A Closer Look at GAN Priors: Exploiting Intermediate Features for Enhanced Model Inversion Attacks

Yixiang Qiu[1,2†], Hao Fang[2†], Hongyao Yu[1†], Bin Chen[1,3,4#], MeiKang Qiu[5], and Shu-Tao Xia[2,4]

[1] Harbin Institute of Technology, Shenzhen
[2] Tsinghua Shenzhen International Graduate School, Tsinghua University
[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[4] Pengcheng Laboratory [5] Augusta University
qiuyixiang@stu.hit.edu.cn, fang-h23@mails.tsinghua.edu.cn
yuhongyao@stu.hit.edu.cn, chenbin2021@hit.edu.cn
qiumeikang@yahoo.com, xiast@sz.tsinghua.edu.cn

A. Additional Experimental Results

In this section, we report extra results for our experiment, which are not shown in the main part of the paper.

A.1 More Comparison with PLGMI [16]

While the reconstructed images from PLGMI have low fidelity and high FID [5] metric, they achieve relatively high accuracy. To make a comprehensive comparison, we evaluate the attack performance of our method against the state-of-the-art PLGMI under more experimental settings. Table 1 states the results on different target models, including ResNet-152 [4], ResNeSt-101 [17], and DenseNet-169 [6]. Our method maintains superiority in all metrics compared to PLGMI under the aforementioned settings, aligning with the conclusion that our method achieves new state-of-the-art attack performance.

A.2 More Results on Different Combinations of Datasets and Models

For the facial image classification task, We further conduct more extensive experiments under multiple combinations of public and private datasets for overall evaluation. The PPA [14] and PLGMI [16] are selected as the baseline for comparison due to its comprehensive performance in both accuracy and fidelity. The results shown in Table 5 demonstrate the outstanding performance of our method under various scenarios and verify the excellent generalizability for the proposed method.

[†] Equal contribution.
[#] Corresponding author.
This work was done while Yixiang Qiu was pre-admitted to Tsinghua University.

Table 1: Comparison results with PLGMI against different target models trained on FaceScrub [12] with the public dataset being MetFaces [7].

Target Model	Method	↑ Acc@1	↑ Acc@5	↓ δ_{eval}	↓ δ_{face}	↓FID
ResNet-152	PLGMI	0.429	0.708	0.805	243.868	166.614
	Ours	**0.904**	**0.984**	**0.882**	**138.752**	**69.937**
ResNeSt-101	PLGMI	0.571	0.791	0.814	234.669	213.955
	Ours	**0.922**	**0.983**	**0.884**	**132.609**	**76.195**
DenseNet-169	PLGMI	0.390	0.645	0.818	198.095	222.563
	Ours	**0.933**	**0.987**	**0.851**	**125.050**	**82.123**

Table 2: Comparison results against ResNet-152 trained on Stanford Dogs. The GAN prior is pre-trained on AFHQ [1] dataset.

Method	↑ Acc@1	↑ Acc@5	↓ δ_{face}	↓FID
PPA	0.950	0.996	**59.250**	32.040
Ours	**0.982**	**1.000**	70.1956	**27.282**

For the dog breed classification task, Table 2 compares attack performance against the ResNet-152 [4] trained on Stanford Dogs [3]. The encouraging results exhibit the superiority of our method in another task.

Table 3: Comparison of our method with state-of-the-art methods against ResNet-18 trained on FaceScrub in Improved precision and recall, density, and coverage metrics.

Public Dataset	Method	↑ *Precision*	↑ *Recall*	↑ *Density*	↑ *Coverage*
FFHQ	PPA [14]	0.143	0.003	0.336	0.261
	PLGMI [16]	0.005	0.000	0.003	0.004
	Ours	**0.186**	**0.069**	**0.462**	**0.314**
MetFaces	PPA [14]	**0.208**	0.000	**0.421**	0.182
	PLGMI [16]	0.001	0.024	0.000	0.000
	Ours	0.142	**0.041**	0.329	**0.192**

A.3 More Evaluation Metrics

Following PPA, we further compute the improved Precision-Recall [10] together with Density-Coverage [11] on a per-class basis to measure the sample diversity. We utilize the Inception-v3 [15] model to calculate the four metrics and make a comprehensive comparison with previous approaches, as shown in Table 3. Our method achieve highest scores in most comparisons, indicating superior sample diversity among all the methods.

