





# Is Artificial Intelligence Generated Image Detection a Solved Problem?

Ziqiang Li<sup>1</sup> Jiazhen Yan<sup>1</sup> Ziwen He<sup>1</sup> Kai Zeng<sup>2</sup>

Weiwei Jiang<sup>1</sup> Lizhi Xiong<sup>1</sup> Zhangjie Fu<sup>1</sup>

<sup>1</sup>School of Computer Science, Nanjing University of Information Science and Technology

<sup>2</sup>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?

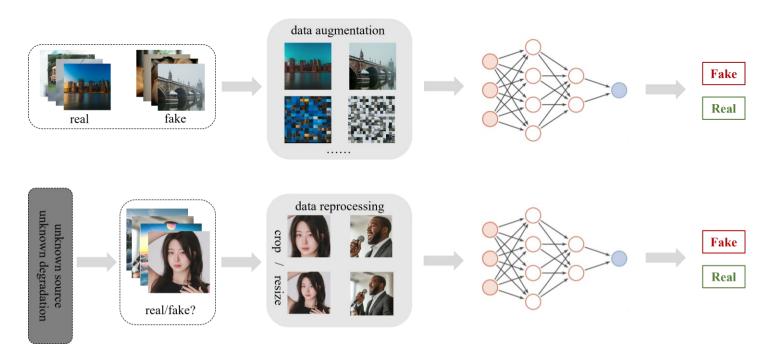


Fig 1. The AIGI Detection Pipeline.

#### **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
- ii) Robustness Assessment
- iii) Data Augmentation Variation Assessment
- iv) Test Data Pre-processing Assessment

Multi-source

Multi-degradation

Identifying the most effective augmentation

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.

$Benchmark \downarrow Dataset \rightarrow$	Generative Methods	~2022	2023	2024~			Image-based & Noise-based		AI-painting Communities
GenImage [29]	8	8	0	0	<b>✓</b>	<b>√</b>	Х	Х	X
AIGCDetction [30]	17	13	4	0	✓	✓	X	X	X
DeepfakeBench [31]	9	9	0	0	X	X	X	X	X
MPBench [32]	11	5	6	0	✓	✓	X	X	X
Diff-Forensics [33]	7	7	0	0	✓	X	X	X	X
WildRF [18]	_	-	-	-	✓	✓	X	✓	X
DF40 [34]	40	27	10	3	X	✓	X	×	X
WildFake [35]	22	17	5	0	✓	✓	✓	X	X
Chameleon [20]	_	-	-	-	✓	✓	X	X	✓
AIGIBench (Ours)	25	9	5	11	<b>✓</b>	✓	<b>√</b>	✓	<b>√</b>

Table 1. Comparison with existing benchmarks on dataset

$\overline{\text{Benchmark} \downarrow \text{Evaluation} \rightarrow}$	Detection Methods	~2022	2023	2024~	Generalization	Robust	Data Augmentation	Test Data Processing
GenImage (NeurIPS 2023) [29]	7	7	0	0	✓	<b>✓</b>	X	Х
AIGCDetction (arXiv 2023) [30]	10	5	5	0	✓	✓	×	×
DeepfakeBench (NeurIPS 2023) [31]	34	30	3	1	✓	✓	×	×
MPBench (NeurIPS 2023) [32]	3	3	0	0	✓	X	×	×
Diff-Forensics (ICCV 2023) [33]	6	5	1	0	✓	X	×	×
WildRF (arXiv 2024) [18]	5	3	1	1	✓	✓	×	×
DF40 (NeurIPS 2024) [34]	7	7	0	0	✓	X	×	×
WildFake (AAAI 2025) [35]	6	2	4	0	✓	✓	×	X
Chameleon (ICLR 2025) [20]	10	4	4	2	✓	✓	×	×
AIGIBench (Ours)	11	3	2	6	✓	<b>✓</b>	✓	$\checkmark$

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 inthe-wild content.

