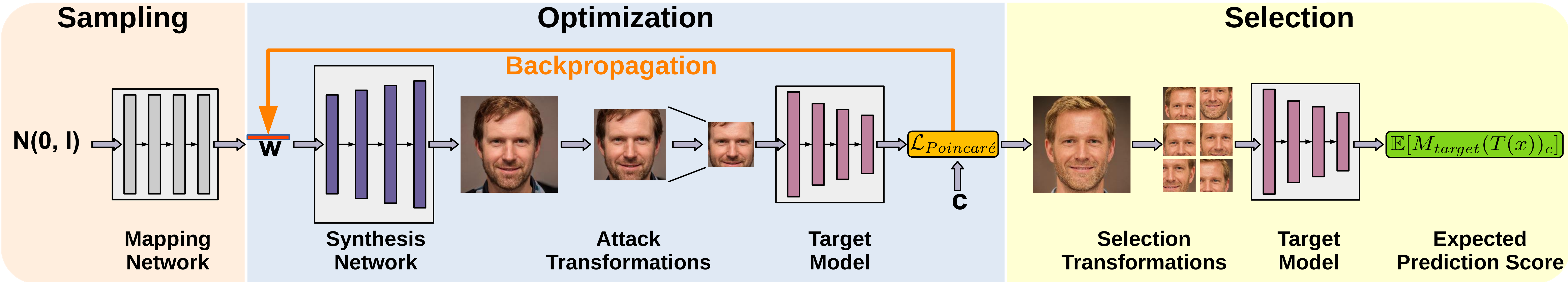
Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks





Lukas Struppek [1], Dominik Hintersdorf [1], Antonio De Almeida Correia [1], Antonia Adler [2], Kristian Kersting [1 3]

[1] Technical University of Darmstadt, Germany [2] Universität der Bundeswehr München, Munich, Germany [3] Hessian Center for AI (hessian.AI), Germany



At a Glance

Model inversion attacks (MIAs) aim to create synthetic images that reflect the class-wise characteristics from a target classifier's private training data by exploiting the model's learned knowledge.

We introduce several novelties to make MIAs robust and flexible:

- **Loosening the connection** between GANs and targets to flexibly exchange both components and even allow the **use of pre-trained GANs**.
- Stabilizing optimization by **random transformations** to facilitate extraction of sensitive features.
- Moving optimization to **hyperbolic spaces** to avoid vanishing gradients and poor local minima.
- Selecting meaningful attack results based on a novel **robustness-based selection process**.

