Open Set Face Forgery Detection via Dual-Level Evidence Collection

Zhongyi Cai[1], Bryce Gernon[2], Wentao Bao[1], Yifan Li[1], Matthew Wright[2], Yu Kong[1]

[1]Michigan State University, East Lansing, USA [2]Rochester Institute of Technology, Rochester, USA

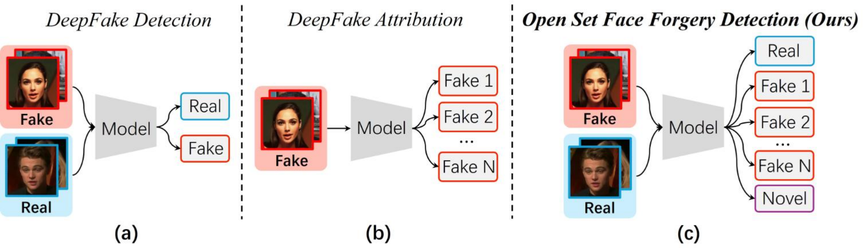
Homepage Linkedin

Background

·Deepfakes, which use deep learning techniques to generate or modify faces, continue to rapidly increase.

·DeepFake Detection conducts binary image classification to distinguish whether the human face is real or faked.

·DeepFake Attribution conducts multi-class image classification to identify the source of fake faces.



Open Set Face Forgery Detection

·Illustration for Fake Categories in OSFFD

-Real faces and fake faces from the seen categories are used to train the model.

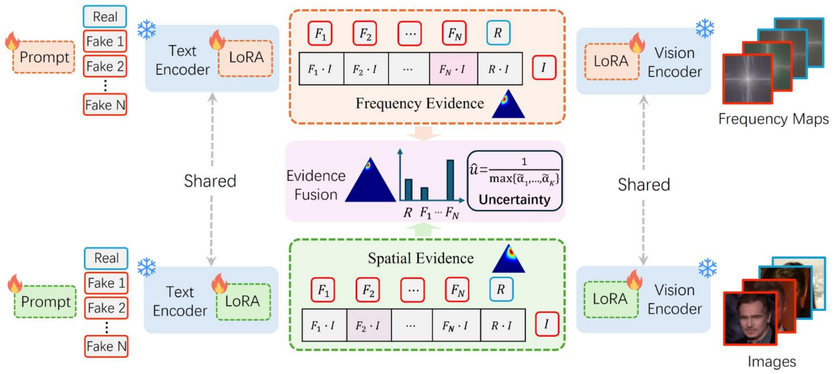
-The model is evaluated on test data that includes both seen classes and previously unseen categories.

·Two Main Challenges:

-Reliable detection on unseen deepfake categories (without the availability of training data from those categories)

-Robust real-vs-fake binary deepfake detection

Dual-Level Evidential Face Forgery Detection



·Dual-Level Evidential face forgery Detection (DLED) exploits Evidential Deep Learning (EDL) through a dual-level evidential architecture that captures category characteristics of facial imagery across the spatial and frequency domains, yielding sufficiently discriminative evidence.

·DLED addresses the evidence aggregation challenge with an uncertainty-guided fusion mechanism and further incorporates an uncertainty-improvement procedure to enhance the reliability of the resulting estimates.

Empirical Results

TABLE I: Comparisons of model performance with diverse baseline methods implemented by ourselves for the OSFFD problem. We use different data configurations for the seen and unseen fake categories. For "FS", "FR", and "EFS", we let each fake category be the unseen category and let the remaining two be seen categories. For "FE & SM", we take FS, FR, and EFS as seen categories and let FE and SM be the unseen categories. The best results are highlighted in **bold**.

Methods		FS		FR		EFS		FE & SM		Avg	
		Acc	DR	Acc	DR	Acc	DR	Acc	DR	Acc	DR
Two-stage	OC-FakeDect [23]	58.16	14.68	60.69	11.43	56.14	9.01	56.74	11.67	57.93	11.70
	SBI [48]	65.15	1.07	64.19	3.00	61.24	0.91	62.27	0.66	63.21	1.41
CNN-based + OSR	Xception [41]	64.60	23.90	53.51	29.06	57.62	22.70	55.28	29.04	57.75	26.17
	SPSL [30]	65.07	16.71	54.10	18.93	59.67	18.12	60.02	25.98	59.71	19.93
	SIA [50]	62.09	13.59	54.62	13.36	56.85	10.99	56.29	22.53	57.46	15.12
	UCF [64]	65.08	0.30	50.98	0.20	52.95	1.28	52.69	1.80	55.42	0.89
	NPR [52]	**75.37**	17.37	64.63	6.75	70.43	4.36	71.45	29.20	70.47	14.42
CLIP-based + OSR	CLIP Closed Set Finetuning	67.24	\	65.19	\	64.53	\	66.24	\	65.80	\
	CLIP Zero-Shot [40]	52.30	0.81	50.36	0.26	46.01	0.38	47.62	0.25	49.07	0.43
	UnivFD [37]	68.81	3.88	64.00	2.48	63.21	0.73	66.34	8.22	65.59	3.83
	CLIPing [24]	66.44	14.38	62.41	6.09	61.29	4.92	66.26	19.27	64.10	11.16
	D^3 [68]	70.46	8.14	64.71	8.90	61.65	1.17	66.33	8.26	65.79	6.62
	Ours	71.37	**33.61**	**66.83**	**34.92**	**75.52**	**34.71**	**74.48**	**82.18**	**72.05**	**46.35**

