JUNE 18-22, 2023
CVPR VANCOUVER, CANADA
Tag: WED-AM-370
ID : 5352

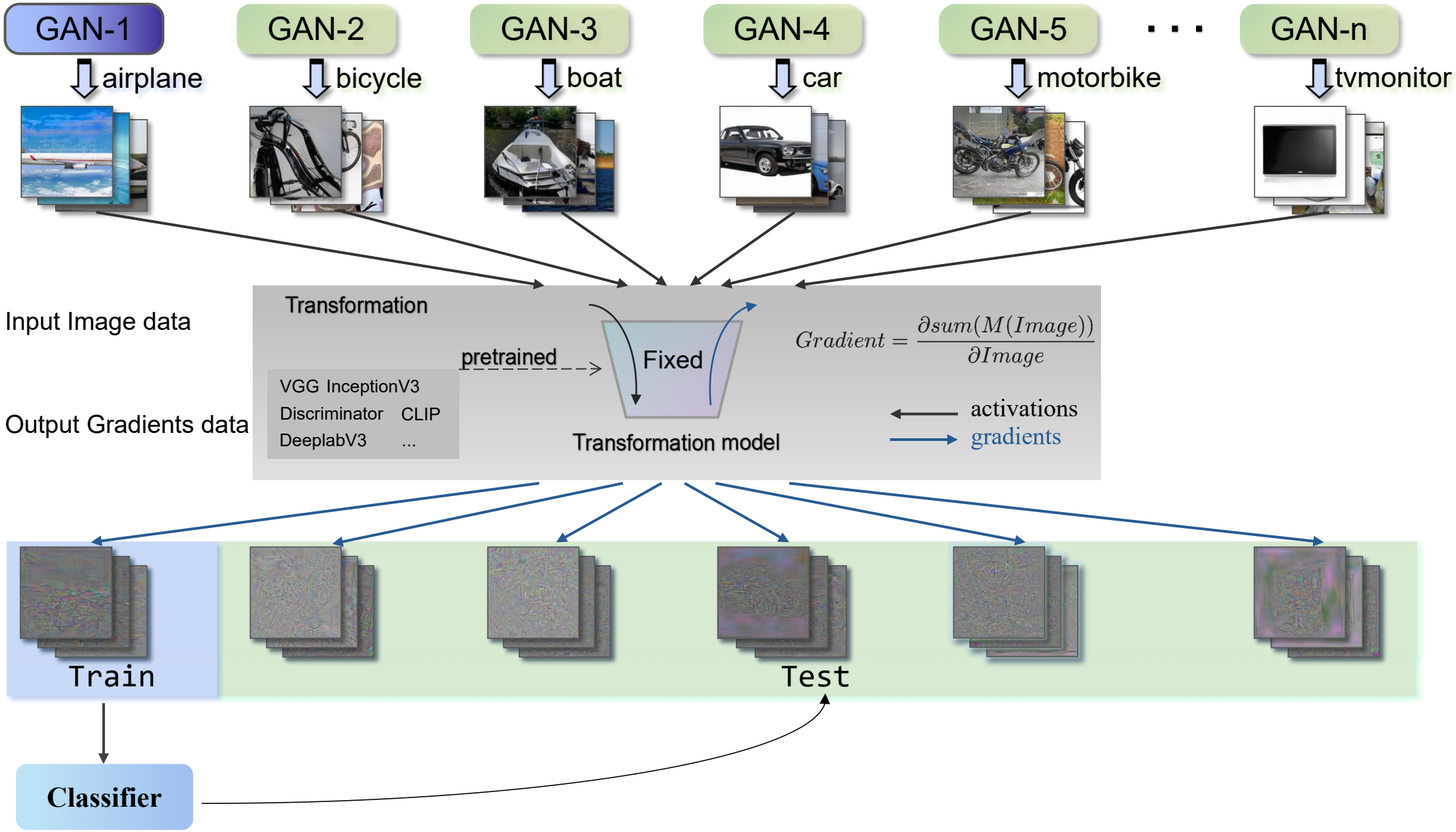
Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection

