method on generating *multimodal* anonymizations and deanonymizations.

Despite this, we want to ensure that our method does no worse than existing methods in terms of anonymization and deanonymization (setting aside the password conditioning capability). To demonstrate this, following [26], we compare to the following baselines: **Ren *et al.*** [26]: a learned face anonymizer that maintains action detection accuracy; **Superpixel** [3]: each pixel's RGB value is replaced with its superpixel's mean RGB value; **Edge** [3]: face regions are replaced with corresponding edge maps; **Blur** [29]: images are downsampled to extreme low-resolution (8×8) and then upsampled back; **Noise**: strong Gaussian noise ($\sigma^2 = 0.5$) is added to the image; **Masked**: face areas ($0.6\times$ of the face image) are masked out.

We also train deanonymizers for each baseline (*i.e.*, to recover the original face), by using the same generator architecture with our reconstruction and photo-realism losses. Please refer to supp Fig. 1 for a qualitative example of the baselines and their anonymizations/deanonymizations.

Fig. 4 shows anonymization vs. deanonymization (recovery) quality on CASIA and LFW using SphereFace and VGGFace2 as our face recognizers. Our ap-

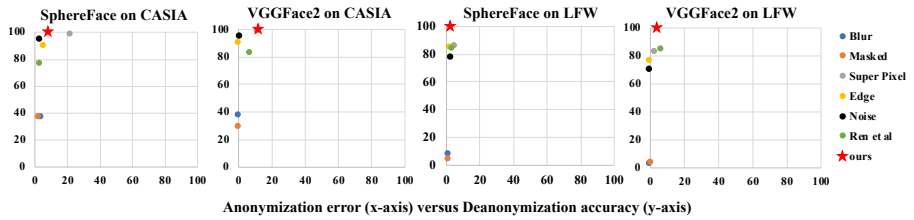


Fig. 4: Anonymization vs. deanonymization quality, measured by face verification error/accuracy on CASIA and LFW. Top-left corner is ideal. This result shows that we don't sacrifice (de)anonymization ability by introducing password conditioning.

	CASIA				LFW			
Method	LPIPS	DSSIM	L_1	L_2	LPIPS	DSSIM	L_1	L_2
Ren et al	0.08	0.07	0.06	0.009	0.08	0.07	0.06	0.010
Superpixel	0.09	0.10	0.06	0.01	0.10	0.11	0.07	0.02
Edge	0.25	0.24	0.25	0.24	0.28	0.26	0.29	0.18
Blur	0.30	0.21	0.12	0.04	0.34	0.24	0.14	0.05
Noise	0.12	0.12	0.07	0.01	0.13	0.12	0.08	0.01
Masked	0.10	0.09	0.07	0.02	0.16	0.13	0.10	0.05
Ours	**0.03**	**0.03**	**0.04**	**0.004**	**0.04**	**0.03**	**0.04**	**0.004**

Table 2: CASIA and LFW reconstruction error. Ours produces best deanonymizations.

proach performs competitively to Ren *et al*. [26], "Superpixel", "Edge", "Blur" , "Noise", and "Masked" when considering both anonymization and deanonymization quality together. This result confirms that we do not sacrifice the ability to anonymize/deanonymize by introducing password-conditioning. In fact, in terms of reconstruction (deanonymization) quality (Table 2), our method outperforms the baselines by a large margin because we train our identity transformer to do anonymization and deanonymization in conjunction in an end-to-end way.

Lastly, we perform AMT perceptual studies to rate our anonymizations and deanonymizations. Specifically, we randomly sample 150 testing images (I), and generate for each image: an anonymized face with a random password (A), a recovered face with correct inverse password (R), and a recovered face with wrong password (WR). We then distribute 600 I *vs* A, I *vs* R, I *vs* WR, and A *vs* WR pairs to turkers and ask "Are they the same person?". For each pair, we collect responses from 3 different turkers and take the majority as the answer to reduce noise.

The turkers reported **4.7**% / **100**% / **0.7**% / **1.3**% on $IvsA$ / $IvsR$ / $IvsWR$ / $AvsWR$. (low, high, low, low is ideal.) This further shows our method obtains the desired password-conditioned anonymization/deanonymization goals. We show all failure pairs for $IvsA$ in supp Sec. 5 and analyze the error there.

