ECCV'20 ONLINE

23-28 AUGUST 2020

16TH EUROPEAN CONFERENCE ON

COMPUTER VISION

WWW.ECCV2020.EU

Black-Box Face Recovery from Identity Features

Anton Razzhigaev[1,2], Klim Kireev[1,2], Edgar Kaziahmedov[1,2], Nurislam Tursynbek[1,2], Aleksandr Petiushko[1,3]

Huawei[1], Skoltech[2], MSU[3]

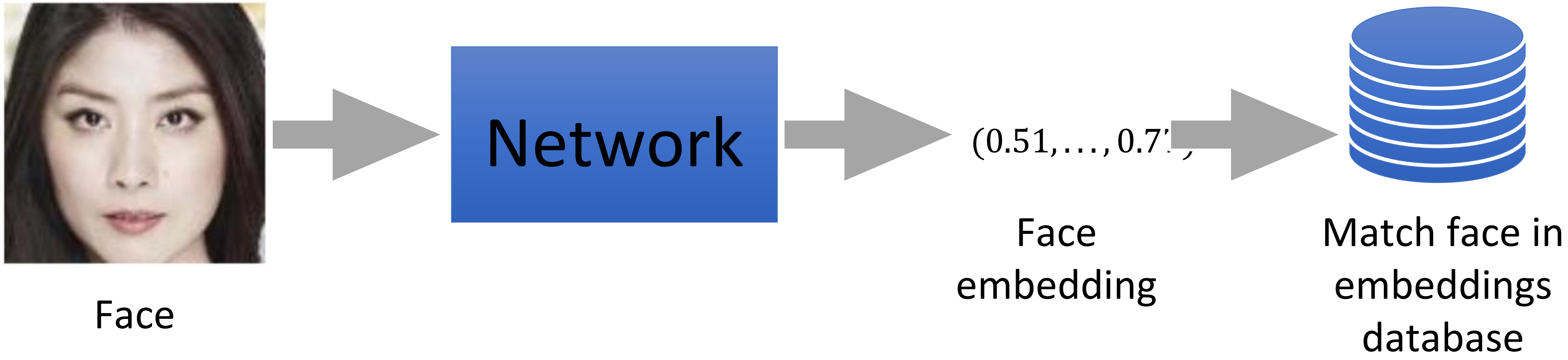
HUAWEI

Skoltech
Skolkovo Institute of Science and Technology

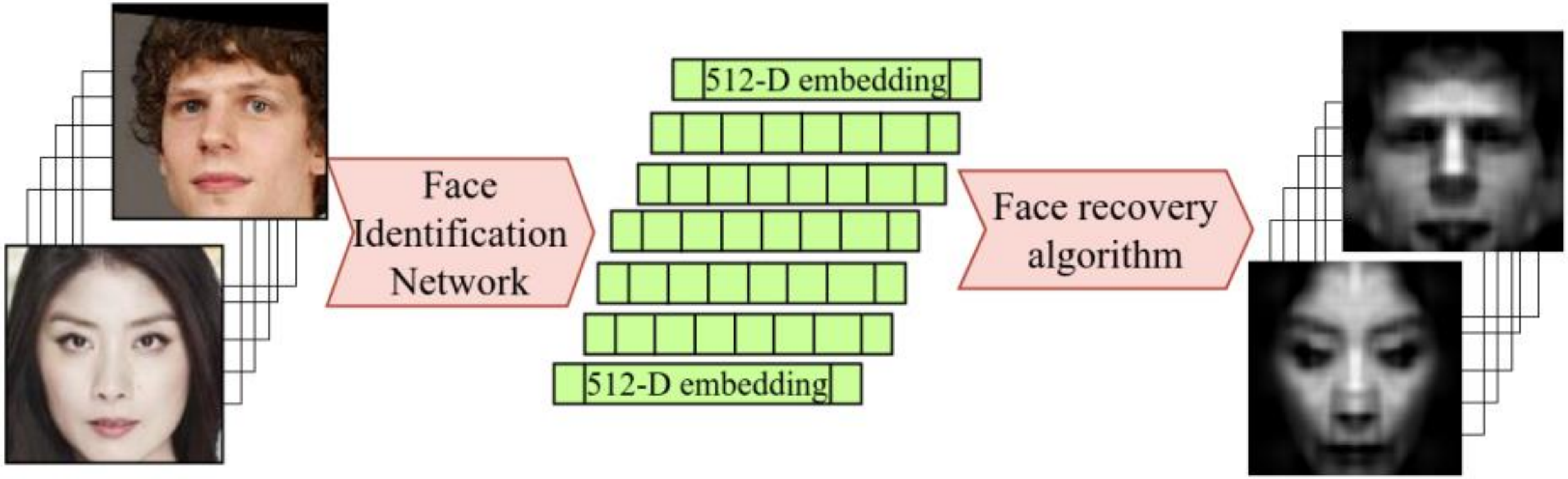
Lomonosov Moscow State University

*Work done at the **Huawei Moscow Research Center, Intelligent Systems & Data Science lab**

Face recognition pipeline



Idea of face recovery



Novelty of our approach

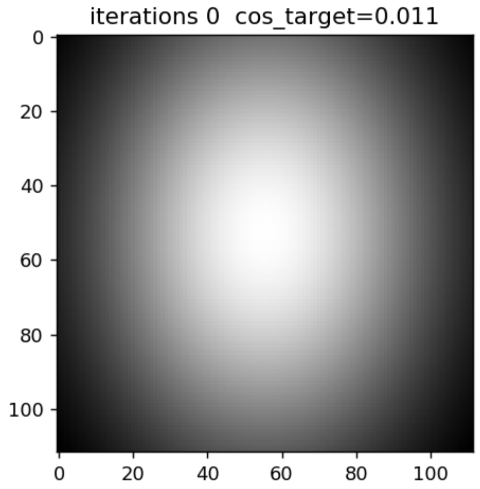
Table 1: Comparison table

Algorithm	Target model	Setting	Dataset-free
Ours	Arcface output	Black-Box	+
NBNet[19]	FaceNet output	Black-Box	-
Cole et. al. [3]	FaceNet intermediate features	White-box	-
CNN[34]	FaceNet output features	White-box	-
Gradient wrt input [18]	Any classifier output	White-box	+

- The **key point** of the proposed algorithm is that it **reconstructs not just faces but** recognizable **identities** without any prior knowledge about how faces look like
- It is not just face generator, it is a way to reveal appearance of a person from the face embedding

Algorithm

Original



Algorithm 1 Face recovery algorithm

INPUT: target face embedding y, black-box model M, loss function L, $N_{queries}$

```
X ← 0
Initialize G_0
for i ← 0 to N_queries do:
    Allocate image batch X
    Sample batch G of random gaussians
    X_j = X + G_0 + G_j
    y' = M(X)
    ind = argmin( L(y'_i, y) )
    X ← X + G_ind
    G_0 ← 0.99 · G_0
    i ← i + batchsize
end for
X ← X + G_0
```

