[Paper]

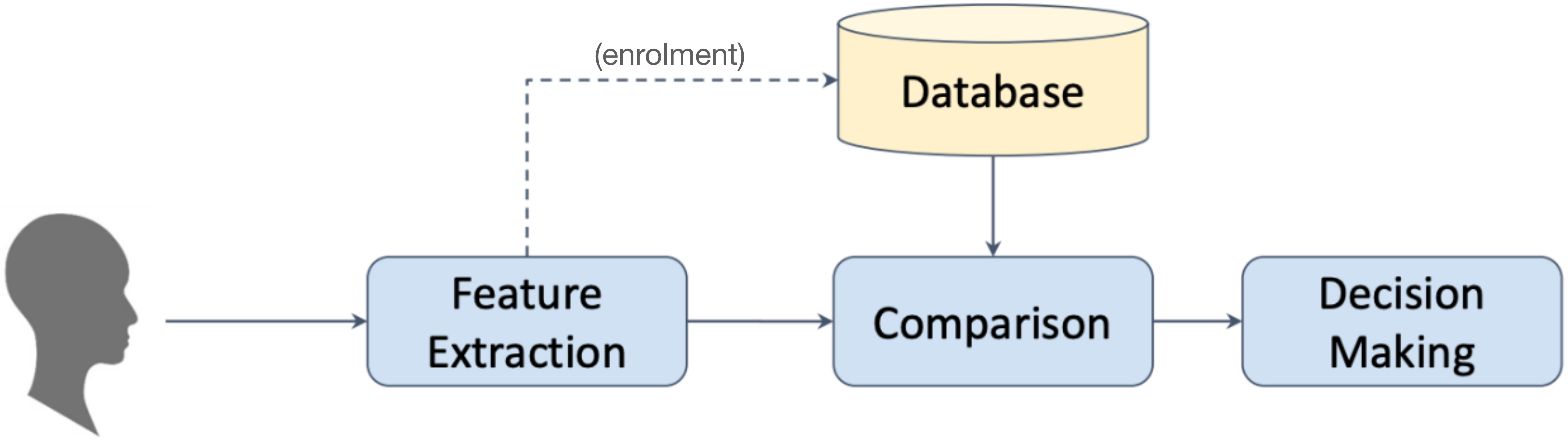
Face Reconstruction from Facial Templates by Learning Latent Space of a Generator Network

Hatef Otroshi Shahreza[1,2], Sébastien Marcel[1,3]

[1] Idiap Research Institute, Martigny, Switzerland
[2] École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3] Université de Lausanne (UNIL), Lausanne, Switzerland

