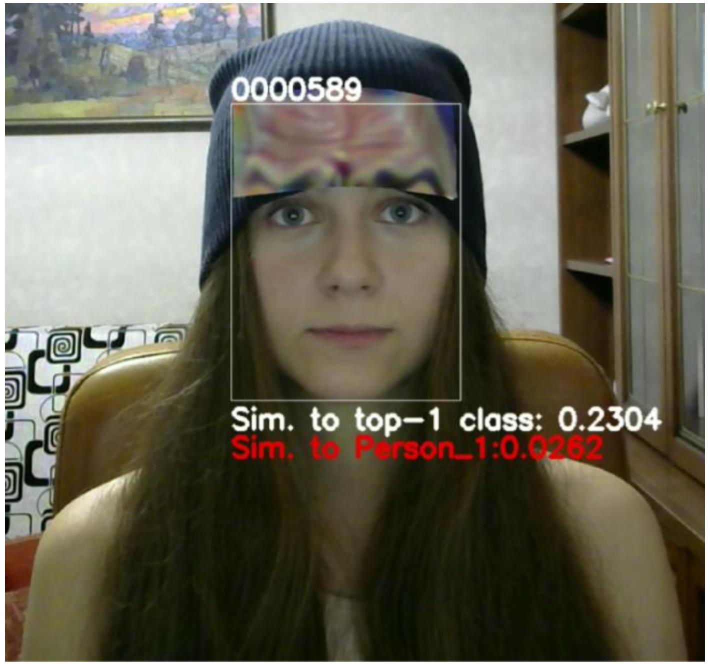
AdvHat: Real-World Adversarial Attack on ArcFace Face ID System

Stepan Komkov and Aleksandr Petiushko

LOMONOSOV MOSCOW STATE UNIVERSITY
FACULTY OF MECHANICS AND MATHEMATICS
HUAWEI

The main idea of the paper

We show that a carefully designed rectangular sticker placed on a hat may fool state of the art solutions in the FaceID domain



Prior Art

Mahmood Sharif *et al.* are the first who created real-world adversarial accessorizes that may fool the FaceID system in real life

Examples of accessories that make a person unrecognizable for the FaceID system from [1]

[1] Sharif M. et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition //Proceedings of the 2016 acm sigsac conference on computer and communications security. – 2016. – С. 1528-1540.

Prior Art

However, there are some drawbacks in [1] in the context of the current level of technique:

- The attacks were prepared for the previous generation of FaceID models
- Models for closed-set recognition were considered only
- Shooting conditions varied weakly (*e.g.* angles of head and lightning condition)
- You have to prepare a complex shape object to reproduce the attack

[1] Sharif M. et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition //Proceedings of the 2016 acm sigsac conference on computer and communications security. – 2016. – C. 1528-1540.

