Dual Contrastive Learning for General Face Forgery Detection

Ke Sun1 , Taiping Yao2 , Shen Chen2 , Shouhong Ding2* , Jilin Li 2 , Rongrong Ji1 *
1Media Analytics and Computing Lab, Department of Artificial Intelligence,
School of Informatics, Xiamen University, 361005, China
2Youtu Lab, Tencent, China
skjack@stu.xmu.edu.cn, rrji@xmu.edu.cn
{taipingyao, kobeschen, ericshding, jerolinli}@tecent.com

2022. 4. 24 • ChongQing

—— AAAI 2022

Reported by Sijin Liu

1.Introduction

2.Method

3.Experiments

Introduction

- Although the previous methods achieve promising results in intra-domain where the data distributions in training set and test set are the same, the performance drops significantly when facing the unseen domain scenario.

- these methods inherited from image classification models emphasize category-level differences rather than the essential discrepancies between real and fake images.

Method

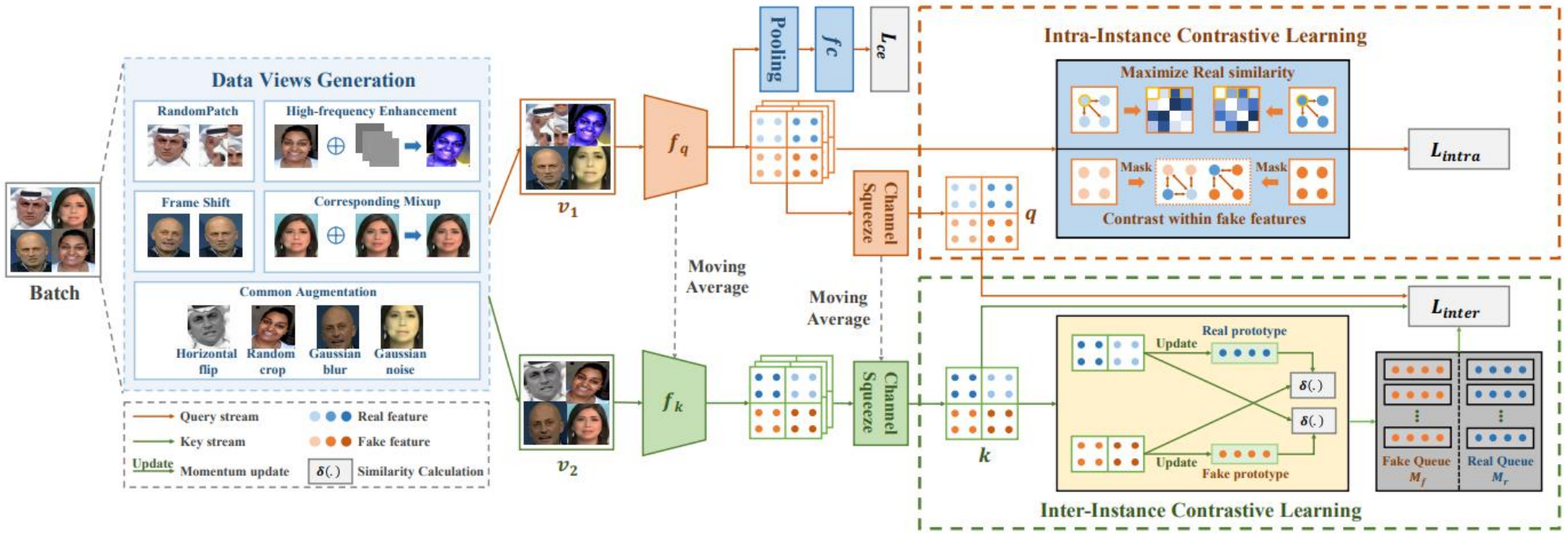
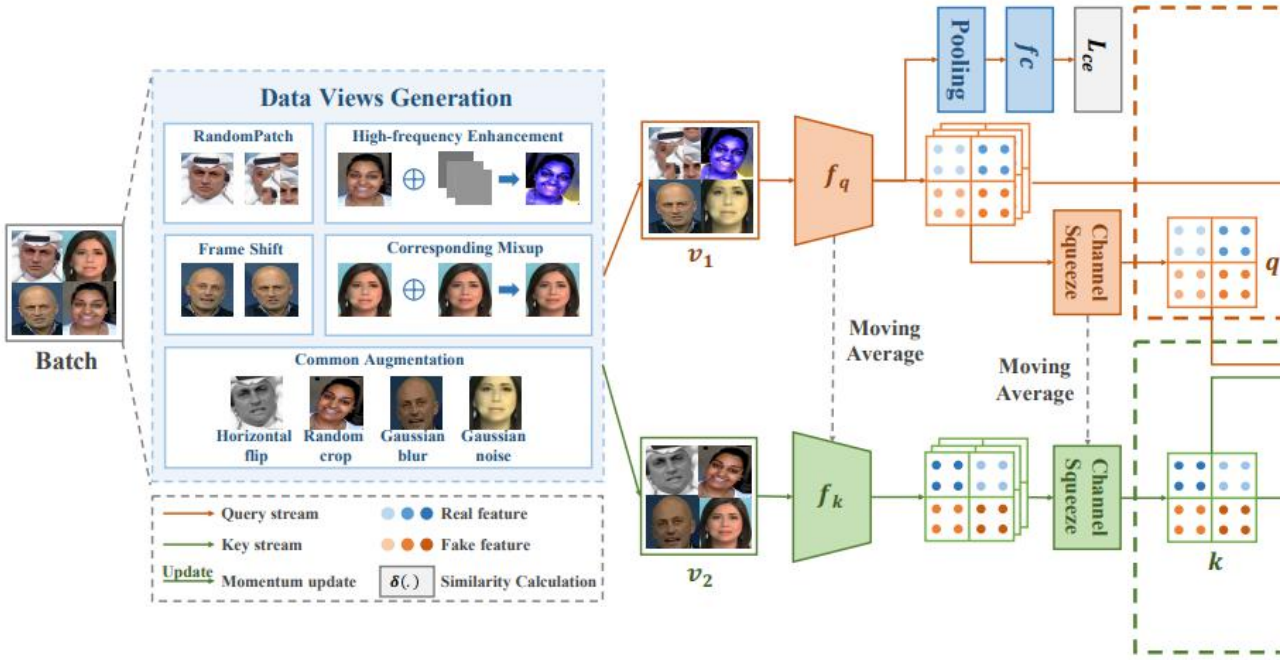