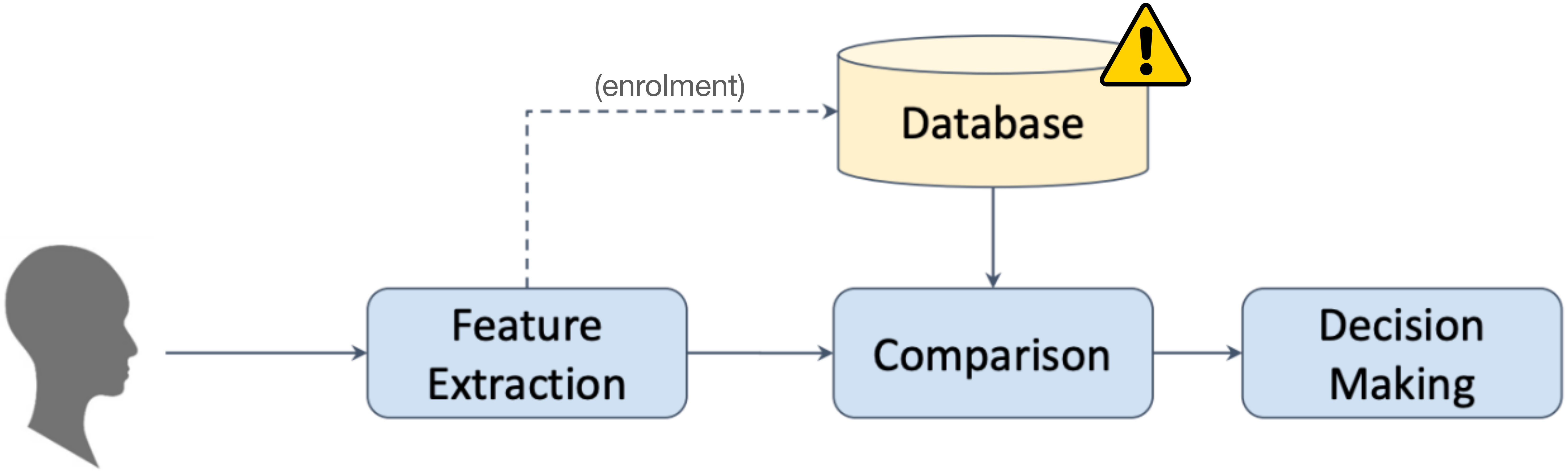
Introduction

- **Problem definition**



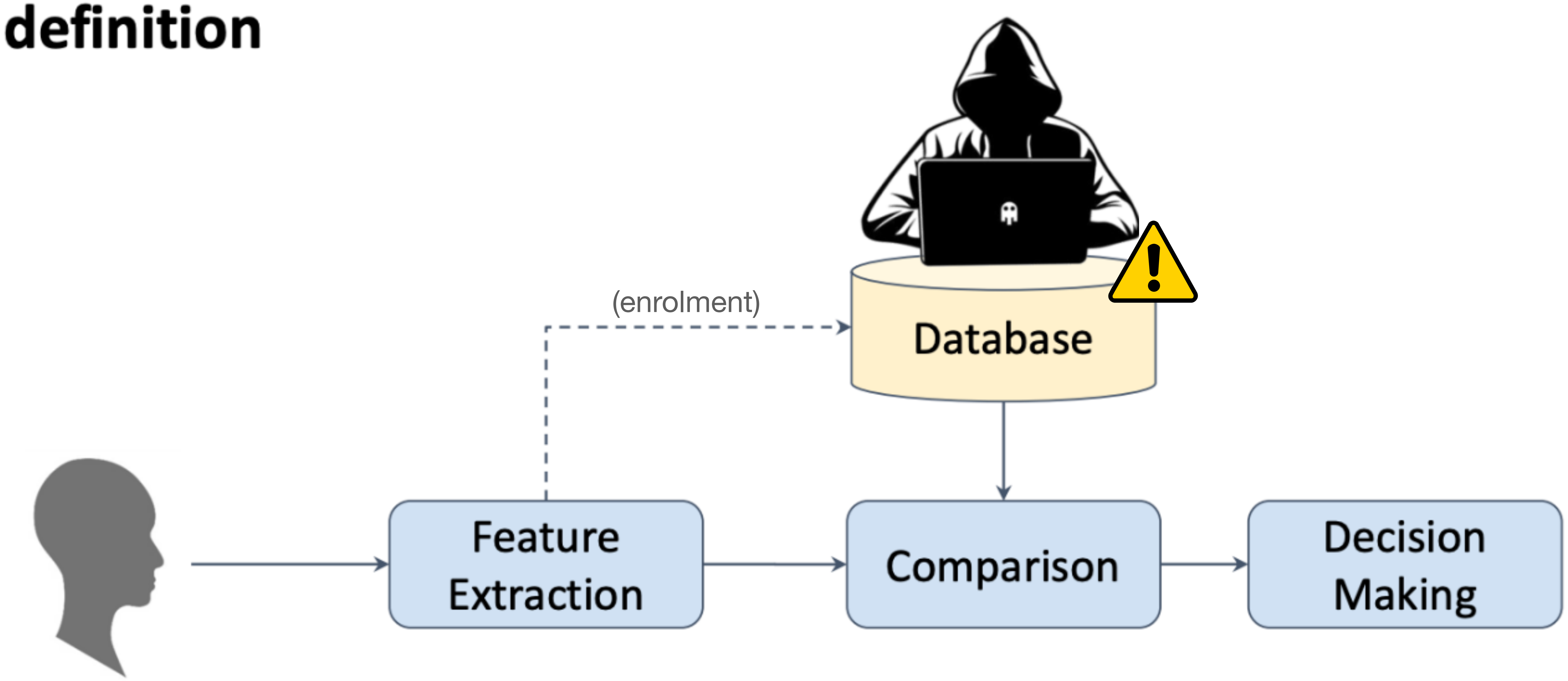
Introduction

- **Problem definition**



Introduction

- **Problem definition**



Threat Model

- **Adversary's goal:** The adversary aims to reconstruct face images from templates stored in the database of a face recognition system (F_{database}), and use the reconstructed face images to enter the same or different face recognition system (F_{target}).