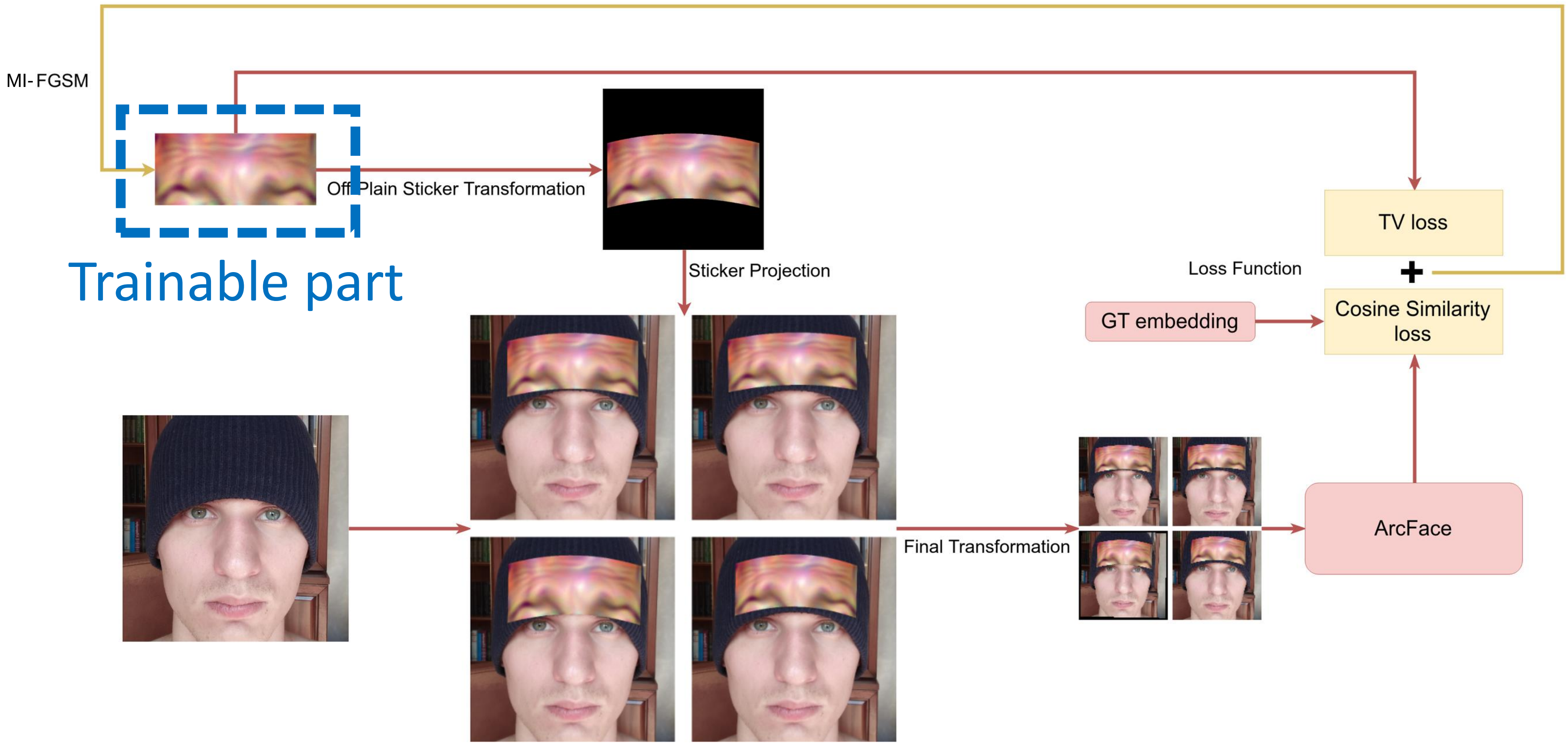
Our Work

- We focus on ResNet 100E-IR, ArcFace@ms1m-refine-v2 model [2]
  - One of the strongest models for face recognition and at the same time publicly available
- We use an open-set scenario, *i.e.* we concentrate on similarities between feature vectors instead of class probabilities
  - A common technique used for Face Recognition
- We estimate our attack in various shooting conditions
- Our method is easy to reproduce
  - See https://github.com/papermsucode/advhat for the instruction for reproduction

[2] Deng J. et al. Arcface: Additive angular margin loss for deep face recognition //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. – 2019. – C. 4690-4699.

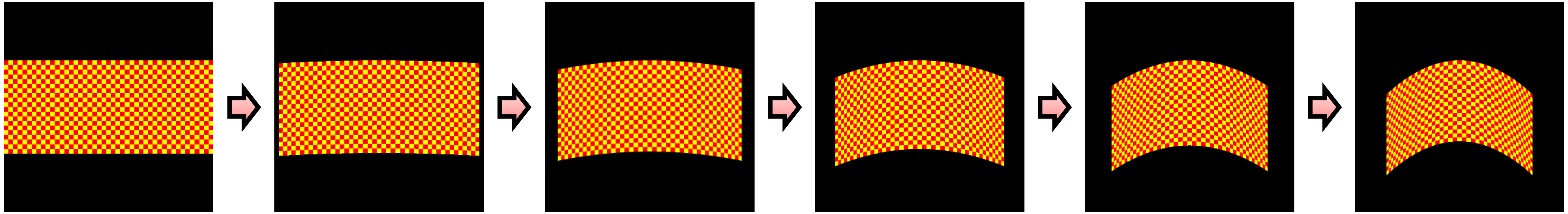
Our Method

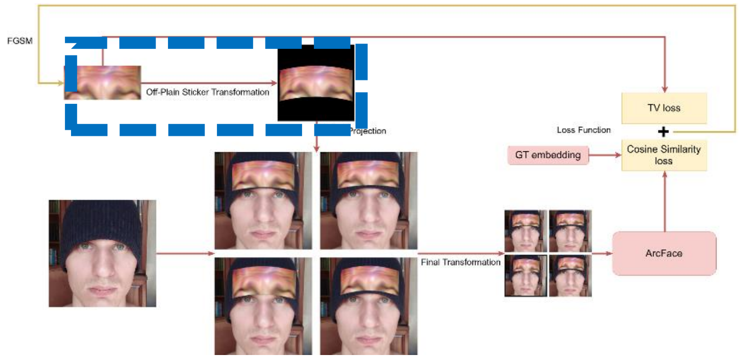
The overall pipeline for sticker preparation is as follows:



Off-Plain Sticker Transformation

- When we put a rectangular sticker on a hat, it bends and rotates:



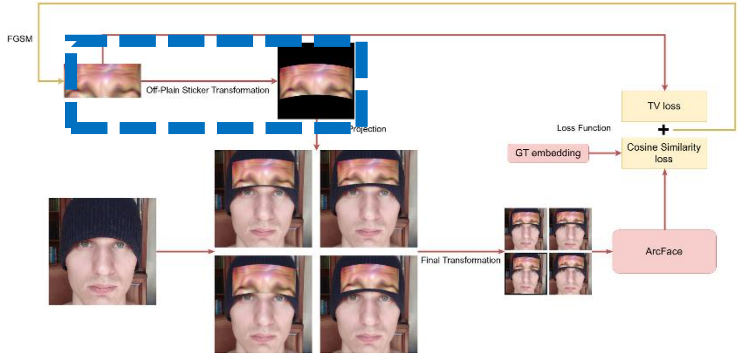


Off-Plain Sticker Transformation

- We simulate bending as a parabolic transformation of the sticker
- The initial coordinates of the flat sticker are changed as follows during bending and rotation:

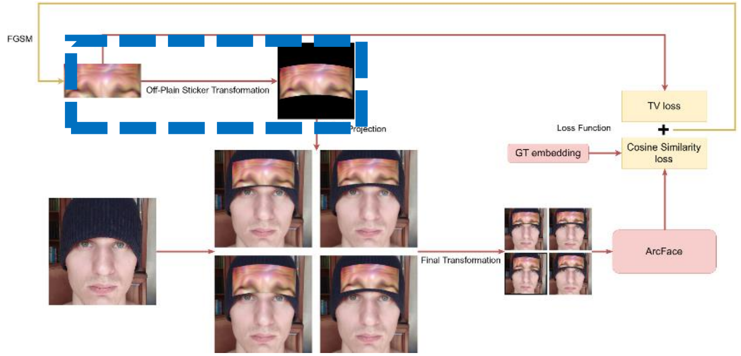
$$\begin{pmatrix} x_0 \\ y_0 \\ 0 \end{pmatrix} \mapsto \begin{pmatrix} x_1 \\ y_1 \\ z_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -sin\theta \\ 0 & sin\theta & \cos\theta \end{pmatrix} \cdot \begin{pmatrix} b \cdot \left(|x_0| \cdot \sqrt{x_0^2 + \frac{1}{4 \cdot b^2}} + \frac{1}{4 \cdot b^2} \cdot \ln\left(|x_0| + \sqrt{x_0^2 + \frac{1}{4 \cdot b^2}} \right) - \frac{1}{4 \cdot b^2} \cdot \ln\frac{1}{2 \cdot b} \right) \\ y_0 \\ b \cdot x_1^2 \end{pmatrix}$$

- Then we render a new sticker image by projecting along the z-axis which is perpendicular to the initial plane of the sticker



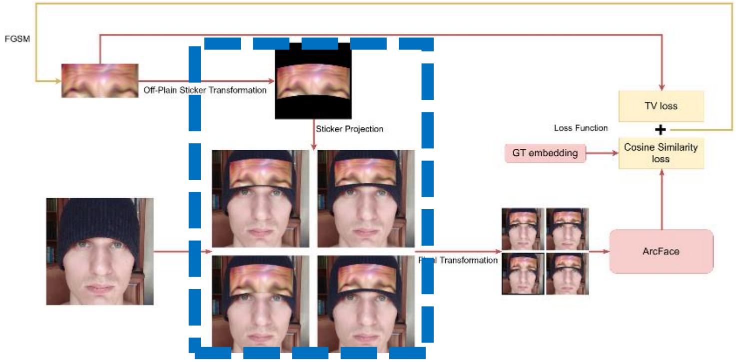
Off-Plain Sticker Transformation

- We implement the proposed transformation using conventional tensor operations in a fully differentiable way
- We change the parabola rate and the angle of rotation a little during the attack preparation to make the attack more robust