Figure 1: Overview of our proposed DCL framework. Given training images, we first transform them into two different views via the Data Views Generation module. Then the Intra-instance contrastive learning module and Inter-instance contrastive learning module are proposed to learn general features.

Method



Data Views Generation

1) RandomPatch

2) High-frequency enhancement

3) Frame shift

4) Corresponding mixup

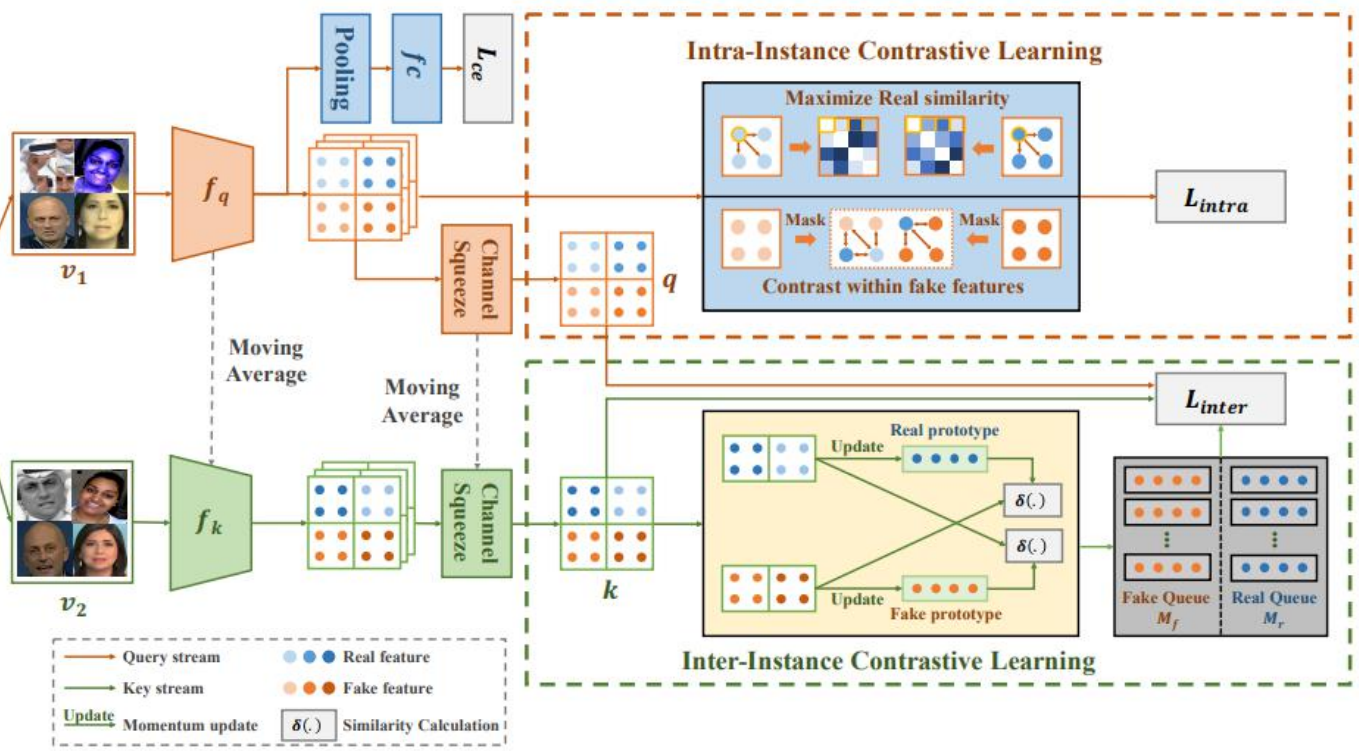
The input data $x_i \in R^{H\times W\times 3}$ with label $y_i \in \{0,1\}$ is firstly transformed into two different views $v_1(x_i)$ and $v_2(x_i)$ via data views generation module

