DF40: Toward Next-Generation Deepfake Detection

Zhiyuan Yan[1], Taiping Yao[2†], Shen Chen[2], Yandan Zhao[2], Xinghe Fu[2], Junwei Zhu[2], Donghao Luo[2], Chengjie Wang[2], Shouhong Ding[2], Yunsheng Wu[2], Li Yuan[1†]

NEURAL INFORMATION PROCESSING SYSTEMS

Dataset, Benchmark, codebase, pre-training weights are released at :

https://github.com/YZY-stack/DF40

Motivation

Existing Problems:

We found the **dataset** (both train and test) can be the "primary culprit" of the limited detection model's generalization performance in the real world.

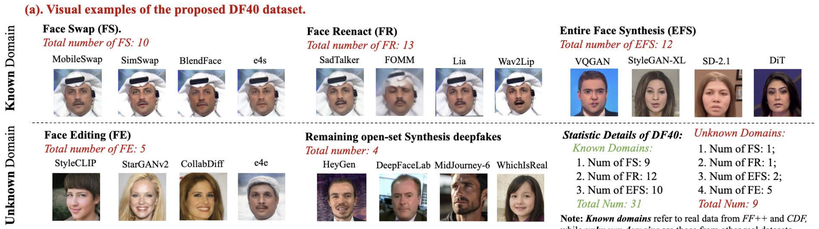
- **(1) forgery diversity:** Deepfake techniques are commonly referred to as both face forgery and entire face synthesis. Most existing datasets only contain partial types of them;
- **(2) forgery realism:** The dominated training dataset, FF++, contains out-of-date forgery techniques from the past four years, making it difficult to guarantee effective detection toward nowadays' SoTA deepfakes;
- **(3) evaluation protocol:** Most detection works perform evaluations on one type, e.g., training and testing on face-swapping types only, which hinders the development of universal deepfake detectors.

Contribution

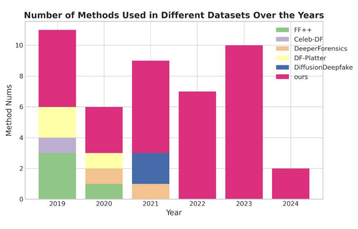
1. We propose a new **highly diverse dataset called DF40** for generalizable deepfake detection, with **40 distinct deepfake techniques** involved.
2. We propose a **comprehensive benchmark** for thorough evaluations and in-depth analysis, leading to 7 new insightful findings and 4 new open questions to the field.

Features of our DF40 dataset

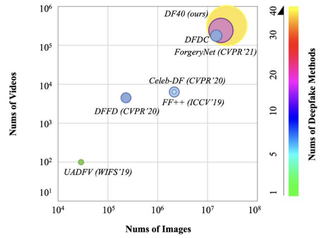
(a). Visual examples of the proposed DF40 dataset.



(b). Statistic comparison of forgery number used in different datasets



(c). Comparison with other dataset in *diversity* and scale.



