CVPR 2023

Implicit Identity Driven Deepfake Face Swapping Detection

Suhyeon Ha

2024.05.27

KAIST

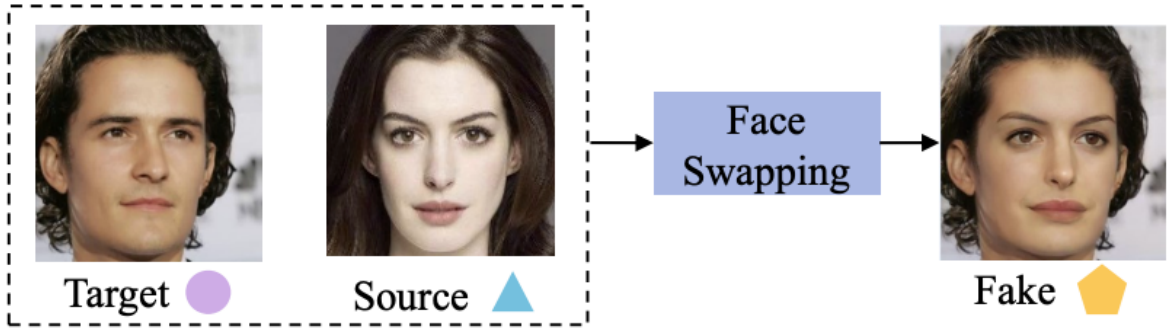
Review

- **One-class self-supervised learning** using real face images only.

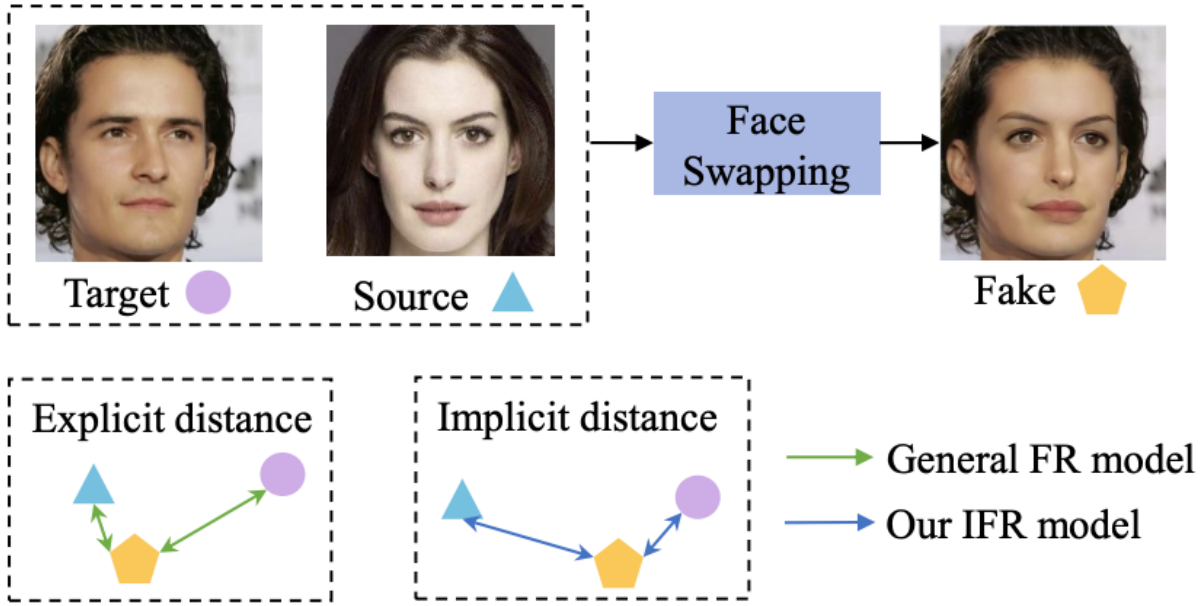
- **Soft discrepancy** : Different local perturbations introduced into real images.
- **Pretext Task**: Through the localization of the soft discrepancy region and the detection of different augmentation methods.