5.2 Photo-realism

To evaluate whether our (de)anonymization affects photo-realism, we conduct AMT user studies. We follow the same perceptual study protocol from [39] and

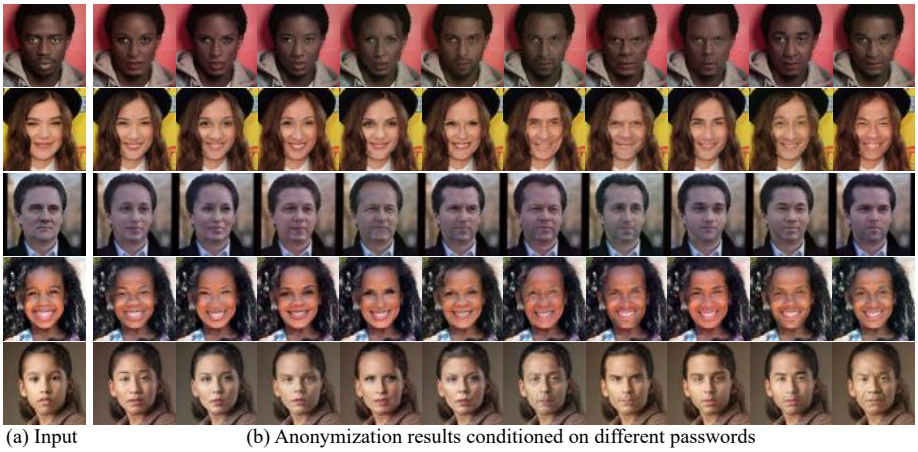


Fig. 5: Multimodality results on CASIA. We observe a wide range of identity changes with different passwords.

test on both anonymizations and wrong recoveries. For each test, we randomly generate 100 "real *vs*. fake" pairs. For each pair, we average responses from 10 unique turkers. Turkers label our anonymizations as being more real than a real face **28**.**9**% of the time, and label our wrong reconstructions as more real than a real face **15**.**4**% of the time. (Chance performace is 50%.) This shows that our generated images are quite photo-realistic.

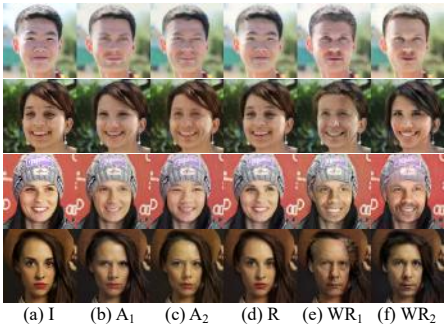
5.3 Multimodality

We next evaluate our model's ability to create different faces given different passwords. Fig. 5 shows qualitative results. Our transformer successfully changes the identity into a broad spectrum of different identities, from women to men, from young to old, *etc*.

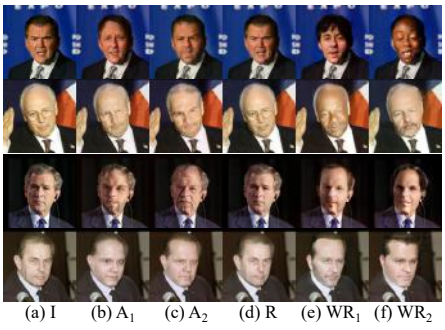
We quantitatively evaluate multimodality through an AMT perceptual study. We ask AMT workers to compare 150 A_1 *vs* A_2 and 150 WR_1 *vs* WR_2 pairs (pairs of anonymized / wrong-recovered faces with different passwords generated from the same input image) and ask "are they the same person?". The turkers reported "yes" only **12**.**2**% and **2**.**7**% of the time, respectively (lower is better). The results show that our transformer does quite well in generating different identities given different passwords.

5.4 Generalization and difficult cases

Fig. 6 shows generalization results on FFHQ and LFW using our model trained on CASIA. Without any fine-tuning, our model achieves good generalization performance on both the high quality FFHQ dataset and the LFW dataset where resolution is usually lower.



(a) FFHQ



(b) LFW

Fig. 6: FFHQ and LFW generalization results. I: original image, $A_{1,2}$: anonymized faces using different passwords, $R/WR_{1,2}$: recovered faces with correct/wrong passwords.

Fig. 7: Hard cases on CASIA. See Fig. 6 caption for key.

Fig. 8: Typical failures of each ablation. See Fig. 6 caption for key.

Fig. 7 shows hard-case qualitative results on CASIA. Our method works well even if the faces are with occlusions (sunglasses), with extreme poses, vague, under dim light, etc. We provide more qualitative results in supp.

5.5 Applying CV algorithms on transformed faces

Unlike most traditional anonymization algorithms [3,29], our choice of achieving photo-realism on the (de)anonymizations makes it possible to apply existing computer vision algorithms directly on the transformed faces. To demonstrate this, we apply an off-the-shelf MTCNN [37] face bounding box and keypoint detector on the transformed faces. Qualitative detection results (see supp Fig. 5) are good. Quantitatively, although we do not have the ground truth annotations for transformed faces, we observe that our (de)anonymizations mostly do not

Avg spatial coordinate difference	CASIA	LFW	FFHQ
Bounding boxes	1.81	1.62	1.91
Keypoints	0.94	0.76	0.89

Table 3: Average pixel difference in detected coordinates of face bounding boxes and 5 keypoints between transformed faces (A, R, WR) and input face (I).

change the head/keypoints' positions from the input faces so we can compare the detection results between the input faces and the transformed faces. Results are shown in Table 3, which shows that a face detection algorithm trained on real images performs accurately on our transformed faces.