Chuangchuang Tan[1,2], Yao Zhao[1,2*], Shikui Wei[1,2], Guanghua Gu[3,4], Yunchao Wei[1,2]

[1]Institute of Information Science, Beijing Jiaotong University,
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology,
[3] School of Information Science and Engineering, Yanshan University ,
[4] Hebei Key Laboratory of Information Transmission and Signal Processing

https://github.com/chuangchuangtan/LGrad

Generalizable Deepfake Detection



JUNE 18-22, 2023
CVPR VANCOUVER, CANADA
Tag: WED-AM-370
ID : 5352

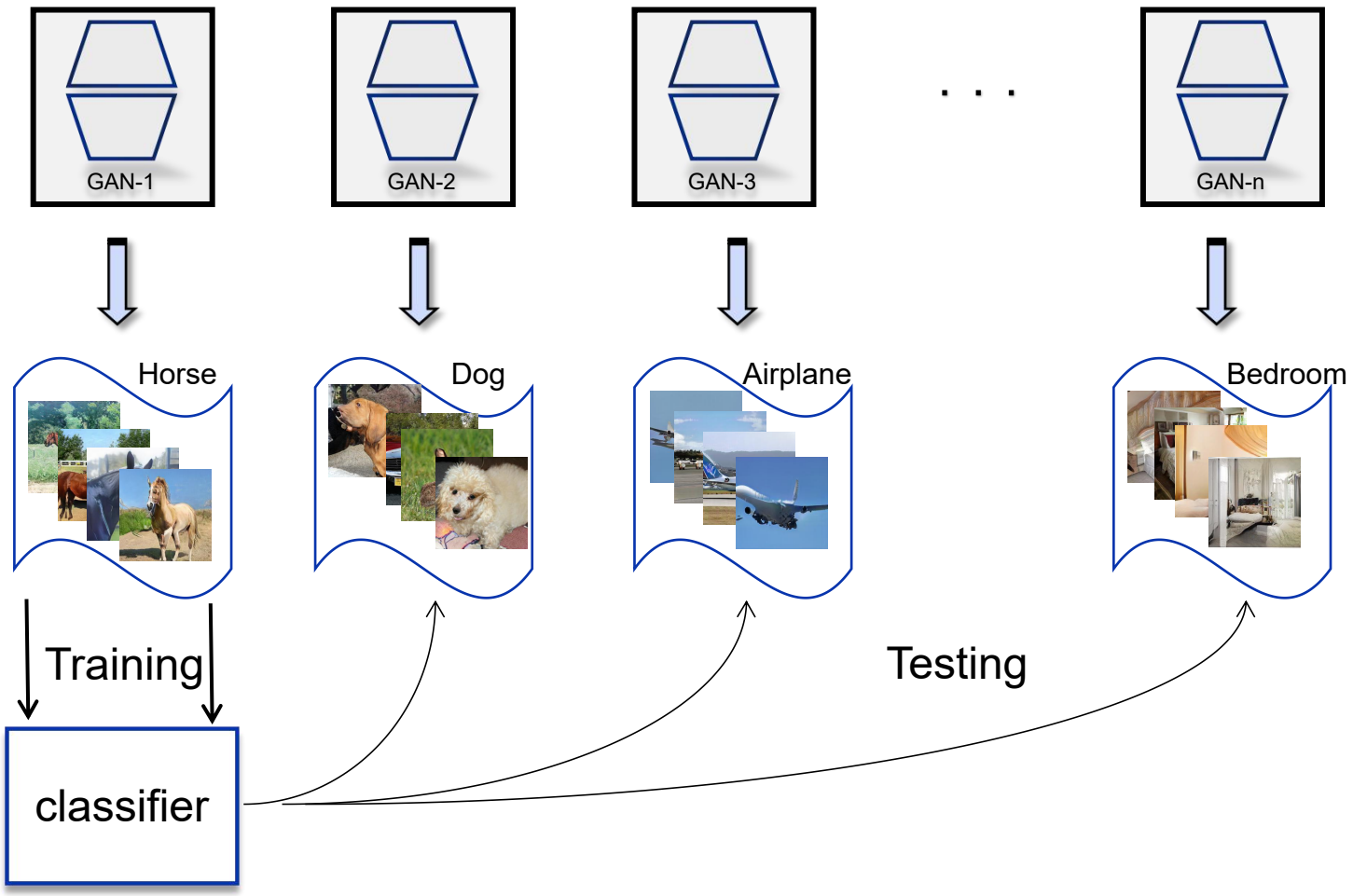
Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection

Chuangchuang Tan[1,2], Yao Zhao[1,2*], Shikui Wei[1,2], Guanghua Gu[3,4], Yunchao Wei[1,2]

[1]Institute of Information Science, Beijing Jiaotong University,
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology,
[3] School of Information Science and Engineering, Yanshan University ,
[4] Hebei Key Laboratory of Information Transmission and Signal Processing

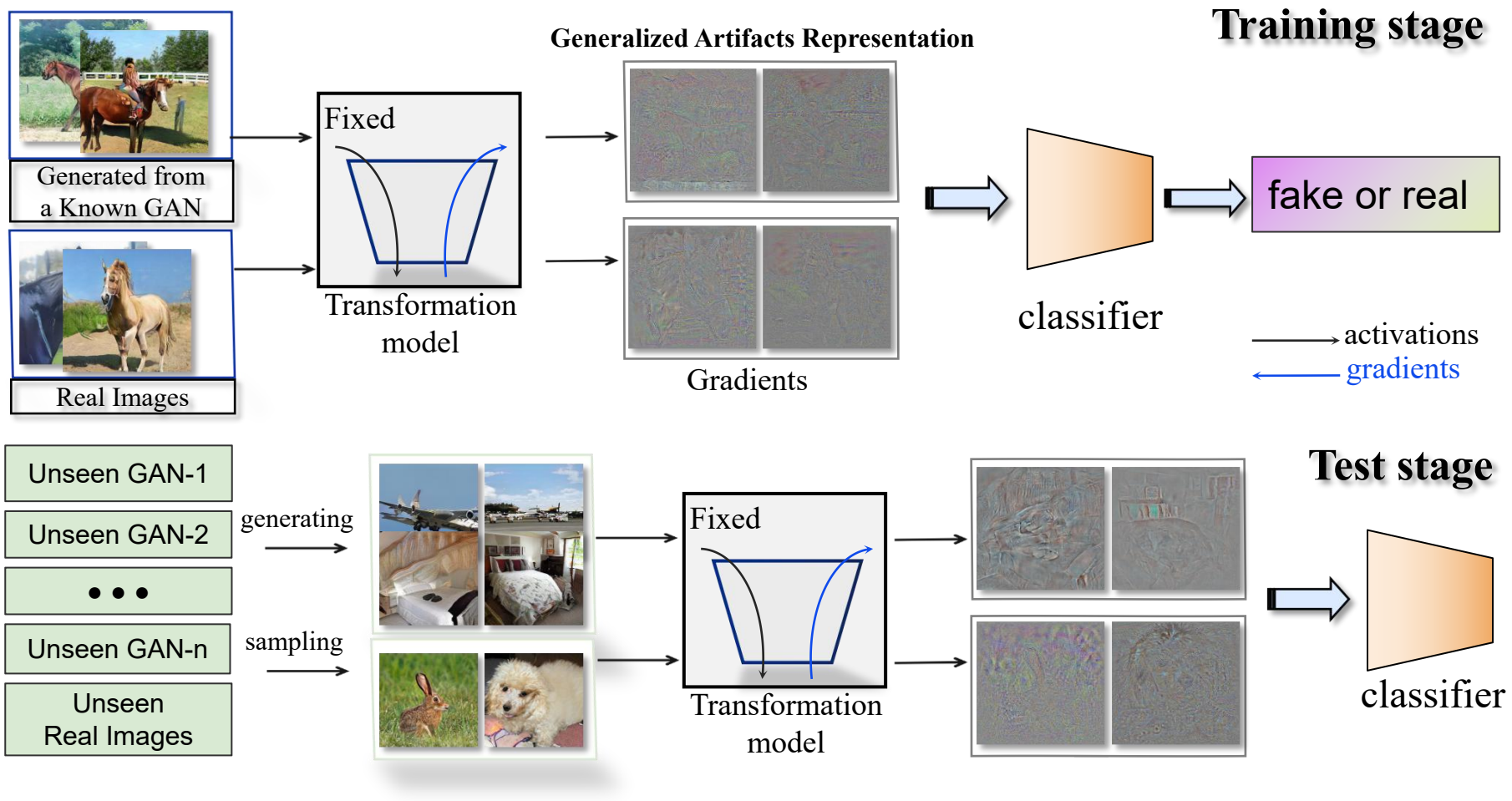
https://github.com/chuangchuangtan/LGrad