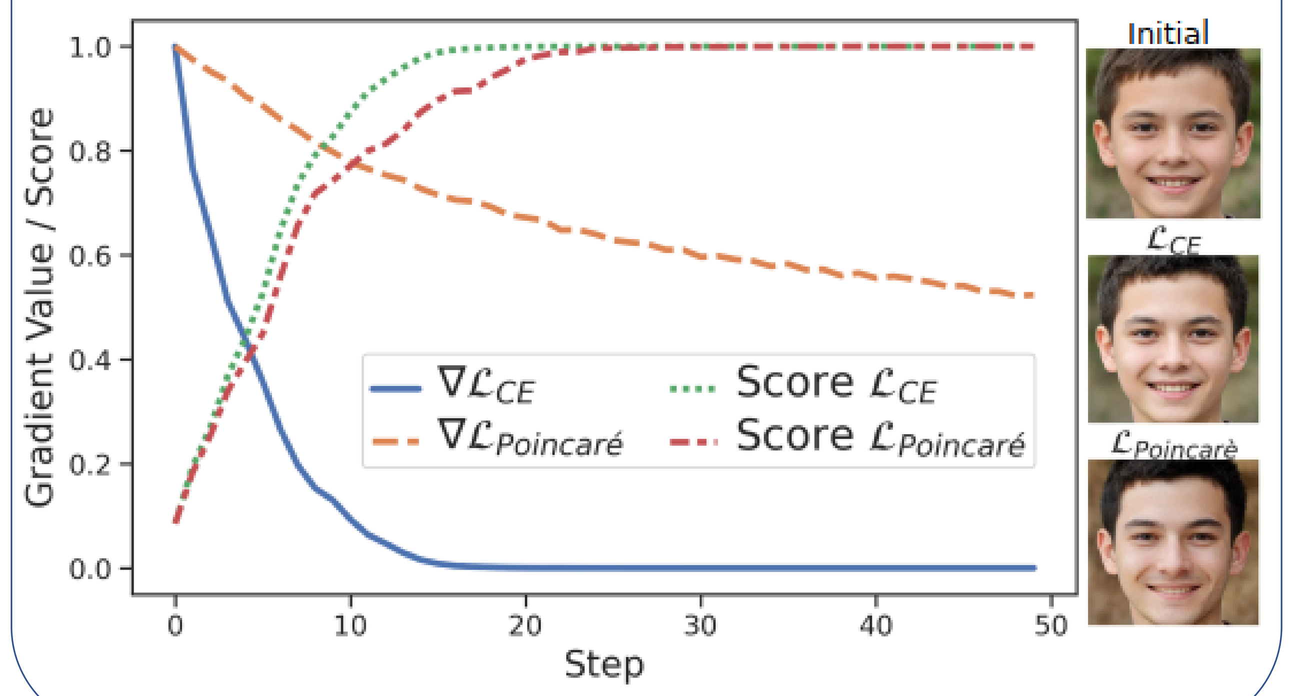
Overcoming Vanishing Gradients

How to avoid vanishing gradients and support characteristic feature extraction?

Solution: We move the optimization to hyperbolic, non-Euclidean spaces and use the Poincaré distance to guide the attack:

$$\mathcal{L} = \operatorname{arccosh}\left(1 + \frac{2\,\|u - v\|_2^2}{(1 - \|u\|_2^2)(1 - \|v\|_2^2)}\right)$$

We set u to be the normalized output logits $u = \frac{o}{\|o\|_1}$ and v as the one-hot encoded target vector, replacing the 1 by 0.9999.



Increasing Flexibility

How to make attacks less time- and resource-consuming and more flexible?

Solution: We developed our approach with the use of pre-trained GANs in mind:

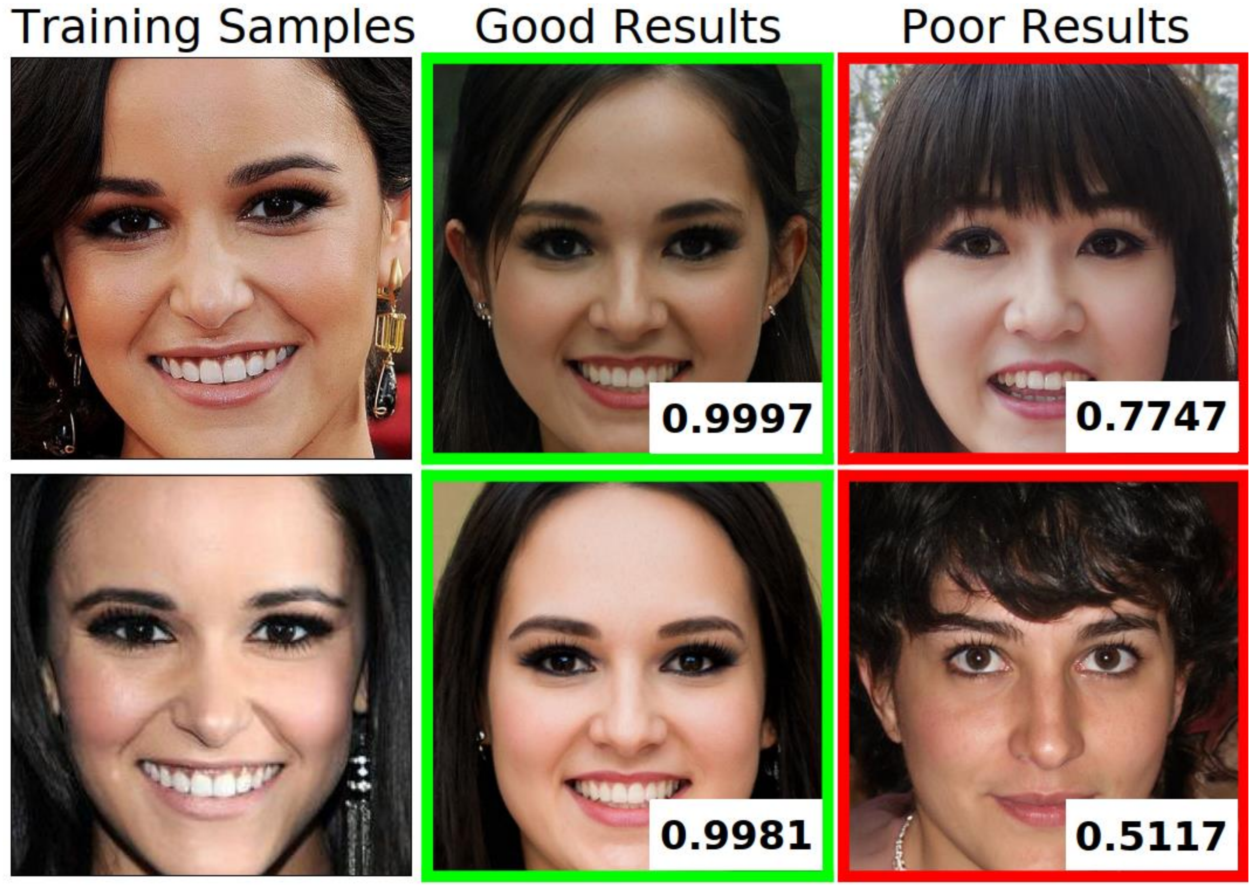
- Generator from the same domain as the target distribution is sufficient to perform the attacks.
- Usage of publicly available models, such as StyleGAN2 or BigGAN, is possible.