5.6 Ablation studies

Finally, we evaluate the contribution of each component and loss in our model. Here, original image (I), anonymized face with two different passwords ($A_{1,2}$), recovered face with correct inverse password (R), and recovered faces with wrong passwords ($WR_{1,2}$):
w/o $\mathcal{L}_{\mathbf{dis}}$: We remove feature dissimilarity loss on (A_1, A_2) and (WR_1, WR_2).
w/o WR: We do not explicitly train to produce wrong reconstructions.
w/o $\mathcal{L}_{\mathbf{aux}}$: We remove the password-predicting auxiliary network Q, but still embed the passwords.
w/o $\mathcal{L}_{\mathbf{rec\_cls}}$: We remove the face classification loss on the reconstruction.

Fig. 8 shows the typical drawbacks of each ablation model. w/o $\mathcal{L}_{dis}$ shows that $\mathcal{L}_{dis}$ is necessary to achieve semantic-level multimodality on both anonymization and wrong reconstruction. w/o WR shows that without training for wrong reconstructions, the transformer fails to conceal identities when given incorrect passwords. w/o $\mathcal{L}_{aux}$ verifies the importance of the auxiliary network, which helps improve photo-realism and we also observe it helps with multimodality. Without $\mathcal{L}_{rec\_cls}$, the reconstruction quality suffers because of unbalanced losses.

6 Discussion

We presented a novel privacy-preserving face identity transformer with a password embedding scheme, multimodal identity change, and a multi-task learning objective. We feel that this paper has shown the promise of password-conditioned face anonymization and deanonymization to address the privacy versus accessibility tradeoff. Although relatively rare, we sometimes notice artifacts that look similar to general GAN artifacts. They tend to arise due to the difficulty of image generation itself – we believe they can be greatly reduced with more advances in image synthesis research, which can be (orthogonally) plugged into our system.

Acknowledgements. This work was supported in part by NSF IIS-1812850, NSF IIS-1812943, NSF CNS-1814985, NSF CAREER IIS-1751206, AWS ML Research Award, and Google Cloud Platform research credits. We thank Jason Ren, UC Davis labmates, and the reviewers for constructive discussions.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: CCS (2016)
2. Bao, J., Chen, D., Wen, F., Li, H., Hua, G.: Towards open-set identity preserving face synthesis. In: CVPR (2018)
3. Butler, D.J., Huang, J., Roesner, F., Cakmak, M.: The privacy-utility tradeoff for remotely teleoperated robots. In: ICHRI (2015)
4. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG (2018)
5. Cao, Y., Liu, B., Long, M., Wang, J.: Hashgan: Deep learning to hash with pair conditional wasserstein gan. In: CVPR (2018)
6. Chen, J., Wu, J., Konrad, J., Ishwar, P.: Semi-coupled two-stream fusion convnets for action recognition at extremely low resolutions. In: WACV (2017)
7. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: NeurIPS (2016)
8. Erkin, Z., Franz, M., Guajardo, J., Katzenbeisser, S., Lagendijk, I., Toft, T.: Privacy-preserving face recognition. In: PETS (2009)
9. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., Wernsing, J.: Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In: ICML (2016)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: NeurIPS (2014)
11. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in 'Real-Life' Images (2008)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
13. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. arXiv:1812.04948 (2018)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
15. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. arXiv:1512.09300 (2015)
16. Li, T., Lin, L.: Anonymousnet: Natural face de-identification with measurable privacy. In: CVPR Workshops (2019)
17. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR (2017)
18. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: ICCV (2017)
19. Mathieu, M., Couprie, C., LeCun, Y.: Deep multi-scale video prediction beyond mean square error. arXiv:1511.05440 (2015)
20. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv:1411.1784 (2014)
21. Odena, A., Dumoulin, V., Olah, C.: Deconvolution and checkerboard artifacts. Distill (2016). https://doi.org/10.23915/distill.00003, `http://distill.pub/2016/deconv-checkerboard`
22. Perarnau, G., Van De Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. arXiv:1611.06355 (2016)