OUTPUT: reconstructed face X

$$L(y, y') = \lambda \cdot (\|y\| - \|y'\|)^2 - s(y, y'),$$

where,

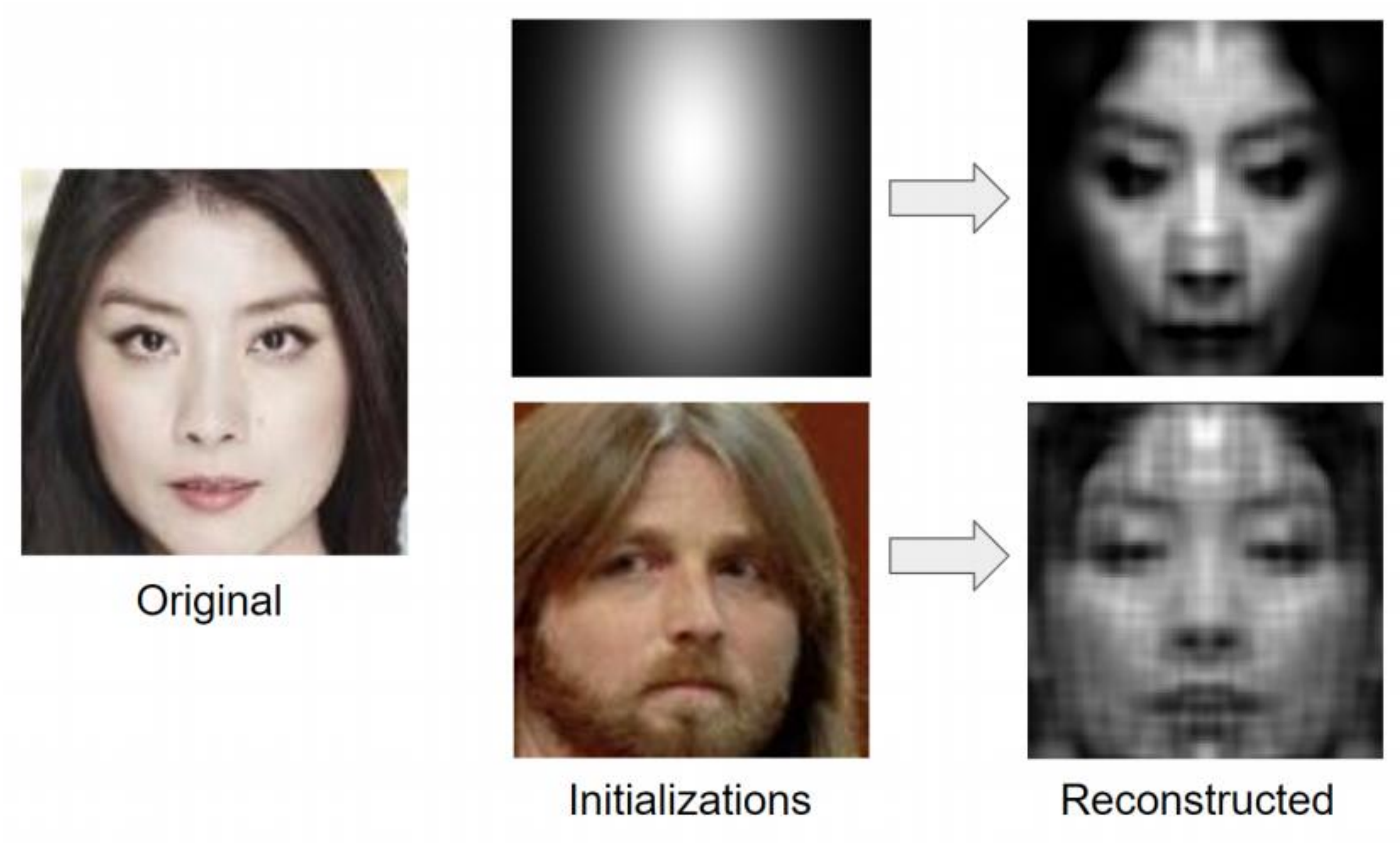
s - cosine similarity function,

$\|y\|$ - L_2 norm of the target embedding,

$\|y'\|$ - L_2 norm of the embedding of a reconstructed image,

$\lambda = 0.0025$, empirically found hyperparameter

Different initialization techniques



Initialization with an **optimal Gaussian blob** leads to the **proper proportions** of the reconstructed face, **but requires a norm of the face embedding** in the loss

Initialization with a **random face** leads to a **collapse to a shape of the initial face**, but does **not require norm of the embedding** in the loss

Function space for “drawing”