Dataset	Publication	Latest Fake	Methods	FS	FR	EFS	FE	Fake Videos	Fake Images	Pretraining
DF-TIMIT [37]	ArXiv'18	faceswap-GAN [64] (2018)	2	2	-	-	-	640	-	-
UADFV [91]	ICASSP'19	Unknown	1	1	-	-	-	49	252	-
FaceForensics++ [62]	ICCV'19	NeuralTextures [71]	4	2	2	-	-	4K	-	-
DeepFakeDetection [17]	None	Unknown	5	5	-	-	-	3068	-	-
CDF [42]	CVPR'20	Unknown	1	1	-	-	-	5,639	-	-
DFFD [13]	CVPR'20	StyleGAN [34] (2018)	7	7	-	-	-	3000	0.2M+	-
DeeperForensics-1.0 [30]	CVPR'20	DF-VAE [30] (2020)	1	1	-	-	-	10K	-	-
DFDC [18]	ArXiv'20	StyleGAN [34] (2018)	8	5	1	2	-	0.1M+	-	-
ForgeryNet [22]	CVPR'21	StarGANv2 [11] (2020)	15	6	4	2	3	0.1M+	1M+	✓
FakeAVCeleb [36]	NeurIPS'21	Wav2Lip [55] (2021)	4	2	2	-	-	9.5K	-	-
KoDF [38]	ICCV'21	Wav2Lip [55] (2021)	6	3	3	-	-	0.1M+	-	-
FFIW [94]	CVPR'21	FSGAN [52] (2019)	3	3	-	-	-	10K	-	-
DF3 [32]	TMM'22	StyleGAN3 [33] (2021)	6	-	-	6	-	15k+	-	-
DeepFakeFace [69]	ArXiv'23	Stable-Diffusion [78] (2021)	3	1	-	2	-	-	90K	-
DF-Platter [51]	CVPR'23	FaceShifter [39] (2020)	3	3	-	-	-	0.1M+	-	-
DiffusionDeepfake [4]	ArXiv'24	Stable-Diffusion [78] (2021)	2	-	-	2	-	-	0.1M+	-
DF40 (Ours)	-	PixArt-α [8] (2024)	**40**	**10**	**13**	**12**	**5**	**0.1M+**	**1M+**	✓

Main Evaluation of four standard Protocols

Protocol-1: Cross method evaluation

Training Set	Model	Testing Set (FF) FS (FF)	FR (FF)	EFS (FF)	Avg. (FF)
FS (FF)	Xception [12]	0.991	0.892	0.810	0.898
	CLIP [57]	0.996	0.908	0.837	0.914
	SRM [48]	0.988	0.867	0.703	0.853
	SPSL [44]	0.987	0.849	0.735	0.857
	RECCE [6]	0.991	0.855	0.758	0.868
	RFM [75]	0.992	0.884	0.821	0.899
FR (FF)	Xception [12]	0.838	0.996	0.670	0.835
	CLIP [57]	0.932	0.999	0.798	0.910
	SRM [48]	0.893	0.998	0.698	0.863
	SPSL [44]	0.901	0.998	0.695	0.865
	RECCE [6]	0.865	0.997	0.716	0.859
	RFM [75]	0.892	0.999	0.776	0.889
EFS (FF)	Xception [12]	0.665	0.807	0.999	0.824
	CLIP [57]	0.688	0.889	0.999	0.859
	SRM [48]	0.596	0.776	0.999	0.790
	SPSL [44]	0.659	0.811	0.999	0.823
	RECCE [6]	0.691	0.801	0.999	0.830
	RFM [75]	0.653	0.795	0.999	0.816
BI (FF)	SBI [65]	0.810	0.714	0.678	0.734