23. Pumarola, A., Agudo, A., Martinez, A.M., Sanfeliu, A., Moreno-Noguer, F.: Ganimation: Anatomically-aware facial animation from a single image. In: ECCV (2018)
24. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv:1605.05396 (2016)
25. Regmi, K., Borji, A.: Cross-view image synthesis using conditional gans. In: CVPR (2018)
26. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: ECCV (2018)
27. Rivest, R.L., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Communications of the ACM (1978)
28. Ryoo, M.S., Kim, K., Yang, H.J.: Extreme low resolution activity recognition with multi-siamese embedding learning. In: AAAI (2018)
29. Ryoo, M.S., Rothrock, B., Fleming, C., Yang, H.J.: Privacy-preserving human activity recognition from extreme low resolution. In: AAAI (2017)
30. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: CVPR (2017)
31. Sun, Q., Ma, L., Joon Oh, S., Van Gool, L., Schiele, B., Fritz, M.: Natural and effective obfuscation by head inpainting. In: CVPR (2018)
32. Wang, Z., Chang, S., Yang, Y., Liu, D., Huang, T.S.: Studying very low resolution recognition using deep networks. In: CVPR (2016)
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., et al.: Image quality assessment: from error visibility to structural similarity. IEEE TIP **13**(4), 600–612 (2004)
34. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv:1411.7923 (2014)
35. Yonetani, R., Naresh Boddeti, V., Kitani, K.M., Sato, Y.: Privacy-preserving visual learning using doubly permuted homomorphic encryption. In: ICCV (2017)
36. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: ICCV (2017)
37. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (2016)
38. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
39. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)
40. Zhu, J.Y., Zhang, R., Pathak, D., Darrell, T., Efros, A.A., Wang, O., Shechtman, E.: Toward multimodal image-to-image translation. In: NeurIPS (2017)
41. Zhu, S., Urtasun, R., Fidler, S., Lin, D., Change Loy, C.: Be your own prada: Fashion synthesis with structural coherence. In: ICCV (2017)

Appendices

A Additional details

Fig. 9 shows a qualitative example of the baselines and their anonymizations/deanonymizations.

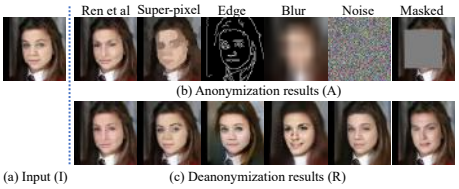


Fig. 9: Baselines. Super-pixel, Edge, Blur, Noise, Masked sacrifice photo-realism for anonymization.

We use batch normalization and our transformer T is based on the 9-block Resnet generator from [39]. We also replace the transformer T's fractionally-strided convolution layers with the resize-convolution layers in [21] to alleviate checkerboard artifacts.

For auxiliary network Q that predicts the embedded passwords, since there are a total of 2^N passwords, it is not ideal to have a 2^N-way classifier when N is large. Instead, we set up $N/4$ 16-way classifiers, with each classifier responsible for classifying its corresponding 4 bits into 2^4 classes.

Let $p_i \in \{0, ..., 15\}$ denote a 4-bit chunk of p and $\hat{p}_i$ denote the chunk predicted by Q. $Q(I, T_p I) = (f_1, , f_{N/4})$, where f_i is a 16-dim vector (logit). $Prob(\hat{p}_i = j) = Softmax(f_i)_j$.

$$\mathcal{L}_{aux}(T, Q) = -\sum_{i=1}^{N/4} \log(Prob(\hat{p}_i = p_i)). \tag{18}$$

For Q's architecture, we modify PatchGAN by switching the last convolutional layer to an average pooling layer followed by $N/4$ parallel fully-connected layers that predict the passwords.

The face recognition model F (SphereFace [17]) is trained on aligned and cropped faces, so during training, we use the same manner of aligning face by facial landmarks before inputting any faces to F as in [26]. The facial landmarks are detected by MTCNN [37]. For the VGGFace2 [4] face recognition model, we follow the same setting as the original paper: We use MTCNN [37] for face detection. The bounding boxes are then expanded by a factor 1.3x to include the whole head, which are used as network inputs.

All networks in our architecture were trained from scratch with a learning rate of 0.0001 for 15 epochs except the pre-trained face recognition model which used a learning rate of 0.00001. We use Adam solver [14] and a total batch size of 48 on 4 GPUs.

For the AMT photo-realism test, we do not include the synthesized images in which a man's face is with hair that obviously belongs to a woman; in such cases, Turkers may attribute fakeness to prior experience (it is uncommon to see a man with a woman's hairstyle) rather than photo-realism. This could be resolved by training separate face identity transformers for each gender.

B Discussion on reverse engineering

Threat models to our model are either white-box (have complete knowledge of T) or black-box (get input-output pairs from T).