< Examples of faces with soft-discrepancies >

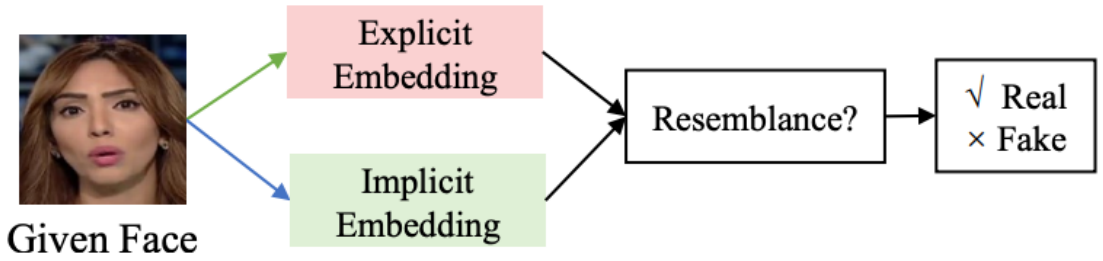
Introduction



Introduction



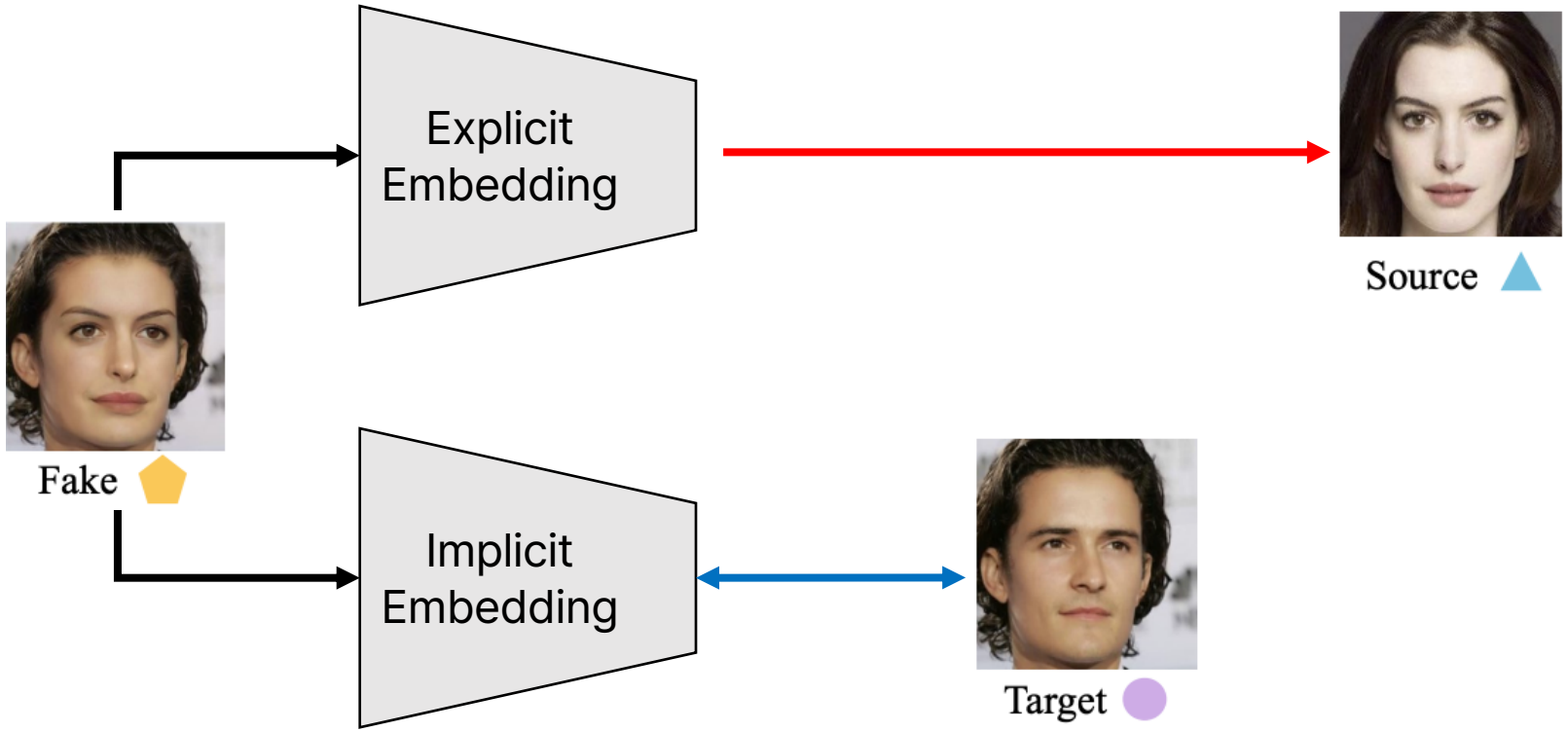
Introduction



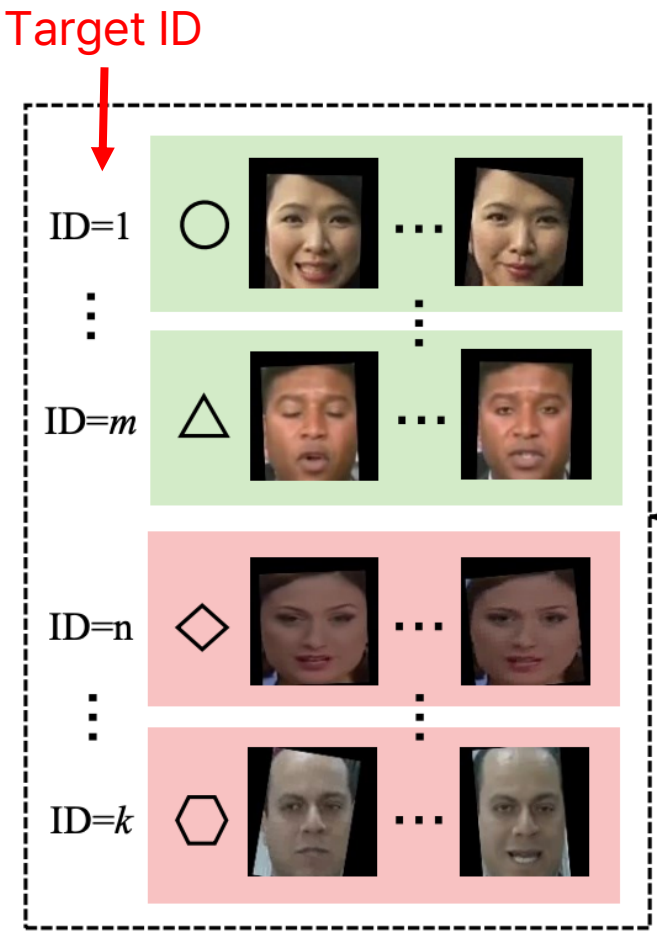
Introduction

- We propose the implicit identity driven framework for face swapping detection, which explores the implicit identity of fake faces. This enhances the deep network to distinguish fake faces with unknown manipulations.
- We specially design explicit identity contrast (EIC) loss and the implicit identity exploration (IIE) loss. EIC aims to pull real samples closer to their explicit identities and push fake samples away from their explicit identities. IIE is margin-based and guides fake faces with known target identities to have small intraclass distances and large inter-class distances.
- Extensive experiments and visualizations demonstrate the superiority of our method over the state-of-the-art approaches.