$$f_q(v_1(x_i)) \in R^{C\times H'\times W'}$$
$$f_k(v_2(x_i)) \in R^{C\times H'\times W'}$$

the parameters of key encoder θ' is updated via exponential moving-average strategy from query encoder parameters θ:

$$\theta' = \beta\theta' + (1-\beta)\theta, \quad (1)$$
$$L_{ce} = y\log y' + (1-y)\log(1-y'), \quad (2)$$

Method



Inter-Instance Contrastive Learning

Specifically, we maintain two feature queues: real queue M_r and fake queue M_f to construct the negative sample of the corresponding query.

Our key idea can be derived as: *the more real the fake face is, the more it can be defined as a difficult sample.* Specifically, as shown in yellow dotted frame, we defined two prototypes P_{real} and P_{fake} for real and fake features respectively and updated using EMA scheme defined as:

$$P_{fake} = \alpha P_{fake} + (1-\alpha)k_{fake}, \tag{4}$$

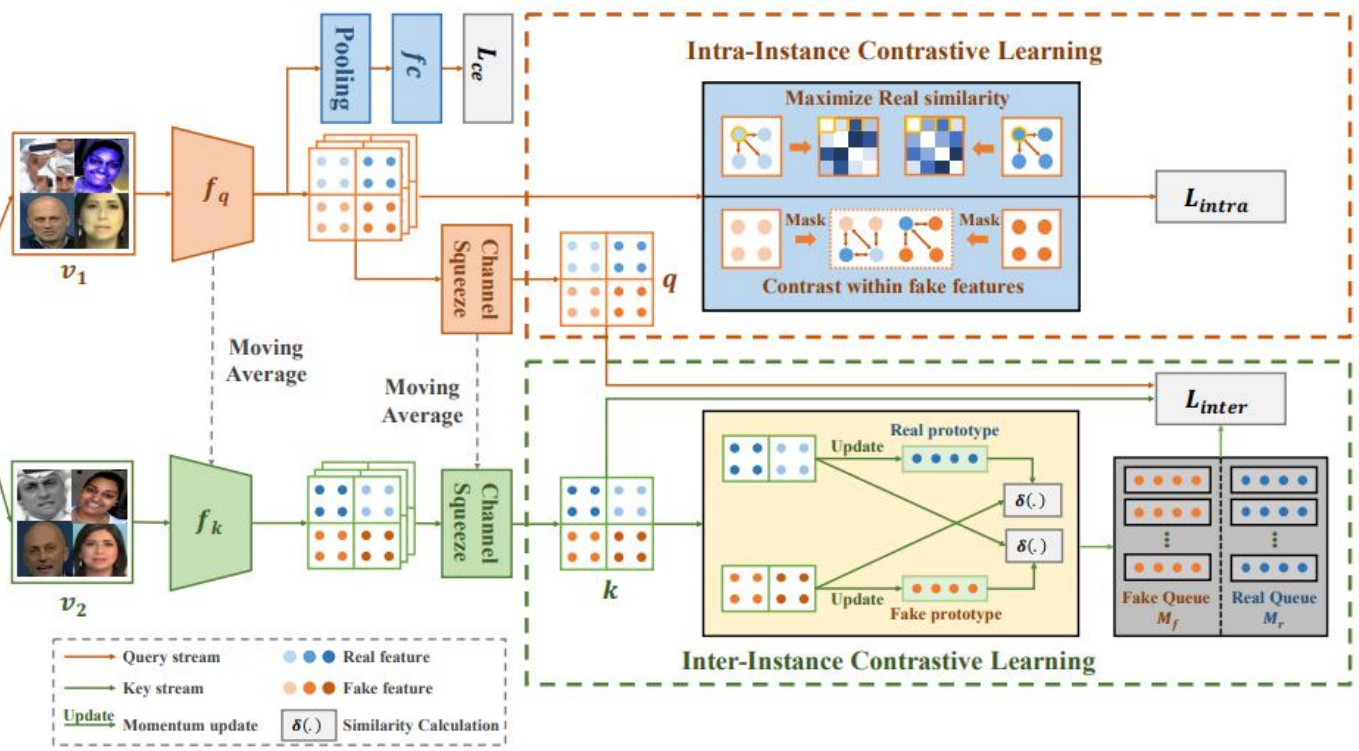
$$P_{real} = \alpha P_{real} + (1-\alpha)k_{real}. \tag{5}$$

