Prediction Exposes Your Face: Black-box Model Inversion via Prediction Alignment: Supplementary Material

Yufan Liu[1,2], Wanqian Zhang[1✉], Dayan Wu[1], Zheng Lin[1,2], Jingzi Gu[1], and Weiping Wang[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
{liuyufan,zhangwanqian,wudayan,linzheng,gujingzi,wangweiping}@iie.ac.cn

A Evaluation metrics details

The detailed description of the evaluation metrics we use is as follows:

Attack Acc evaluates the accuracy of the reconstructed image on the target identity via an evaluation classifier, considered as the replacement of human judgment. We use the model proposed by Cheng *et al*. [1] for evaluation, which is pre-trained on MS-Celeb-1M [2] and fine-tuned on the private dataset.

KNN Dist is the average L_2 distance of reconstructed image features and features of the nearest image corresponding to target identity.

Feat Dist is the average L_2 distance of the reconstructed image features and the centroid of image features corresponding to the target identity. Both of them are calculated with features of the layer before fully connected layer in classifier.

LPIPS evaluates the perceptual similarity of reconstructed image and target identity image, which is more in line with human perception.

B More analyses on training data

In order to study the impact of training data on our method, we first assess the impact of synthesized data in the public dataset under the standard setting. The results in Tab. 1 show that the lack of synthesized data lead to a clear decline in attack performance although slightly better perceptual quality. As the previous analysis indicates in [7], synthesized data is generated from real latent codes in the $\mathcal{W}^+$ space, which helps our PAE to align the prediction vector space with the $\mathcal{W}^+$ space. However, real latent codes might be outside the real image domain, thereby causing a slight impact on the LPIPS.

In addition, we randomly select the same amount of data from the FFHQ dataset and synthesized data, instead of our top-n selection from public dataset, to train our encoder. The results in Tab. 2 show that training our encoder with unselected public data results in a 42% decrease from our baseline method. We believe that unselected training data fails to allow our encoder to learn a good

✉Corresponding author

mapping relationship from prediction vectors to $\mathcal{W}^+$ space. In other words, since public images do not belong to any target identity, there may exist some images that have rather even prediction values on each identity, exerting negative effect on optimizing our PAE encoder.

Table 1: Attack performance when synthesized data is not in the public dataset. Target model is VGG16 trained on CelebA. Public dataset also comes from CelebA, whereas there is no identity overlap with the private dataset.

Type	Attack Acc↑	KNN Dist↓	Feat Dist↓	LPIPS↓
w/o synthesized data	0.705	1081.5	973.3	0.241
ours	**0.715**	**1039.1**	**920.8**	0.244

C More analyses on aligned ensemble attack

We compare our baseline with the variant of adapting aligned ensemble attack on images' latent codes, as shown Fig. 1. Specifically, we directly feed the set of top-n public images, classified as the target identity, into the pre-trained image encoder to reconstruct images [7]. As can be seen, if we weight both the latent code in the $\mathcal{W}^+$ space and the feature tensor from $\mathcal{F}$, the attack performance becomes decreased as shown by the orange bar vs. green bar. One possible reason is that weighted ensembling of feature tensors would compromise the integrity of the features. We also keep the input unchanged and only ensemble the latent code in the $\mathcal{W}^+$ space. The results show an improvement in attack accuracy, but there is still a gap with the baseline, as shown by the blue bar vs. green bar.

Table 2: Attack performance when there is not data selection in our method. The target model is Face.evoLVe trained on FaceScrub with FFHQ as the public dataset. The test accuracy of the target model is 96%.

Type	Attack Acc↑	KNN Dist↓	Feat Dist↓	LPIPS↓
w/o data selection	0.34	1294.3	1290.5	0.249
ours	**0.59**	**1243.8**	**1256.0**	**0.246**

D Ablation on StyleGAN

To further study the impact of StyleGAN in our method, we compare two variants with our baseline. On the one hand, we replace the StyleGAN in our pipeline with the conventional GAN in RLB-MI [3]. To keep consistency with our pipeline, we only add a linear layer after our PAE encoder to align dimensions. On the other hand, we switch our PAE encoder to one single linear layer, and replace

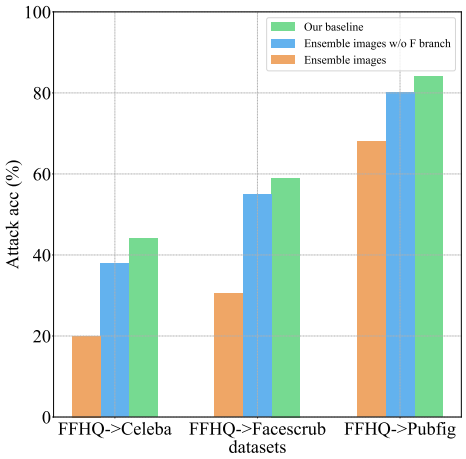

Fig. 1: Comparison on aligned ensemble attack between ensembled latent codes of prediction vectors and those of images.