Protocol-2: Cross domain evaluation

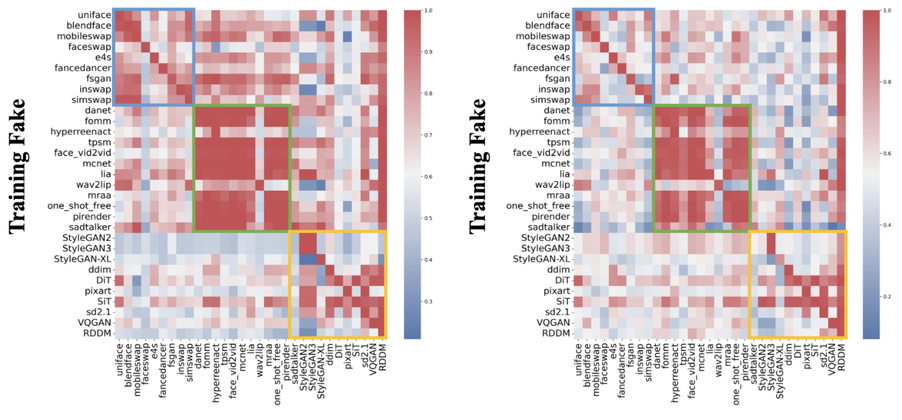
Training Set	Model	Testing Set (CDF) FS (CDF)	FR (CDF)	EFS (CDF)	Avg. (CDF)
FS (FF)	Xception [12]	0.922	0.657	0.642	0.740
	CLIP [57]	0.967	0.744	0.730	0.814
	SRM [48]	0.919	0.621	0.603	0.714
	SPSL [44]	0.938	0.656	0.648	0.747
	RECCE [6]	0.926	0.632	0.610	0.723
	RFM [75]	0.939	0.637	0.628	0.735
FR (FF)	Xception [12]	0.481	0.857	0.369	0.569
	CLIP [57]	0.638	0.933	0.209	0.593
	SRM [48]	0.454	0.869	0.326	0.550
	SPSL [44]	0.479	0.852	0.256	0.529
	RECCE [6]	0.452	0.881	0.332	0.555
	RFM [75]	0.492	0.882	0.359	0.578
EFS (FF)	Xception [12]	0.586	0.594	0.983	0.721
	CLIP [57]	0.617	0.735	0.988	0.780
	SRM [48]	0.589	0.620	0.964	0.724
	SPSL [44]	0.635	0.651	0.975	0.754
	RECCE [6]	0.623	0.603	0.984	0.737
	RFM [75]	0.644	0.666	0.981	0.764
BI (FF)	SBI [65]	0.679	0.609	0.723	0.670

Protocol-3: Cross method and Cross data evaluation

Training Set	Model	Testing Set DeepFaceLab (●)	HeyGen (●)	MidJourney-6 (●)	Whichisreal (●)	StarGAN (●)	StarGAN2 (●)	StyleCLIP (●)	e4c (●)	CollabDiff (●)	Avg.
FS (FF)	Xception [12]	0.882	0.394	0.384	0.535	0.577	0.616	0.426	0.553	0.546	0.546
	CLIP [57]	0.930	0.539	0.540	0.439	0.896	0.746	0.730	0.738	0.674	0.692
	SRM [48]	0.866	0.473	0.298	0.538	0.606	0.617	0.572	0.410	0.699	0.564
	SPSL [44]	0.930	0.370	0.414	0.557	0.559	0.590	0.536	0.574	0.584	0.565
	RECCE [6]	0.899	0.537	0.293	0.509	0.580	0.599	0.399	0.520	0.492	0.536
	RFM [75]	0.918	0.719	0.286	0.496	0.652	0.570	0.705	0.689	0.798	0.648
FR (FF)	Xception [12]	0.705	0.473	0.459	0.323	0.492	0.456	0.006	0.175	0.050	0.349
	CLIP [57]	0.845	0.614	0.632	0.466	0.762	0.436	0.298	0.631	0.611	0.588
	SRM [48]	0.786	0.604	0.510	0.357	0.473	0.434	0.044	0.428	0.080	0.413
	SPSL [44]	0.704	0.543	0.446	0.272	0.348	0.423	0.002	0.585	0.060	0.376
	RECCE [6]	0.724	0.576	0.314	0.278	0.529	0.374	0.005	0.177	0.060	0.337
	RFM [75]	0.739	0.588	0.511	0.325	0.407	0.423	0.009	0.201	0.030	0.360
EFS (FF)	Xception [12]	0.497	0.325	0.472	0.772	0.777	0.677	0.984	0.611	0.997	0.679
	CLIP [57]	0.745	0.506	0.534	0.828	0.946	0.823	0.929	0.923	0.983	0.802
	SRM [48]	0.527	0.358	0.338	0.794	0.769	0.703	0.982	0.509	0.997	0.664
	SPSL [44]	0.641	0.383	0.427	0.694	0.699	0.723	0.922	0.602	0.967	0.673
	RECCE [6]	0.583	0.505	0.442	0.753	0.769	0.724	0.964	0.643	0.979	0.707
	RFM [75]	0.619	0.349	0.551	0.623	0.730	0.636	0.966	0.665	0.979	0.680
BI (FF)	SBI [65]	0.764	0.402	0.342	0.426	0.591	0.586	0.564	0.379	0.570	0.514