Sticker Projection

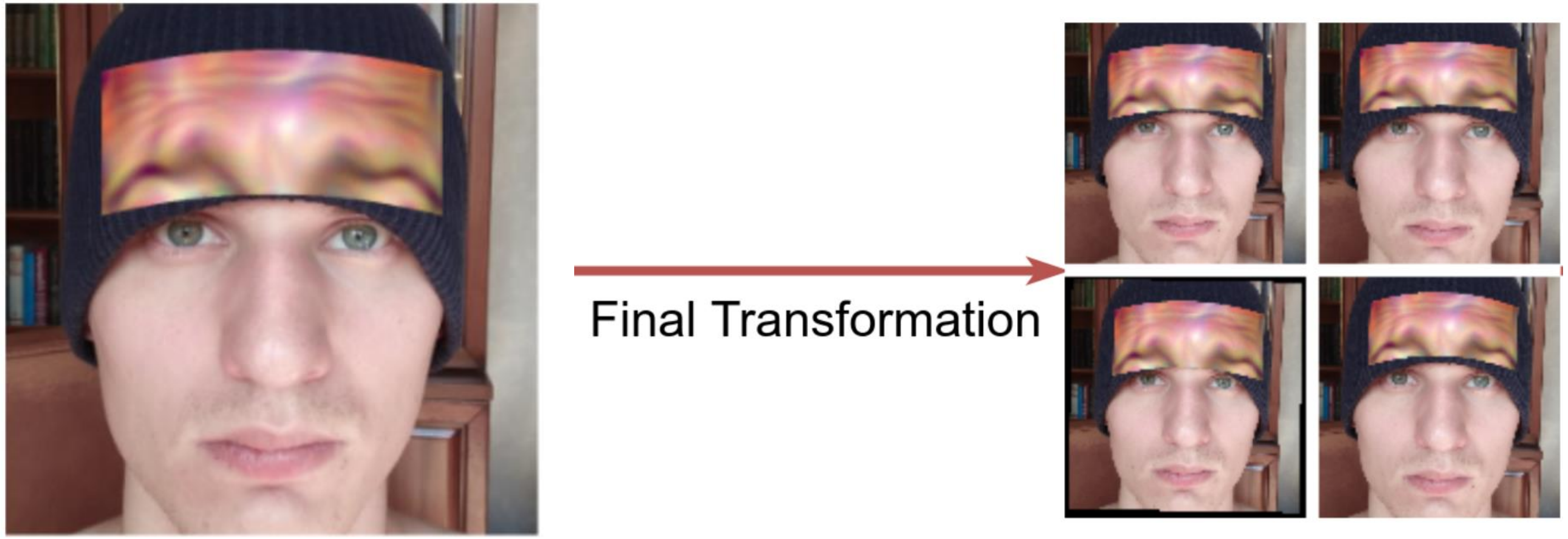
- We use Spatial Transformer Layer (STL) [3] to project the obtained sticker on the image of face
- We also slightly change the parameters of the projection during the attack preparation:

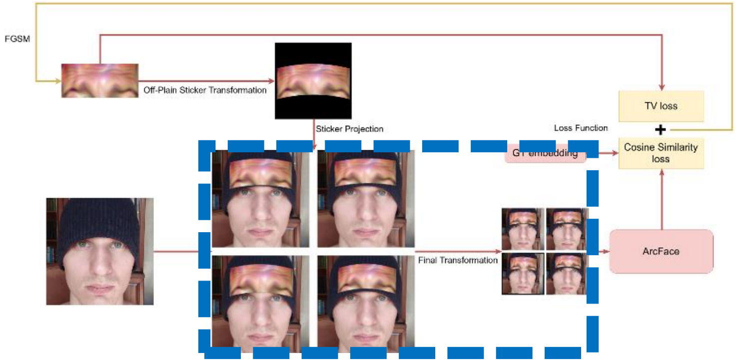


[3] Jaderberg M. et al. Spatial transformer networks //Advances in neural information processing systems. – 2015. – C. 2017-2025.

Final Transformation

- We use STL [3] again to resize the obtained image to the size required by the FaceID model
- We slightly jitter the whole image during this part





[3] Jaderberg M. et al. Spatial transformer networks //Advances in neural information processing systems. – 2015. – С. 2017-2025.