Sample Selection

How to select meaningful attack results?

Solution: We select the results with the most robust prediction scores on the target model M_{target} under strong transformations $T(x)$:

$$E[M_{target}(T(x)) \approx \frac{1}{N}\sum_{i=1}^{N} M_{target}(T(x))_c$$



Increasing Robustness

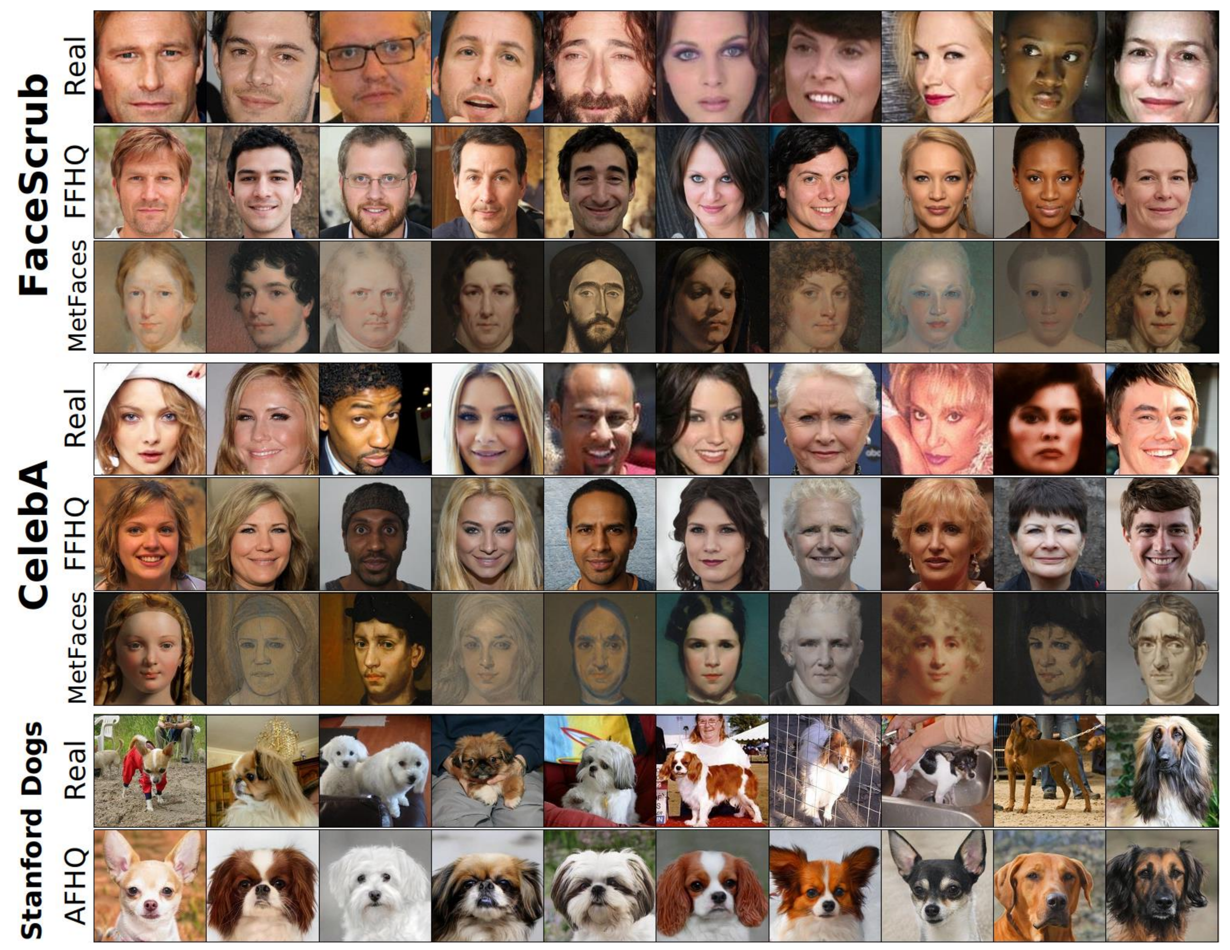
How to avoid the generation of misleading features and overcome distributional shifts?