Problem: Learn from one GAN, and fight with other GANs



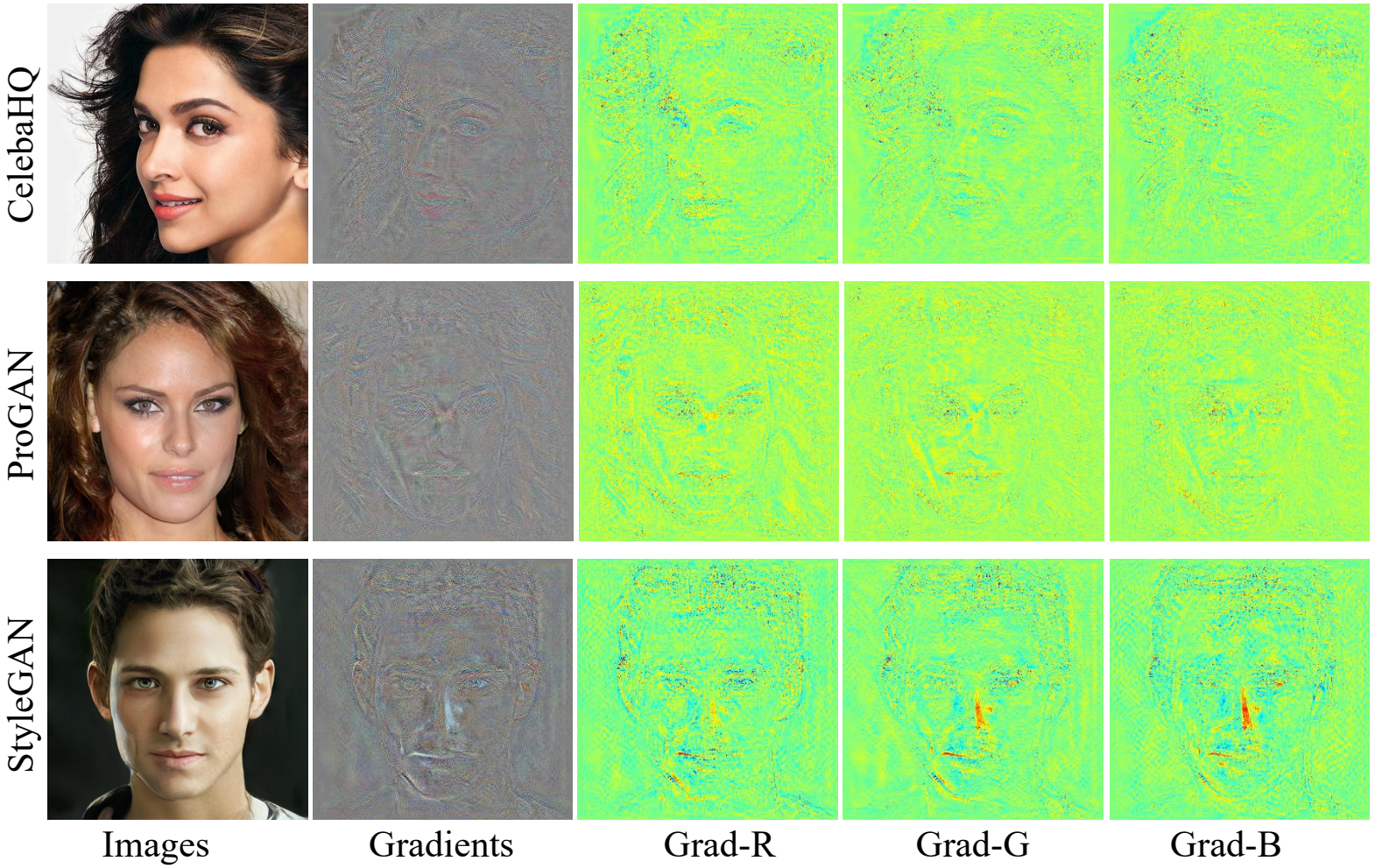
How to train a detector on *one GAN* model of *one class of data* to detect other fake images?