Loss Function

- A batch of images with various projection parameters is fed to the ArcFace FaceID model to calculate their feature vectors
- We minimize cosine similarity between evaluated feature vectors and the ground truth one using the first loss function:

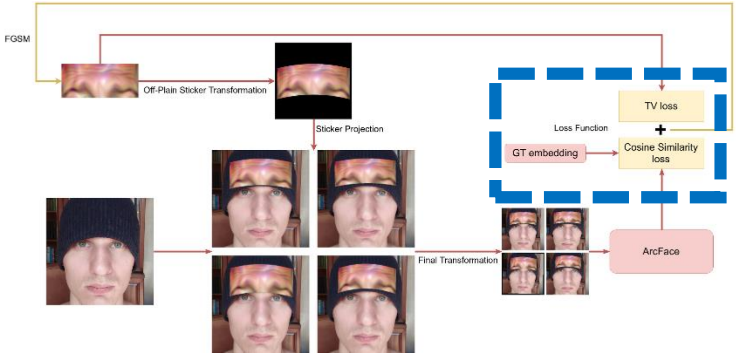
$$\mathcal{L}_{sim}(x, a) = \cos(e_x, e_a)$$

- Total Variation loss makes the sticker image smoother:

$$\mathrm{TV}(x) = \sum_{i,j} \sqrt{\left((x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2 \right)}$$

- The final loss is the sum of the aforementioned losses:

$$\mathcal{L}_{final}(x, a) = \mathcal{L}_{sim}(x, a) + \lambda \cdot TV(x).$$



Some Context

- FaceID is an open-set task *i.e.* the known set of classes during inference can differ from the training ones
- The common procedure of recognition of face:
  1. Calculation of feature vector of the face
  2. Similarity measurement to the stored feature vectors
  3. If similarity to the top-1 class ***A*** is bigger than a threshold then the face is labeled as a face of person ***A***
     - Thresholds are chosen based on the acceptable rate of false recognitions
     - For instance, a threshold for the ArcFace system varies from 0.328 to 0.823 according to the IJB-B benchmark [4]

[4] Whitelam C. et al. Iarpa janus benchmark-b face dataset //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. – 2017. – C. 90-98.