A.4 More Visual Results

We show more qualitative results generated from the StyleGAN2 [9] pre-trained on the MetFaces [7] dataset in Fig 1 and Fig 2. Fig 1 compares the visual samples from different methods when attacking the ResNet-18 [4] trained on FaceScrub. Fig 2 displays the visual images generated from different end layers, showing the gradual change during the intermediate features optimization.

Table 4: Ablation study on $L = 1$. The setup aligns with Sec. 4.5.

Decomposition	$\uparrow Acc@1$	$\uparrow Acc@5$	$\downarrow$FID
$L = 0$	0.713	0.897	46.497
$N = 1$	0.766	0.927	44.553
$N = 2$	0.795	0.940	**44.435**
$N = 3$	0.825	0.956	45.442
$N = 4$	**0.833**	0.955	46.197
$N = 5$	0.832	**0.957**	47.628
$N = 6$	0.830	**0.957**	50.908
$N = 7$	0.829	0.956	57.009
IF-GMI	**0.846**	**0.965**	**42.970**

A.5 More Ablation Studies

To ensure the concrete contributions from each intermediate layer and support the selected number in Sec. 4.2, we conduct further ablation with $L = 1$ where G is decomposed in all possible ways. Following StyleGAN2 [9], we decompose the generator into style blocks, which serve as the basic intermediate layer units. StyleGAN2 comprises a total of 9 style blocks. Given that features from later blocks are higher-dimensional and more expensive to optimize, we only select the first seven blocks as intermediate layers. For ablation with $L = 1$, we define "$N = k$" as the scenario where the blocks before the $(k + 1)$-th block are designated as G_1, and the remaining blocks are designated as G_2. Therefore, the optimized intermediate feature becomes the output of the k-th style block. The settings of $N = k, k \in [1, \ldots, 7]$ cover all possible decomposition ways when $L = 1$. The ablation results are presented in Table 4. $L = 0$ is defined as a special case where only latent vectors are optimized. The last row showcases results under the standard decomposition of $L = 3$.

As illustrated in Table 4 in terms of $Acc@1$, $Acc@5$ and FID, the first four intermediate features achieve a good balance between accuracy and image realism. Thus, we opt the first three layers when $L = 3$ for improved performance.

Table 5: Comparison results against more combinations of different datasets and models. Public→Private denotes that the GAN prior is pre-trained on the public dataset to attack the target model trained on the private dataset.

Public→Private	Target Model	Method	↑ Acc@1	↑ Acc@5	↓ δ_{eval}	↓ δ_{face}	↓FID
FFHQ→FaceScrub	ResNet-152	PPA	0.927	0.989	0.716	**123.250**	46.690
		PLGMI	0.967	0.999	0.772	170.220	211.217
		Ours	**0.981**	**0.999**	**0.691**	131.902	**45.703**
	ResNeSt-101	PPA	0.940	0.992	0.720	**119.790**	46.300
		PLGMI	0.831	0.950	0.873	148.555	177.807
		Ours	**0.987**	**0.998**	**0.660**	121.791	**42.768**
	DenseNet-169	PPA	0.953	0.995	0.687	115.200	46.720
		PLGMI	0.986	0.998	0.659	145.948	134.844
		Ours	**0.986**	**0.998**	**0.653**	**114.682**	**42.896**
FFHQ→CelebA	ResNet-152	PPA	0.806	0.946	0.736	**312.580**	40.430
		PLGMI	0.504	0.739	1.630	876.689	70.991
		Ours	**0.922**	**0.985**	**0.680**	315.543	**30.394**
	ResNeSt-101	PPA	0.830	0.954	0.751	299.730	44.040
		PLGMI	0.871	0.968	1.640	709.896	120.983
		Ours	**0.935**	**0.987**	**0.705**	**298.363**	**35.389**
	DenseNet-169	PPA	0.731	0.905	0.764	**312.320**	43.240
		PLGMI	0.758	0.921	1.622	627.920	115.409
		Ours	**0.840**	**0.955**	**0.726**	314.669	**36.568**
MetFaces→CelebA	ResNet-152	PPA	0.396	0.643	1.063	387.810	**74.030**
		PLGMI	0.183	0.391	1.627	682.042	181.555
		Ours	**0.817**	**0.945**	**0.815**	**334.843**	81.179
	ResNeSt-101	PPA	0.371	0.629	1.124	387.610	**75.070**
		PLGMI	0.711	0.896	1.617	654.594	175.523
		Ours	**0.814**	**0.942**	**0.897**	**319.716**	76.831
	DenseNet-169	PPA	0.309	0.558	1.096	396.810	**81.720**
		PLGMI	0.443	0.704	1.610	615.914	173.141
		Ours	**0.670**	**0.868**	**0.893**	**341.303**	82.597