Threat Model

- **Adversary's goal:** The adversary aims to reconstruct face images from templates stored in the database of a face recognition system (F_{database}), and use the reconstructed face images to enter the same or different face recognition system (F_{target}).

- **Adversary's knowledge:**
  - The leaked face templates enrolled in the database.
  - The *whitebox/blackbox* knowledge of the feature extractor model (F_{database})
    - In case of blackbox scenario, the adversary has a whitebox knowledge of another face recognition model to use in training (F_{loss}).

Threat Model

- **Adversary's goal:** The adversary aims to reconstruct face images from templates stored in the database of a face recognition system (F_{database}), and use the reconstructed face images to enter the same or different face recognition system (F_{target}).

- **Adversary's knowledge:**
  - The leaked face templates enrolled in the database.
  - The *whitebox/blackbox* knowledge of the feature extractor model (F_{database})
    - In case of blackbox scenario, the adversary has a whitebox knowledge of another face recognition model to use in training (F_{loss}).

- **Adversary's capability:** The adversary can inject the reconstructed face image as a query to the target face recognition system (F_{target}).

Threat Model

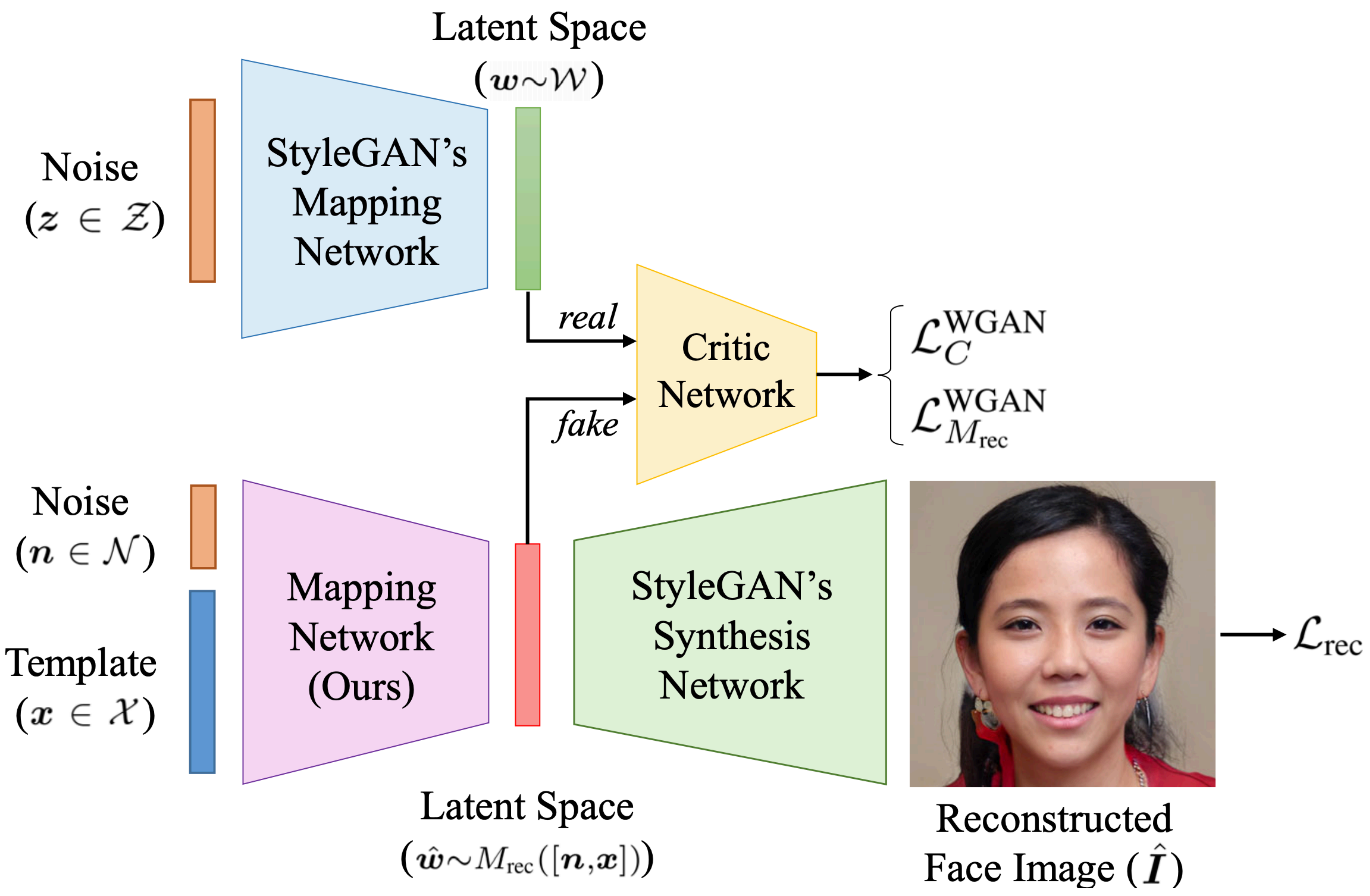
- **Adversary's goal:** The adversary aims to reconstruct face images from templates stored in the database of a face recognition system (F_{database}), and use the reconstructed face images to enter the same or different face recognition system (F_{target}).

