Is Artificial Intelligence Generated Image Detection a Solved Problem?

Ziqiang Li[1] **Jiazhen Yan**[1] **Ziwen He**[1] **Kai Zeng**[2]

Weiwei Jiang[1] **Lizhi Xiong**[1] **Zhangjie Fu**[1]

[1]School of Computer Science, Nanjing University of Information Science and Technology

[2]University of Siena, Siena, Italy

Page: https://github.com/HorizonTEL/AIGIBench

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Background & Motivation

In light of reported detection accuracies exceeding 95%, a critical question emerges:

Is Artificial Intelligence Generated Image detection a solved problem?

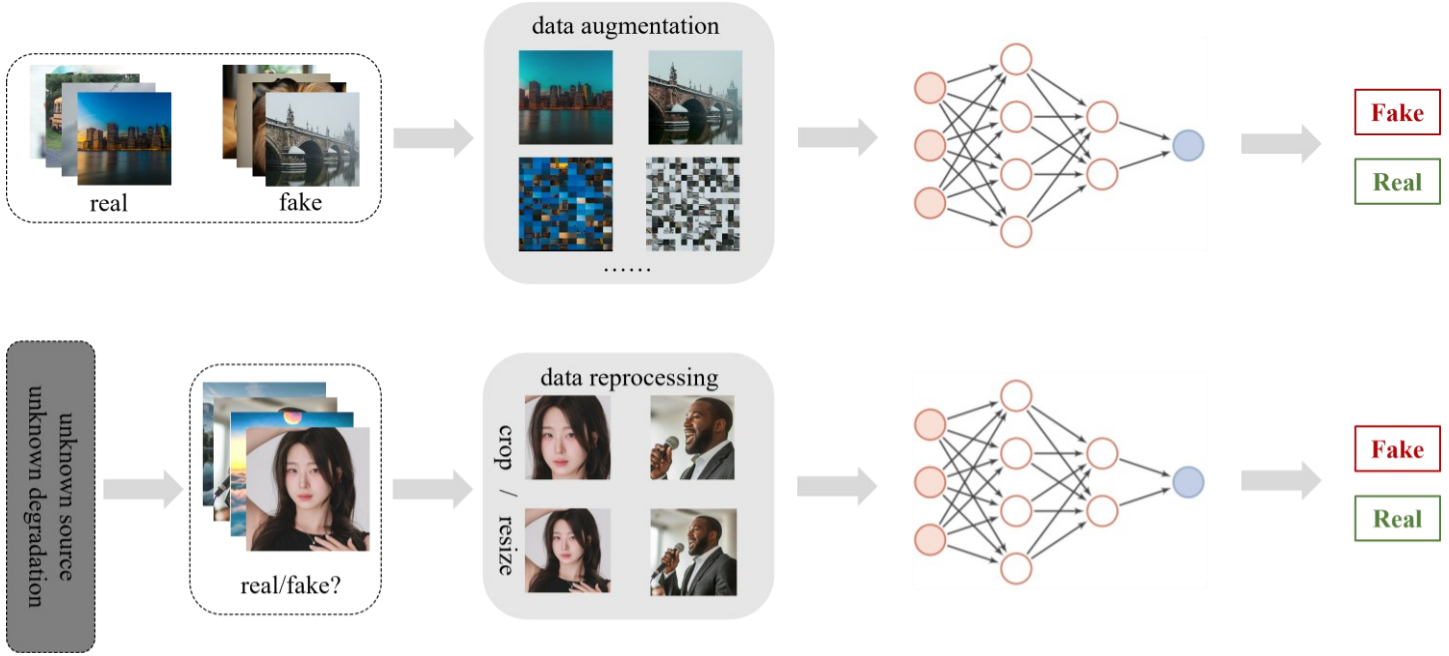

Fig 1. The AIGI Detection Pipeline.

Moivation

We decided to re-examine the entire AIGI detection process, including the training set and data augmentation during training, and the unknown data source and data preprocessing operations during testing.