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

Network   Resert	T- + D- + - +	D . (		Pac		0-1-6	7.4.3.70	0-1-0		0.1.0		***	TID.	DI I	-		540	-	
Rener-50   100   98.1   98.1   98.1   2.0   98.3   98.5   39.5   98.5   99.5   98.5   98.5   98.5   98.5   98.5   99.5   98.5   99.5   98.5   99.5   98.5   99.5	$\textbf{Test Dataset} \rightarrow$	ProC	jAN		iAN		jAN3		AN-XL		swim		·IK		race		E4S		Swap
NR Proper Note No. 19	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.
Gram-net         99,2         91,2         89,6         61,0         91,1         41,1         89,8         60,0         90,3         60,7         78,4         10,0         81,0         21,0         81,0         21,0         10,0         21,0	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
Lamb	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	<u>6.2</u>	98.1	4.1	98.1	1.4
CHIPDERCEION         97,9         98,9         92,9         94,1         75,9         82,0         92,0         94,0         97,0         98,0         91,0	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
Frequency   Fr	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
NPR         100         98.9         93.2         8.4         93.2         63.0         93.2         51.0         87.6         98.2         98.0         87.6         98.0         98.0         88.0         99.0         91.0         90.0         91.0         91.0         91.2         91.0	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
DFFReq         99.9         96.3         88.5         34.6         89.2         51.9         87.6         95.6         88.4         80.6         89.7         55.4         82.6         90.0         83.8         90.0         93.2         90.0         93.2         90.0         83.0         90.0         93.2         90.0         93.0	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
Brain Part Land P	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
All Region	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
SAFE         100.0         99.9         8.6c         91.0         96.0         92.0         96.5         92.0         96.0         97.0         92.0         92.0         93.0	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
Pack Dataset	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
Detectors ↓         R.Acc, R.Acc	SAFE	100.0	99.9	<u>96.6</u>	91.2	<u>96.6</u>	92.9	<u>96.5</u>	89.7	<u>96.3</u>	99.3	100.0	20.7	<u>93.8</u>	0.8	94.8	28.9	<u>97.1</u>	3.3
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	$\textbf{Test Dataset} \rightarrow$	InS	wap	SimS	Swap	FLUX	1-dev	Midjour	ney-V6	GLI	DE	DAL	LE-3	Imag	en3	,	SD3	SI	XL
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         98.4         7.3         98.6         97.6         98.3         98.1         98.2         98.1         98.0         98.2         98.3         98.2         98.3         98.4         91.0         91.0         93.0         98.6         98.3         98.3         98.2         91.0         <	$\mathbf{Detectors} \downarrow$	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.
Gamenet         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         10.0         91.0	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
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         62.4         78.8         86.6         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         70.0         78.2         79.0         64.2         68.2         65.8         81.5         66.6         88.1         66.6         98.9           PFRPR         94.0         0.0         94.8         0.0         98.8         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         41.5         89.2         62.1         89.1         73.4         89.1         88.7         41.5         89.2         62.1         89.1         73.4         89.1         88.7         41.5         89.2         62.1         89.1         79.2         81.1         99.2         99.2         89.3         99.3         89.3         99.2         89	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
Production   Production   Production   Production   Production   Product   Product	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
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         99.3         77.2         83.4         92.4         81.8         90.5         97.7         90.5         92.6         91.1         99.0         91.1         99.3         77.2         83.4         92.4         81.8         90.5         97.7         90.5         92.6         91.1         99.0         91.1         99.3         97.9         90.2         97.7         90.5         92.6         91.1         99.0         91.1         99.0         91.1         90.2         91.1         99.0         91.1         99.0         92.2         91.1         92.0         91.1         92.0         91.1         92.0         92.2 </td <td>LGrad</td> <td>86.2</td> <td>1.3</td> <td>85.8</td> <td>2.3</td> <td>84.1</td> <td>76.6</td> <td>78.7</td> <td>41.5</td> <td>85.9</td> <td>78.9</td> <td>84.3</td> <td>29.7</td> <td>83.9</td> <td>40.2</td> <td>84.4</td> <td>42.4</td> <td>84.4</td> <td>62.7</td>	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
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         97.0         92.6         91.1         99.0         91.1         98.3         97.0         95.6         91.1         99.0         91.1         98.3         97.0         96.6         91.1         99.0         91.1         98.3         97.0         95.6         91.1         99.0         99.2         98.3         99.2         99.8         99.8         99.8         99.8         91.0         97.2         87.8         86.1         1.8         96.0         88.0         88.0         98.0         98.0         98.0         88.0 <td>CLIPDetection</td> <td>78.4</td> <td>8.2</td> <td>78.8</td> <td><u>8.6</u></td> <td>73.3</td> <td>86.6</td> <td>50.0</td> <td>80.6</td> <td>78.2</td> <td>75.2</td> <td>74.9</td> <td>75.2</td> <td>73.6</td> <td>84.2</td> <td>78.4</td> <td>90.6</td> <td>78.4</td> <td>91.0</td>	CLIPDetection	78.4	8.2	78.8	<u>8.6</u>	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
