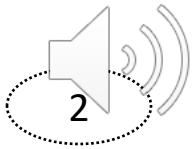
Generating Master Faces for Use in Performing Wolf Attacks on Face Recognition Systems

Huy H. Nguyen, Junichi Yamagishi,
Isao Echizen, and Sébastien Marcel

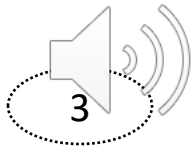
IJCB 2020

Overview

1. Motivation
2. Related Work
3. Proposed Method
4. Evaluation
5. Summary



1. Motivation



1. Motivation

“Wolf” sample in biometrics: an input which can be falsely accepted as a match with multiple templates [Une 2007]

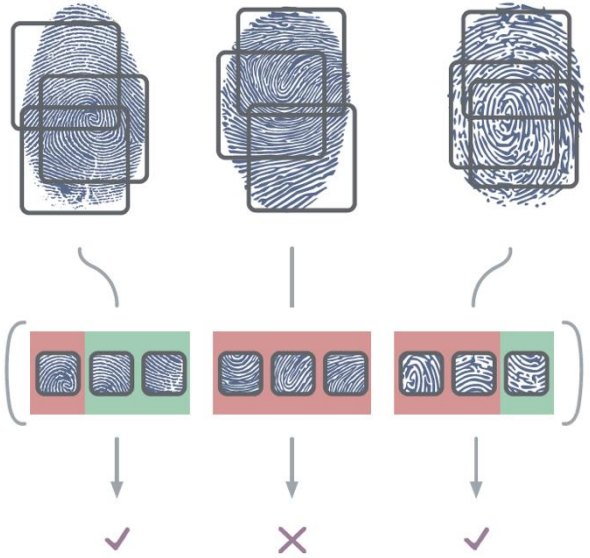
Master key → Master face (or wolf face)

Source: Internet

1. Motivation

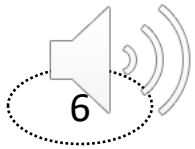
Face morphing
Need victim's face
Credit: Seibold, C (2017)

Deep master faces
No knowledge of the enrolled IDs
High-quality facial images
Credit: GoT



Deep master prints
Bontrager, P (2018)

2. Related Work



2.1. Face Generation

VAEs family

Kingma, P (2013)

White, T (2016)

GANs/WGANs family

Goodfellow, I (2014)

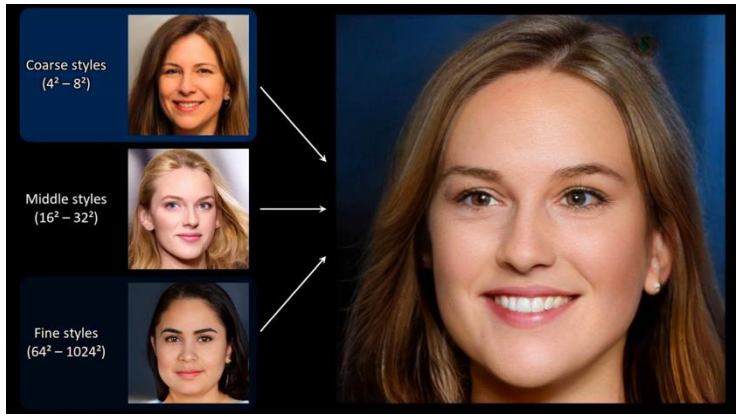
Arjovsky, M (2017)

Wu, J (2017)

- Low resolution & low-quality generated images (except the recent BigGAN)
- VAEs: Trade of between quality & disentangleability
- GANs: Difficult to train

ProgressiveGAN

Karras, T (2017)