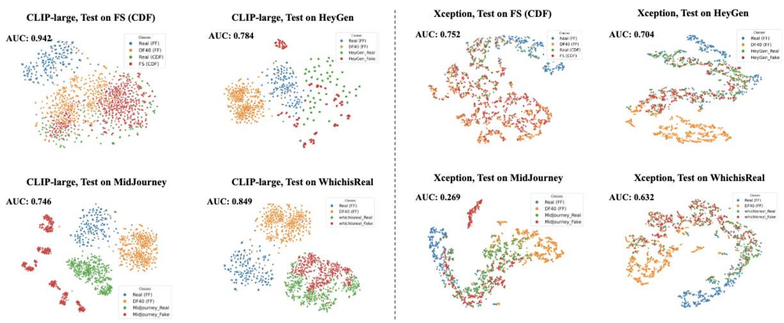
Protocol-4: One-Verse-All (OvA) evaluation

***Same Data Domain* (H_1)** **Cross Data Domain (H_2)**



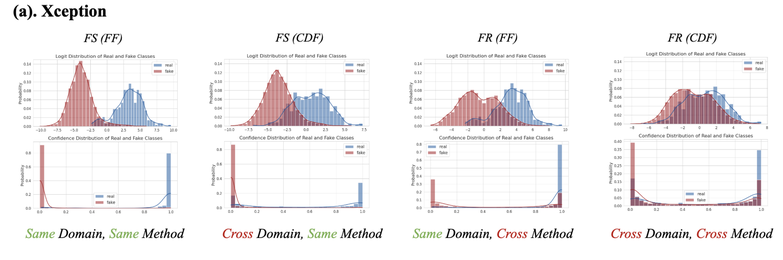
Same Domain, Same Method *Cross Domain, Same Method* *Same Domain, Cross Method* *Cross Domain, Cross Method*

t-SNE analysis for CLIP and Xception



Logits and confidence analysis

(a). Xception



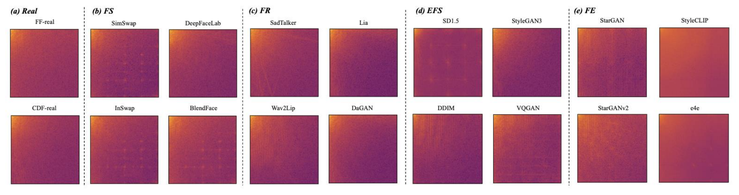
Generalizing non-face deepfakes

Training Set	Model	Testing Set ADM	BigGAN	GLide	MidJourney	SD-v4	SD-v5	Vqdm	Wukong	Avg.
DF40 (FF)	Xception	0.723	0.529	0.514	0.558	0.490	0.494	0.469	0.505	0.535
	CLIP-base	0.940	0.850	0.666	0.447	0.494	0.494	0.682	0.542	0.639
	CLIP-large	0.911	0.967	0.736	0.571	0.630	0.614	0.882	0.660	0.746

Is super-resolution deepfake?

Test / Train	FS FSGAN	FS BlendFace	FR LIA	FR Wav2Lip	EFS DiT	EFS DDIM	FE e4e
SRI	0.772	0.835	0.746	0.564	0.687	0.713	0.543
SRI + Super-Resolution	0.983	0.825	0.988	0.833	0.997	0.946	0.978

Frequency analysis



Evaluation of existing detectors on previous datasets and our DF40.

***CFDA (ECCV'24)* achieves significant advantages over other models in previous datasets, but perform comparable or even worser than other models.**

Email and contact us

- yanzhiyuan1114@gmail.com
- zhiyuanyan@stu.pku.edu.cn