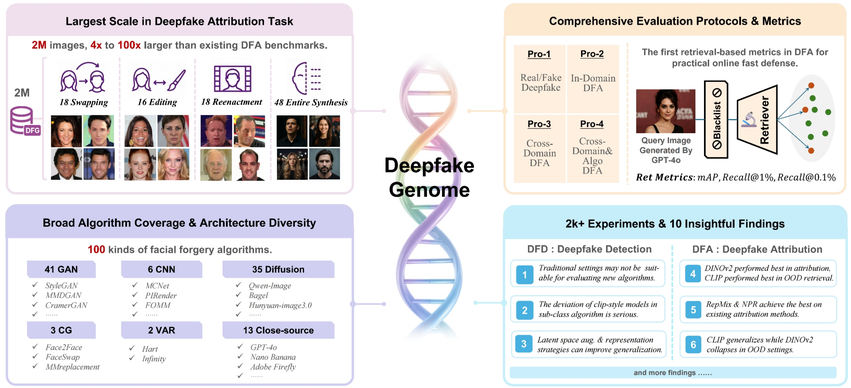
DeepfakeGenome: Toward Next-Generation Deepfake Attribution

Shunli Wang[1*], **Xinyu Zhou**[2*], **Yandan Zhao**[1], **Wenbin Liang**[2,3], **Taiping Yao**[1]
Ke-Yue Zhang[1], **Zhiyuan Yan**[4], **Shouhong Ding**[1†], **Lizhuang Ma**[2†]

[1] Tencent Youtu Lab [2] Shanghai Jiao Tong University
[3] East China Normal University [4] Peking University
[*] Equal contribution [†] Corresponding author.

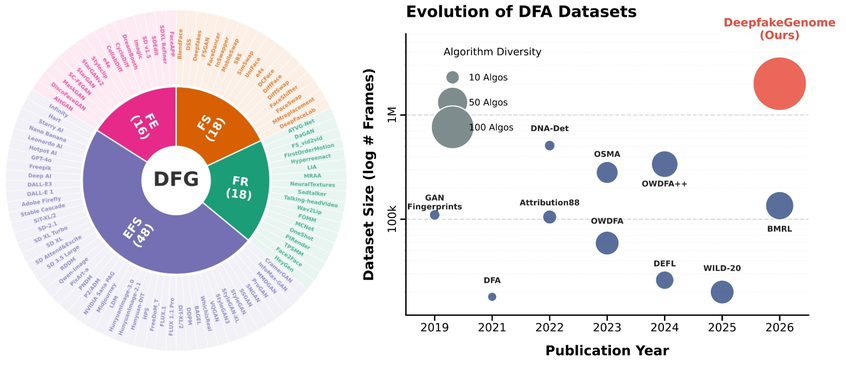
Project Page: https://huggingface.co/spaces/DeepfakeGenome/DFG_Page

A comprehensive deepfake detection and attribution benchmark named DeepfakeGenome (DFG).



- ✓ The DFG contains **100** facial forgery algorithms and **2M** images in total. It achieves **4× to 100× larger** than prior DFA benchmarks.
- ✓ We designed **4 protocols** for practical evaluation, including a novel retrieval-based attribution paradigm.
- ✓ Over **2K experimental** evaluations are conducted, and **10 insightful findings** are derived.

DeepfakeGenome attains the largest scale and highest richness.



- ✓ The proposed DFG benchmark comprehensively outperforms previous DFA datasets in terms of both **data scale** and **algorithm diversity**.

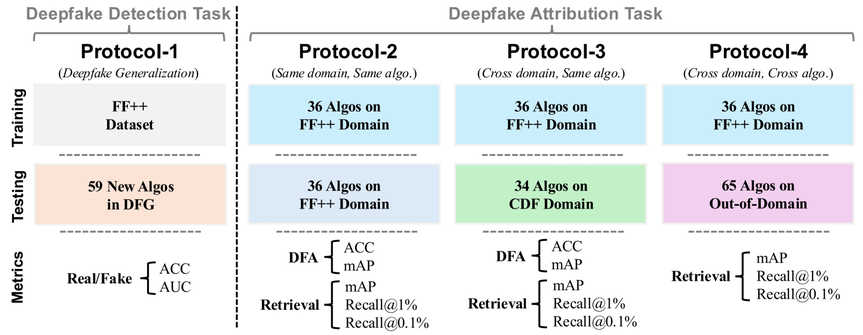
4 Forgery Types

- ➢ 18 * Face Swapping (FS)
- ➢ 18 * Face Reenactment (FR)
- ➢ 16 * Face Editing (FE)
- ➢ 48 * Entire Face Synthesis (EFS)

6 Algorithm Architectures

- ➢ 41 * GAN
- ➢ 35 * Diffusion
- ➢ 6 * CNN
- ➢ 3 * CG
- ➢ 2 * VAR
- ➢ 13 * Closed-source

4 Protocols are designed to comprehensively assess deepfake detection and attribution performance.



- ✓ **Protocol-1**. Traditional deepfake settings.
- ✓ **Protocol-2**. Same domain, Same algorithm.
- ✓ **Protocol-3**. Cross domain, Same algorithm.
- ✓ **Protocol-4**. Cross domain, Cross algorithm. 36 algorithms in FF++ domain as training set, and conduct retrieval test for 65 algorithms out of FF++ domain, which is suitable for **real-world application scenarios**.