Method & Evaluation

A novel and comprehensive dataset and benchmark: ***AIGIBench***

Benchmark Four core tasks that mirror practical challenges often overlooked in idealized test environments.

i)	**Generalization Assessment**	Multi-source
ii)	**Robustness Assessment**	Multi-degradation
iii)	**Data Augmentation Variation Assessment**	Identifying the most effective augmentation
iv)	**Test Data Pre-processing Assessment**	Identifying the most effective pre-processing

Dataset 23 subsets covering both advanced and widely adopted image generation techniques.

a) GAN-based noise-to-image generation
b) Diffusion for text-to-image generation
c) GANs for deepfake
d) Diffusion for personalized generation
e) Collected from social media platforms

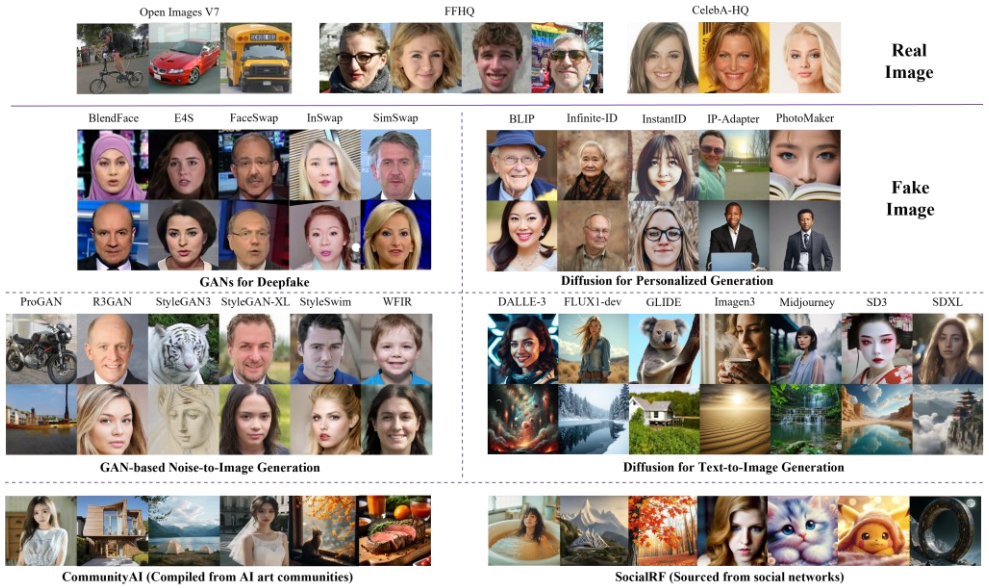

Fig 2. Visualizations of our datasets of AIGIBench.

Experiments & Discussion

Training Setting-II Training on 144K images generated by both SD-v1.4 and ProGAN, covering the same four object categories.

Evaluation Four tasks above.

Benchmark ↓ Dataset →	Generative Methods	~2022	2023	2024~	General Content	GAN & Diffusion	Image-based & Noise-based	Social Networks	AI-painting Communities
GenImage [29]	8	8	0	0	✓	✓	✗	✗	✗
AIGCDetction [30]	17	13	4	0	✓	✓	✗	✗	✗
DeepfakeBench [31]	9	9	0	0	✗	✗	✗	✗	✗
MPBench [32]	11	5	6	0	✓	✓	✗	✗	✗
Diff-Forensics [33]	7	7	0	0	✓	✗	✗	✗	✗
WildRF [18]	-	-	-	-	✓	✓	✗	✓	✗
DF40 [34]	40	27	10	3	✗	✓	✗	✗	✗
WildFake [35]	22	17	5	0	✓	✓	✓	✗	✗
Chameleon [20]	-	-	-	-	✓	✓	✗	✗	✓
AIGIBench (Ours)	25	9	5	11	✓	✓	✓	✓	✓

Table 1. Comparison with existing benchmarks on dataset

Benchmark ↓ Evaluation →	Detection Methods	~2022	2023	2024~	Generalization	Robust	Data Augmentation	Test Data Processing
GenImage (NeurIPS 2023) [29]	7	7	0	0	✓	✓	✗	✗
AIGCDetction (arXiv 2023) [30]	10	5	5	0	✓	✓	✗	✗
DeepfakeBench (NeurIPS 2023) [31]	34	30	3	1	✓	✓	✗	✗
MPBench (NeurIPS 2023) [32]	3	3	0	0	✓	✗	✗	✗
Diff-Forensics (ICCV 2023) [33]	6	5	1	0	✓	✗	✗	✗
WildRF (arXiv 2024) [18]	5	3	1	1	✓	✓	✗	✗
DF40 (NeurIPS 2024) [34]	7	7	0	0	✓	✗	✗	✗
WildFake (AAAI 2025) [35]	6	2	4	0	✓	✓	✗	✗
Chameleon (ICLR 2025) [20]	10	4	4	2	✓	✓	✗	✗
AIGIBench (Ours)	11	3	2	6	✓	✓	✓	✓

Table 2. Comparison with existing benchmarks on evaluation.

