GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image

Eight generators are used to generate the images. The number of fakes images are the same as the real images from ImageNet

Dataset	Image Content	Generator Category		Public Avalibility	Real Images	Fake Images
		GAN	Diffusion			
UADFV [30]	Face	✓	✗	✗	241	252
FakeSpotter [25]	Face	✓	✗	✗	6,000	5,000
DFFD [3]	Face	✓	✗	✓	58,703	240,336
APFDD [6]	Face	✓	✗	✗	5,000	5,000
ForgeryNet [9]	Face	✓	✗	✓	1,438,201	1,457,861
DeepArt [27]	Art	✗	✓	✓	64,479	73,411
CNNSpot [26]	General	✓	✗	✓	362,000	362,000
IEEE VIP Cup [24]	General	✓	✓	✗	7,000	7,000
DE-FAKE [22]	General	✗	✓	✗	20,000	60,000
CiFAKE [1]	General	✗	✓	✓	60,000	60,000
GenImage	**General**	✓	✓	✓	**1,331,167**	**1,350,000**

1000 classes images

Over 1 million generated images

Containing GAN and Diffusion Generator

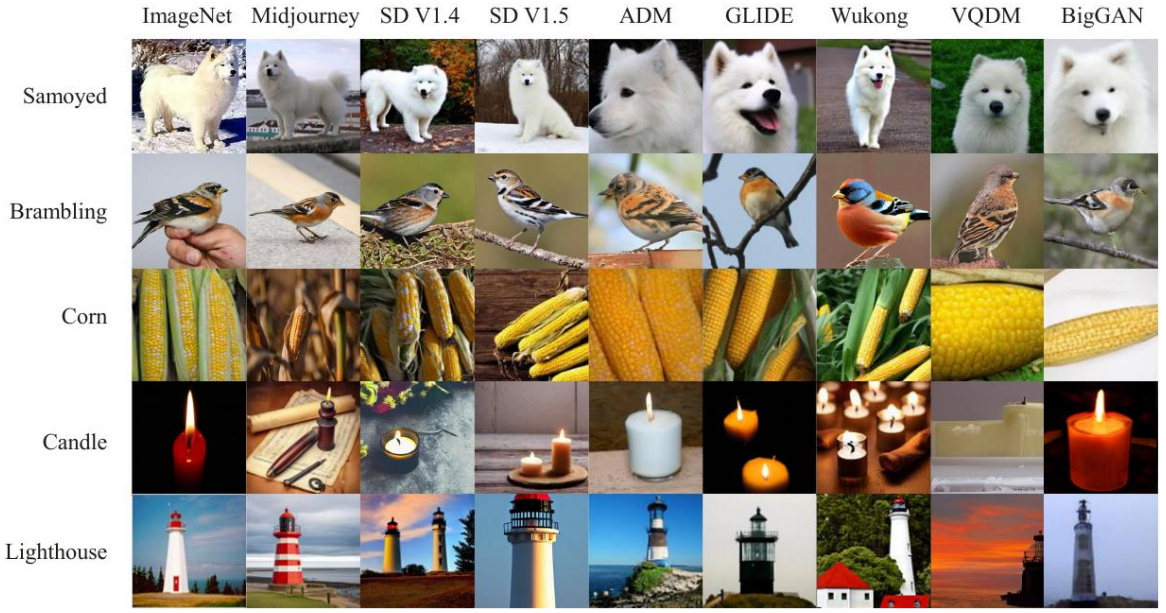
GitHub

https://github.com/GenImage-Dataset/GenImage

Project Address

https://genimage-dataset.github.io/

Image Visulization



Human eye cannot classify real and fake images.

The content is rich.

The variance of each class issignificant.

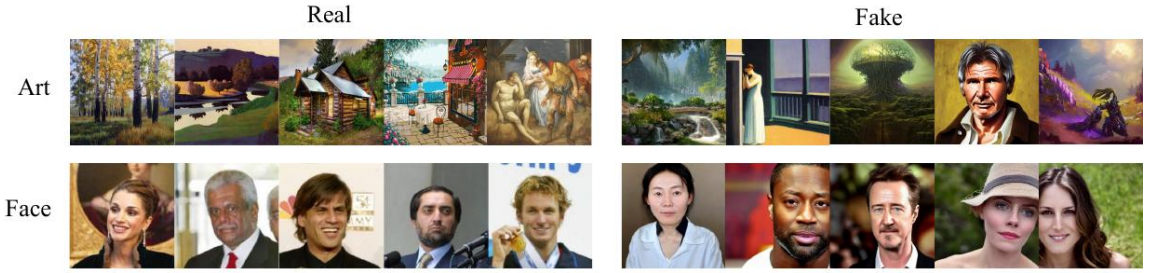
GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image

Verify the significance of increasing the number of images.Increasing the classes or the number of each class can efficiently increase accuracy.