$$AP_j = \frac{1}{K_j}\sum_{i=1}^{K_j}\frac{i}{Rank_j(i)},$$

$$mAP = \frac{1}{M}\sum_{j=1}^{M} AP_j,$$

$$Recall@\alpha = \frac{K_{recall}}{K_j},$$

$$K_{recall} = \sum_{x\in\mathcal{S}_j}\mathbb{I}\left(sim(f_{query}, f_x) \geq \tau_\alpha\right),$$

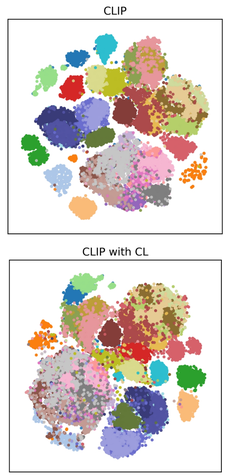
10 highly insightful findings in deepfake detection and attribution after 2K experimental evaluations.

- ➢ **Finding-1**. Traditional Deepfake training settings may not be suitable for evaluating emerging algorithms.
- ➢ **Finding-2**. The deviation of clip-style models in sub-class algorithm is serious.
- ➢ **Finding-3**. Latent space augmentation and representation learning strategies can significantly improve the generalization on new attacks.
- ➢ **Finding-4**. DINOv2 performed best in attribution, while CLIP performed best in OOD retrieval.
- ➢ **Finding-5**. RepMix and NPR achieve the best performance among the existing attribution methods.
- ➢ **Finding-6**. CLIP generalizes while DINOv2 collapses in OOD settings.
- ➢ **Finding-7**. Frequency and structural traces generalize better than specific spatial features.
- ➢ **Finding-8**. Contrastive learning biases the feature space and hurts OOD generalization performance.
- ➢ **Finding-9**. The OOD generalization ability of CLIP-ViTs relies heavily on pre-trained semantic priors.
- ➢ **Finding-10**. Low-level and high-level features are complementary for OOD generalization.

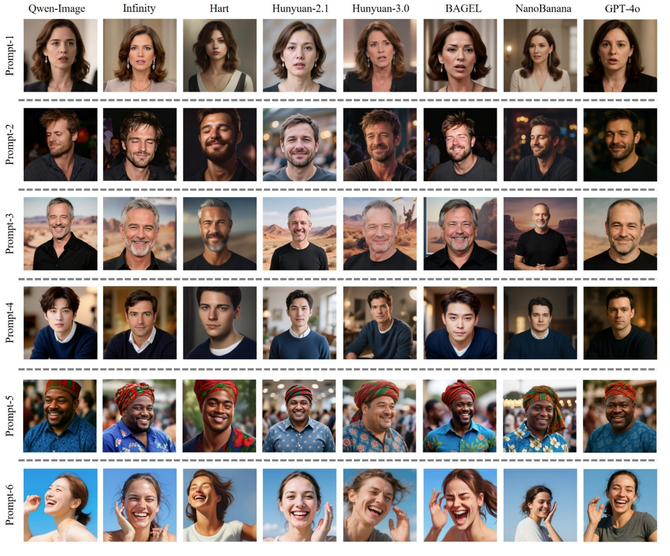
Experiments are conducted for retrieval interception scenarios.

Methods	Process	Ret-mAP	Ret-Recall @1%	Ret-Recall @0.1%	Avg
CLIP-ViT	-	0.5507	0.8355	0.7358	0.7073
CLIP-ViT	*w/o* FFT	0.2758	0.6220	0.4892	0.4623
Vanilla-ViT	-	0.2553	0.4050	0.2199	0.2934
CLIP-ViT	*w.* CL	0.4233	0.7692	0.6458	0.6127
CLIP-ViT	*w.* LSDA [67]	0.4306	0.7162	0.6070	0.5846
CLIP-ViT	*w.* aug	0.3672	0.6676	0.5332	0.5226
Effort [3]	*w.* aug	**0.5849**	**0.8649**	**0.7915**	**0.7471**
CLIP-ViT	*w.* ps=14	0.5428	0.8196	0.7456	0.7026
CLIP-ViT	*w.* ps=7	0.4349	0.7390	0.6459	0.6066
Effort	*w.* ps=14	0.4234	0.6059	0.4582	0.4958

Method1	Method2	Ret-mAP	Ret-Recall @1%	Ret-Recall @0.1%	Avg
CLIP-ViT	-	0.5507	0.8355	0.7358	0.7073
CLIP-ViT†	CLIP-ViT	0.6138	0.8651	0.8019	0.7602
CLIP-ViT	Effort‡	0.6294	**0.9070**	0.8469	0.7944
CLIP-ViT†	Effort‡	**0.6536**	0.9056	**0.8598**	**0.8063**
DINOv2	Effort‡	0.5248	0.8101	0.6946	0.6765
DINOv2	CLIP-ViT	0.5034	0.8078	0.7186	0.6766
DINOv2	CLIP-ViT†	0.5263	0.8329	0.7639	0.7077



Visualization of 8 synthetic algorithms in DeepfakeGenome.



All images are AIGC-generated.
No real individuals or personal information are included.

Contact

wangshunli@fysics.ai xinyuzhourain@sjtu.edu.cn