$$\begin{cases} \delta(k_{fake}, P_{real}) > \theta, & M_f \leftarrow k_{fake} \\ \delta(k_{real}, P_{fake}) > \theta, & M_r \leftarrow k_{real}, \end{cases} \tag{6}$$

$$L_{inter} = -\log \frac{e^{\delta(q,k)/\tau}}{e^{\delta(q,k)/\tau} + \sum_{k_m \in K_q^-} e^{\delta(q,k_m)/\tau}}, \tag{3}$$

$$\delta(u,v) = \frac{u}{||u||} \cdot \frac{v}{||v||}$$

Method



$$L_{all} = \phi(L_{inter} + L_{intra}) + (1-\phi)L_{ce}, \qquad (10)$$

Intra-Instance Contrastive Learning

Specifically, given forgery image x_{fi}, we first generate the pixel-level mask $m_i \in R^{H\times W}$ by subtracting its corresponding real image $x_{ri}^{'}$: $m_i = |x_i - x_i^{'}|$. Then we resize the m_i into the same spatial size as the feature map $f_q(v_1(x_{fi}))$ denoted as $m_i^{'} \in R^{H^{'}\times W^{'}}$. Subsequently, we segment the $f_q(v_1(x_{fi}))$ into real parts $P_r \in \{p_{r1}, p_{r2}, ...p_{rn}\}$ and forgery parts $P_f \in \{p_{f1}, p_{f2}, ..., p_{fk}\}$ using $m_i^{'}$, where $p_f, p_r \in R^C$ and n, k denote the number of real and fake parts thus $n + k = H^{'}W^{'}$. Then the intra-instance contrastive loss for forgery features L_{intra}^f is calculated based upon InfoNCE as follows:

$$L_{intra}^f = -\log \frac{\sum_{i,j=1}^{n} e^{\delta(p_{ri}, p_{rj})/\tau}}{\sum_{i,j}^{n} e^{\delta(p_{ri}, p_{rj})/\tau} + \sum_{i=1}^{n}\sum_{j=1}^{k} e^{\delta(p_{fi}, p_{rj})/\tau}}, \qquad (7)$$