- **Adversary's knowledge:**
    - The leaked face templates enrolled in the database.
    - The *whitebox/blackbox* knowledge of the feature extractor model (F_{database})
        - In case of blackbox scenario, the adversary has a whitebox knowledge of another face recognition model to use in training (F_{loss}).

- **Adversary's capability:** The adversary can inject the reconstructed face image as a query to the target face recognition system (F_{target}).

- **Adversary's strategy:** The adversary trains a face reconstruction model to invert the leaked facial templates. Then, the adversary can use the reconstructed face images and inject as a query to the target face recognition system (F_{target}).

Proposed Method



- **WGAN loss**

$$\mathcal{L}_C^{\text{WGAN}} = \mathbb{E}_{\boldsymbol{w} \sim \mathcal{W}}[C(\boldsymbol{w})] - \mathbb{E}_{\hat{\boldsymbol{w}} \sim M_{\text{rec}}([\boldsymbol{n}, \boldsymbol{x}])}[C(\hat{\boldsymbol{w}})]$$

$$\mathcal{L}_{M_{\text{rec}}}^{\text{WGAN}} = \mathbb{E}_{\hat{\boldsymbol{w}} \sim M_{\text{rec}}([\boldsymbol{n}, \boldsymbol{x}])}[C(\hat{\boldsymbol{w}})]$$

- **Reconstruction loss**

$$\mathcal{L}_{\text{rec}} = \mathcal{L}_{\text{pixel}} + \mathcal{L}_{\text{ID}}$$

$$\mathcal{L}_{\text{pixel}} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{X}}[\|\boldsymbol{I} - G(\boldsymbol{x})\|_2^2],$$

$$\mathcal{L}_{\text{ID}} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{X}}[\|F_{\text{loss}}(\boldsymbol{I}) - F_{\text{loss}}(G(\boldsymbol{x}))\|_2^2].$$

Proposed Method

	$\mathbf{F}_{\mathbf{database}}$	$\mathbf{F}_{\mathbf{loss}}$	**Evaluation**	**Adversary's Knowledge of Original and Target Systems**	**Difficulty of Attack**
Attack 1	whitebox	F_{database}	same system	whitebox knowledge of F_{database} and F_{target}	very easy
Attack 2	whitebox	F_{database}	different system (transferability)	whitebox knowledge of F_{database}	easy
Attack 3	blackbox	adversary's own	same system	blackbox knowledge of F_{database} and F_{target}	difficult
Attack 4	blackbox	F_{target}	different system (transferability)	blackbox knowledge of F_{database} and whitebox knowledge of F_{target}	difficult
Attack 5	blackbox	adversary's own	different system (transferability)	only blackbox knowledge of F_{database}	very difficult

Experiments

- **Sample Reconstructed Face Images**

Experiments

Table 4: Evaluation of attacks with *whitebox* knowledge of the system from which the template is leaked (i.e., $F_{\text{loss}} = F_{\text{database}}$) against SOTA FR models in terms of adversary's success attack rate (SAR) using our proposed method on the MOBIO and LFW datasets. The values are in percentage and correspond to the threshold where the target system has FMR $= 10^{-3}$. Cells are color coded according the type of attack as defined in Section 2 for attack 1 (light gray) and attack 2 (dark gray).