LGrad: Learning on Gradients



Our framework turns the data-driven problem into a transformation-model-driven problem.

Visualization of generalized representation: Gradients



In the gradients,
1) the content of images is filtered out,
2) the discriminative pixels are retained.

Transformation models

Trans. Model	Test Models																	
	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
VGG16	96.0	99.5	65.5	88.8	74.7	93.8	73.0	78.1	77.8	86.1	99.8	100.0	60.7	63.1	60.4	67.8	76.0	84.7
InceptionV3	64.9	74.4	58.8	66.9	65.4	73.8	50.9	52.1	59.1	68.4	52.6	58.5	54.0	56.5	49.8	50.1	56.9	62.6
Resnet50	86.4	95.0	81.0	92.2	83.7	93.5	57.2	56.2	68.3	75.8	96.4	99.5	51.2	52.7	63.4	70.4	73.4	79.4
CLIP-Resnet50	87.6	95.8	80.2	90.0	78.9	91.0	60.1	61.1	84.2	87.9	88.5	95.1	72.6	71.6	64.9	64.4	77.1	82.1
ViT	51.2	71.1	51.7	65.7	52.4	67.9	50.6	53.3	54.1	75.2	51.6	64.0	50.6	61.1	50.0	53.7	51.5	64.0
DeeplabV3	81.6	91.6	68.7	80.4	70.6	84.5	54.5	55.7	66.2	71.0	87.9	94.7	51.7	53.1	59.3	58.9	67.6	73.7
Idinvert	97.4	99.8	71.6	95.3	71.2	95.4	86.6	94.8	78.7	85.7	97.4	99.7	72.0	82.1	60.1	72.1	79.4	90.6
ProGAN-bedroom	98.4	99.9	82.6	95.6	83.3	98.4	76.2	81.8	82.3	90.6	99.7	100.0	71.7	75.0	52.8	57.8	80.9	87.4
ProGAN-bridge	97.8	99.7	86.4	97.5	85.7	97.3	72.5	78.7	76.8	87.5	94.1	99.9	62.5	75.8	53.2	61.3	78.6	87.2
StyleGAN-bedroom	99.4	99.9	96.0	99.6	93.8	99.4	79.5	88.9	84.7	94.4	99.5	100.0	70.9	81.8	66.7	77.9	**86.3**	**92.7**
StyleGAN-cats	97.4	99.7	83.4	97.3	77.4	96.4	69.8	74.6	79.3	90.2	97.8	99.8	68.0	77.4	65.9	72.9	79.9	88.5
StyleGAN2-church	99.1	100.0	88.2	97.7	91.9	99.6	70.1	71.7	80.6	89.1	95.6	99.8	60.8	68.9	72.7	76.5	82.4	87.9