Defectors   Parce	FreqNet	72.6	0.9	72.0	0.6	64.7	92.4	25.3	<u>83.6</u>	71.8	79.7	64.2	<u>68.2</u>	65.8	81.5	66.6	88.1	66.6	98.9
LabeDa         93.0         0.0         92.6         0.0         99.0         92.3         77.2         83.4         92.4         81.8         90.5         97.0         92.6         91.1         99.0         91.1         98.3         97.0         97.0         98.6         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         97.0         96.6         91.7         96.6         91.7         96.6         97.0         11.2         98.8         91.0         97.2         87.8         96.1         1.8         96.4         97.0         96.6         91.7         96.6         99.9           Test Dataset →         R.Acc.         FAcc.         R.Acc.         FAcc.         R.Acc.         FAcc.         R.Acc.         FAcc.         R.Acc.         FAcc.         R.Acc.	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
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 99.3 99.5 99.5 99.5 99.5 99.5 9	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
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         99.2         99.2         70.0         1.8         96.1         97.0         96.6         91.7         96.6         99.2         99.2         97.0         1.0         97.0         96.6         91.7         96.6         98.0         8.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2         9.2	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
Test Dataset →         B⊥r         Infinitiri ID         Insimilari ID         IP-August ID         Photomater ID         R.Acc.         R.Acc. <t< td=""><td>AIDE</td><td>89.6</td><td><u>11.4</u></td><td>88.2</td><td>21.5</td><td>86.0</td><td>90.0</td><td>73.0</td><td>79.8</td><td>88.4</td><td>98.4</td><td>85.7</td><td>24.5</td><td>85.7</td><td>93.9</td><td>89.3</td><td>99.3</td><td>89.3</td><td>97.6</td></t<>	AIDE	89.6	<u>11.4</u>	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
Detectors ↓         R.Acc.         F.Acc.         R.Acc.         F	SAFE	<u>97.7</u>	56.8	<u>97.0</u>	1.1	<u>96.3</u>	99.8	<u>91.0</u>	97.2	<u>97.2</u>	<u>87.8</u>	<u>96.1</u>	1.8	<u>96.4</u>	97.0	<u>96.6</u>	91.7	<u>96.6</u>	99.9
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 <b>98.3</b> 1.1 <b>98.3</b> 8.1 <b>97.9</b> 6.0 <u>98.4</u> 1.7         94.7         7.5         97.3         5.3 <b>98.2</b> 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<	$\textbf{Test Dataset} \rightarrow$	В	LIP	Infin	ite-ID	Ins	tantID	IP-	Adapter	Phot	toMaker	S	ocialRF	Com	munityA	AI.		Mean	
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         34.6         83.7         22.2         98.9         11.4         85.8         39.6         62.9         66.6           CLIPDetection         87.7         100.0         65.4         92.7         65.8         93.9         65.8         93.0         65.5         88.6         68.5 <th><b>Detectors</b> ↓</th> <th>R.Acc.</th> <th>F.Acc.</th> <th>R.Acc.</th> <th>F.Acc.</th> <th>R.Acc.</th> <th>F.Acc.</th> <th>R.Acc</th> <th>. F.Acc.</th> <th>. R.Acc</th> <th>. F.Acc</th> <th>c. R.Ac</th> <th>c. F.Acc</th> <th>c. R.Acc</th> <th>c. F.Ac</th> <th>cc. R.</th> <th>Acc. F.A</th> <th>cc. Ac</th> <th>. A.P.</th>	<b>Detectors</b> ↓	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc.	F.Acc.	R.Acc	. F.Acc.	. R.Acc	. F.Acc	c. R.Ac	c. F.Acc	c. R.Acc	c. F.Ac	cc. R.	Acc. F.A	cc. Ac	. A.P.
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 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	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	1 9	5.7 27	.9 61.	9 69.3
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 <t< td=""><td>CNNDetection</td><td>98.0</td><td>56.5</td><td>98.3</td><td>1.1</td><td>98.3</td><td>8.1</td><td>97.9</td><td>6.0</td><td>98.4</td><td>1.7</td><td>94.7</td><td>7.5</td><td>97.3</td><td>5.3</td><td>3 9</td><td>8.2 11</td><td>.6 54.</td><td>9 67.0</td></t<>	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	3 9	8.2 11	.6 54.	9 67.0
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         <	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	2 9	0.5 26	.6 58.	6 62.4
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         93.5         85.9         97.5         97.2         18.4         99.0         9.3 </td <td>LGrad</td> <td>89.4</td> <td>96.6</td> <td>84.8</td> <td>17.0</td> <td>84.0</td> <td>61.0</td> <td>85.5</td> <td>54.9</td> <td>85.0</td> <td>34.6</td> <td>83.7</td> <td>22.2</td> <td>98.9</td> <td>11.</td> <td>4 8</td> <td>5.8 39</td> <td>.6 62.</td> <td>9 66.6</td>	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 8	5.8 39	.6 62.	9 66.6
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         93.5         85.9         97.5         97.2         18.4         99.0         9.3 </td <td>CLIPDetection</td> <td>85.1</td> <td>92.1</td> <td>75.2</td> <td>93.8</td> <td>74.0</td> <td>96.9</td> <td>73.3</td> <td>92.0</td> <td>73.3</td> <td>65.2</td> <td>53.3</td> <td>55.5</td> <td>82.8</td> <td>51.3</td> <td>2 7</td> <td>3.3 <b>7</b>1</td> <td>.5 72.</td> <td>5 75.6</td>	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	51.3	2 7	3.3 <b>7</b> 1	.5 72.	5 75.6
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	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 6	5.9 66	.4 66.	
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	•	98.4	99.9	93.1	34.6	93.5	34.1	93.1		92.6	3.6	96.3		.		_	3.8 41	.9 67.	73.9
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				_															
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	SAFE	99.4	100.0	96.3	99.8	96.5	99.9	95.9	89.8	96.0	98.0	'					_		_

