Plug & Play Attacks

Towards Robust and Flexible Model Inversion Attacks

Lukas Struppek
TU Darmstadt

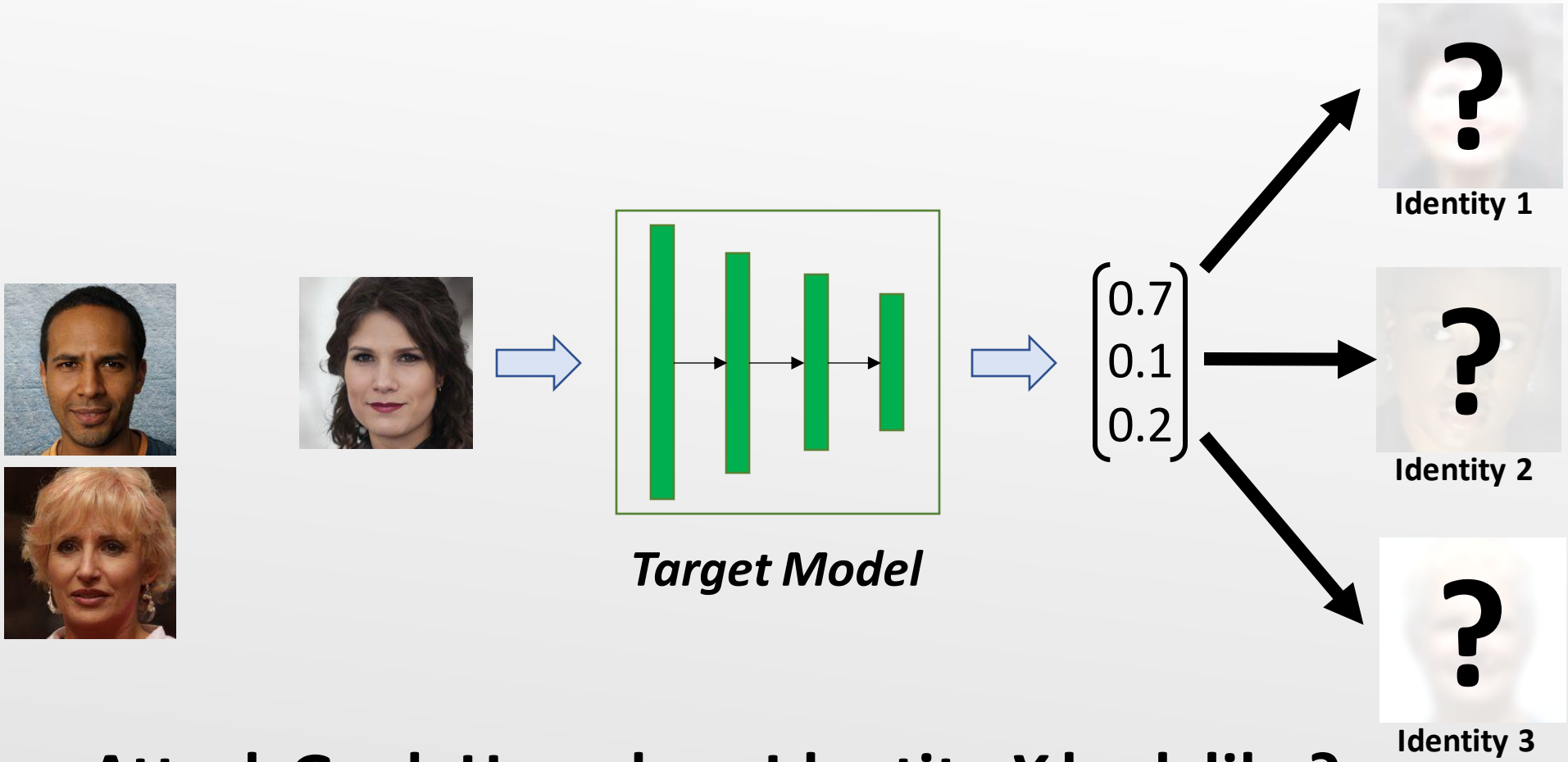
Dominik Hintersdorf
TU Darmstadt

Antonio De Almeida Correia
TU Darmstadt

Antonia Adler
Universität der Bundeswehr

Kristian Kersting
TU Darmstadt
Hessian AI

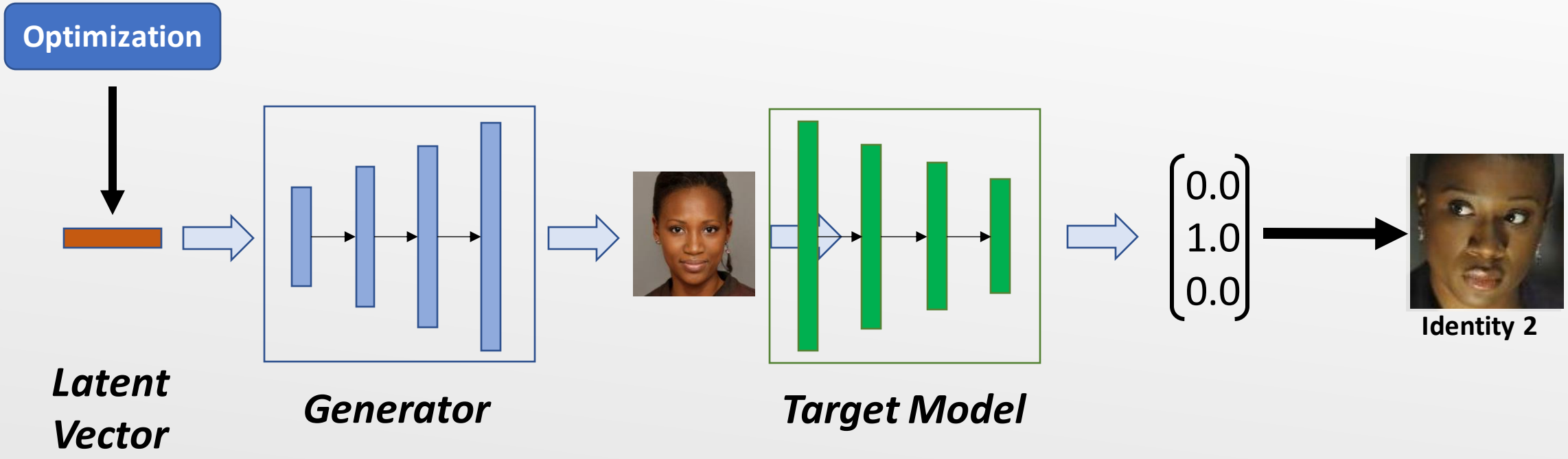