Theoretically speaking, assuming all desiderata are achieved:

- Since every password leads to a unique photorealistic identity, without prior knowledge, a brute-force adversary cannot decide which one is correct.
- Adversaries $\mathcal{V}$ in the form of $\mathcal{V}_1(T_pI) = \hat{p}$ or $\mathcal{V}_2(T_pI) = \hat{I}$ wont work. We can use any password p to anonymize and deanonymize T_pI and still get T_pI, but in this case $\mathcal{V}_1$ should output $-p$:

$$\forall p, \hat{p} = \mathcal{V}_1(T_pI) = \mathcal{V}_1(T_{-p}(T_pT_pI)) = -\hat{p}, \textit{contradict}! \tag{19}$$

  Similar argument applies to $\mathcal{V}_2$. Note that different from adversaries, our auxiliary network Q also takes the original face as input.

In pracitce, due to existing artifacts in GANs, the desiderata are not perfectly achieved. And thus our current model cannot achieve this theoretical robustness against adversaries. We believe 1) Orthogonally plugging in better image synthesizing techniques; 2) Explicitly introducing robustness against adversaries are the future directions to pit against reverse engineering.

C Discussion on wrong reconstruction better hides identity

Both qualitative results and AMT studies show that Wrongly Recovered faces (*WR*s) better hide identities. We believe this is happening because:

- *WR* has less constraints to satisfy compared to Anonymized faces (A) in our loss formulation. Our training process could lead *WR* to become more optimized for the face classification loss as it does not need to care about the reconstruction loss, while A does need to be optimized to allow reconstruction of Recovered faces (R).
- *WR* is a result of two transformations from the input face rather than one. while A is a result of only a single transformation. More transformations lead to more identity changes (though we also notice more artifacts in *WR* than A).

Increasing the weight of the face classification loss $\mathcal{L}_{adv}$ applied to A may make A hide identity better.

D Why do we update the face classifier during training?

This is an adversarial learning setting that makes the transformer more robust. During each generators stage, we train T to make T_pI have a different identity from I. During each discriminators stage, we train F to correctly classify I as well as classify T_pI as y_I, i.e., see through the disguise of T_pI. T and F compete against each other so that our anonymization has certain robustness under the

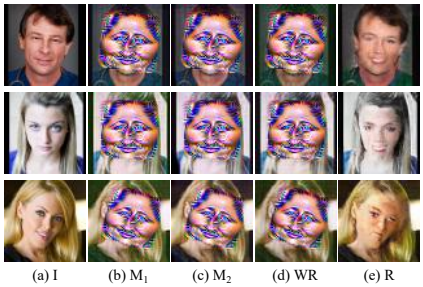


Fig. 10: Ablation study on CASIA-WebFace trained with non-adversarial face classification loss Eq. 20 , which shows that this loss dominates the multi-task learning objective quickly, so adversarial training on face classification is necessary.

attack of finetuning F. We dont want to disturb the pretraining of F too much, so we set a much lower learning rate for F, see Sec. 1.

We also did an ablation study where the face flassifier F is fixed during training in a non-adversarial training manner, i.e., we replace Eq. 10 in the main paper with:

$$\begin{aligned}\mathcal{L}_{non\_adv}(T) = &- \mathbb{E}_{(I,p)}\mathcal{L}_{CE}(F(T_pI), y_I) \\ &- \mathbb{E}_{(I,p'\neq -p)}\mathcal{L}_{CE}(F(T_{p'}T_pI), y_I),\end{aligned} \tag{20}$$

Fig. 10 shows the common failure pattern: the anonymizations are no longer photorealistic but all have a common very fake face and reconstruction also suffers. These results indicate that this setting does not work. As shown from the loss curve, the misclassification loss quickly turns into large magnitude and dominates the full objective. On the other hand, the adversarial training makes the misclassification loss not easily satisfied and not dominating.

E Additional results

In Fig. 11, we show all 7 out of 150 pairs (4.7%) that turkers report as the input and anonymized faces belonging to the same person. Even though the turkers reported "yes", our transformer still works to some extent – it changes color of skin/eyes, shape of eyes/nose/mouth/facial muscles. The same background and the same hair styles may have confused the turkers. In addition, they are mostly hard cases: dim light, side faces, heavy paints, and grayscale images. For these cases we do not have enough samples in the training set. If we collect more samples of these cases, we expect the model to perform better.

The quantitative reconstruction results on FFHQ [13] is 0.0602/0.0471/0.0509/0.0057 for LSIPS/SSIM/L1/L2, as a supplement for Table 2 in the main paper, which indicates that our transformer generalizes well on the deanonymization task on FFHQ, a dataset with plentiful variation in age, ethnicity and image background.

We show more qualitative results on CASIA [34] in Fig. 12. For faces of different hair styles, poses, and ages, our model produces high-quality results.

Fig. 11: All pairs of inputs (top) & anonymizations (bottom) turkers reported as same person. Our model still works to some extent.