For real image x_{ri}, since all the features belongs to real, we expect for the self-similarity of $f_q(v_1(x_{ri}))$ become homogeneous. Thus, the intra-instance contrastive loss for real features can be obtained by:

$$L_{intra}^r = -\log sum(e^{f_q(v_1(x_{ri}))\odot f_q(v_1(x_{ri}))^{'T}/\tau}), \qquad (8)$$

$$L_{intra} = L_{intra}^r + L_{intra}^f, \qquad (9)$$

Experiments

Method	*FF++*		DFD		DFDC		Wild Deepfake		Celeb-DF	
	AUC	EER	AUC	EER	AUC	EER	AUC	EER	AUC	EER
Xception	99.09	3.77	87.86	21.04	69.80	35.41	66.17	40.14	65.27	38.77
EN-b4	99.22	3.36	87.37	21.99	70.12	34.54	61.04	45.34	68.52	35.61
Face X-ray	87.40	-	85.60	-	70.00	-	-	-	74.20	-
MLDG	98.99	3.46	88.14	21.34	71.86	34.44	64.12	43.27	74.56	30.81
F3-Net	98.10	3.58	86.10	26.17	72.88	33.38	67.71	40.17	71.21	34.03
MAT(EN-b4)	99.27	3.35	87.58	21.73	67.34	38.31	70.15	36.53	76.65	32.83
GFF	98.36	3.85	85.51	25.64	71.58	34.77	66.51	41.52	75.31	32.48
LTW	99.17	3.32	88.56	20.57	74.58	33.81	67.12	39.22	77.14	29.34
Local-relation	**99.46**	**3.01**	89.24	20.32	76.53	32.41	68.76	37.50	78.26	29.67
Ours	99.30	3.26	**91.66**	**16.63**	**76.71**	**31.97**	**71.14**	**36.17**	**82.30**	**26.53**

Table 1: Cross-database evaluation from FF++(HQ) to DFD, DFDC, Wild Deepfake and Celeb-DF in terms of AUC and EER. The FF++ belongs to the intra-domain results while others represent to the unseen-domain.

Experiments

Method	FF++	Celeb-DF
Meso-4	84.70	54.80
Mesoinception4	83.00	53.60
FWA	80.10	56.90
Xception	95.50	65.50
Multi-task	76.30	54.30
SMIL	96.80	56.30
Two Branch	93.18	73.41
EN-b4	96.39	71.10
Multi-Attention	96.41	72.50
GFF	95.73	74.12
SPSL	96.91	76.88
Ours	**96.97**	**81.00**

Table 2: Cross-dataset evaluation from FF++(LQ) to deepfake class of FF++ and Celeb-DF in terms of AUC.

Train	Method	DF	F2F	FS	NT
DF	EN-b4	*99.97*	76.32	46.24	72.72
	MAT	*99.92*	75.23	40.61	71.08
	GFF	*99.87*	76.89	47.21	72.88
	Ours	***99.98***	**77.13**	**61.01**	**75.01**
F2F	EN-b4	84.52	*99.20*	58.14	63.71
	MAT	86.15	*99.13*	60.14	64.59
	GFF	89.23	*99.10*	**61.30**	64.77
	Ours	**91.91**	***99.21***	59.58	**66.67**
FS	EN-b4	69.25	67.69	*99.89*	48.61
	MAT	64.13	66.39	*99.67*	50.10
	GFF	70.21	68.72	*99.85*	49.91
	Ours	**74.80**	**69.75**	***99.90***	**52.60**
NT	EN-b4	85.99	48.86	73.05	*98.25*
	MAT	87.23	48.22	75.33	*98.66*
	GFF	88.49	49.81	74.31	*98.77*
	Ours	**91.23**	**52.13**	**79.31**	***98.97***

Table 3: Cross-manipulation evaluation in terms of AUC. Diagonal results indicate the intra-domain performance.

Experiments

Method	GID-DF (HQ)		GID-DF (LQ)		GID-F2F (HQ)		GID-F2F (LQ)	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
EfficientNet	82.40	91.11	67.60	75.30	63.32	80.1	61.41	67.40
Focalloss	81.33	90.31	67.47	74.95	60.80	79.80	61.00	67.21
ForensicTransfer	72.01	-	68.20	-	64.50	-	55.00	-
Multi-task	70.30	-	66.76	-	58.74	-	56.50	-
MLDG	84.21	91.82	67.15	73.12	63.46	77.10	58.12	61.70
LTW	85.60	92.70	69.15	75.60	65.60	80.20	65.70	72.40
Ours	**87.70**	**94.9**	**75.90**	**83.82**	**68.40**	**82.93**	**67.85**	**75.07**

Table 4: Performance on multi-source manipulation evaluation, the protocols and results are from (Sun et al. 2021). GID-DF means traning on the other three manipulated methods of FF++ and test on deepfakes class. The same for the others.