Solution: We apply (random) image transformations on the GAN outputs in each optimization step to

1. Adjust images to the target distribution.
2. Stabilize the optimization process.
3. Reduce risk of misleading images.
4. Support extraction of characteristic features for targeted classes.

Qualitative Results

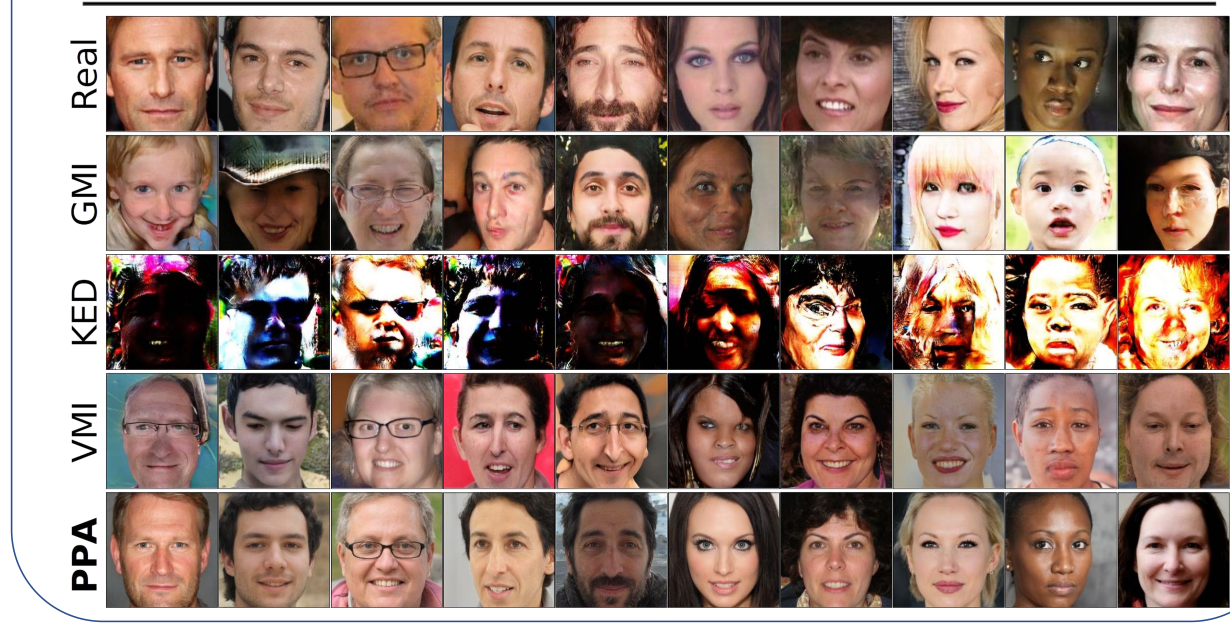
Qualitative results of *Plug & Play Attacks* performed with **publicly available, pre-trained StyleGAN2** models (FFHQ, MetFaces and AFHQ Dogs) against ResNeSt-101 models trained on FaceScrub, CelebA and Stanford Dogs.



Comparison To Existing Attacks

Our *Plug & Play Attacks* significantly outperform previous approaches.

	↑Acc@1	↓δ_{face}	↓FID
GMI (Zhang et al., CVPR 2020)	13.11%	1.2600	77.80
KED (Chen et al., ICCV 2021)	05.72%	1.4366	207.11
VMI (Wang et al., NeurIPS 2021)	61.63%	0.9545	63.27
Plug and Play Attacks (Ours)	**88.46%**	**0.7441**	**41.73**



Conclusion

- Our proposed *Plug and Play Attacks* are a novel **state-of-the-art** model inversion attack.
- Work under strong **distributional shifts** between GAN and target distributions.
- Can make use of **publicly available GANs**, so no additional training or data is required.
- **Reduce risk** of generating misleading or fooling attack results.

Code: https://github.com/LukasStruppek/Plug-and-Play-Attacks

Contact: Lukas Struppek
Technical University of Darmstadt
lukas.struppek@cs.tu-darmstadt.de
@LukasStruppek