Fig. 13 shows qualitative face detection results when applying an off-the-shelf face detector (MTCNN [37]) on the transformed images, see Table 3 in the main paper for quantitative results. The good performance demonstrates that normal computer vision algorithms developed on real images can be directly applied on our transformed faces, which is a great advantage over traditional face anonymization approaches.

F Image in the wild

In Fig. 14, we show that with the help of an off-the-shelf face detector, MTCNN [37], our system works well on images in the wild. The anonymized and deanonymized face areas fit well into the original image. Please also check our uploaded video at `https://youtu.be/FrYmf-CL4yk`, which demonstrates that our model can be consistent in time.

G Further exploration of the password scheme

We further investigate how our password scheme works and what the transformer learns. Since the 16-bit password space has a total of 65,536 different passwords, which is a very large space to explore, we trained an additional model with 8-bit password scheme for experiments in this section.

We show the modifications associated with all the passwords for the exemplar input images (Fig. 15) in Fig. 16, 17, 18 respectively, where Fig. 15(a) and Fig. 15(b) are both children and Fig. 15(c) is more different in age and appearance.

From the qualitative results we observe that similar original faces lead to similar modifications when given the same password. Interestingly, our transformer achieves gender equality – half of the passwords transform to female identities and the remaining half transform to males regardless of the inputs' genders. And all the transformed faces satisfy our anonymization goal. These qualitative results also show that more diverse passwords lead to more diverse anonymized faces.

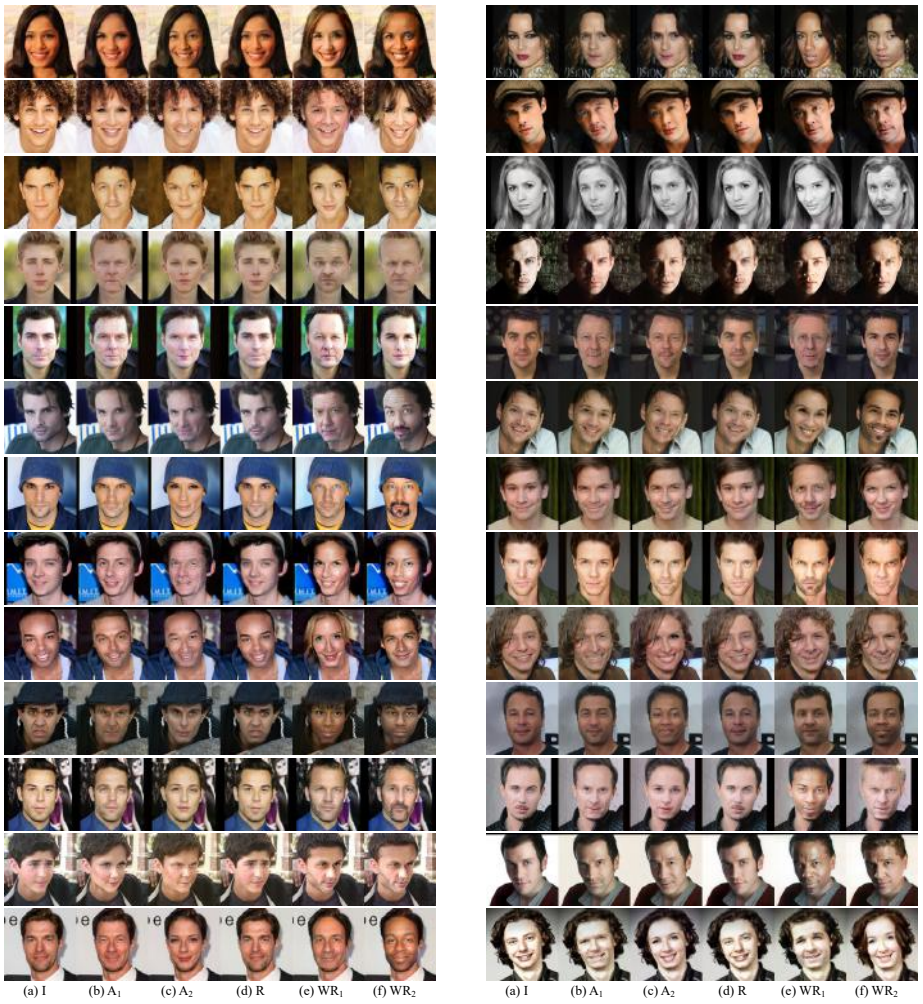
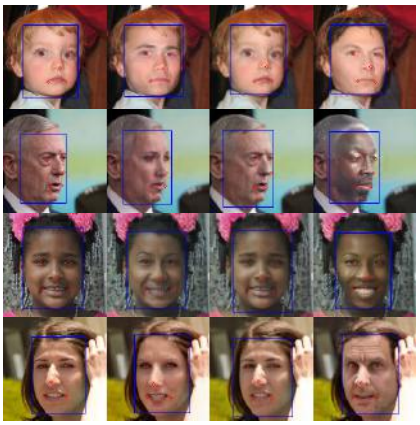


Fig. 12: Additional qualitative results on CASIA. I: original image, $A_{1,2}$: anonymized faces conditioned on different passwords, $R/WR_{1,2}$: recovered faces with correct/wrong passwords.