TABLE II: Comparisons of prediction accuracy with diverse baselines implemented by ourselves for the Real-vs-Fake detection task. Data configurations are the same as those in OSFFD. All baseline models are implemented following their original algorithms.

Methods	FS	FR	EFS	FE & SM	Avg
OC-FakeDect [23]	48.09	48.45	48.18	47.16	47.97
SBI [48]	50.13	50.36	50.07	49.96	50.13
Xception [41]	71.73	67.98	67.19	67.49	68.60
SPSL [30]	72.29	65.87	70.34	69.57	69.52
SIA [50]	69.45	64.13	66.91	64.64	66.28
UCF [64]	71.10	64.78	65.18	67.98	67.26
NPR [52]	80.76	75.73	77.67	77.21	77.84
CLIP Zero-Shot [40]	52.96	53.20	53.12	56.62	53.97
UnivFD [37]	77.64	76.83	79.33	81.31	78.78
CLIPing [24]	78.46	77.15	79.58	81.09	79.07
D^3 [68]	78.56	77.00	79.67	79.81	78.76
Ours	**87.22**	**85.93**	**83.52**	**84.97**	**85.41**

TABLE III: Ablation Study of DLED. The table presents DR results under the same data configuration as used in the main OSFFD experiments.

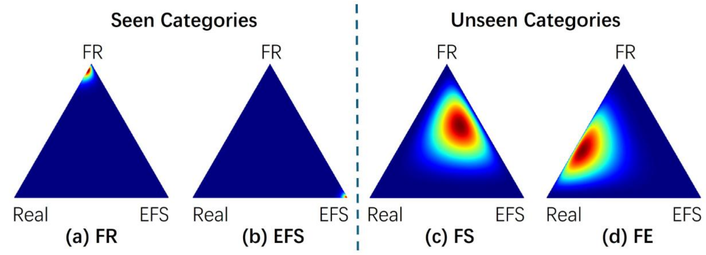
	Models	FS	FR	EFS	FE & SM	Avg
Spatial Branch	Zero-Shot with MaxLogit	0.81	0.26	0.38	0.25	0.42
	Zero-Shot with EDL	1.58	0.58	0.68	0.63	0.87
	Finetuning with EDL	13.02	30.94	8.33	50.59	25.71
Frequency Branch	Zero-Shot with MaxLogit	3.85	2.35	6.98	0.53	3.43
	Zero-Shot with EDL	4.71	2.51	6.06	0.55	3.46
	Finetuning with EDL	14.34	8.49	7.69	**90.36**	30.22
Two Branches	Evidence Fusion	32.42	**36.16**	32.56	79.74	45.22
	Full DLED	**33.61**	34.92	**34.71**	82.18	**46.36**

Visualization

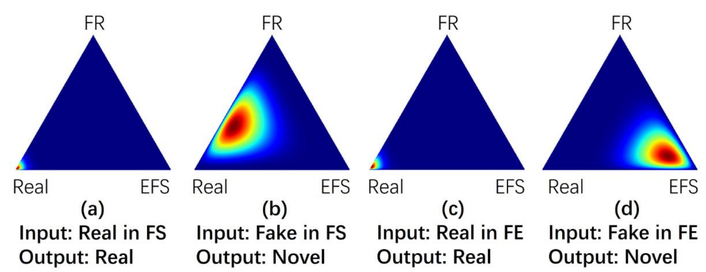
·Visualization of Attention Map

DLED attends to category-specific semantic cues. In the case of Face Swapping (a), the model highlights edge regions indicative of face transplantation, whereas for Face Editing (b), it focuses on manipulated areas such as sunglasses, hairbands, and hair.

·Visualization of Evidence Distribution



The evidence for seen fake categories FR and EFS is condensed in their corresponding corner with low uncertainty, while the evidence for novel fake categories FS and FE is sparse with higher uncertainty.



Visualization of the Evidence Distribution for novel real and fake faces. The prediction confidence for new realfaces remains high, whereas novel fake faces exhibit low confidence accompanied by high prediction uncertainty.