Supplementary Material: 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

1 Additional details

Fig. 1 shows a qualitative example of the baselines and their anonymizations/deanonymizations.

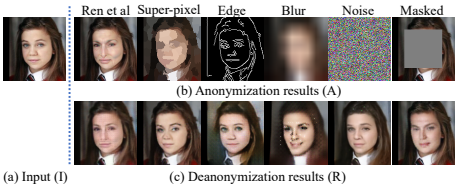


Fig. 1: Baselines. Super-pixel, Edge, Blur, Noise, Masked sacrifice photo-realism for anonymization.

We use batch normalization and our transformer T is based on the 9-block Resnet generator from [9]. We also replace the transformer T's fractionally-strided convolution layers with the resize-convolution layers in [5] 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. $Q(I, T_p I) = (f_1, \ldots, f_{N/4})$, where f_i is a 16-dim vector (logit). $Prob(\hat{p}_i = j) = Softmax(f_i)_j$.

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

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 [4]) 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 [6]. The facial landmarks are detected by MTCNN [8]. For the VGGFace2 [1] face recognition model,

we follow the same setting as the original paper: We use MTCNN [8] 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 [3] 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.

2 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 $\mathcal{V}_2(T_pI) = \hat{I}$ won't 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 $-p'$:

$$\forall p', \hat{p} = \mathcal{V}_1(T_pI) = \mathcal{V}_1(T_{-p'}(T_{p'}T_pI)) = -\hat{p}', contradict! \tag{2}$$

  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.

3 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).

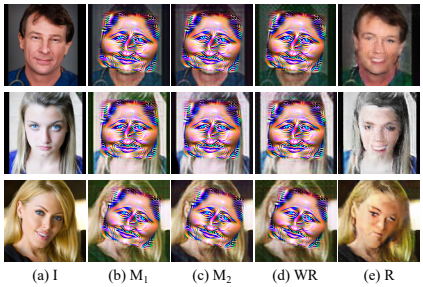


Fig. 2: Ablation study on CASIA-WebFace trained with non-adversarial face classification loss Eq. 3 , which shows that this loss dominates the multi-task learning objective quickly, so adversarial training on face classification is necessary.

– *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.

4 Why do we update the face classifier during training?

This is an adversarial learning setting that makes the transformer more robust. During each generator's stage, we train T to make T_pI have a different identity from I. During each discriminator's 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 attack of finetuning F. We don't 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:

$$\begin{aligned}\mathcal{L}_{non\_adv}(T) = &- \mathbb{E}_{(I,p)}\mathcal{L}_{CE}(F(T_pI), y_I) \\ &- \mathbb{E}_{(I,p'\neq -p)}\mathcal{L}_{CE}(F(T_{p'}T_pI), y_I),\end{aligned} \tag{3}$$

Fig. 2 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.

Fig. 3: All pairs of inputs (top) & anonymizations (bottom) turkers reported as same person. Our model still works to some extent.

5 Additional results

In Fig. 3, 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 [2] 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 [7] in Fig. 4. For faces of different hair styles, poses, and ages, our model produces high-quality results.

Fig. 5 shows qualitative face detection results when applying an off-the-shelf face detector (MTCNN [8]) 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.

6 Image in the wild

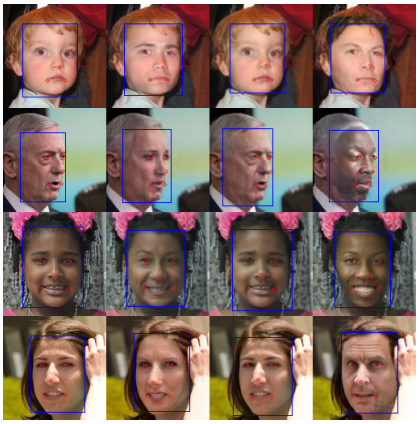
In Fig. 6, we show that with the help of an off-the-shelf face detector, MTCNN [8], our system works well on images in the wild. The anonymized and deanonymized face areas fit well into the original image. Please also check our uploaded video at `https://youtu.be/FrYmf-CL4yk`, which demonstrates that our model can be consistent in time.