(a) I (b) A (c) R (d) WR

(a) CASIA



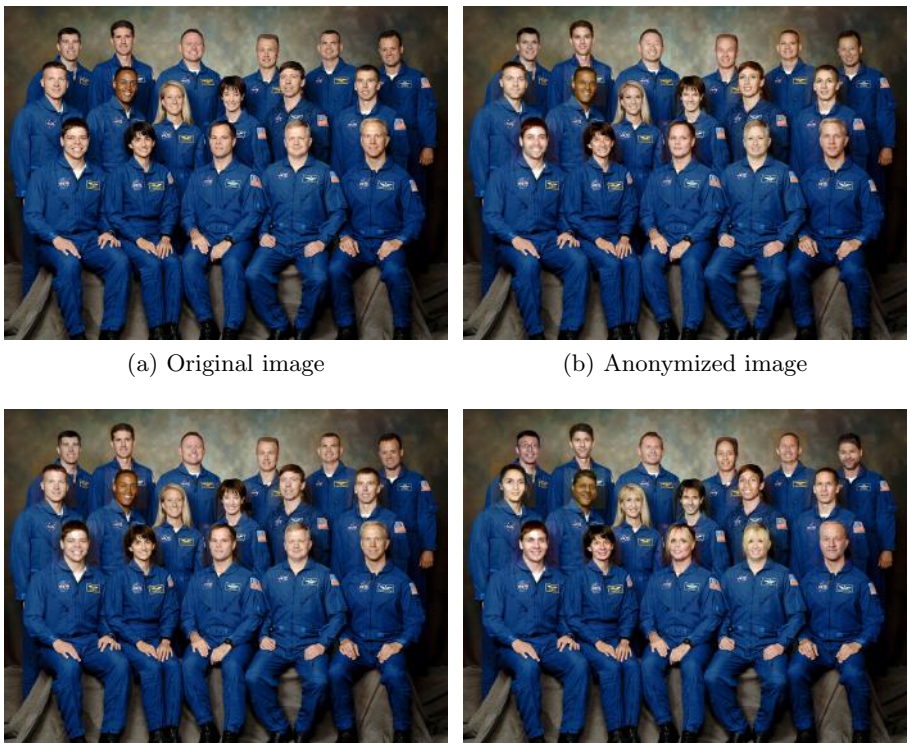
(a) I (b) A (c) R (d) WR

(b) FFHQ

(a) I (b) A (c) R (d) WR

(c) LFW

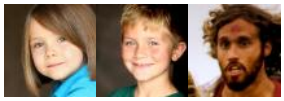
Fig. 13: Qualitative face detection results on transformed images. Photo-realism makes existing computer vision algorithms work on our transformed images directly.



(a) Original image (b) Anonymized image

(c) Deanonymized image with correct password (d) Deanonymized image with wrong password

Fig. 14: Image in the wild example.



(a) (b) (c)

Fig. 15: Original images. (a) and (b) are similar. (c) is more different from (a) and (b).