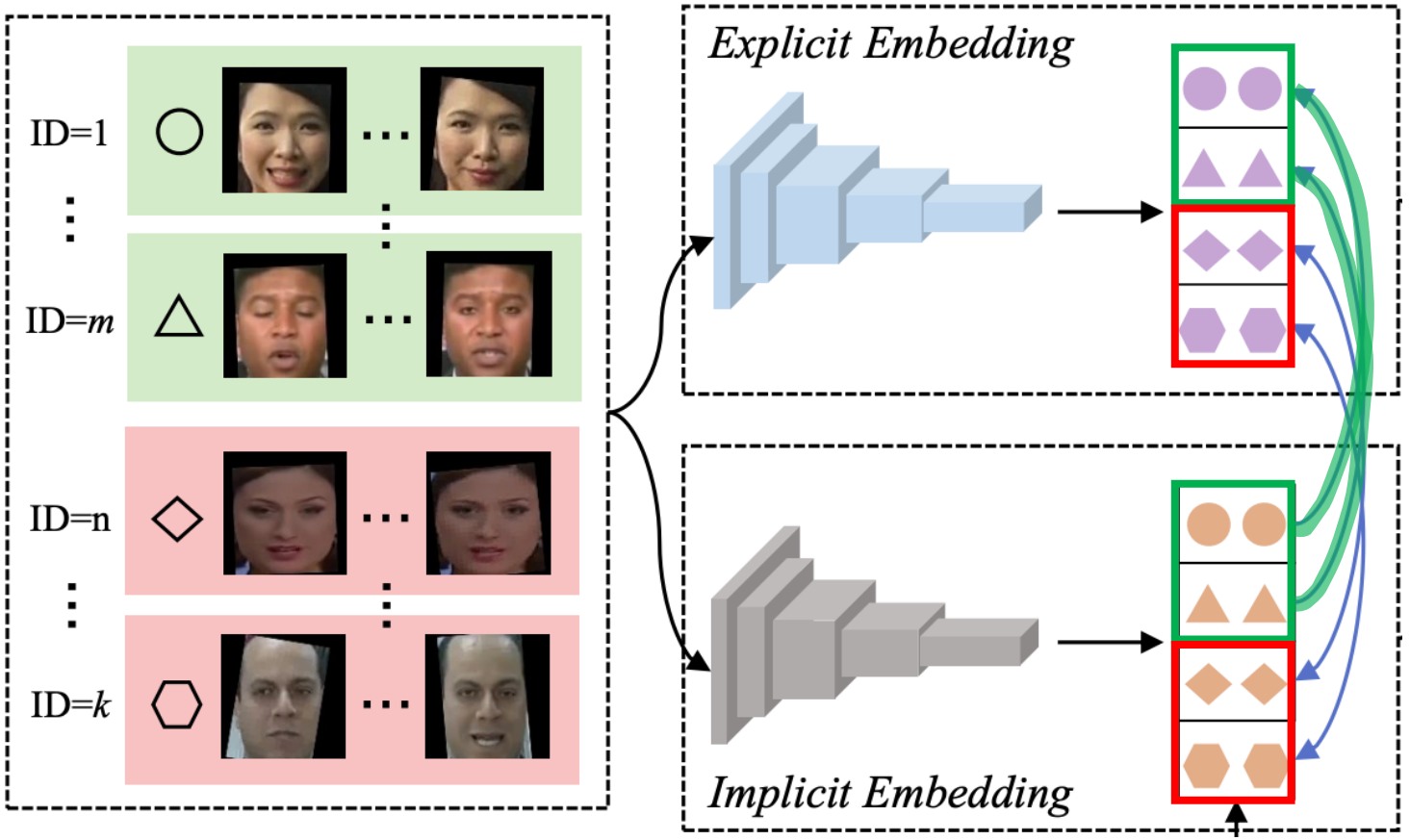
Introduction



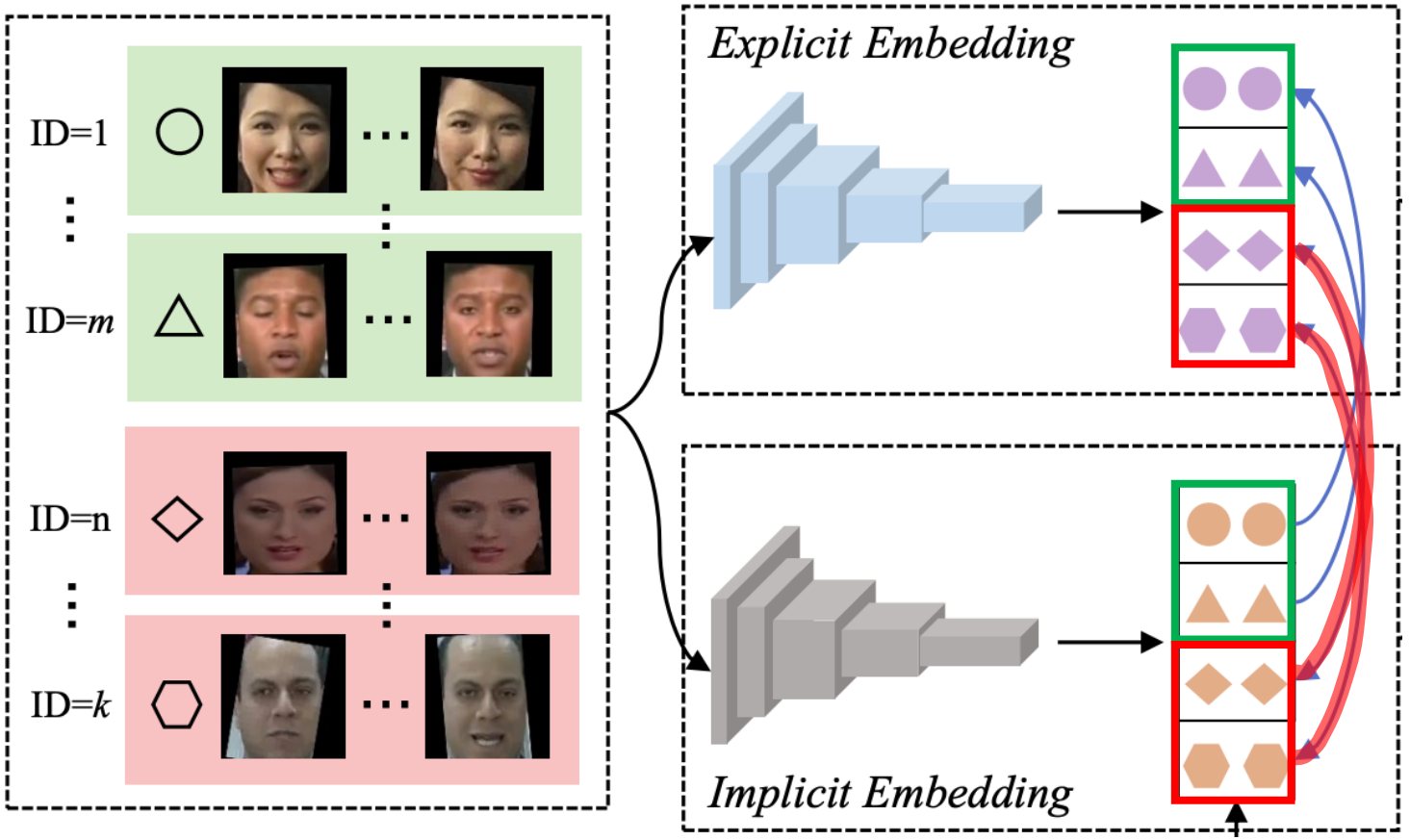
Introduction



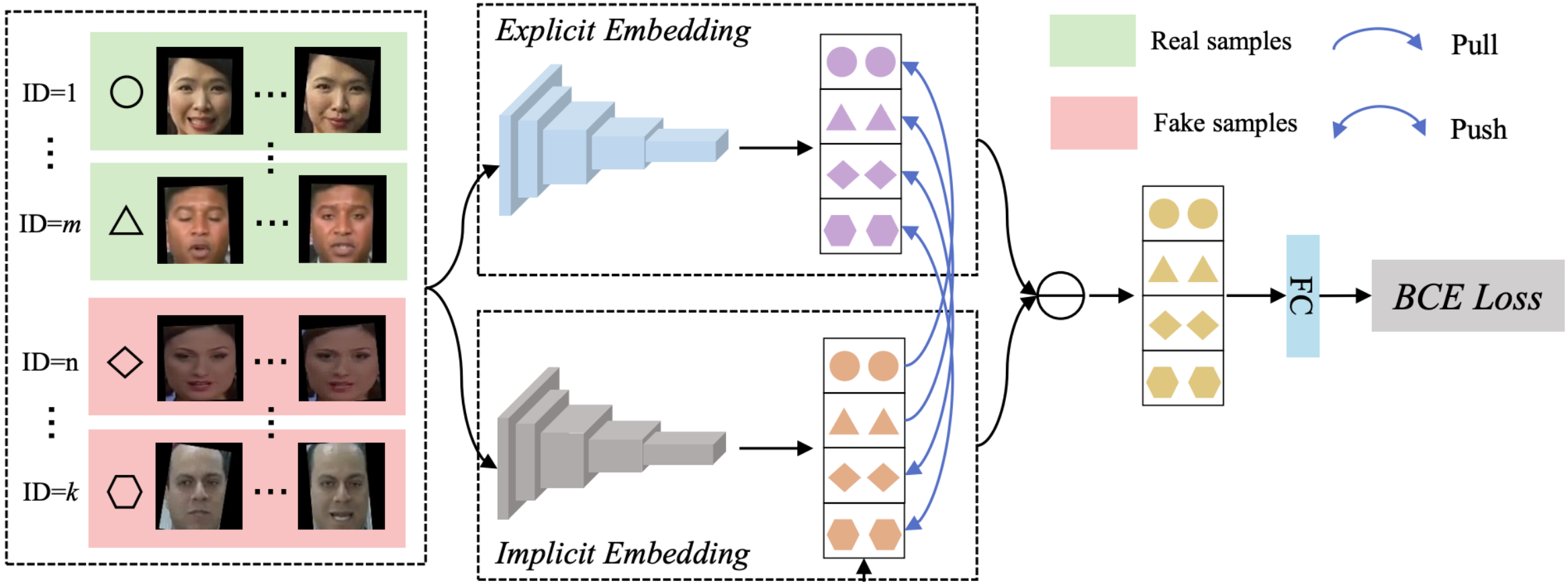
Introduction



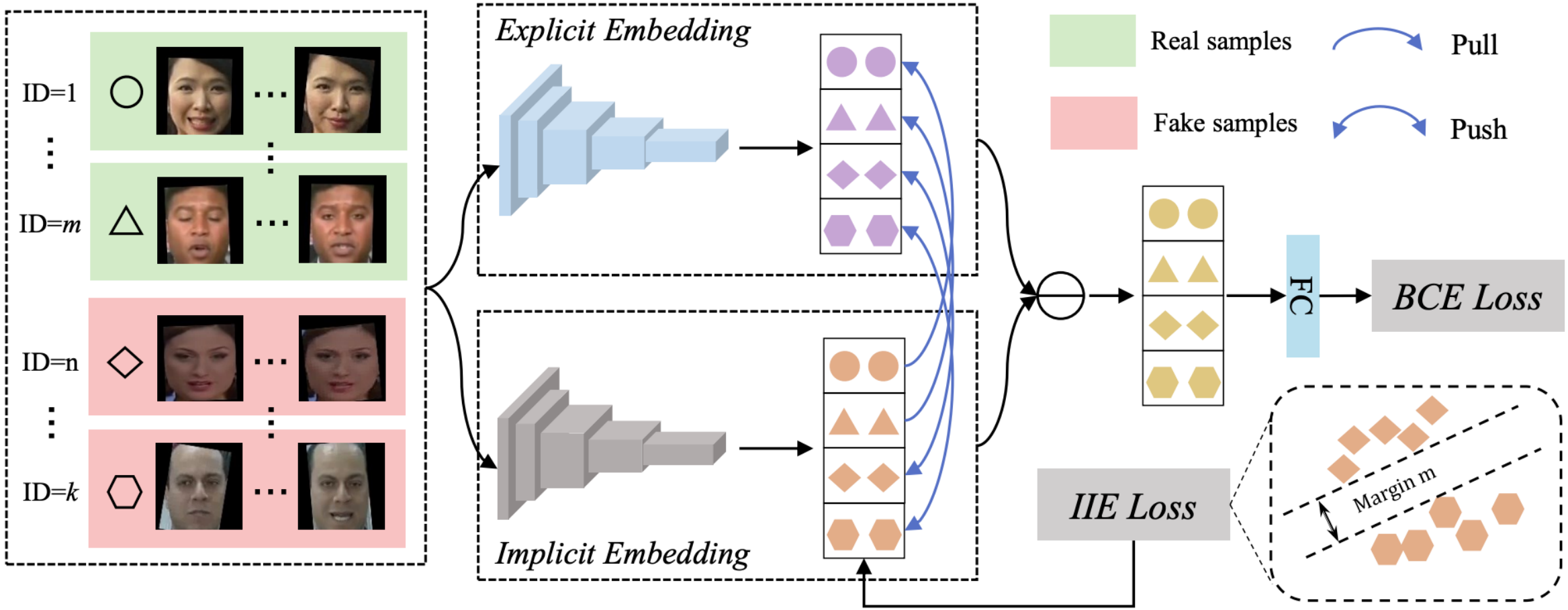
Introduction



Introduction



Introduction



Loss Function

$$\mathcal{L}_{iie} = \mathcal{L}^{+}_{iie} + \mathcal{L}^{-}_{iie}.$$

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda_1 \mathcal{L}_{eic} + \lambda_2 \mathcal{L}_{iie},$$

Explicit Identity Contrast