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.

## Task 3: Data Augmentation Variation Assessment

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

## Task 4: Test Data Pre-processing Assessment

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

$\textbf{Detectors} \rightarrow \qquad \qquad \text{Resnet-50}$			)	CN	NDetection	on	C	ram-net			LGrad		CLI	PDetection	n		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
$\textbf{Detectors} \rightarrow$		N	IPR			DFFre	q		La	aDeDa			AIDE			S	AFE	
Robust Settings \	R.A	Acc. 1	F.Acc.	A.P.	R.Acc.	F.Ac	c. A.I	P.   R.A	Acc.	F.Acc.	A.P.	R.Acc	. F.Ac	c. A.F	. R.	Acc.	F.Acc.	A.P.
Origin	93	3.8	41.9	73.9	89.6	51.	9 75.	7   91	1.7	54.9	79.3	88.1	67.0	82.	7   9	06.8	63.0	82.6
JPEG Compressio	n 10	0.0	0.2	59.2	100.0	0.1	58.	8   10	0.0	0.0	61.6	98.9	1.5	50	3   1	0.00	0.0	48.7
Gaussian Noise	98	3.5	6.2	68.5	86.3	32.	2 69.	0   98	3.8	2.6	68.5	93.0	22.4	72.:	5   1	0.00	1.2	46.9
Up-down Samplin	ıg 94	1.8	34.3	81.0	91.8	41.	9 75.	3 92	2.2	46.6	84.5	74.8	27.4	55.	1 1	0.00	16.2	73.5
Mean	96	5.8	20.7	70.7	91.9	31.	5 69.	7   95	5.7	26.0	73.5	88.7	29.6	65	2 9	99.2	20.1	62.9

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

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

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

$\overline{\textbf{Detectors}} \rightarrow$	CLIPDetection		FreqNet		NPR		DFFreq		LaDeDa		SAFE	
<b>Process</b> ↓	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. ©