7 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,

Fig. 4: 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) CASIA



(b) FFHQ

(c) LFW

Fig. 5: Qualitative face detection results on transformed images. Photo-realism makes existing computer vision algorithms work on our transformed images directly.

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. 7) in Fig. 8, 9, 10 respectively, where Fig. 7(a) and Fig. 7(b) are both children and Fig. 7(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.

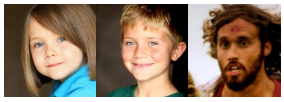
(a) Original image

(b) Anonymized image

(c) Deanonymized image with correct password

(d) Deanonymized image with wrong password

Fig. 6: Image in the wild example.



(a) (b) (c)

Fig. 7: Original images. (a) and (b) are similar. (c) is more different from (a) and (b).

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. 8: Modifications associated with all the passwords whose original face image is Fig. 7(a).

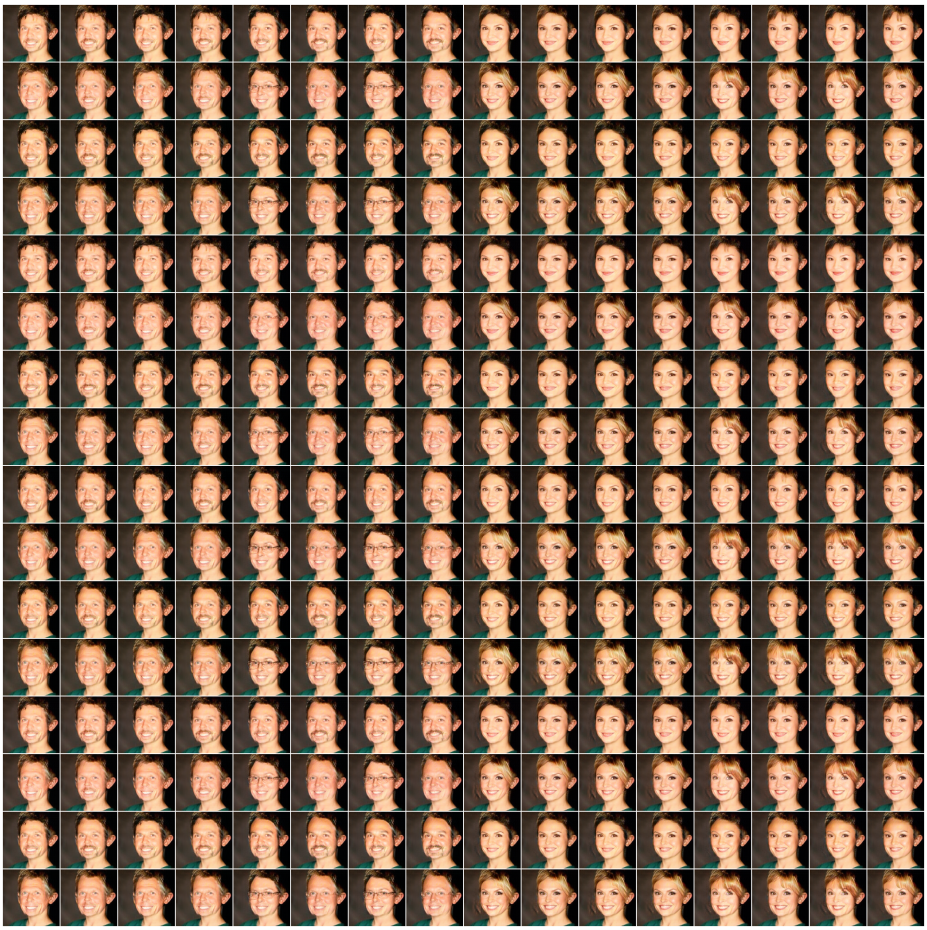


Fig. 9: Modifications associated with all the passwords whose original face image is Fig. 7(b).

Fig. 10: Modifications associated with all the passwords whose original face image is Fig. 7(c).

References

1. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: FG (2018)
2. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. arXiv:1812.04948 (2018)
3. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
4. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: CVPR (2017)
5. 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`
6. Ren, Z., Lee, Y.J., Ryoo, M.S.: Learning to anonymize faces for privacy preserving action detection. In: ECCV (2018)
7. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv:1411.7923 (2014)
8. 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)
9. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV (2017)