Experiments & Discussion

Task 1: Generalization Assessment

- Suffer notable performance degradation on **real-world manipulations** such as DeepFakes and in-the-wild content.

- No single method consistently outperforms others across all generative scenarios, underscoring the **difficulty of developing generalizable detectors.**

Test Dataset →	ProGAN		R3GAN		StyleGAN3		StyleGAN-XL		StyleSwim		WFIR		BlendFace		E4S		FaceSwap	
Detectors ↓	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.
Resnet-50	100.0	98.1	95.1	2.0	95.3	46.3	95.0	25.1	95.7	70.7	95.4	3.6	92.7	0.0	93.4	0.0	96.5	0.0
CNNDetection	99.9	95.3	**98.6**	2.3	**99.3**	9.1	**98.2**	0.7	**98.3**	6.9	99.4	0.2	**98.7**	6.2	**98.1**	4.1	**98.1**	1.4
Gram-net	99.8	97.2	89.6	6.1	91.1	40.1	89.8	56.0	90.3	60.7	78.4	10.7	84.6	0.0	85.0	0.0	93.0	0.0
LGrad	99.2	94.1	84.8	23.6	88.2	52.6	84.1	74.1	85.6	77.0	83.4	17.2	80.9	2.4	82.1	0.5	86.4	2.5
CLIPDetection	97.9	98.9	72.9	94.1	76.5	82.6	72.5	**96.7**	74.7	98.1	48.0	**91.9**	64.6	5.5	67.0	**46.9**	78.4	**27.3**
FreqNet	99.3	99.4	64.7	59.9	68.0	**98.2**	64.3	95.5	64.7	97.1	30.2	89.3	46.2	0.3	50.3	1.1	73.4	6.2
NPR	100.0	98.9	93.2	8.4	93.2	63.6	92.4	28.2	93.7	77.7	95.3	7.9	89.0	0.0	89.9	0.0	95.0	0.0
DFFreq	99.9	96.3	88.5	34.6	89.2	51.9	87.6	59.6	88.4	80.6	89.7	55.4	82.6	0.0	83.8	0.0	91.4	0.3
LaDeDa	100.0	99.7	90.2	19.5	91.6	93.2	90.5	80.5	90.8	97.3	97.8	19.2	84.8	0.0	85.9	0.0	93.2	0.0
AIDE	99.1	95.3	86.7	**99.0**	85.1	91.1	85.7	91.7	85.4	82.0	99.9	42.9	79.7	**23.2**	82.1	6.6	89.0	14.3
SAFE	**100.0**	**99.9**	96.6	91.2	96.6	92.9	96.5	89.7	96.3	**99.3**	**100.0**	20.7	93.8	0.8	94.8	28.9	97.1	3.3

Test Dataset →	InSwap		SimSwap		FLUX1-dev		Midjourney-V6		GLIDE		DALLE-3		Imagen3		SD3		SDXL	
Detectors ↓	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.
Resnet-50	96.6	0.0	96.2	0.0	95.6	69.1	89.5	15.5	96.2	62.4	95.5	11.8	95.8	27.9	95.3	33.1	95.3	50.4
CNNDetection	**98.3**	9.7	**98.0**	6.2	**98.5**	16.3	**98.9**	5.8	**97.6**	4.6	**98.1**	9.8	**98.2**	4.2	**98.4**	13.3	**98.4**	7.3
Gram-net	92.9	0.3	93.3	0.1	89.9	39.0	78.2	9.6	92.3	50.8	90.5	16.4	90.7	10.5	91.0	14.0	91.0	36.6
LGrad	86.2	1.3	85.8	2.3	84.1	76.6	78.7	41.5	85.9	78.9	84.3	29.7	83.9	40.2	84.4	42.4	84.4	62.7
CLIPDetection	78.4	8.2	78.8	8.6	73.3	86.6	50.0	80.6	78.2	75.2	74.9	**75.2**	73.6	84.2	78.4	90.6	78.4	91.0
FreqNet	72.6	0.9	72.0	0.6	64.7	92.4	25.3	83.6	71.8	79.7	64.2	68.2	65.8	81.5	66.6	88.1	66.6	98.9
NPR	94.6	0.0	94.8	0.0	93.3	97.2	83.9	53.8	94.8	70.3	93.0	21.2	93.5	78.2	94.2	89.7	94.2	79.0
DFFreq	91.0	0.0	90.8	0.0	88.9	64.1	74.9	54.0	91.0	86.0	88.5	14.5	89.2	62.1	89.1	73.4	89.1	88.7
LaDeDa	93.0	0.0	92.6	0.0	90.0	99.3	77.2	83.4	92.4	81.8	90.5	9.7	90.5	92.6	91.1	99.0	91.1	98.3
AIDE	89.6	11.4	88.2	**21.5**	86.0	90.0	73.0	79.8	88.4	**98.4**	85.7	24.5	85.7	93.9	89.3	**99.3**	89.3	97.6
SAFE	97.7	**56.8**	97.0	1.1	96.3	**99.8**	91.0	**97.2**	97.2	87.8	96.1	1.8	96.4	**97.0**	96.6	91.7	96.6	**99.9**