Deep Neural Networks are Changing the World – But there are Privacy Concerns



Attack Goal: How does Identity *X* look like?

(Generative) Model Inversion Attacks

Fredrikson et al. (CCS 2015); Zhang et al. (CVPR 2020)



Attack Goal: How does Identity *X* look like?

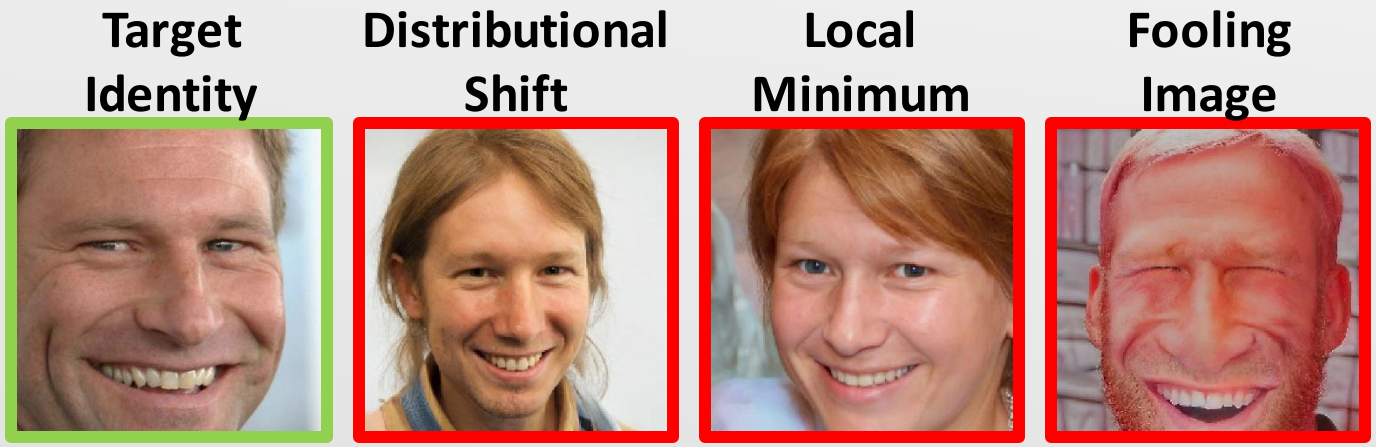
Model Inversion Attacks Face Several Challenges