Original | Initialization | 30k | 60k | 300k

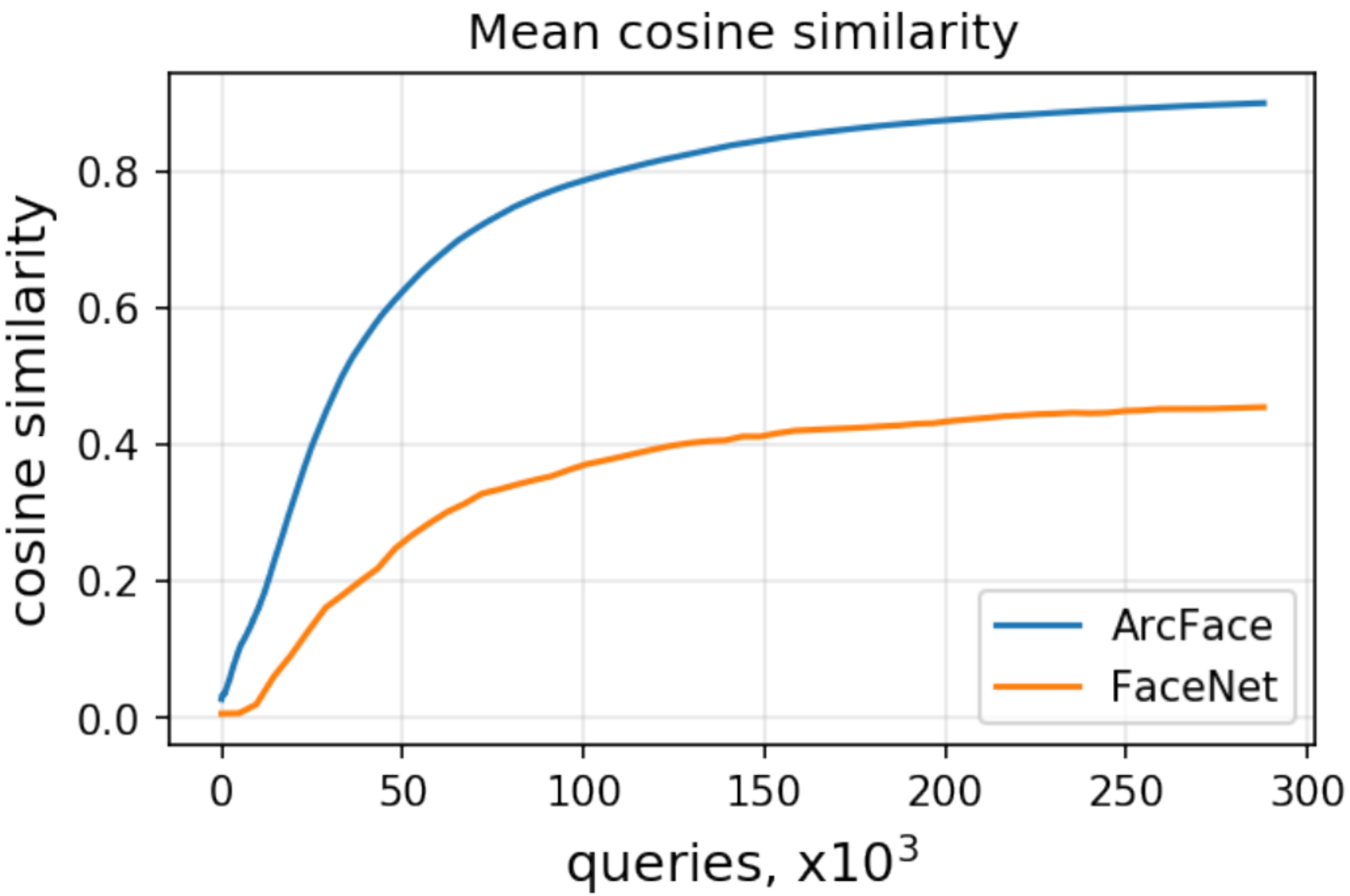
$$G(x,y) = A \cdot \exp \frac{(x-x_0)^2}{2\sigma_1^2} \exp \frac{(y-y_0)^2}{2\sigma_2^2}$$

x, y - pixel coordinates in the image,
x_0, y_0 - coordinates of a center of gaussian,
σ_1, σ_2 - vertical and horizontal standard deviations,
A - amplitude

Gaussian 2D functions:

1. Semi-local
2. Frequency control through σ

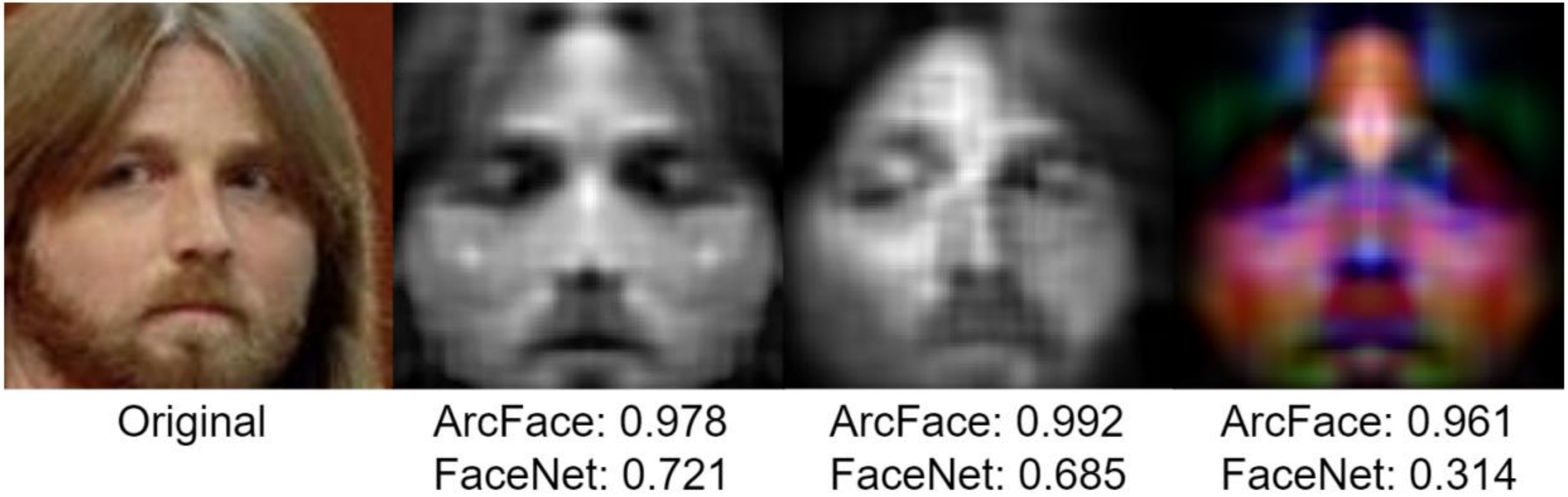
Evaluation



- Attacking: **ArcFace** network
- Evaluating: using an "**independent critic**" – **FaceNet** network