$\mathcal{L}_{total}$ with $\mathcal{L}_{mse}$. Also, during the attack phase, we use one-hot vectors as input instead of our aligned ensemble attack. By doing this, we only replace the generative network of LB-MI [6] with that of StyleGAN, do the minimal modifications for adjusting the pipeline, and keep other parts unchanged.

From the results of Tab. 3, we can draw the following three conclusions:

1. Due to the inferior priors of a regular GAN compared to StyleGAN, and its lack of the disentangled latent space, it performs worse than StyleGAN in the same pipeline.

2. For the LB-MI pipeline, using the more powerful StyleGAN does not improve the attack performance. We believe that in the LB-MI pipeline, there is a lack of a good encoder to map prediction vectors into the latent space of the generator, thus the powerful prior of StyleGAN cannot be fully utilized. Also, the one-hot vector leads to too much loss of information, which is insufficient for high-fidelity reconstructions.

3. Comparing two variants with the results from our baseline, it underlines that the key characteristics of our pipeline and StyleGAN are complementing each other. Concretely, StyleGAN possesses strong prior knowledge. Meanwhile, our specially designed PAE encoder can effectively utilize StyleGAN's disentangled space. Moreover, our loss function is better equipped to guide the alignment of the prediction vector space and $\mathcal{W}^+$ space. During the attack phase, our proposed alignment integrated attack is also based on the disentangled characteristics of StyleGAN's $\mathcal{W}^+$ space.

E User study

To further quantify the accuracy of the images reconstructed using our method, we introduce human study to evaluate the effectiveness of the attack from hu-

Table 3: Ablation on StyleGAN. The target model is VGG16 trained on CelebA (also as the public dataset). **Variant1**: replace the StyleGAN in our pipeline with the conventional GAN in RLB-MI. **Variant2**: combine the StyleGAN with LB-MI.

Type	Attack Acc↑	KNN Dist↓	Feat Dist↓	LPIPS↓
variant1	0.125	1629.0	1565.1	0.304
variant2	0.005	2347.7	2306.2	0.527
ours	**0.715**	**1039.1**	**920.8**	**0.244**

Table 4: User study results. User preferences indicate which option users prefer between two methods of inversion images. The results show that users prefer the outcomes produced by our method.

Method	User Preference↑
RLB-MI	25.60%
ours	**74.40%**

man subjective perspective. We design our experiment according to the setup in previous work [4]. As shown in Fig. 2, we randomly select five real images for each target identity from the private dataset to serve as references for the users. The users need to choose between two options, deciding which one image matches the reference identity better. The two options are the images obtained from attacks on the reference identity using RLB-MI and our method respectively. The target model is Face.evoLVe trained on CelebA (also as the public dataset). We randomly select 50 identities and hand them over to 20 users for experimentation.

The results in Tab. 4 show that user preferences for the inversion results of our method reach 74.40%, while RLB-MI only with 25.60%. This result indicates that users subjectively prefer our results, therefore our method also has superiority in visual quality.

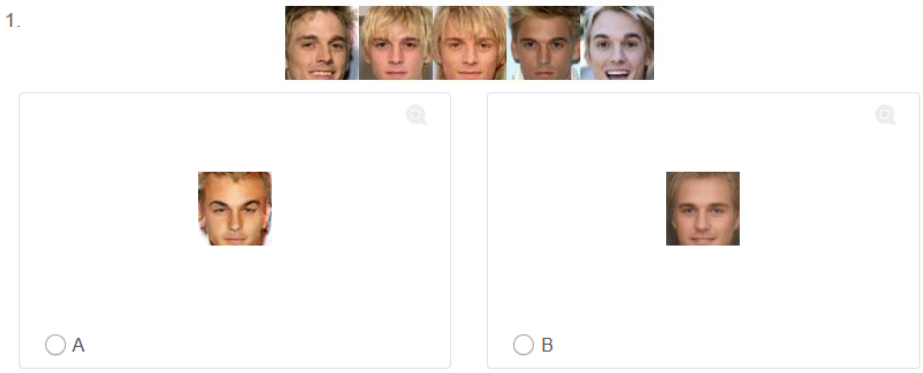

Fig. 2: User study interface

Table 5: Comparison of implicit attribute recovery. We compare the state-of-the-art methods with our method, and the attack success rate is measured by the attribute classifier trained on CelebA.