B. Additional Evaluation on robustness

To evaluate the robustness of the proposed method, we select the BiDO [13] as the defense strategy to protect the target model from MI attacks while inducing negligible utility loss. Following the default settings in [13], we train a ResNet-152 on FaceScrub as the defensive target model. Without loss of generality, we evaluate our method and baselines on the first 100 classes of the FaceScrub dataset. See Table 6 for quantitative results. Our method maintains the superior performance and achieves the least decrease of 14.1% in the Acc@1 metric after the defense, indicating the distinguished robustness against the BiDO defense.

C. Additional Experimental Details

For the target and evaluation models, we use those provided by PPA [14]. The test accuracy of each model is shown in Table 7.

Table 6: Robustness evaluation against the BiDO defense strategy with the ResNet-152 trained on FaceScrub. The GAN prior is pre-trained on MetFaces.

Method	$\uparrow Acc@1$	$\uparrow Acc@5$	$\downarrow \delta_{eval}$	$\downarrow \delta_{face}$	$\downarrow$FID
PPA	0.619	0.873	1.010	149.621	**69.263**
Ours	**0.906**	**0.985**	**0.861**	**134.634**	76.108
PPA+BiDO	0.356	0.639	1.119	167.318	**70.759**
Ours+BiDO	**0.765**	**0.924**	**0.957**	**146.949**	74.197

Table 7: The test accuracy of target and evaluation models in the experiments. Note that the models and the values in the table are from PPA [14]

	FaceScrub	**CelebA**	**Stanford Dogs**
ResNet-18	94.22%	-	-
ResNet-152	93.74%	86.78%	71.23%
DenseNet-169	95.49%	85.39%	74.39%
ResNeSt-101	95.35%	87.35%	75.07%
Inception-v3	96.20%	93.28%	79.79%

For initial selection stage, we select 50 latent vectors from a large batch of candidates. Following PPA, we set the number of candidates as 2000 for attacking FaceScrub models and 5000 for attacking CelebA models. For each intermediate feature optimization, we utilize Adam optimizer with a learning rate of 0.005 and $\beta = (0.1, 0.1)$ to optimize the intermediate features $\mathbf{f}$ and extended latent vectors $\mathbf{w}$. When attacking the FaceScrub models, we set the optimization steps list as $[50, 10, 10, 10]$ for each intermediate feature. When attacking the CelebA models, the steps list is set as $[70, 25, 25, 25]$. Guided by the theory [2] that ascended radii of the l_1 ball lead to better results, we set the sequence of radius as $[1000, 2000, 3000, 4000]$ for optimization of both $\mathbf{f}$ and extended $\mathbf{w}$.

Table 8: FID scores between various datasets. Note that the values in the table are from PPA [14].

Dataset 1	**Dataset 2**	**FID**
FFHQ	FaceScrub	77.90
FFHQ	CelebA	59.48
MetFaces	FaceScrub	104.33
MetFaces	CelebA	93.64

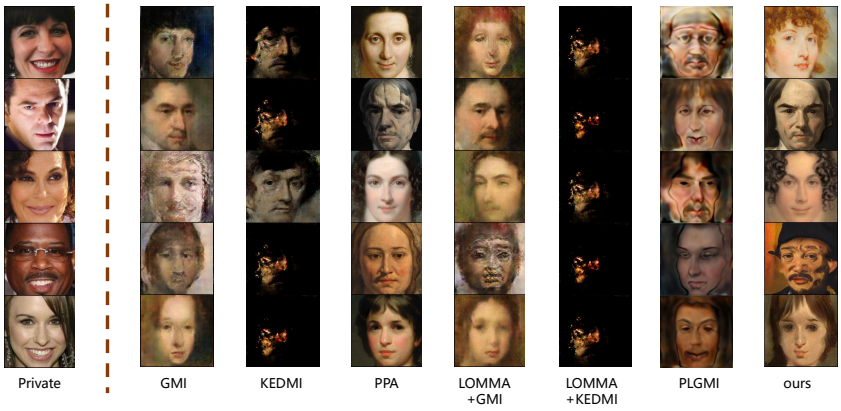

Fig. 1: Visual comparison of reconstructed images from different methods against the ResNet-18 [4] trained on FaceScrub. The GAN prior is pre-trained on MetFaces. The first column shows ground truth images of the target class in the private dataset.