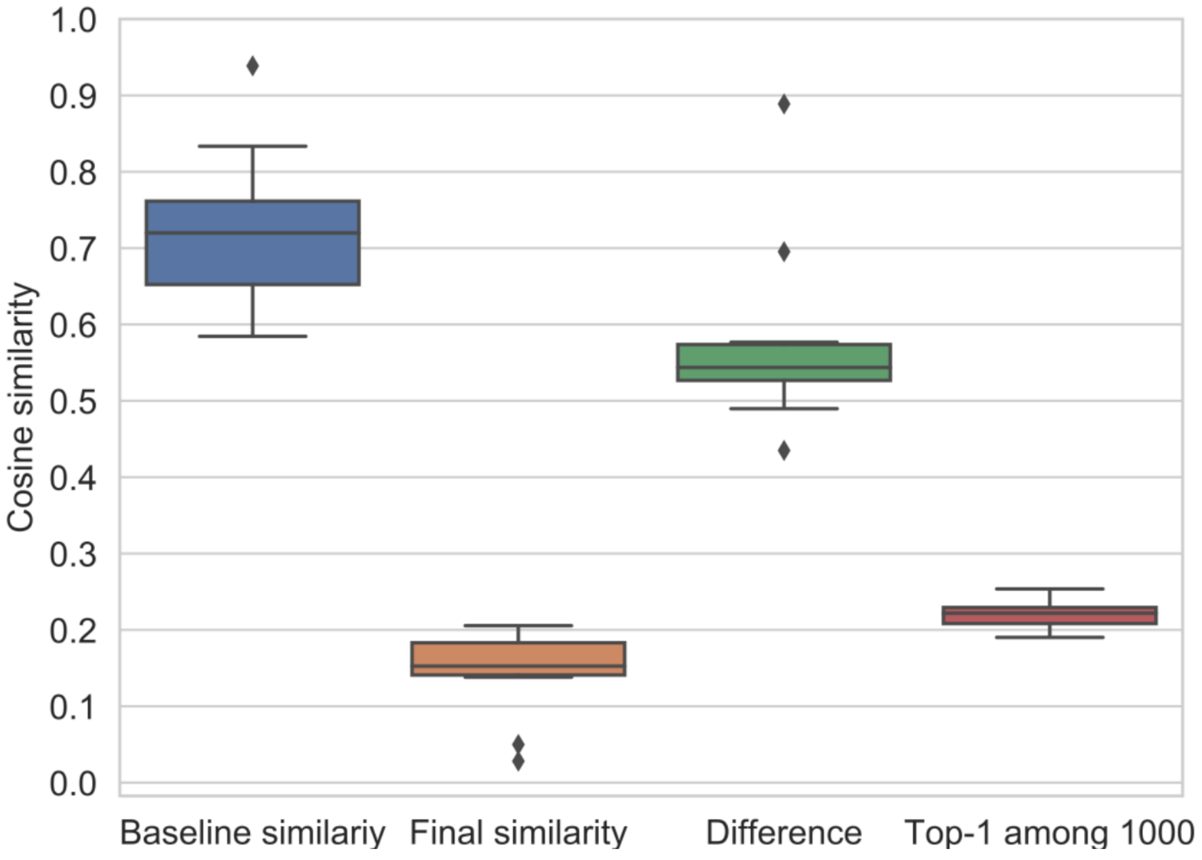
Testing Scenario

- The fooling rate depends on the chosen threshold crucially. That is why we show distributions of similarities to the ground-truth feature vector instead of some specific fooling rates
- We use the first 1000 classes from the CASIA [5] face dataset as known classes
  - We calculate one average feature vector for each of these 1000 classes
- Note that it is harder to fool the recognition system and make the top-1 class incorrect when the number of classes is small
- We evaluate our approach for 10 people of different age and gender: four females of age 30, 23, 16, 5 and six males of age 36, 32, 29, 24, 24, 8

[5] Yi D. et al. Learning face representation from scratch //arXiv preprint arXiv:1411.7923. – 2014.

Experiments with fixed conditions

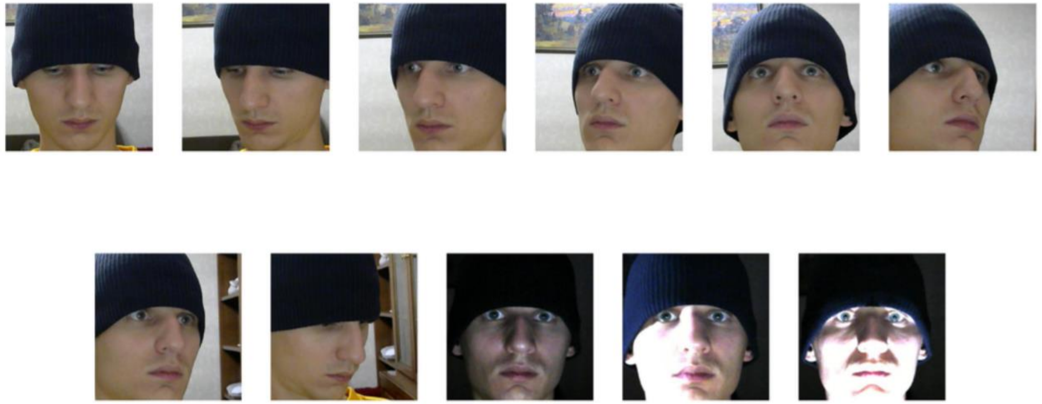
- Blue: Cosine similarities between anchor images and images with a hat
- Orange: Cosine similarities between anchor images and images with an adversarial sticker
- Green: Differences between aforementioned similarities
- Red: The top-1 similarity to the first 1000 classes from CASIA