Experiments

Inter	Views	Hard	Intra	Celeb-DF	DFD
	✓			74.12	88.32
✓				76.81	88.03
✓	✓			79.34	89.24
✓		✓		78.84	89.89
✓	✓	✓		80.30	90.12
✓	✓	✓	✓	**82.30**	**91.66**

Table 5: Ablation study on the influence of different components. Specifically, "Inter" means inter-instance contrastive learning module, "views" represents our special designed data views generation strategy, "hard" indicate the hard sample generation, and "Intra" is short for the Intra-instance contrastive learning module.

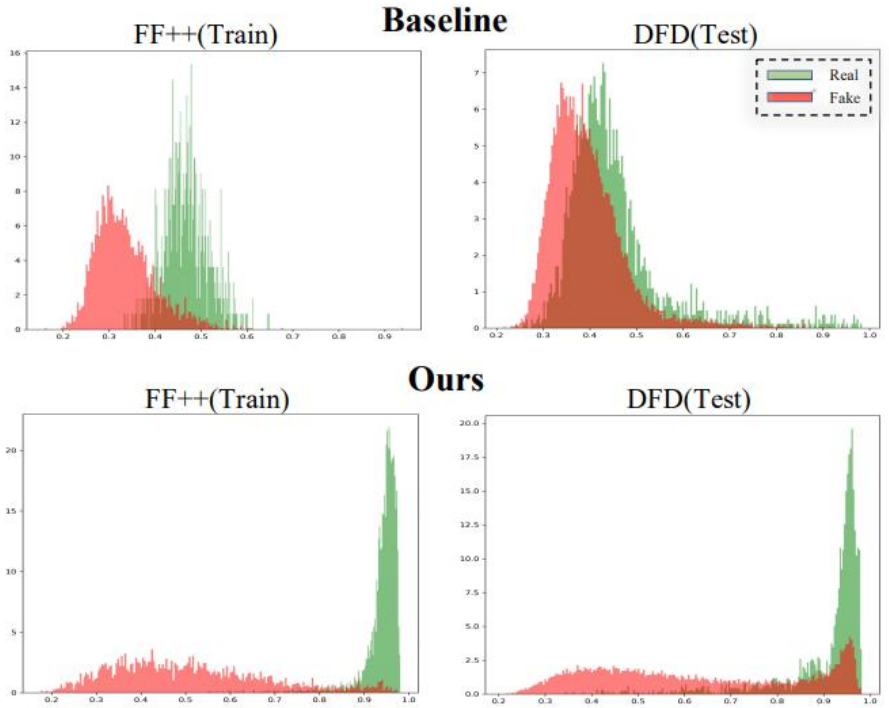
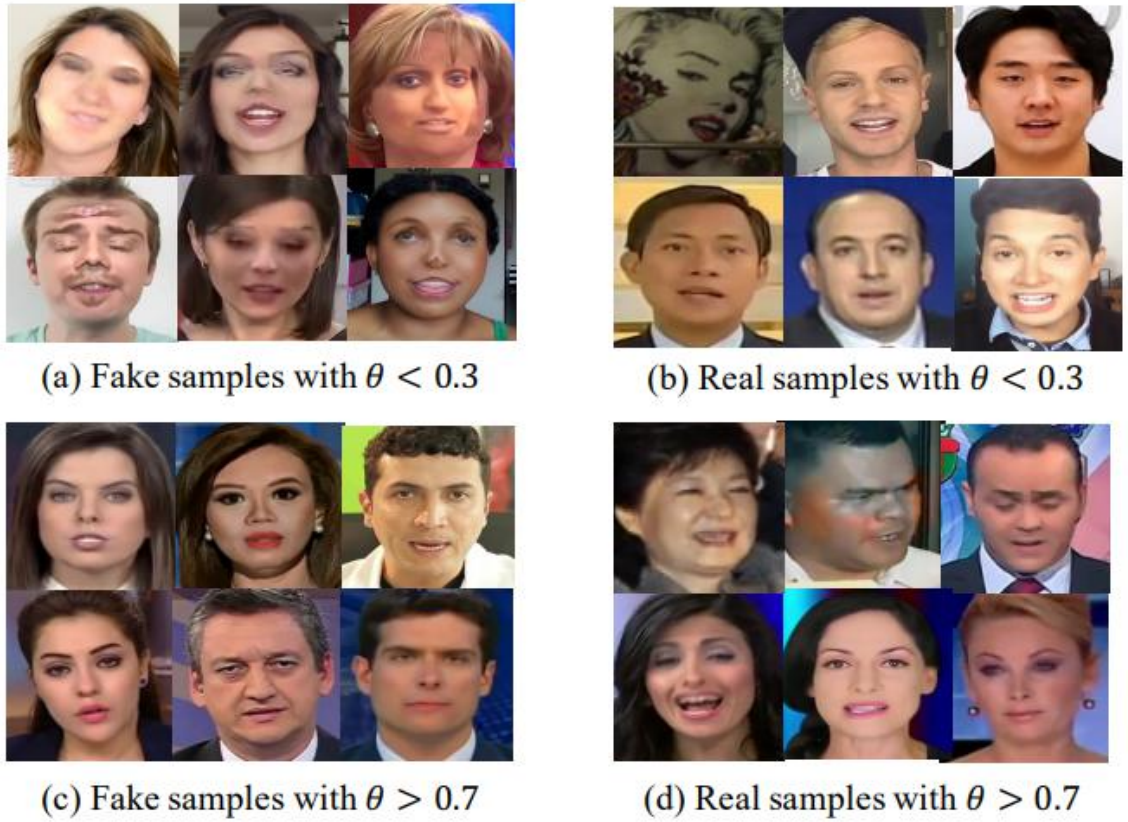