Degradation Factors

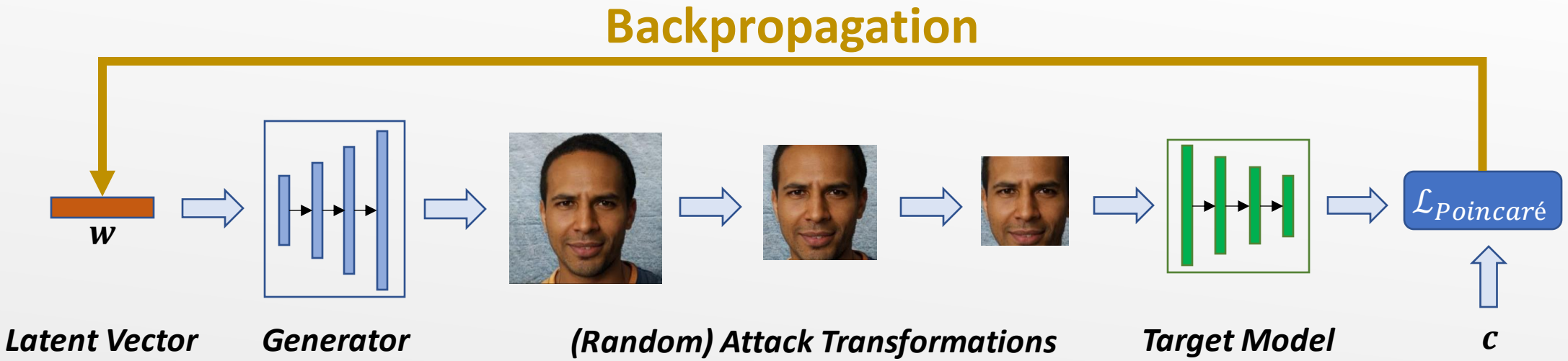
- Distributional Shifts
- Complex Optimization Landscape
- Fooling Images

Limitations of Previous MIAs

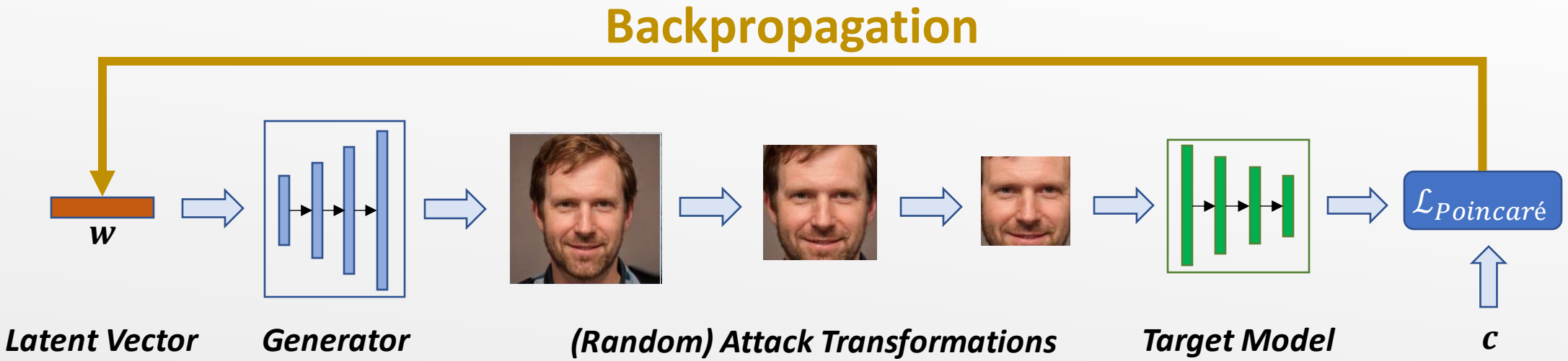
- Tailored on single target
- Time and resource intensive
- Additional input information required