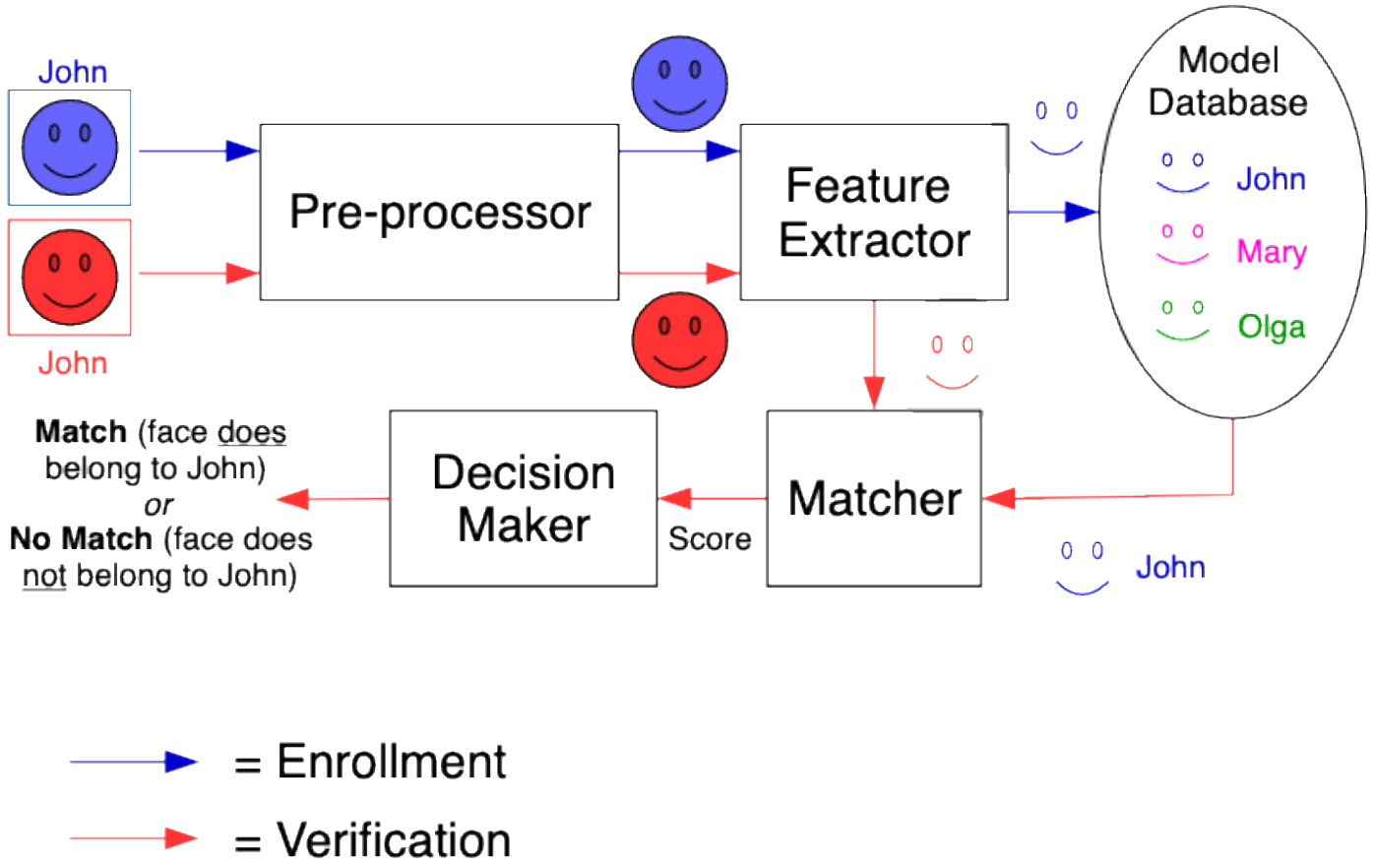
StyleGAN,
StyleGAN 2

Karras, T (2018, 2019)

- High-quality generated images
- Better disentangleability

[Karras, T (2018)] Karras, T., Laine, S., & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. CVPR (pp. 4401-4410).

2.2. Face Recognition (FR)



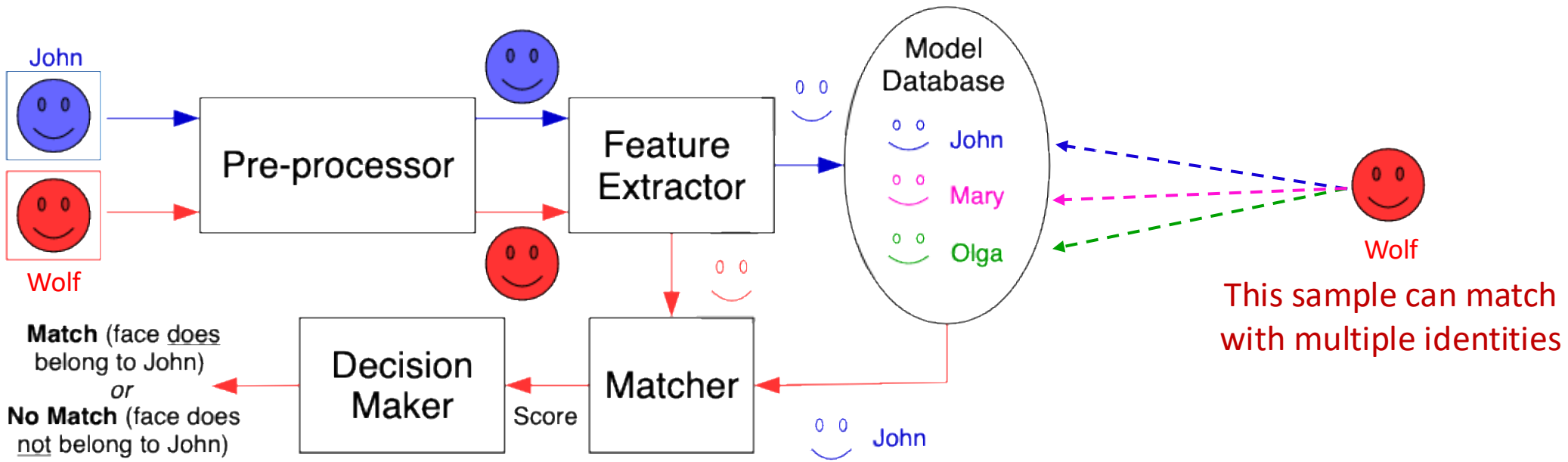
Enrollment and verification in a typical biometric verification system.

Source: Idiap bob toolbox

Structure of a face recognition system:

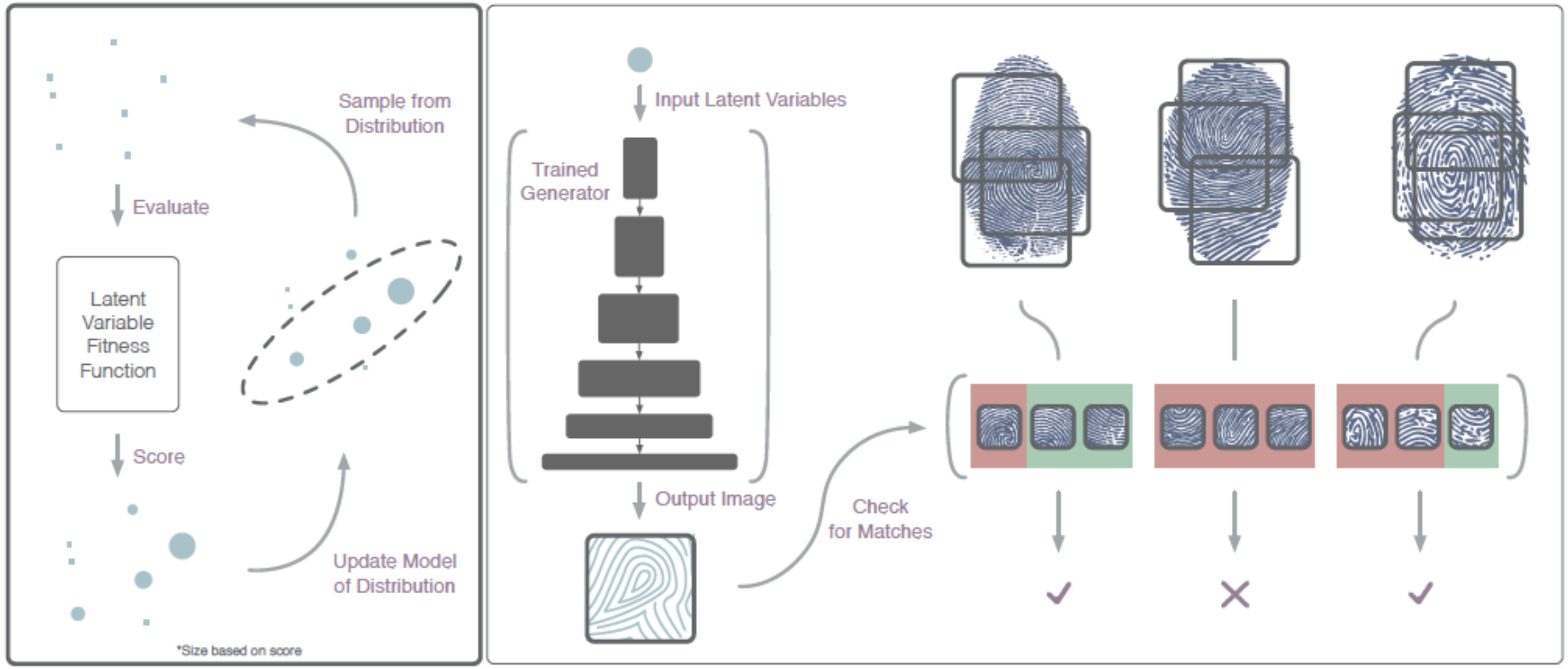
- **Pre-processor**: Face detection & cropping
- **Feature Extractor**: Usually some CNN
- **Matcher**: using some distance (cosine)
- **Decision Maker**: usually based on EER calculated on dev set

2.3. Wolf Attack



Wolf attack does not require the knowledge about the enrolled user' templates!

2.4. Latent Variable Evolution (LVE)



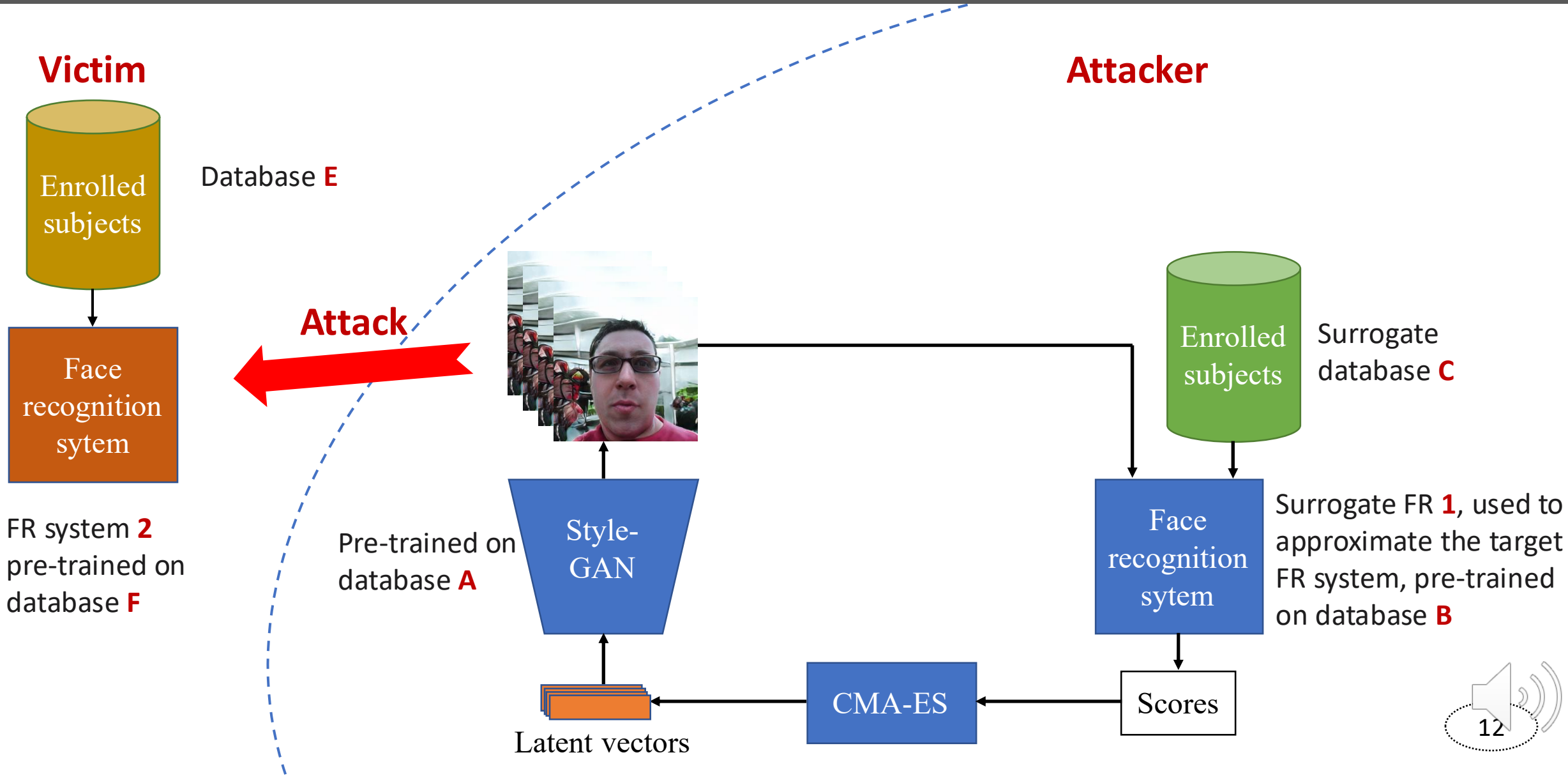
CMA-ES Optimization Latent Variable Fitness Function