Cross-model & category Performance

Methods	Settings		Test Models																	
	Input	#class	ProGAN		StyleGAN		StyleGAN2		BigGAN		CycleGAN		StarGAN		GauGAN		Deepfake		Mean	
			Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.	Acc.	A.P.
Wang	Image	1	50.4	63.8	50.4	79.3	68.2	94.7	50.2	61.3	50.0	52.9	50.0	48.2	50.3	67.6	50.1	51.5	52.5	64.9
Frank	Freq	1	78.9	77.9	69.4	64.8	67.4	64.0	62.3	58.6	67.4	65.4	60.5	59.5	67.5	69.1	52.4	47.3	65.7	63.3
Durall	Freq	1	85.1	79.5	59.2	55.2	70.4	63.8	57.0	53.9	66.7	61.4	99.8	99.6	58.7	54.8	53.0	51.9	68.7	65.0
BiHPF	Freq	1	82.5	81.4	68.0	62.8	68.8	63.6	67.0	62.5	75.5	74.2	90.1	90.1	73.6	92.1	51.6	49.9	72.1	72.1
FrePGAN	Image	1	95.5	99.4	80.6	90.6	77.4	93.0	63.5	60.5	59.4	59.9	99.6	100.0	53.0	49.1	70.4	81.5	74.9	79.3
LGrad(ProGAN-bedroom)	Grad	1	98.4	99.9	82.6	95.6	83.3	98.4	76.2	81.8	82.3	90.6	99.7	100.0	71.7	75.0	52.8	57.8	80.9	87.4
LGrad(StyleGAN-bedroom)	Grad	1	99.4	99.9	96.0	99.6	93.8	99.4	79.5	88.9	84.7	94.4	99.5	100.0	70.9	81.8	66.7	77.9	**86.3**(11.4↑)	**92.7**(13.4↑)
Wang	Image	2	64.6	92.7	52.8	82.8	75.7	96.6	51.6	70.5	58.6	81.5	51.2	74.3	53.6	86.6	50.6	51.5	57.3	79.6
Frank	Freq	2	85.7	81.3	73.1	68.5	75.0	70.9	76.9	70.8	86.5	80.8	85.0	77.0	67.3	65.3	50.1	55.3	75.0	71.2
Durall	Freq	2	79.0	73.9	63.6	58.8	67.3	62.1	69.5	62.9	65.4	60.8	99.4	99.4	67.0	63.0	50.5	50.2	70.2	66.4
BiHPF	Freq	2	87.4	87.4	71.6	74.1	77.0	81.1	82.6	80.6	86.0	86.6	93.8	80.8	75.3	88.2	53.7	54.0	78.4	79.1
FrePGAN	Image	2	99.0	99.9	80.8	92.0	72.2	94.0	66.0	61.8	69.1	70.3	98.5	100.0	53.1	51.0	62.2	80.6	75.1	81.2
LGrad(ProGAN-bedroom)	Grad	2	99.5	100.0	85.8	99.3	83.5	99.4	78.9	87.7	78.8	89.0	99.6	100.0	70.5	77.6	51.9	52.7	81.1	88.2
LGrad(StyleGAN-bedroom)	Grad	2	99.8	100.0	94.8	99.7	92.4	99.6	82.5	92.4	85.9	94.7	99.7	99.9	73.7	83.2	60.6	67.8	**86.2**	**92.2**
Wan	Image	4	91.4	99.4	63.8	91.4	76.4	97.5	52.9	73.3	72.7	88.6	63.8	90.8	63.9	92.2	51.7	62.3	67.1	86.9
Frank	Freq	4	90.3	85.2	74.5	72.0	73.1	71.4	88.7	86.0	75.5	71.2	99.5	99.5	69.2	77.4	60.7	49.1	78.9	76.5
Durall	Freq	4	81.1	74.4	54.4	52.6	66.8	62.0	60.1	56.3	69.0	64.0	98.1	98.1	61.9	57.4	50.2	50.0	67.7	64.4
BiHPF	Freq	4	90.7	86.2	76.9	75.1	76.2	74.7	84.9	81.7	81.9	78.9	94.4	94.4	69.5	78.1	54.4	54.6	78.6	77.9
FrePGAN	Image	4	99.0	99.9	80.7	89.6	84.1	98.6	69.2	71.1	71.1	74.4	99.9	100.0	60.3	71.7	70.9	91.9	79.4	87.2
LGrad(ProGAN-bedroom)	Grad	4	99.7	100.0	87.8	99.1	91.7	99.7	80.9	89.3	78.2	89.0	99.8	100.0	73.5	78.6	53.1	55.0	83.1	88.8
LGrad(StyleGAN-bedroom)	Grad	4	99.9	100.0	94.8	99.9	96.0	99.9	82.9	90.7	85.3	94.0	99.6	100.0	72.4	79.3	58.0	67.9	**86.1**	**91.5**