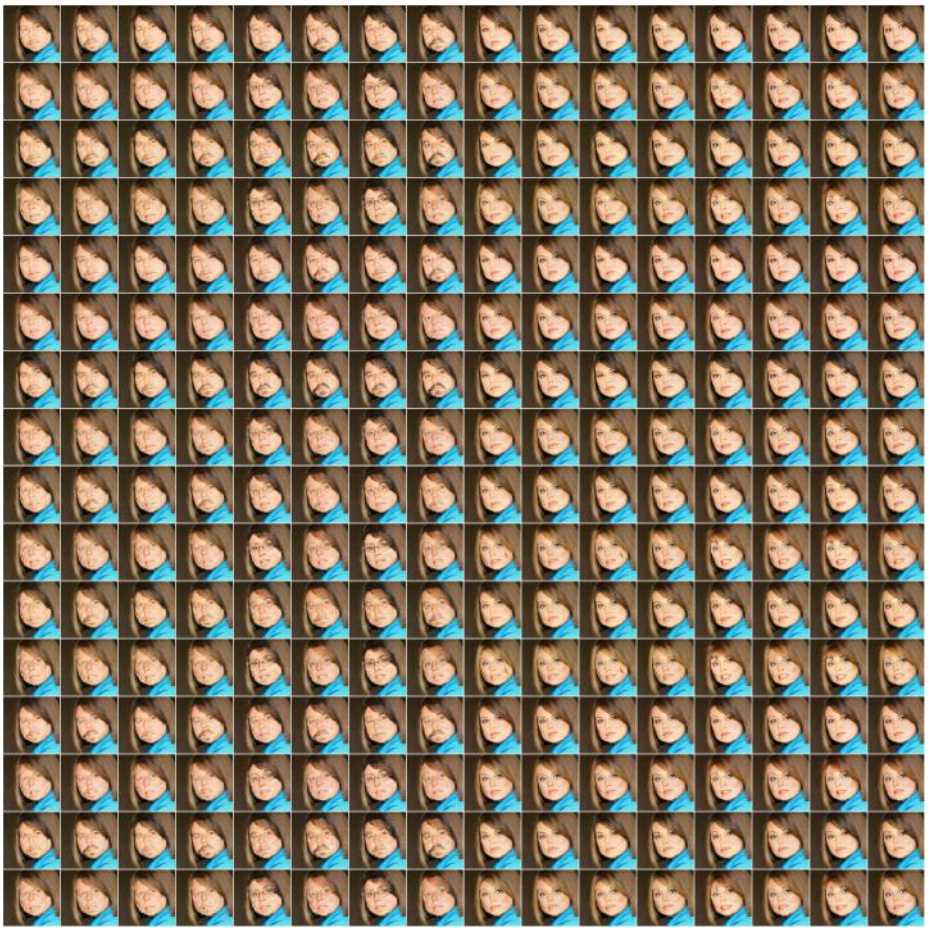


Fig. 16: Modifications associated with all the passwords whose original face image is Fig. 15(a).

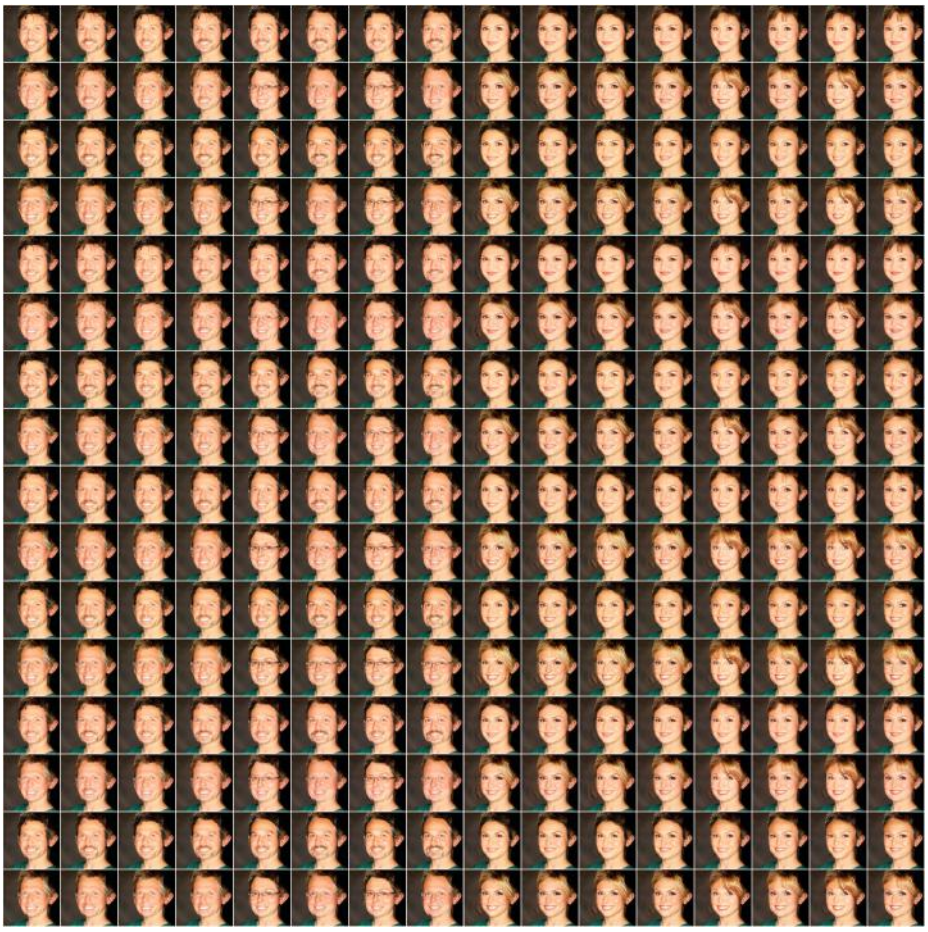


Fig. 17: Modifications associated with all the passwords whose original face image is Fig. 15(b).

Fig. 18: Modifications associated with all the passwords whose original face image is Fig. 15(c).