LVE with a trained network appied on a **partial** fingerprint recognition system.

Bontrager, P (2018)

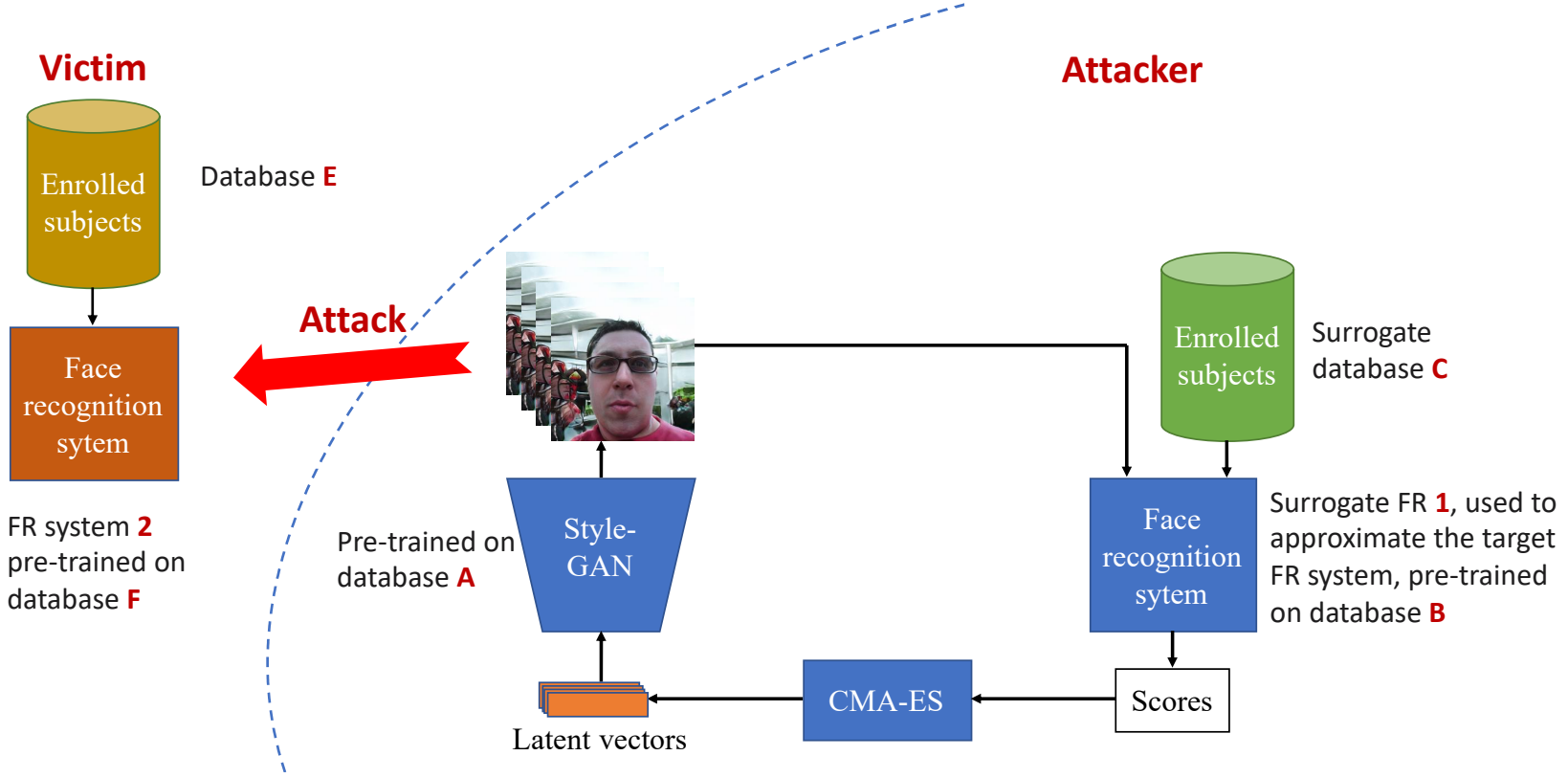
3. Proposed Method

3.1. Overview of the Proposed Method

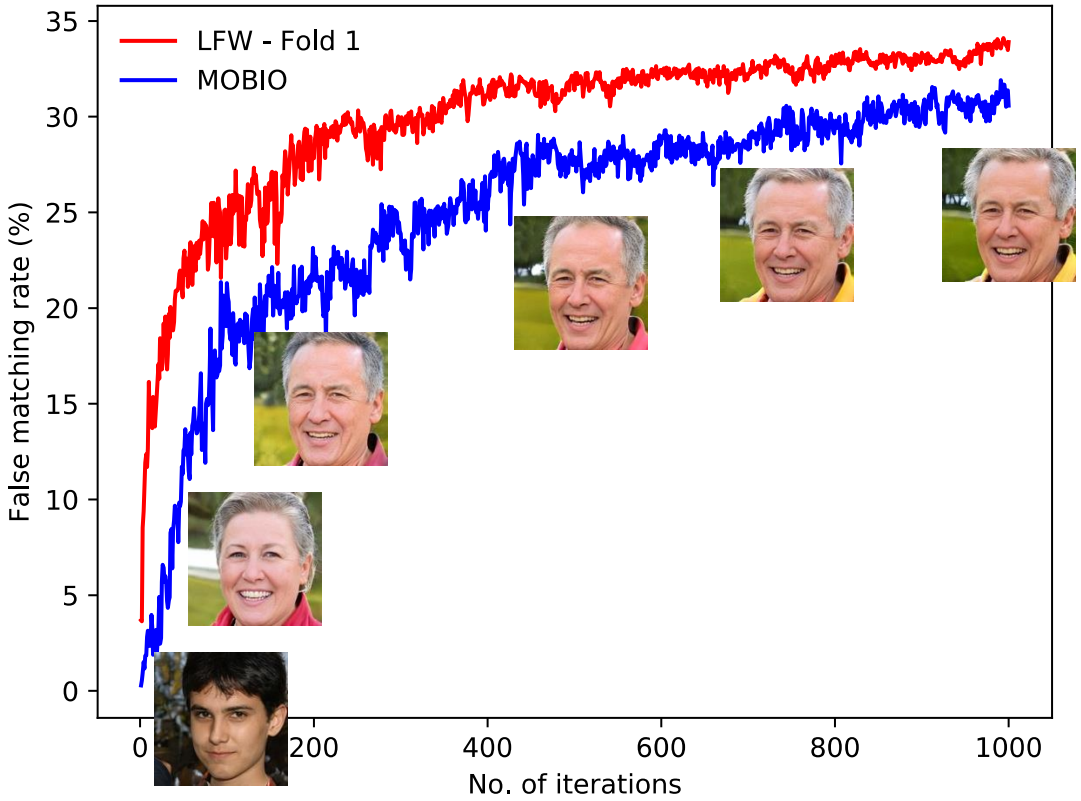


3.2. Contributions

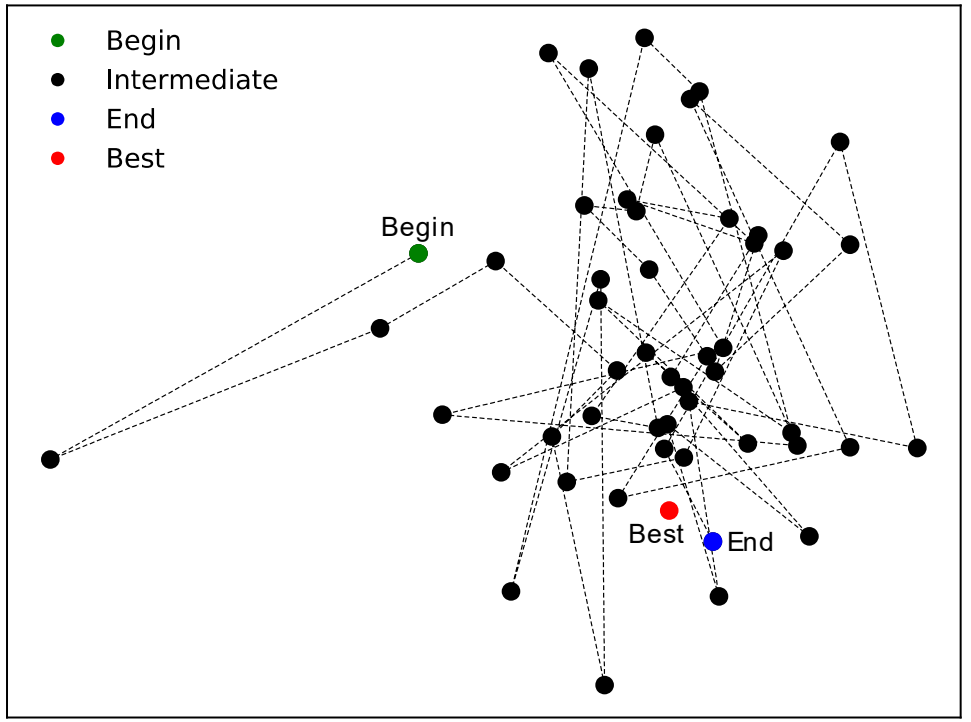
- First research to investigate wolf attack on facial domain
- Focused on both gray-box & black-box attacks
- Improved the LVE algorithm by changing the way of scoring
- Performed analysis about the matched IDs (gender, race, appearances)



3.3. Training

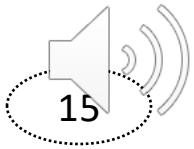


False matching rates (FRMs) are increasing when training the proposed method on LFW database [Huang, G 2007] and MOBIO database [McCool, C 2012].



T-SNE visualization of master faces obtained every 20 iterations (1000 in total) on LFW database.