$$\mathcal{L}_{\text{eic}} = \frac{1}{N_F}\sum_{i\in F}\delta\left(F_{im}\left(x_i\right), F_{em}\left(x_i\right)\right) - \frac{1}{N_R}\sum_{i\in R}\delta\left(F_{im}\left(x_i\right), F_{em}\left(x_i\right)\right),$$

- x_i: face image
- δ: cosine similarity
- F_{im}: implicit identity embedding network
- F_{em}: generic explicit face recognition network
- R: a set of real samples
- F: a set of fake samples
- N_R: the number of R
- N_F : the number of F

Implicit Identity Exploration

w/ known implicit identity

$$\mathcal{L}_{iie}^{+} = -\mathbb{E}_{x_i, y_i \sim \mathcal{K}} \left[\log \frac{e^{s(\cos(\theta y_i) - m)}}{e^{s(\cos(\theta_{y_i}) - m)} + \sum_{j \neq y_i} e^{s \cos \theta_j}} \right]$$

- $\mathcal{K}$: real & fake samples with known implicit identities
- x_i: face image
- y_i : implicit identity
- θ_j: angle between $F_{im}(x_i)$ and proxy of j-th identity
- s: feature rescale hyperparemeter
- m: margin hyperparameter

Implicit Identity Exploration

w/ unknown implicit identity

$$\mathcal{L}_{iie}^{-} = -\mathbb{E}_{x_i, y_i^* \sim \mathcal{U}} \left[\log \frac{e^{\left(v_{y_i^*}^T F_{im}(x_i)/\tau \right)}}{\sum_{j=1}^{Q} e^{\left(v_j^T F_{im}(x_i)/\tau \right)}} \right]$$

- $\mathcal{U}$: unknown fake samples
- x_i: face image
- $y_i{}^*$: unknown implicit identity
- $V \in \mathbb{R}^{D \times Q}$: lookup table
- τ: temperature

Loss Function

$$\mathcal{L}_{iie} = \mathcal{L}^{+}_{iie} + \mathcal{L}^{-}_{iie}.$$

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda_1 \mathcal{L}_{eic} + \lambda_2 \mathcal{L}_{iie},$$