Total	Image Classes	Images Per Class	Identical-Generator Acc. (%)	Cross-Generator Acc. (%)
1600	10	160	73.6	60.9
1600	100	16	81.8	62.7
16000	100	160	97.7	68.7
16000	1000	16	96.3	68.4
162000	1000	162	99.9	72.1

Training the model on GenImage, and testing it on face and art images.

Image Content	Image Source		Acc. (%)
	Real	Fake	
Face	LFW	SD V1.4	99.9
Art	Laion-Art	SD V1.4	95.0



Verify the significance of collecting a large number of image categories
Using image category subsets for training can effectively generalize to other categories

Image Classes	Fake Images Per Class	Acc.(%)
10	1280	78.8
50	256	80.3
100	128	81.2
100	1300	98.4

Detectors trained with GenImage
can accurately detect AI generated images from the internet

GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image

Task1： Cross-generator Evaluation

Examining the generalization of detectors
under different generator scenarios
Stable Diffusion V1.4 is used for training, and eight different generators are used for testing

Method	Testing Subset								Avg
	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Acc.(%)
ResNet-50 [8]	54.9	**99.9**	99.7	**53.5**	61.9	98.2	56.6	52.0	72.1
DeiT-S [23]	55.6	**99.9**	99.8	49.8	58.1	98.9	56.9	53.5	71.6
Swin-T [13]	**62.1**	**99.9**	99.8	49.8	**67.6**	99.1	**62.3**	**57.6**	**74.8**
CNNSpot [26]	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2
Spec [31]	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8
F3Net [19]	50.1	**99.9**	**99.9**	49.9	50.0	**99.9**	49.9	49.9	68.7
GramNet [14]	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9

Different generators are used for training and testing

Method	Testing Subset								Avg
	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Acc.(%)
ResNet-50 [8]	59.0	72.3	72.4	59.7	73.1	71.4	60.9	66.6	66.9
DeiT-S [23]	60.7	74.2	74.2	59.5	71.1	73.1	61.7	66.3	67.6
Swin-T [13]	**61.7**	**76.0**	**76.1**	61.3	**76.9**	**75.1**	**65.8**	**69.5**	**70.3**
CNNSpot [26]	58.2	70.3	70.2	57.0	57.1	67.7	56.7	56.6	61.7
Spec [31]	56.7	72.4	72.3	57.9	65.4	70.3	61.7	64.3	65.1
F3Net [19]	55.1	73.1	73.1	**66.5**	57.8	72.3	62.1	56.5	64.6
GramNet [14]	58.1	72.8	72.7	58.7	65.3	71.3	57.8	61.2	64.7

Cross validation using ResNet-50 on different generation models
Using the same generator for training and testing results in high accuracy.
Using different generators for training and testing results in low accuracy

Training Subset	Testing Subset								Avg
	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Acc.(%)
Midjourney	**98.8**	76.4	76.9	64.1	78.9	71.4	52.5	50.1	71.1
SD V1.4	54.9	**99.9**	99.7	53.5	61.9	98.2	56.6	52.0	**72.1**
SD V1.5	54.4	99.8	**99.9**	52.7	60.1	98.5	56.9	51.3	71.7
ADM	58.6	53.1	53.2	**99.0**	97.1	53.0	61.5	88.3	70.4
GLIDE	50.7	50.0	50.1	56.0	**99.9**	50.3	51.0	74.0	60.2
Wukong	54.5	99.7	99.6	51.4	58.3	**99.9**	58.7	50.9	71.6
VQDM	50.1	50.0	50.0	50.7	60.1	50.2	**99.9**	66.8	59.7
BigGAN	49.9	49.9	49.9	50.6	68.4	49.9	50.6	**99.9**	58.6

Task 2: Image degradation verification

The test uses low resolution, compressed, and blurred images.
Examining the Generalization of Detectors for Degenerated Images

Method	Testing Subset						Avg
	LR (112)	LR (64)	JPEG (q=65)	JPEG (q=30)	Blur (σ=3)	Blur (σ=5)	Acc(%)
ResNet-50 [8]	96.2	57.4	51.9	51.2	**97.9**	69.4	70.6
DeiT-S [23]	97.1	54.0	55.6	50.5	94.4	67.2	69.8
Swin-T [13]	97.4	54.6	52.5	50.9	94.5	52.5	67.0
CNNSpot [26]	50.0	50.0	**97.3**	**97.3**	97.4	77.9	78.3
Spec [31]	50.0	49.9	50.8	50.4	49.9	49.9	50.1
F3Net [19]	50.0	50.0	89.0	74.4	57.9	51.7	62.1
GramNet [14]	**98.8**	**94.9**	68.8	53.4	95.9	**81.6**	**82.2**