4. Evaluation



4.1. Settings

Databases:

- Flickr-Faces-HQ (FFHQ)
  → train StyleGAN [Karras, T 2018]

- CASIA-WebFace [Yi, D 2014], MS-Celeb [Guo, Y 2016], MultiPIE [Gross, R 2016]
  → train FR systems

- LFW [Huang, G 2007] and MOBIO [McCool, C 2012]
  → Generate & evaluate master faces

4.1. Settings

Face recognition systems (used bob toolkit):

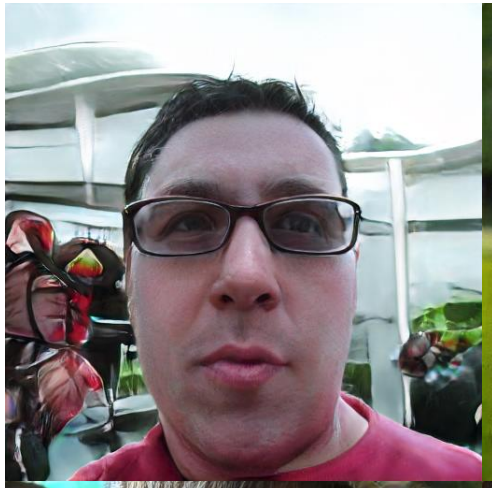
- Inception-ResNet-v2 based system [Freitas Pereira, T 2018] trained on:
  - CASIA-WebFace → used to run the LVE algorithm
  - MS-Celeb

- FaceNet [Schroff, F 2015] trained on MS-Celeb by David Sandberg

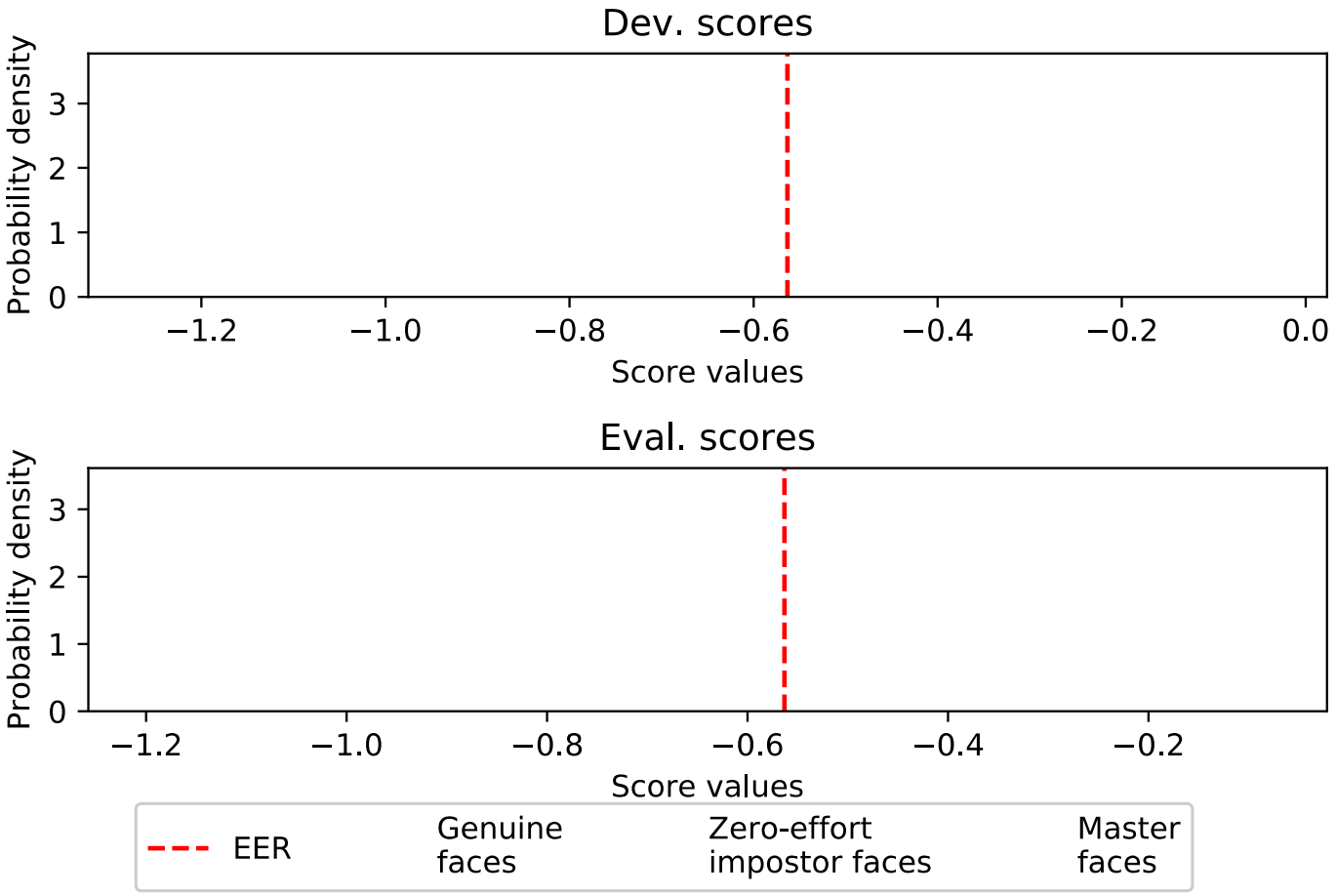
- DR-GAN [Tran, L 2017] trained on Multi-PIE & Casia-WebFace