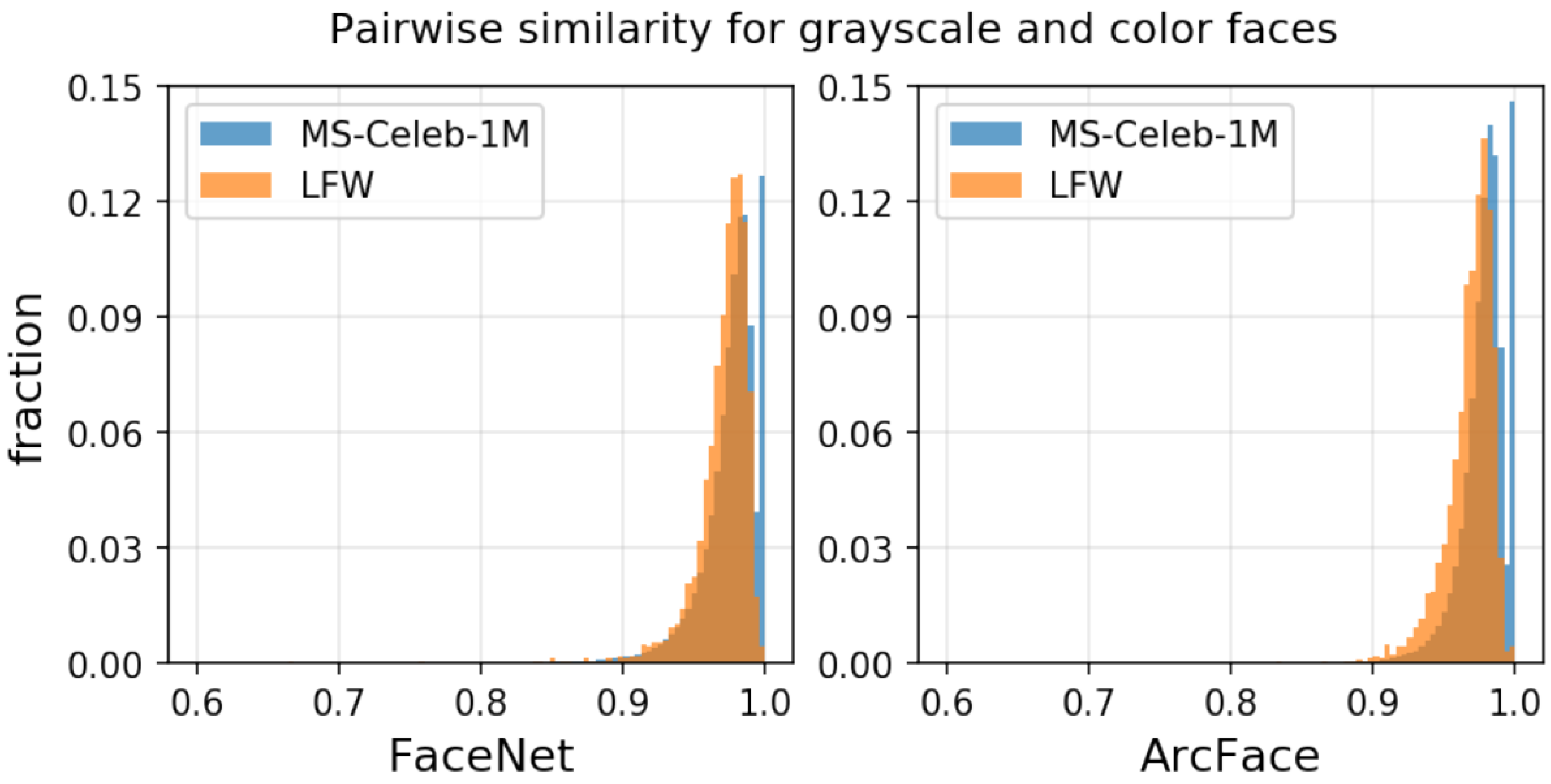
Symmetry VS asymmetry VS color

Symmetry, black&white | Asymmetry, black&white | Symmetry, rgb

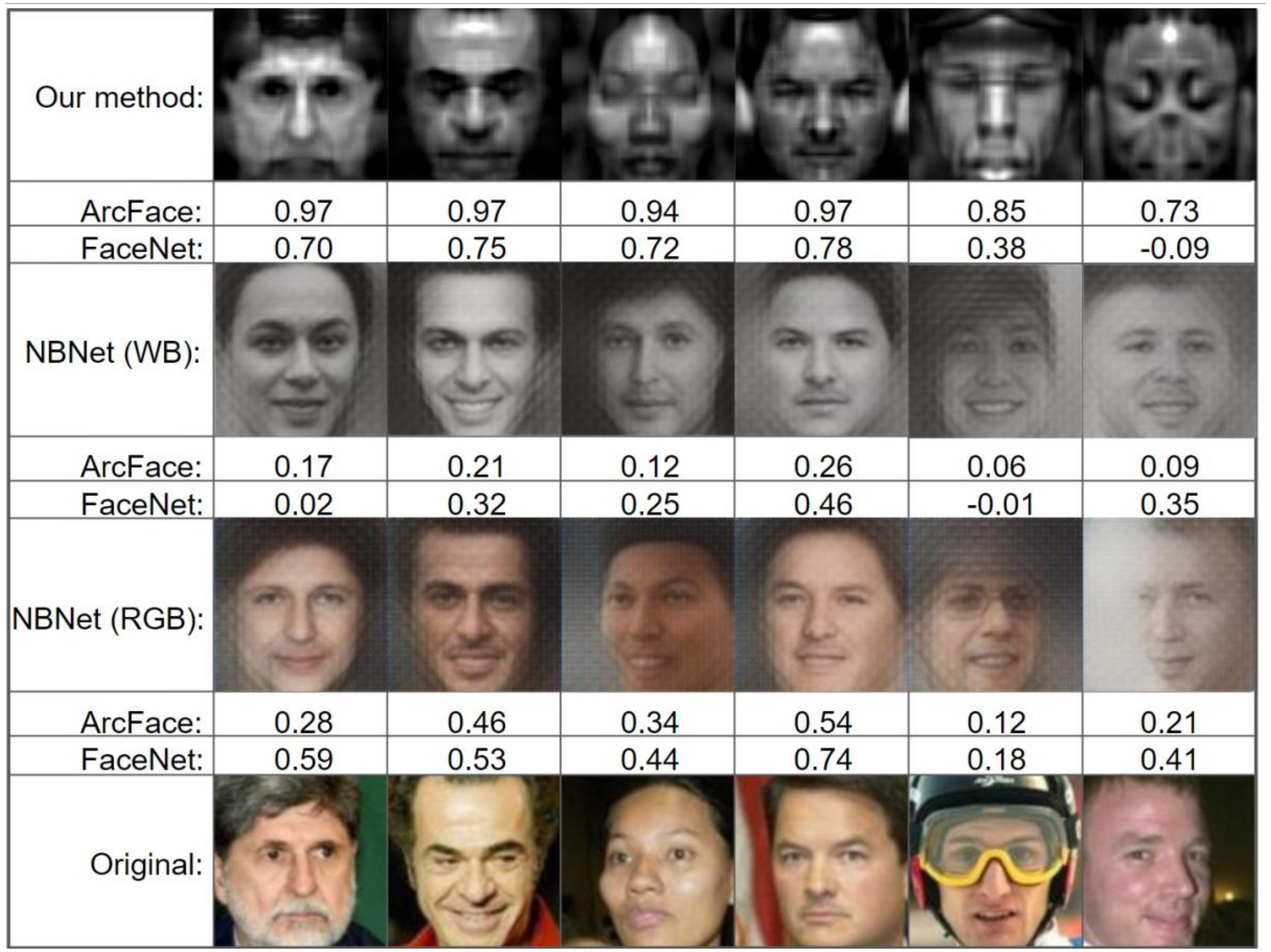


Best results with **symmetry** and **black & white** constraints

Identity features do not contain much information about color



Comparison with others



Our method:

- produces **less realistic faces**,
- but more likely **preserves identity of a person** which **correlates** with similarities of **"independent critic"** network

NBNet: Mai, G., Cao, K., Yuen, P.C., Jain, A.K.: *On the reconstruction of face images from deep face templates*. IEEE transactions on pattern analysis and machine intelligence, 2018

Comparison with others

Method	ArcFace	FaceNet	# of queries
(Ours) Symmetric gauss, LFW (wb)	**0.92**	**0.46**	**300**k
(Ours) Asymmetric gauss, LFW (wb)	0.85	0.42	400k
NBNet, LFW (RGB)	0.25	0.34	3M
NBNet, LFW (wb)	0.19	0.27	3M
(Ours) Symmetric gauss, MS1M-ArcFace (wb)	**0.91**	**0.44**	**300**k
NBNet, MS1M-ArcFace (RGB)	0.26	0.38	3M
NBNet, MS1M-ArcFace (wb)	0.20	0.32	3M

Table 2: Average cosine similarity by ArcFace and FaceNet (independent critic) between embedding of a reconstructed image and embedding of target image for subsets of 1000 images from LFW and MS1M-ArcFace and corresponding number of queries.

Conclusions

- We demonstrate that it is **possible** to **recover recognizable faces** from deep **feature vectors** of a face-recognition model in a **black-box mode** with **no prior knowledge**.
- The **proposed method outperforms** current solutions in terms of the **average cosine similarity** of embeddings produced by the **attacked model** and an **independent critic**.
- The **proposed method** requires a significantly **smaller number of queries** compared to previous solutions and **does not need** prior information such as **proper training dataset**.