$\mathbf{F}_{\text{database}}$	MOBIO					LFW				
	ArcFace	**ElasticFace**	**HRNet**	**AttentionNet**	**Swin**	**ArcFace**	**ElasticFace**	**HRNet**	**AttentionNet**	**Swin**
ArcFace	92.38	81.90	71.43	70.48	74.29	86.82	74.20	36.57	36.40	58.86
ElasticFace	78.10	87.62	64.29	64.76	69.05	78.25	82.52	41.80	40.25	61.09

Table 5: Evaluation of attacks (with *blackbox* knowledge of the system from which the template is leaked i.e., F_{database}) against SOTA FR models in terms of adversary's success attack rate (SAR) using different methods on the MOBIO and LFW datasets. The values are in percentage and correspond to the threshold where the target system has FMR $= 10^{-3}$. **M1**: NbNetB-M [Mai et al., 2018], **M2**: NbNetB-P [Mai et al., 2018], **M3**: [Dong et al., 2021], **M4**: [Vendrow and Vendrow, 2021], and **M5**: [Dong et al., 2023]. Cells are color coded according the type of attack as defined in Section 2 for attack 3 (lightest gray), attack 4 (middle dark gray), and attack 5 (darkest gray).

$\mathbf{F}_{\text{database}}$	$\mathbf{F}_{\text{loss}}$	$\mathbf{F}_{\text{target}}$	MOBIO						LFW					
			M1	**M2**	**M3**	**M4**	**M5**	**Ours**	**M1**	**M2**	**M3**	**M4**	**M5**	**Ours**
ArcFace	**ElasticFace**	**ArcFace**	1.90	15.24	2.38	28.10	58.57	**81.90**	10.68	40.25	12.91	58.88	75.31	**77.16**
		ElasticFace	1.43	11.43	4.29	15.24	37.61	**73.81**	8.36	34.39	6.35	29.10	50.17	**68.06**
		HRNet	0.95	6.19	2.86	10.00	30.48	**57.14**	1.30	7.78	1.75	9.20	24.72	**28.45**
		AttentionNet	0	6.67	3.33	4.29	26.67	**54.29**	1.33	7.17	2.29	9.17	24.16	**28.87**
		Swin	1.43	13.33	3.81	10.95	40.00	**67.14**	4.27	23.85	5.97	21.75	41.27	**48.28**
ElasticFace	**ArcFace**	**ArcFace**	2.38	18.57	2.86	16.19	48.09	**87.14**	15.33	48.67	11.81	37.45	65.40	**83.20**
		ElasticFace	3.81	43.81	4.76	43.33	72.38	**89.05**	21.44	58.16	11.59	52.88	74.08	**83.43**
		HRNet	0.48	20.00	1.43	10.48	42.86	**73.81**	3.46	18.36	2.74	11.82	32.99	**49.02**
		AttentionNet	1.90	18.10	3.33	9.05	40.00	**71.90**	2.89	16.31	2.91	10.95	31.15	**46.63**
		Swin	0.95	26.19	2.86	15.24	46.67	**75.24**	9.22	38.79	8.26	24.62	51.20	**66.89**