Attributes	Attack Acc↑			
	PLG-MI	RLB-MI	BREP-MI	**ours**
Bald	0.985	0.982	0.980	**0.986**
Big Lips	0.424	0.421	0.373	**0.435**
Brown Hair	0.516	0.509	0.451	**0.565**
Chubby	0.876	0.825	0.882	**0.884**
Double Chin	0.873	0.877	0.863	**0.884**
Eyeglasses	0.905	0.877	0.922	**0.928**
Gray Hair	0.935	0.895	0.941	**0.942**
Heavy Makeup	0.517	0.439	0.451	**0.551**
High Cheekbones	0.667	0.667	0.725	**0.899**
Narrow Eyes	0.902	0.930	0.882	**0.942**
Wearing Necktie	0.772	0.825	0.824	**0.855**
All	0.695	0.702	0.712	**0.726**

F Attributes recovery

In this section, we evaluate the performance of our method in attribute restoration to further verify the effectiveness of image reconstruction. Specifically, we train 40 attribute classifiers on the CelebA dataset. Since different images of the same identity may have different attribute features, it is reasonable that the same identity may be different due to different occasions, dress ups and other factors. We take the union of the attributes of different images belonging to the same identity, to form the feature attribute set for calculating the performance of attribute restoration. We calculate the attack accuracy of identity feature on images that are successfully attacked using different methods.

The results in Tab. 5 show that our method is the best in terms of attack performance on all 40 attributes of the CelebA dataset, with an improvement of 3.4% over the black-box method RLB-MI and even 4.5% higher than the white-box method PLG-MI. We also select some of these attributes and calculate the attack accuracy on individual attributes. On the selected individual attributes, our attack performance is still higher than that of other methods. These experimental results demonstrate the advantages of our method in attribute restoration, further proving that our PAE encoder aligns the prediction vector space with the more disentangled $\mathcal{W}^+$ space. Furthermore, the aligned ensemble attack contributes to aggregate different feature attributes of the same identity together.

G Effect of attribute editing on MI attack

We also explore the potential for more targeted manipulation of facial attributes within $\mathcal{W}^+$ space. To integrate attribute editing into the inversion process, given a target identity, we obtain its latent code $\mathcal{W}_{ens}$, visualize its heatmap using Grad-CAM, and utilize InterFaceGAN [5] to find editing directions in $\mathcal{W}^+$ space of StyleGAN2. As in Figure 3, we select two editing directions according to the heatmap: $\mathcal{W}_{golden\_hair}$ and $\mathcal{W}_{bushy\_eyebrows}$. Thus, we can conduct the interpolation as: $\mathcal{W}_{edit}$=$\mathcal{W}_{ens}$+$\alpha\mathcal{W}_{golden\_hair}$(or $\mathcal{W}_{bushy\_eyebrows}$) to edit $\mathcal{W}_{ens}$.

Fig. 3: Heatmap of target identity on target model.

Figure 4 shows the effect of attribute editing on Model Inversion attack: 1) Even though our PAE encoder is trained using the prediction vectors as input, we can still effectively conduct edits using the editing directions obtained from the image-to-image reconstruction pipeline [5]. This *coincidentally* proves that our encoder has achieved alignment between the predictions and the $\mathcal{W}^+$ space. 2) We are astonished to discover that manipulating latent code in a specific direction indeed boosts the classification accuracy of inversion images in our pipeline. This idea holds immense potential for further improving the efficacy of our method. 3) In the MI attacks, we argue that manipulating facial attributes still needs to address two problems: What are the properties closely related to the classification? How to determine the amount of attribute editing to improve the attack accuracy? We will explore these in the future.

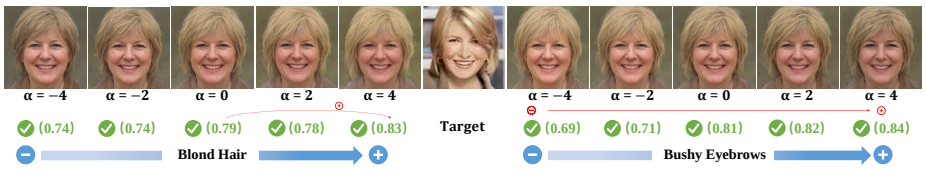

Fig. 4: We attack the target identity and manipulate the obtained $\mathcal{W}_{ens}$ to edit attributes.

References

1. Cheng, Y., Zhao, J., Wang, Z., Xu, Y., Karlekar, J., Shen, S., Feng, J.: Know you at one glance: A compact vector representation for low-shot learning. In: ICCV Workshops. pp. 1924–1932. IEEE Computer Society (2017)
2. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV (3). Lecture Notes in Computer Science, vol. 9907, pp. 87–102. Springer (2016)
3. Han, G., Choi, J., Lee, H., Kim, J.: Reinforcement learning-based black-box model inversion attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20504–20513 (2023)
4. Nguyen, B.N., Chandrasegaran, K., Abdollahzadeh, M., Cheung, N.M.M.: Label-only model inversion attacks via knowledge transfer. Advances in Neural Information Processing Systems **36** (2024)
5. Shen, Y., Gu, J., Tang, X., Zhou, B.: Interpreting the latent space of gans for semantic face editing. In: CVPR (2020)
6. Yang, Z., Zhang, J., Chang, E., Liang, Z.: Neural network inversion in adversarial setting via background knowledge alignment. In: CCS. pp. 225–240. ACM (2019)
7. Yao, X., Newson, A., Gousseau, Y., Hellier, P.: A style-based gan encoder for high fidelity reconstruction of images and videos. In: European conference on computer vision. pp. 581–597. Springer (2022)