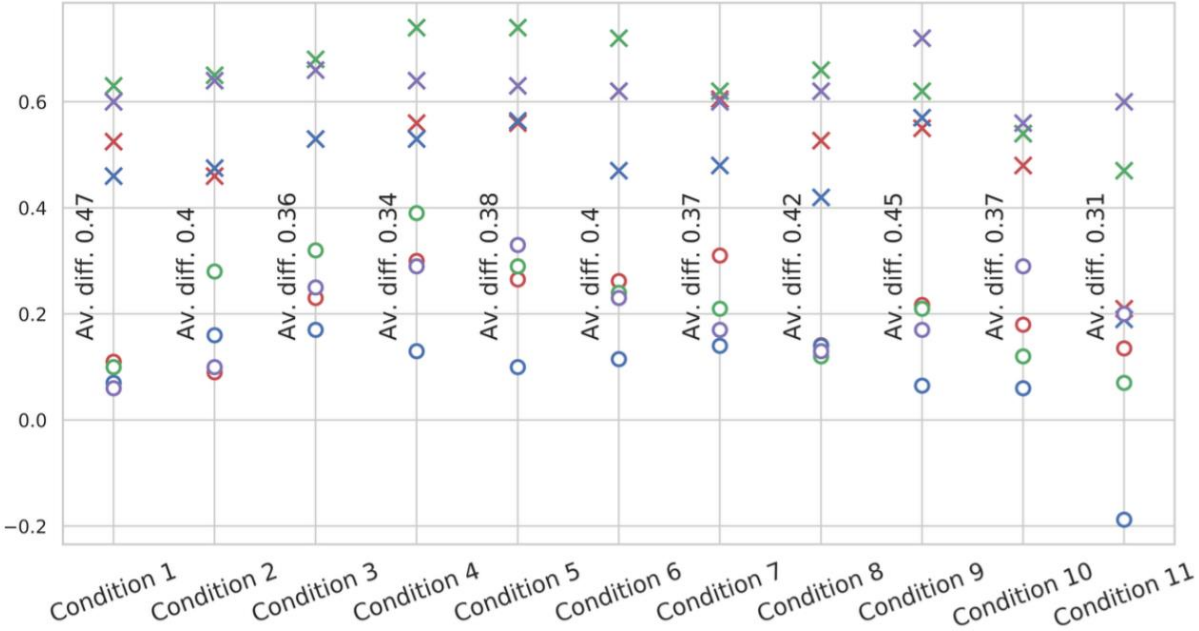


Experiments with various conditions

LOMONOSOV MOSCOW STATE UNIVERSITY
FACULTY OF MECHANICS AND MATHEMATICS
HUAWEI

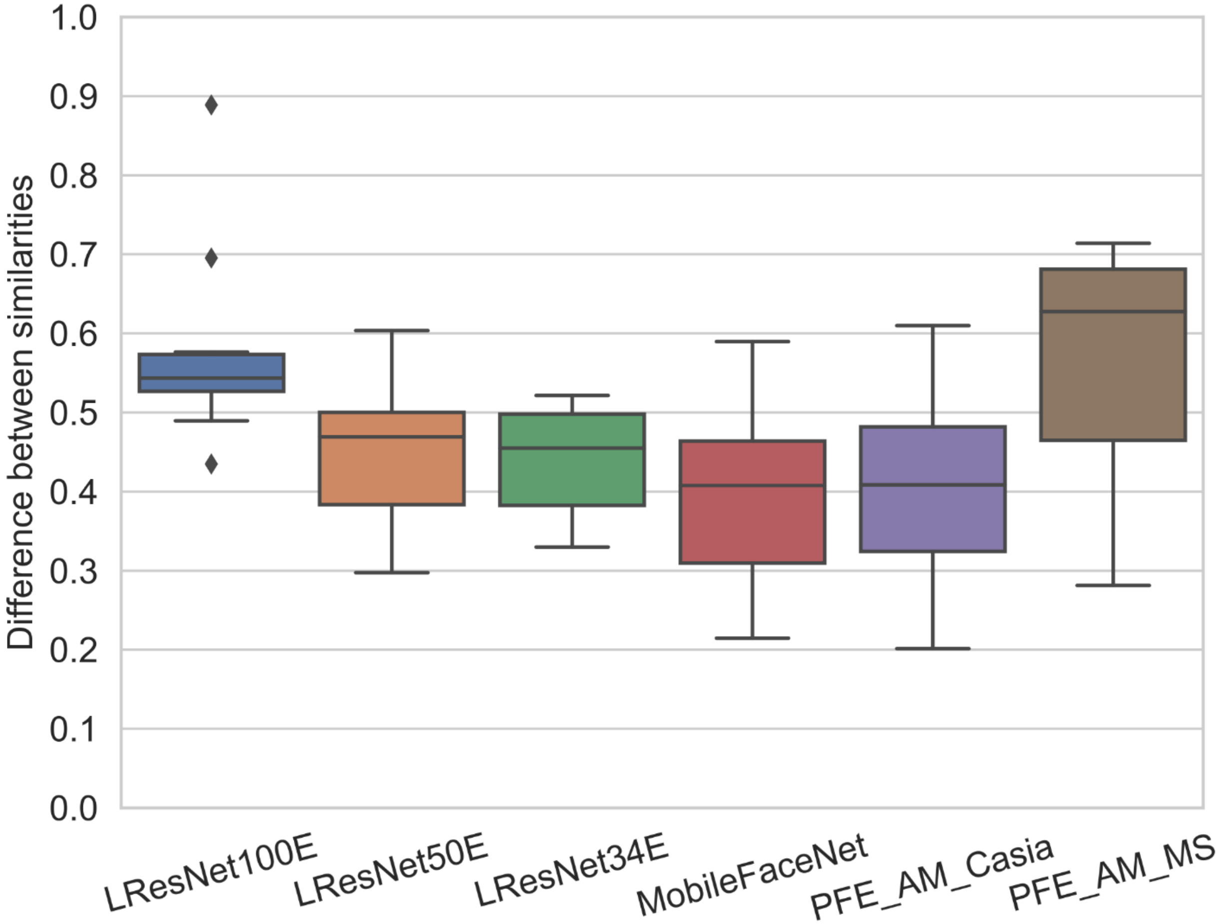
- We also perform experiments with 11 non-trivial conditions to examine the robustness of our approach
  - The conditions are depicted on the right
- Similarities to the ground-truth with (**o**) and without (**x**) adversarial sticker for these conditions