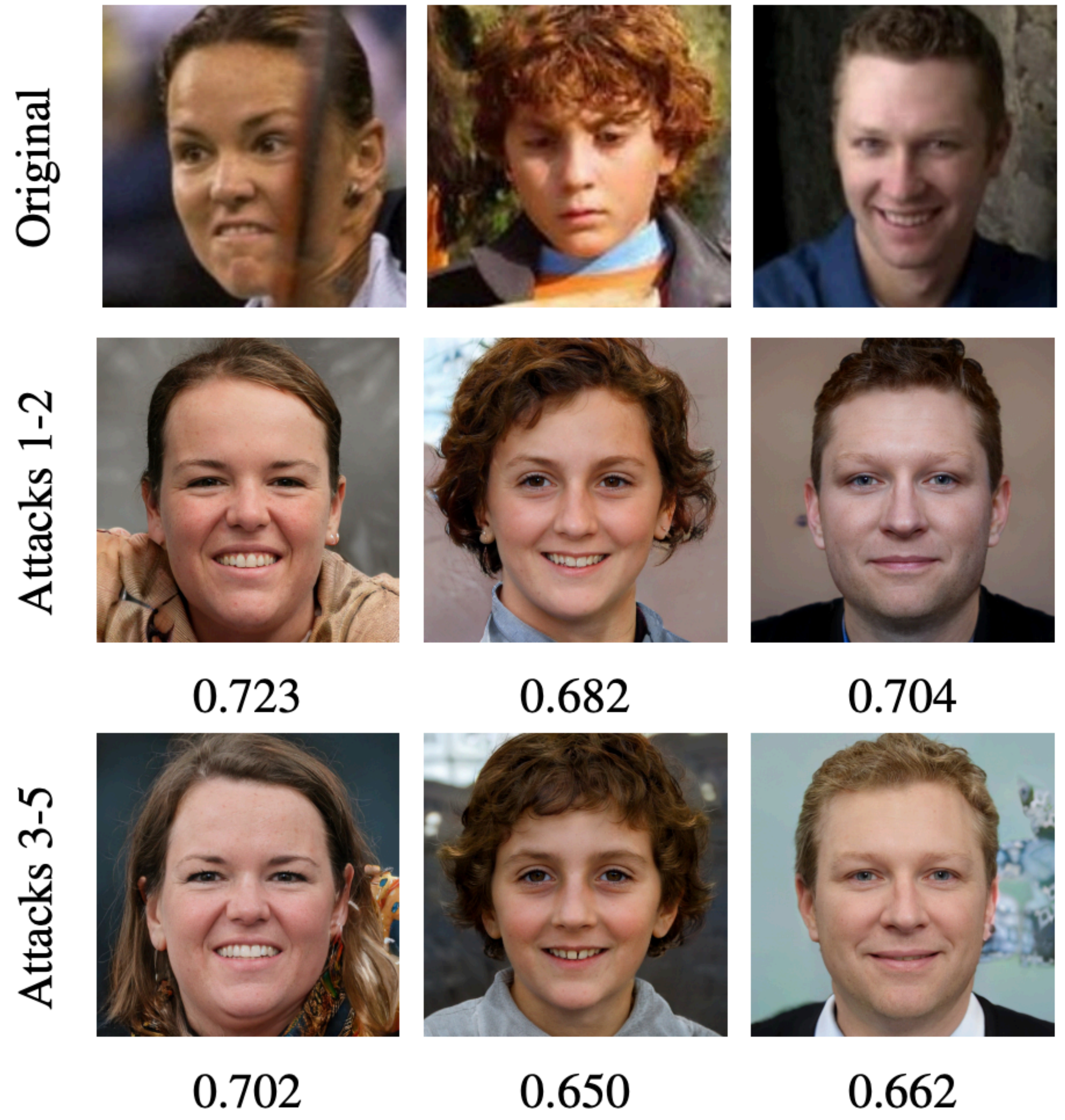
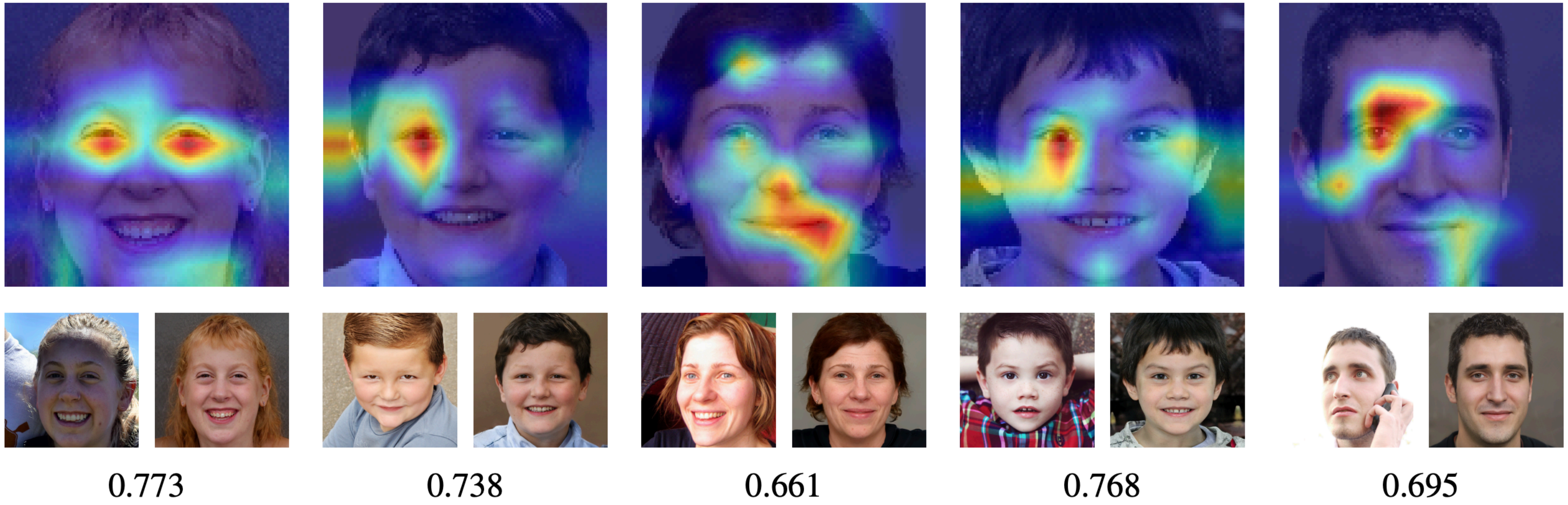