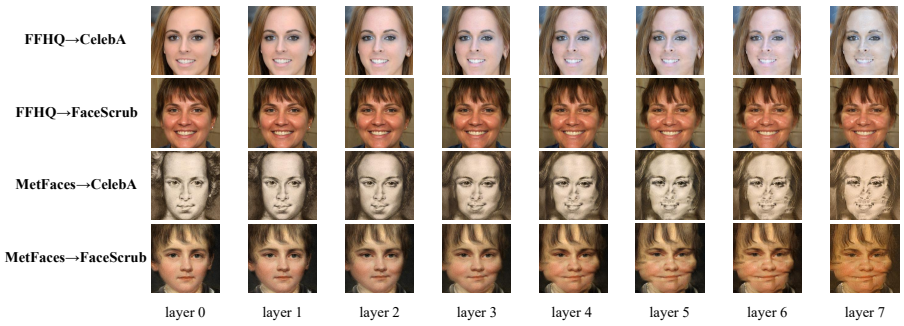

Fig. 2: Visual results generated from different end layers. $DatasetA \rightarrow DatasetB$ denotes that the $DatasetA$ is the public dataset and $DatasetB$ is the private dataset.

D Distributional Shift Between Datasets

Following PPA [14], we employ FID scores to measure distributional distance and the corresponding results are summarized in Table 8. A higher value indicates a greater distributional disparity between datasets, revealing that FFHQ [8] is an easier OOD scenario compared to the more challenging MetFaces [7]. Our method is confirmed to be effective across various OOD scenarios, particularly under tougher scenarios.

E. Limitations & Future work.

Despite the superior performance in most metrics under multiple experimental settings, we find that the generated images achieve relatively high FID scores.

The potential reason is that we directly perform backpropagation on the intermediate features that are higher-dimensional than the latent code while utilizing the same loss originally designed for latent vectors. In the future work, we will explore a better solution to design an appropriate optimization strategy for intermediate features.

References

1. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8188–8197 (2020)
2. Daras, G., Dean, J., Jalal, A., Dimakis, A.G.: Intermediate layer optimization for inverse problems using deep generative models. arXiv preprint arXiv:2102.07364 (2021)
3. Dataset, E.: Novel datasets for fine-grained image categorization. In: First Workshop on Fine Grained Visual Categorization, CVPR. Citeseer. Citeseer. Citeseer (2011)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
7. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. Advances in neural information processing systems **33**, 12104–12114 (2020)
8. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019)
9. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
10. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems **32** (2019)
11. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: International Conference on Machine Learning. pp. 7176–7185. PMLR (2020)
12. Ng, H.W., Winkler, S.: A data-driven approach to cleaning large face datasets. In: 2014 IEEE international conference on image processing (ICIP). pp. 343–347. IEEE (2014)
13. Peng, X., Liu, F., Zhang, J., Lan, L., Ye, J., Liu, T., Han, B.: Bilateral dependency optimization: Defending against model-inversion attacks. In: SIGKDD (2022)
14. Struppek, L., Hintersdorf, D., Correira, A.D.A., Adler, A., Kersting, K.: Plug & play attacks: Towards robust and flexible model inversion attacks. In: ICML (2022)

15. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
16. Yuan, X., Chen, K., Zhang, J., Zhang, W., Yu, N., Zhang, Y.: Pseudo label-guided model inversion attack via conditional generative adversarial network. In: AAAI (2023)
17. Zhang, H., Wu, C., Zhang, Z., Zhu, Y., Lin, H., Zhang, Z., Sun, Y., He, T., Mueller, J., Manmatha, R., et al.: Resnest: Split-attention networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2736–2746 (2022)