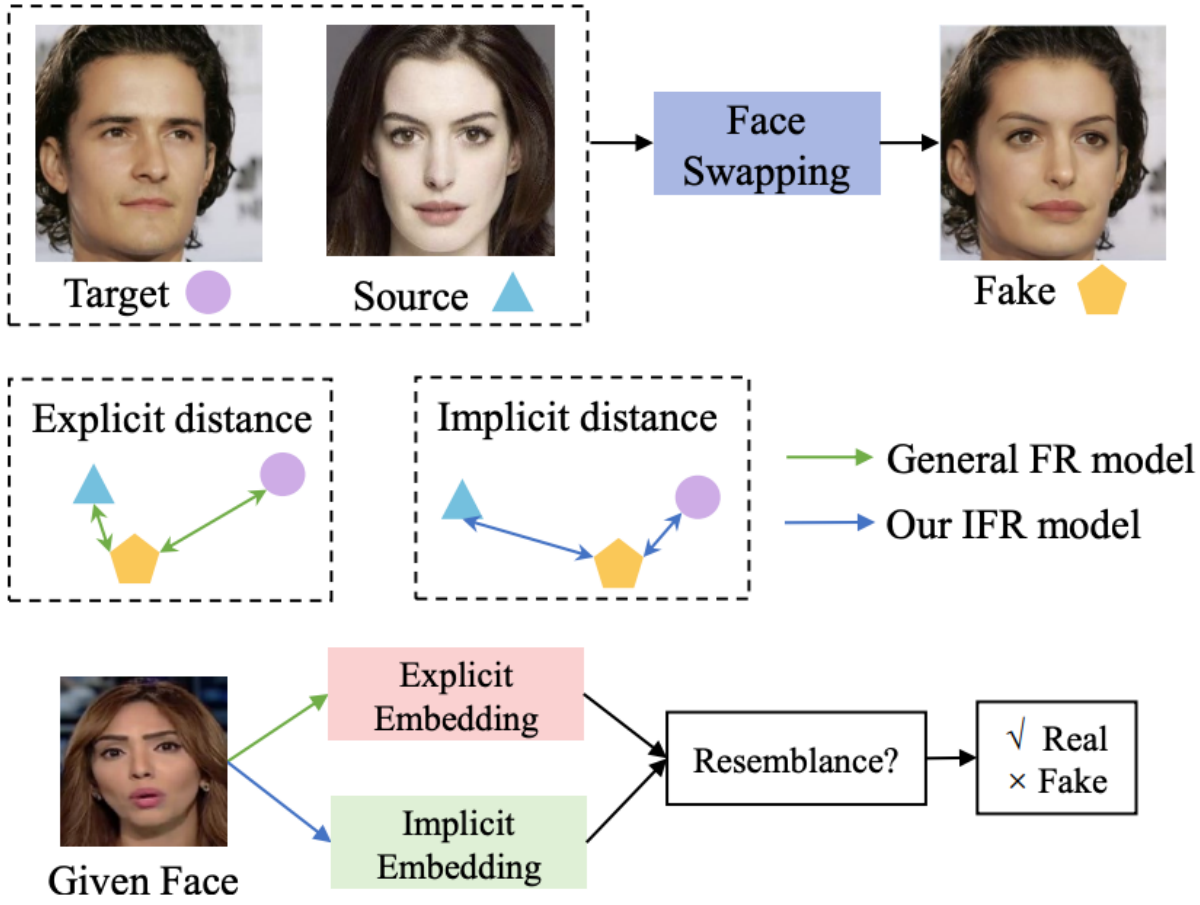
Qualitative Results

Method	FF++		Celeb-DF		DFD		DFDC	
	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)	AUC (%)	EER (%)
Xception [42]	99.09	3.77	65.27	38.77	87.86	21.04	69.90	35.41
EN-b4 [47]	99.22	3.36	68.52	35.61	87.37	21.99	70.12	34.54
Face X-ray [27]	87.40	-	74.20	-	85.60	-	70.00	-
MLDG [24]	98.99	3.46	74.56	30.81	88.14	21.34	71.86	34.44
F3-Net [52]	98.10	3.58	71.21	34.03	86.10	26.17	72.88	33.38
MAT(EN-b4) [53]	99.27	3.35	76.65	32.83	87.58	21.73	67.34	38.31
GFF [32]	98.36	3.85	75.31	32.48	85.51	25.64	71.58	34.77
LTW [45]	99.17	3.32	77.14	29.34	88.56	20.57	74.58	33.81
Local-relation [7]	**99.46**	3.01	78.26	29.67	89.24	20.32	76.53	32.41
DCL [46]	99.30	3.26	82.30	26.53	91.66	16.63	76.71	31.97
UIA-ViT [55]	99.33	-	82.41	-	**94.68**	-	75.80	-
Ours	99.32	**2.99**	**83.80**	**24.85**	93.92	**14.01**	**81.23**	**26.80**

Table 2. Cross-database evaluation from FF++(C23) to Celeb-DF, DFD, and DFDC in terms of AUC and EER. The FF++ belongs to the intra-testing results while others represent to the unseen dataset testing.

KAIST

Recap


Target
Source
Face Swapping
Fake
Explicit distance
Implicit distance
General FR model
Our IFR model
Given Face
Explicit Embedding
Implicit Embedding
Resemblance?
√ Real
× Fake

Recap

Strength

- Simple and effective idea
- Generalizability

Weakness

- Lookup table

Thank you

KAIST