Test Dataset →	BLIP		Infinite-ID		InstantID		IP-Adapter		PhotoMaker		SocialRF		CommunityAI		Mean			
Detectors ↓	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	Acc.	A.P.
Resnet-50	99.2	99.8	95.7	4.0	95.3	26.8	95.5	30.2	95.4	2.7	97.3	13.4	100.0	5.1	95.7	27.9	61.9	69.3
CNNDetection	98.0	56.5	**98.3**	1.1	**98.3**	8.1	**97.9**	6.0	98.4	1.7	94.7	7.5	97.3	5.3	**98.2**	11.6	54.9	67.0
Gram-net	98.0	99.2	90.7	10.6	90.6	59.6	90.4	18.7	90.1	10.0	92.6	11.5	99.0	6.2	90.5	26.6	58.6	62.4
LGrad	89.4	96.6	84.8	17.0	84.0	61.0	85.5	54.9	85.0	34.6	83.7	22.2	98.9	11.4	85.8	39.6	62.9	66.6
CLIPDetection	85.1	92.1	75.2	93.8	74.0	96.9	73.3	92.0	73.3	65.2	53.3	**55.5**	82.8	**51.2**	73.3	**71.5**	72.5	75.6
FreqNet	87.7	100.0	65.4	92.7	65.8	93.9	65.8	93.0	65.5	88.6	68.5	39.3	98.9	12.2	65.9	66.4	66.2	70.1
NPR	98.4	99.9	93.1	34.6	93.5	34.1	93.1	71.8	92.6	3.6	96.3	21.9	99.9	8.2	93.8	41.9	67.9	73.9
DFFreq	96.5	99.4	89.7	50.9	88.5	95.3	88.3	78.1	88.5	87.4	96.2	17.5	99.9	7.3	89.6	51.9	71.1	75.7
LaDeDa	98.1	100.0	90.8	32.2	90.7	82.4	91.0	90.6	90.2	66.7	97.8	19.4	100.0	9.0	91.7	54.9	73.4	79.3
AIDE	92.8	100.0	87.1	97.5	86.6	97.0	86.6	**93.5**	85.9	97.5	97.2	18.4	99.0	9.3	88.1	67.0	77.6	**82.7**
SAFE	**99.4**	**100.0**	96.3	**99.8**	96.5	**99.9**	95.9	89.8	**96.0**	**98.0**	**99.6**	16.4	**100.0**	8.5	96.8	63.0	**79.9**	82.6

Table 2. Comparison with existing benchmarks on evaluation.

Experiments & Discussion

Task 2: Robustness Assessment

- While maintaining high R.Acc. under perturbations, their **F.Acc. drops sharply**, indicating reduced detection reliability in practical settings.
- Large pre-trained models or frequency-domain representations perform better.

Detectors →	Resnet-50			CNNDetection			Gram-net			LGrad			CLIPDetection			FreqNet		
Robust Settings ↓	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.
Origin	95.7	27.9	69.3	98.2	11.6	67.0	90.5	26.6	62.4	85.8	39.6	66.6	73.3	71.5	75.6	65.9	66.4	70.1
JPEG Compression	100.0	0.1	60.1	94.3	17.2	63.7	99.6	1.2	55.8	95.9	7.3	54.6	91.1	33.0	71.6	99.5	1.4	53.0
Gaussian Noise	98.8	4.2	66.1	97.7	2.6	47.0	95.4	10.6	60.5	91.9	17.5	60.0	78.3	58.7	72.2	73.7	48.5	66.2
Up-down Sampling	96.3	26.5	71.5	99.8	1.8	56.7	91.2	25.1	63.9	86.5	57.2	80.3	77.0	66.6	75.0	74.7	63.1	73.2
Mean	97.7	14.7	66.8	97.5	8.3	58.6	94.2	15.9	60.7	90.0	30.4	65.4	79.9	57.4	73.6	78.5	44.9	65.6