4.2. Generated Master Faces

Database Used for LVE Strategy

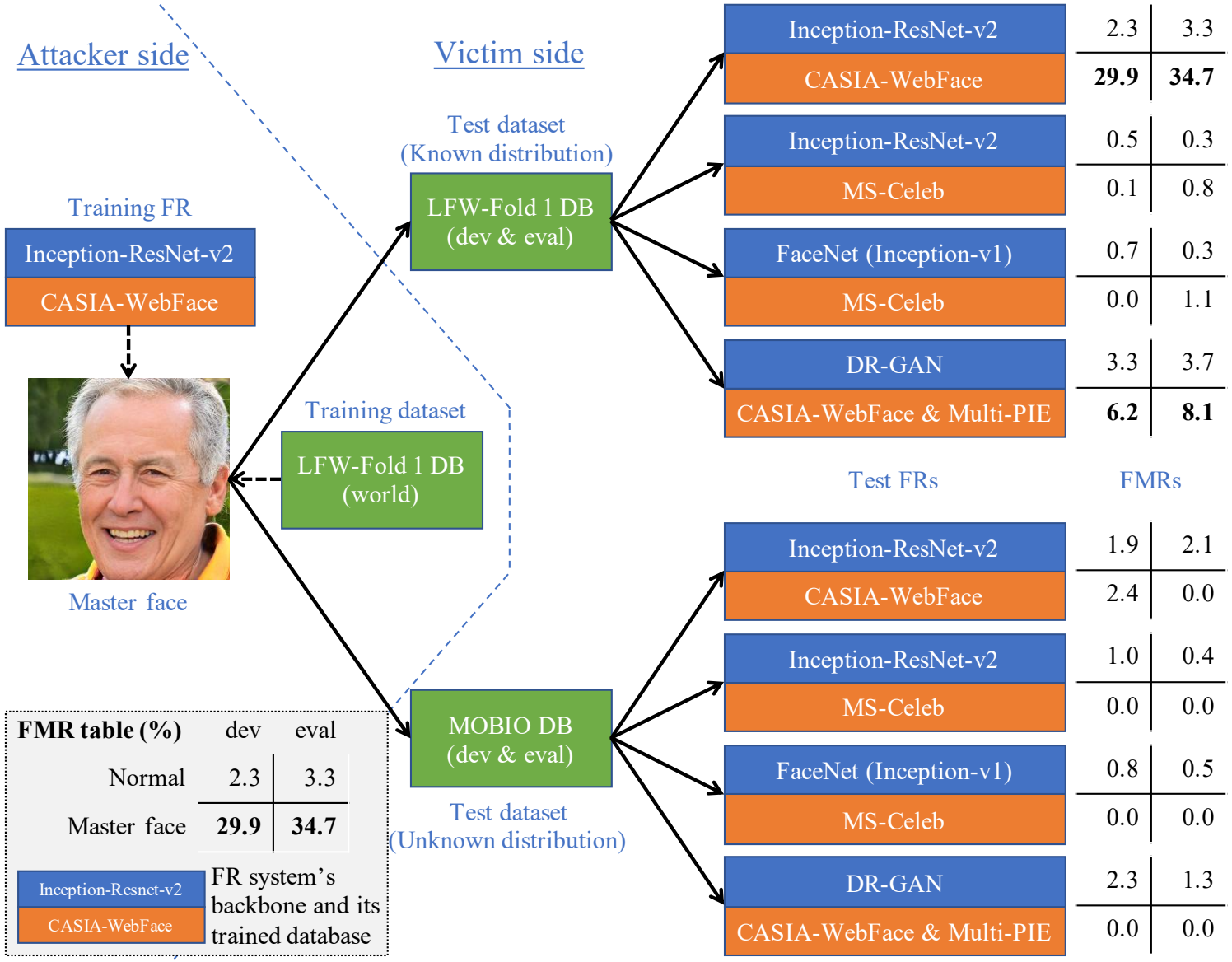
FR Sytem:	MOBIO Database	LFW Database
Inception-ResNet-v2 trained on CASIA-WebFace database		
Inception-ResNet-v2 trained on MS-Celeb database		