Figure 2: Histogram of the average of self-similarity for intra-domain dataset (FF++) and unseen domain (DFD). The first row indicates the histogram of the baseline model (En-b4) while the second row represents that of our DCL.

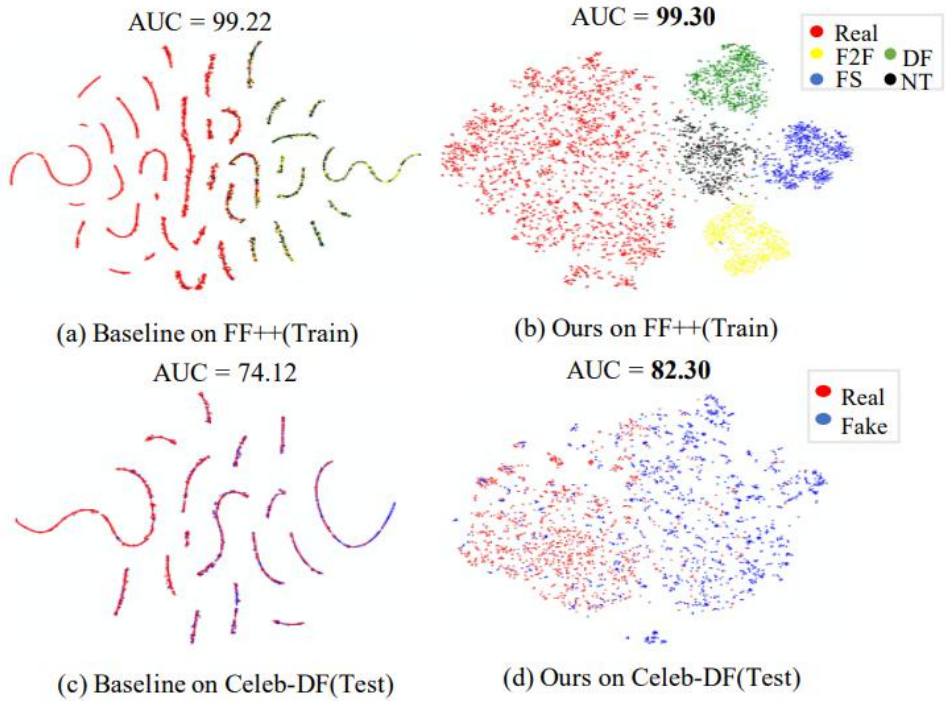
Experiments



(a) Fake samples with $\theta < 0.3$ (b) Real samples with $\theta < 0.3$

(c) Fake samples with $\theta > 0.7$ (d) Real samples with $\theta > 0.7$

Figure 3: Visualization of the hard sample strategy with low and high threshold.



(a) Baseline on FF++(Train) (b) Ours on FF++(Train)

(c) Baseline on Celeb-DF(Test) (d) Ours on Celeb-DF(Test)

Figure 4: Feature distribution of baseline model (En-b4) and DCL on the intra-domain dataset (FF++) and unseen domain dataset (Celeb-DF) via t-SNE.