Detectors →	NPR			DFFreq			LaDeDa			AIDE			SAFE		
Robust Settings ↓	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.	R.Acc.	F.Acc.	A.P.
Origin	93.8	41.9	73.9	89.6	51.9	75.7	91.7	54.9	79.3	88.1	67.0	82.7	96.8	63.0	82.6
JPEG Compression	100.0	0.2	59.2	100.0	0.1	58.8	100.0	0.0	61.6	98.9	1.5	50.3	100.0	0.0	48.7
Gaussian Noise	98.5	6.2	68.5	86.3	32.2	69.0	98.8	2.6	68.5	93.0	22.4	72.5	100.0	1.2	46.9
Up-down Sampling	94.8	34.3	81.0	91.8	41.9	75.3	92.2	46.6	84.5	74.8	27.4	55.1	100.0	16.2	73.5
Mean	96.8	20.7	70.7	91.9	31.5	69.7	95.7	26.0	73.5	88.7	29.6	65.2	99.2	20.1	62.9

Table 3. The robust performance of AI-generated image detectors.

Task 3: Data Augmentation Variation Assessment

- Common data augmentation provide **limited** benefits in improving detector performance and may even introduce performance **trade-offs**.

Data augmentation			**CLIPDetection**		**FreqNet**		**NPR**		**DFFreq**		**SAFE**	
Rotation	Jitter	Mask	R.Acc./F.Acc.	Acc./A.P.	R.Acc./F.Acc.	Acc./A.P.	R.Acc./F.Acc.	Acc./A.P.	R.Acc./F.Acc.	Acc./A.P.	R.Acc/F.Acc.	Acc./A.P.
			73.3/**71.5**	**72.5**/75.6	65.9/**66.4**	66.2/70.1	93.8/41.9	67.9/73.9	89.6/51.5	71.1/75.7	94.3/64.6	79.5/84.5
✓			**86.1**/54.9	70.5/75.7	76.9/58.9	**68.0/71.5**	93.3/**44.0**	**68.7**/74.1	**92.1/52.6**	**72.7/77.4**	82.7/**69.5**	76.1/82.0
	✓		79.1/63.7	71.4/75.6	**89.7**/36.8	63.5/70.2	92.7/38.5	65.6/70.8	90.0/40.1	65.5/70.5	**99.4**/50.1	74.8/**84.8**
		✓	72.8/64.1	68.4/73.5	74.4/59.4	66.9/70.2	96.4/37.0	66.7/73.2	89.9/52.2	71.4/76.2	96.4/60.9	78.7/84.5
✓	✓		80.6/62.2	71.4/**76.6**	76.4/51.8	64.2/67.7	94.3/36.8	65.6/70.6	86.4/45.3	66.2/71.2	93.5/65.0	79.3/81.5
✓	✓	✓	79.6/61.3	70.5/75.8	62.4/62.5	62.4/64.8	**98.1**/32.5	65.3/**75.6**	86.0/47.8	67.3/72.4	96.8/63.0	**79.9**/82.6

Table 4. Evaluating the impact of different data augmentation on AIGI detectors.

Task 4: Test Data Pre-processing Assessment

- F.Acc. often remains unaffected or even degrades.

Detectors →	CLIPDetection		FreqNet		NPR		DFFreq		LaDeDa		SAFE	
Process ↓	R.Acc./F.Acc	Acc./A.P.	R.Acc./F.Acc	Acc./A.P.	R.Acc./F.Acc	Acc./A.P.	R.Acc./F.Acc	Acc./A.P.	R.Acc./F.Acc	Acc./A.P.	R.Acc./F.Acc	Acc./A.P.
Resize	73.3/71.5	72.5/75.6	65.9/66.4	66.2/70.1	93.8/41.9	67.9/73.9	89.6/51.9	71.1/75.7	91.7/54.9	73.4/79.3	63.3/66.5	64.9/68.6
Crop	76.9/56.1	66.5/68.4	84.6/63.5	74.2/80.0	99.3/36.9	68.2/81.9	96.1/51.7	74.4/81.1	98.9/56.1	77.5/82.5	96.8/63.0	79.9/82.6

Table 5. Evaluating the impact of different data pre-processing strategies on AIGI detectors.

Thank you for listening!

All resources are open-source.

Paper:	https://openreview.net/forum?id=N52U2h9k9o
Code :	https://github.com/HorizonTEL/AIGIBench
Datasets:	https://huggingface.co/datasets/HorizonTEL/AIGIBench

Email: 247918horizon@gmail.com

Welcome for any good discussion and cooperation.😇

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.