4.3. Master face Trained on LFW-Fold 1 DB

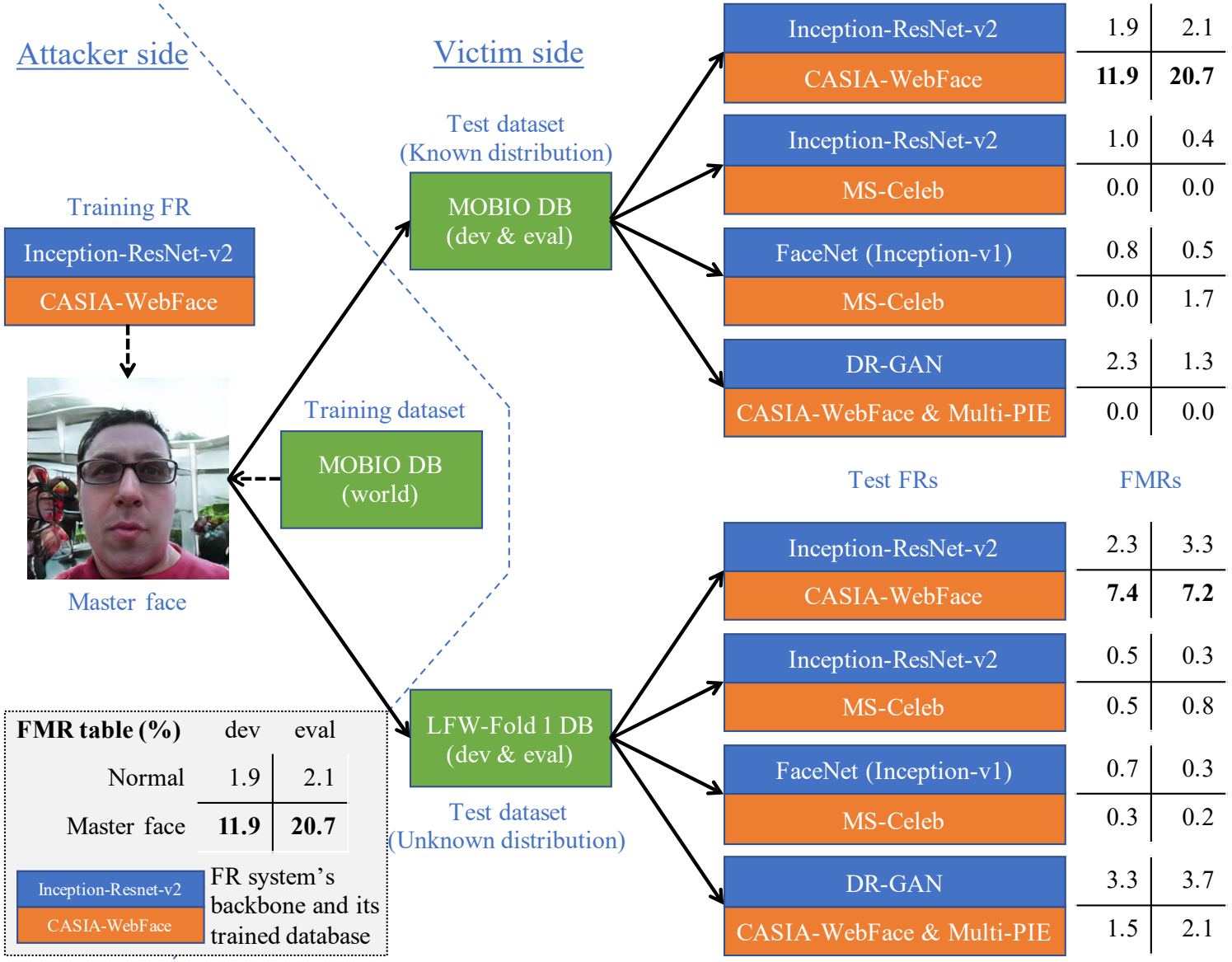


Histogram of scores calculated using Inception-ResNet-v2 based FR system.

4.3. Master face Trained on LFW-Fold 1 DB



4.4. Master face Trained on MOBIO DB



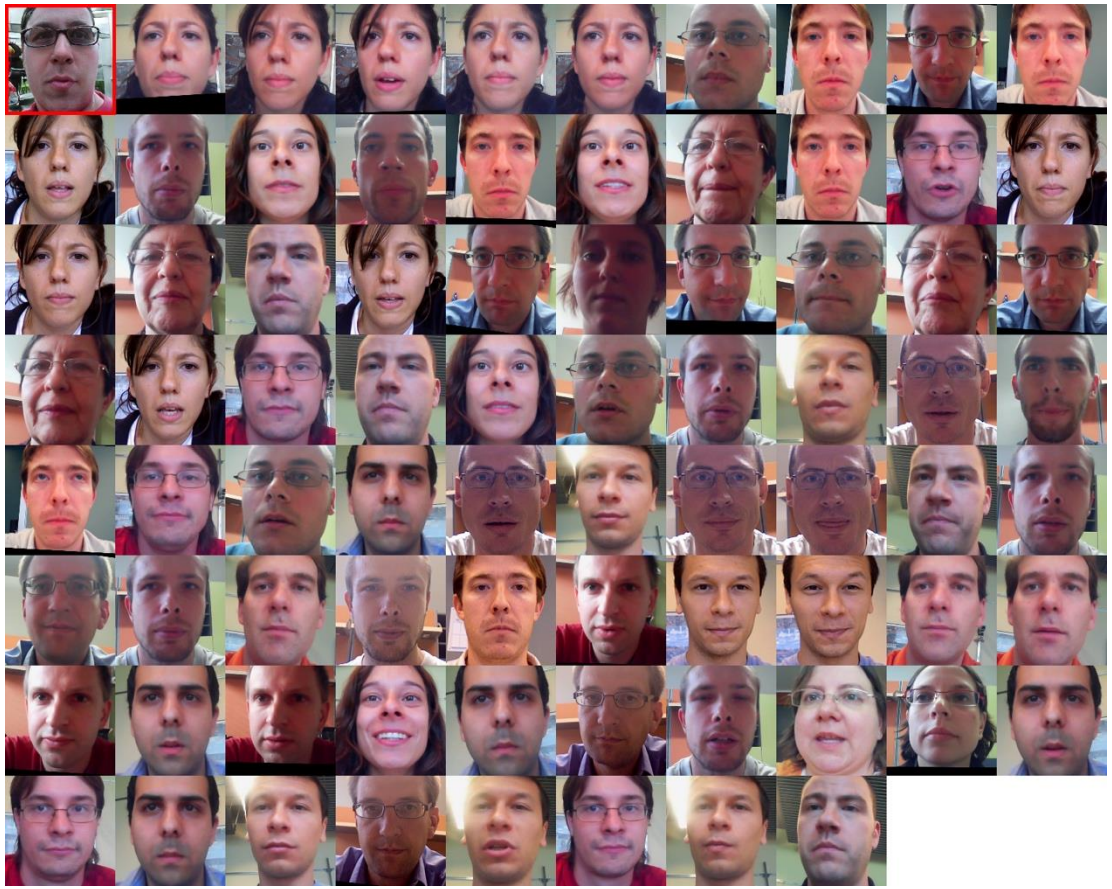
4.5. Summary

Target FR Setting	Scenario 1		Scenario 2	
	Known DB	Unknown DB	Known DB	Unknown DB
Same Arch. – Same DB	**1**	0	**1**	**1**
Same Arch. – Different DB	0	0	0	0
Different Arch. – Same DB	**1**	0	0	0
Different Arch. – Different DB	0	0	0	0

Summary of successful attacks for scenario 1 and 2 with different FR system settings and databases.

4.6. Examples of Matched Enrolled IDs

Master face



MOBIO – eval set

Master face

LFW-Fold 1 – dev set

5. Summary

5. Summary

- The proposed method is **simple but efficient**:
  - Only uses avaiable resources: a pre-trained StyleGAN, a pre-train face recognition system, a face database
  - Only needs a conventional PC without a GPU
  - Runs in less than 24 hours

  →Very easy for attackers to create master faces

- The master faces can be generalized in some cases

- The properties of the master faces can provide clues for understading and improving face recognition systems.

Thank you very much!