Figure 4: Sample face images from the LFW dataset (first raw) and their corresponding reconstructed images using our template inversion method from ArcFace templates in different attacks, attacks 1-2 (second raw) and attacks 3-5 (second raw, using ElasticFace for F_{loss}). The values below each image show the cosine similarity between the corresponding ArcFace templates of original and reconstructed face images.

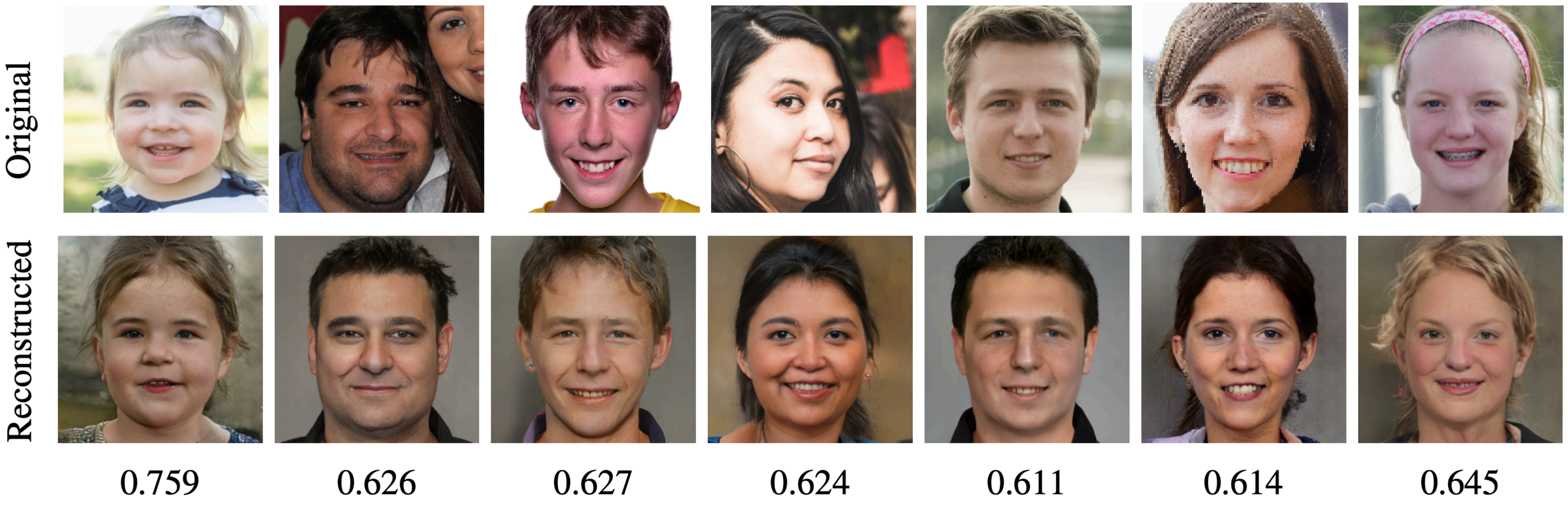
Experiments

- **Important Areas in the Reconstructed Face Images**



Experiments

- **Using a Different Face Generator Network (StyleSwin)**



Thanks for your attention!

[Paper]

[Source Code]