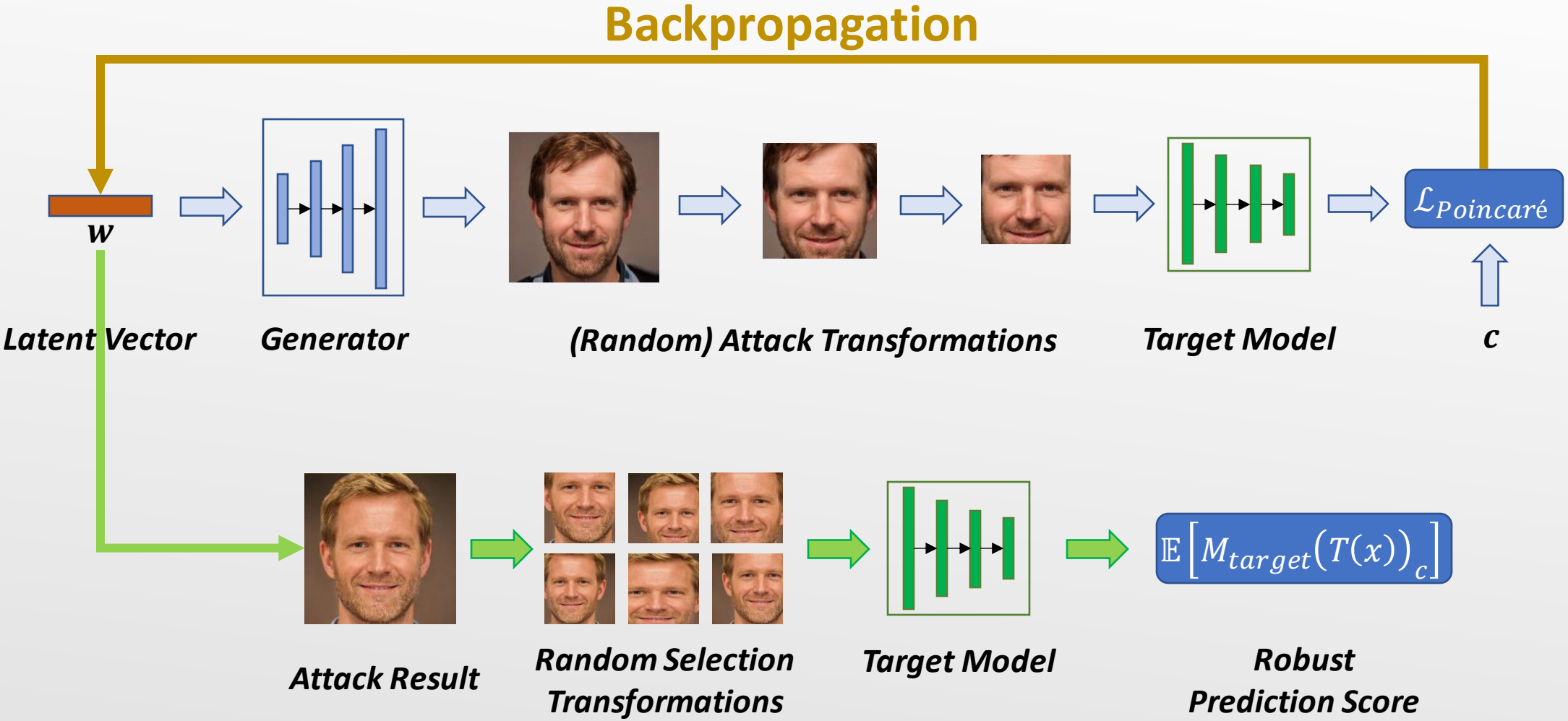
Robust & Flexible Model Inversion Attacks



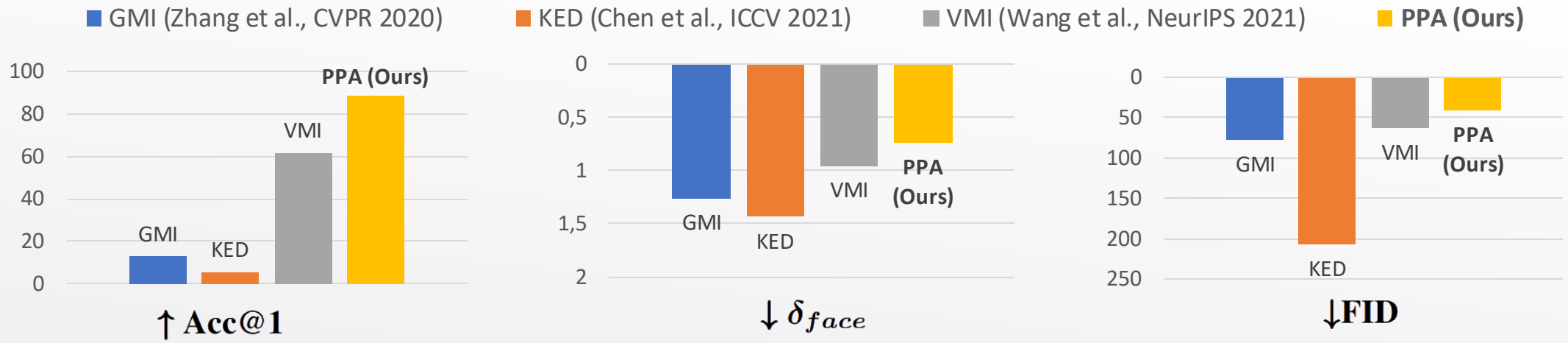
Robust & Flexible Model Inversion Attacks