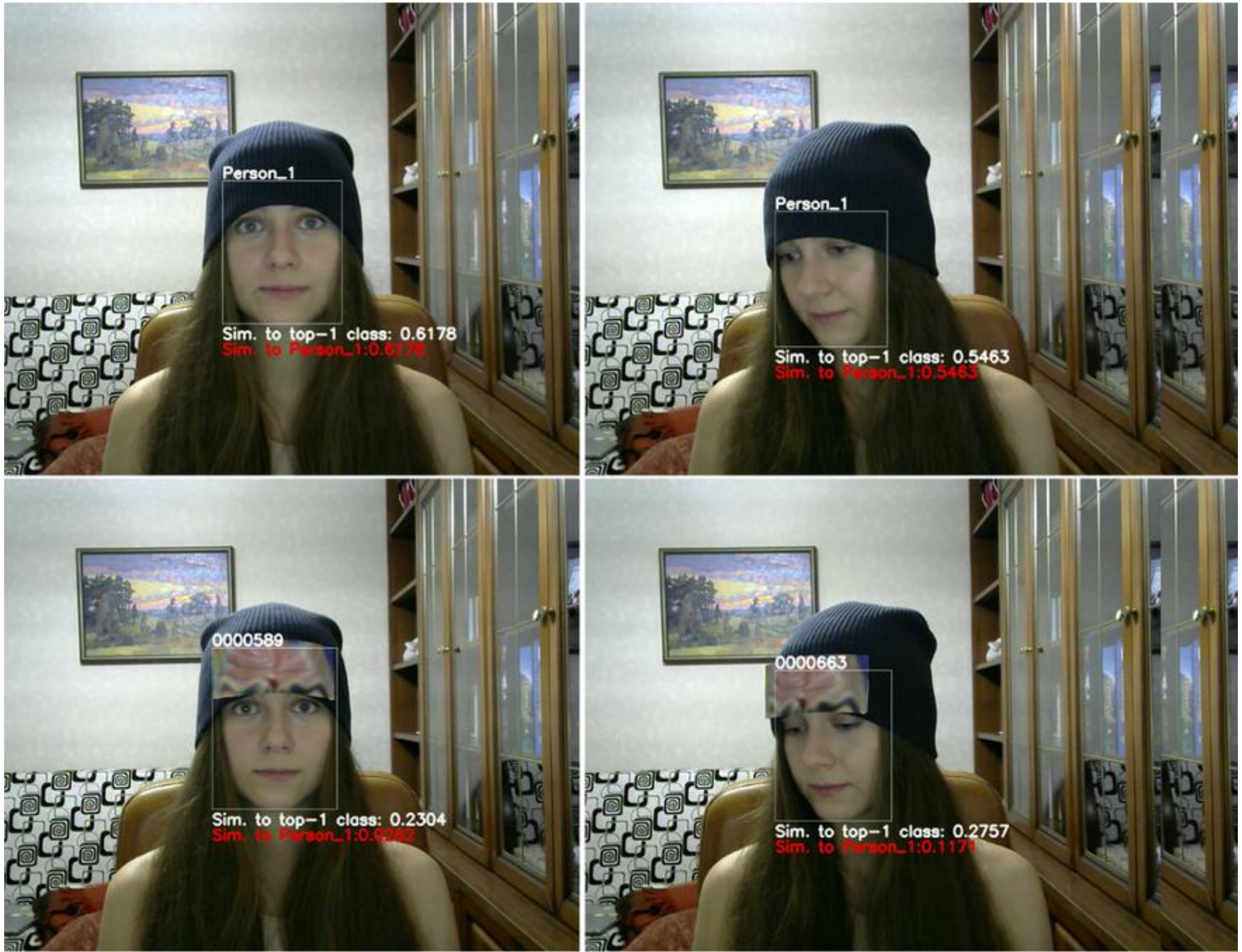


Experiments with transferability

- We examine the transferability of the prepared attacks by using other FaceID models for embedding calculation
  - All the attacks are prepared using LResNet100E
- Distributions of the differences between similarities before and after attacks are on the right



Demonstration



The full video demonstration is available on https://www.youtube.com/watch?v=a4iNg0wWBsQ

LOMONOSOV MOSCOW STATE UNIVERSITY

FACULTY OF MECHANICS AND MATHEMATICS

HUAWEI

Thank you!