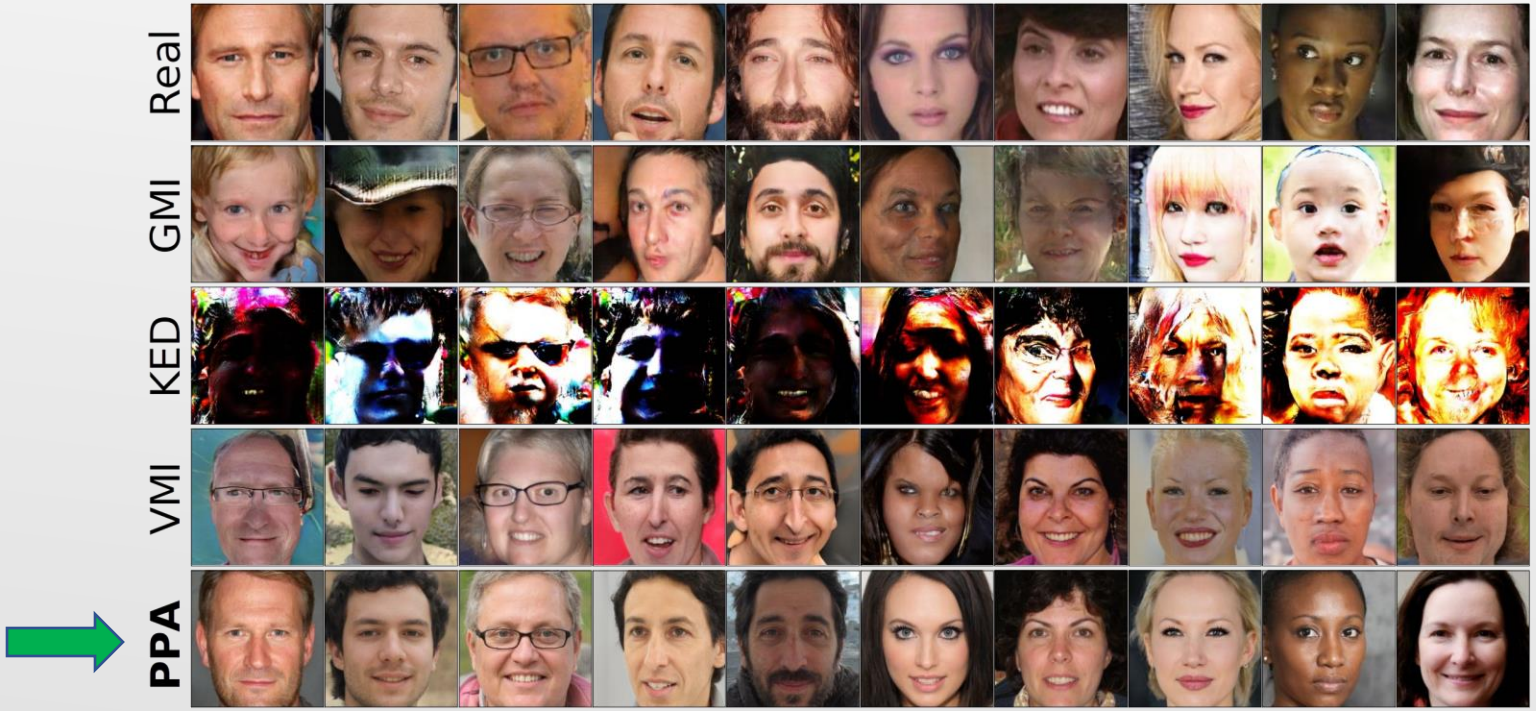


Robust & Flexible Model Inversion Attacks

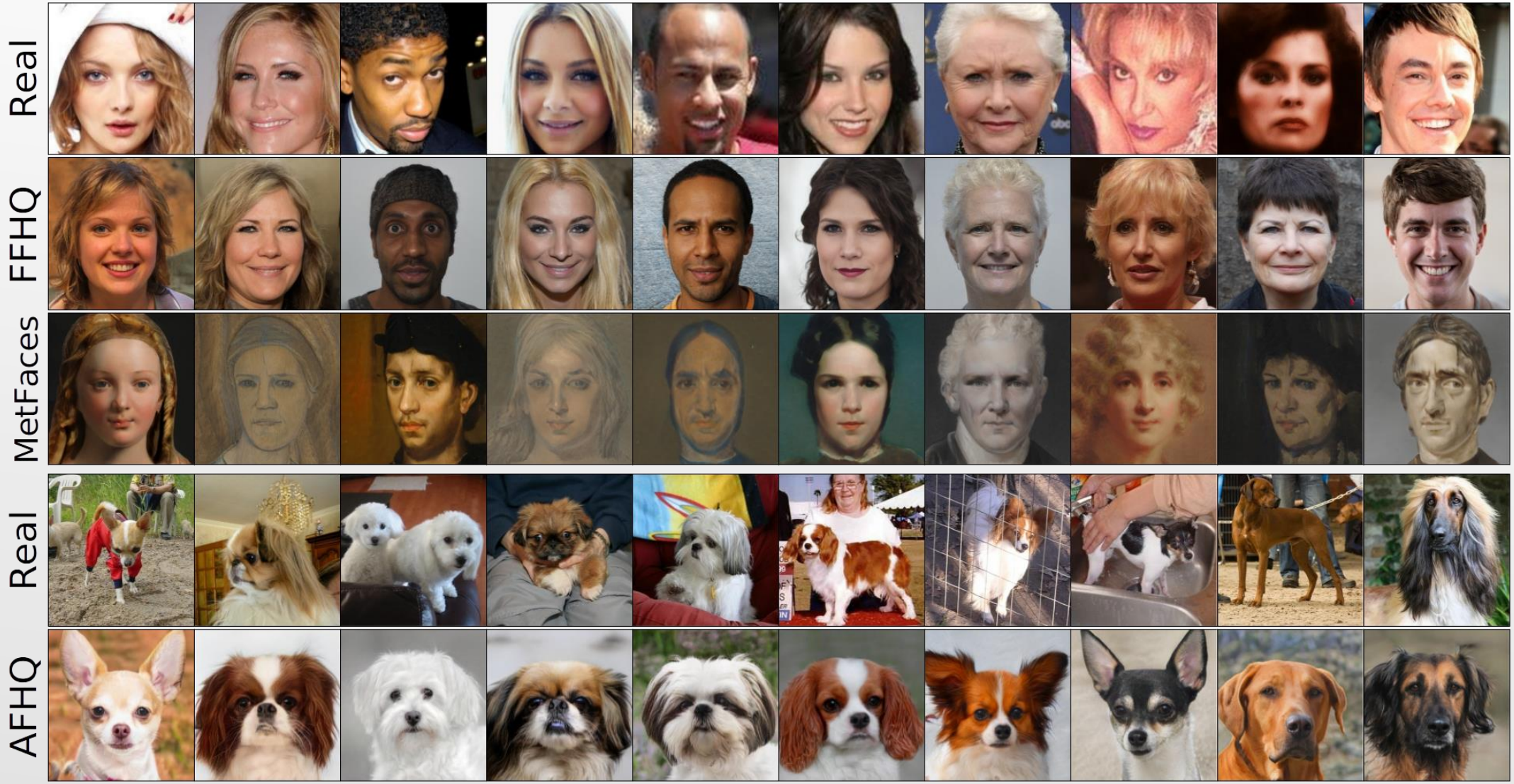


Plug & Play Attacks Outperform Previous Attacks





Plug & Play Attacks Overcome Distributional Shifts



The First Plug & Play Model Inversion Attack

Robust attack optimization

- Random transformations stabilize attacks
- Poincaré loss avoids vanishing gradients
- Selection process filters out poor attack outcomes

Flexible exchange of models

- Usage of pre-trained GANs possible
- Attack is independent of a specific target model

Contact Information:
Lukas Struppek
TU Darmstadt
lukas.struppek@cs.tu-darmstadt.de
@LukasStruppek

Authors:
Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, Kristian Kersting

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