Breaking Semantic Artifacts for Generalized AI-generated Image Detection

NeurIPS 2024

Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, Chao Shen
School of Cyber Science and Engineering,
Xi'an Jiaotong University

Background & Agenda

Task: AI-generated image Detection

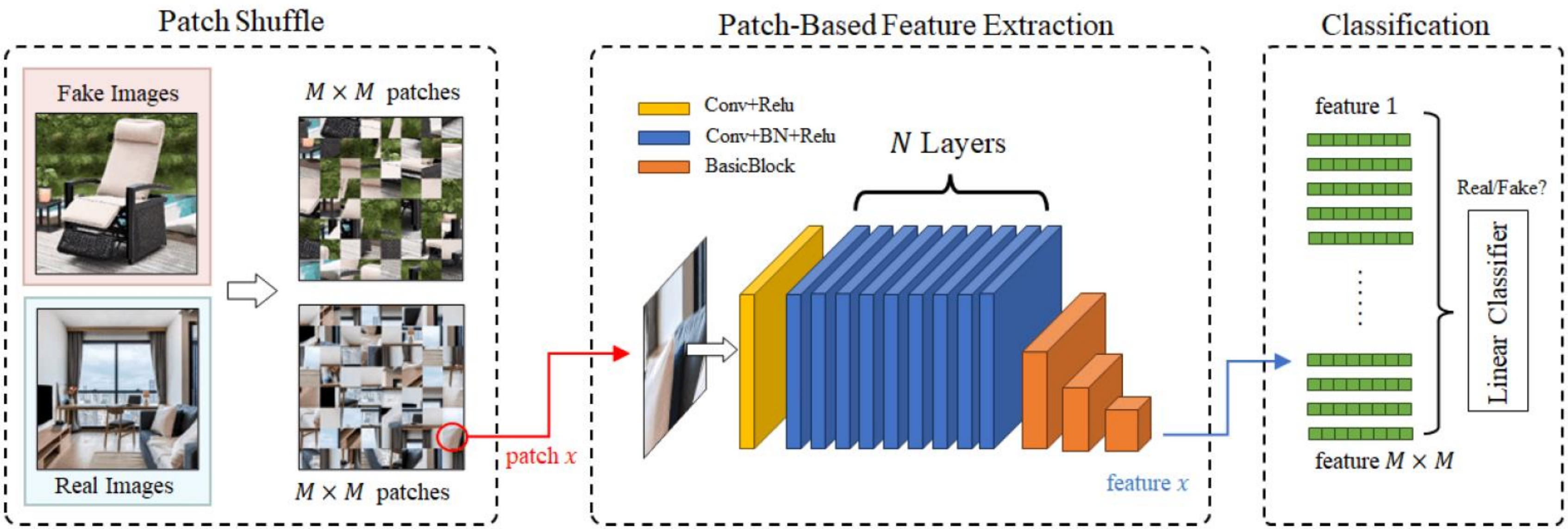
Real or Synthetic?

Agenda

- Motivation
  - Generator Artifacts
  - Semantic Artifacts
- Methodology
  - Patch-based Detection
- Experiment Results
  - Effect of pre-processing
  - Open-World Evaluation

Image Source: arXiv:1912.11035

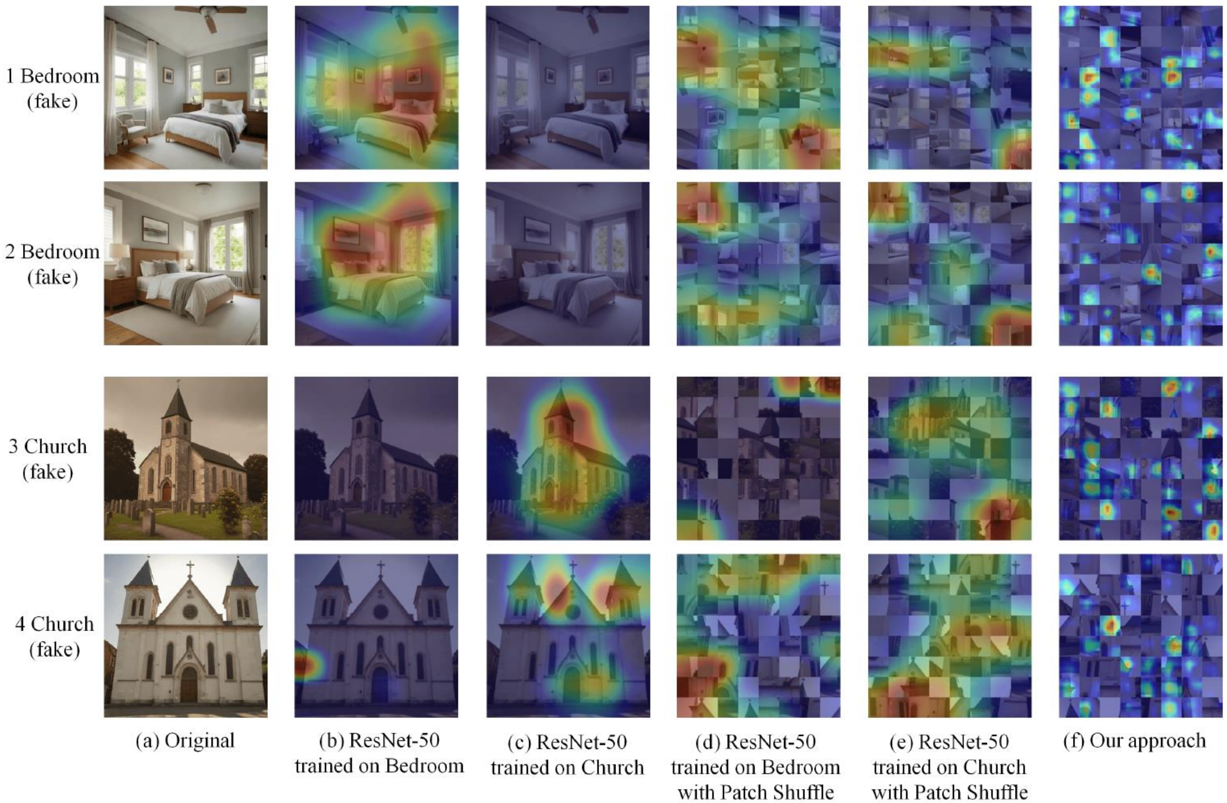
Overall Contributions



- Identify “semantic artifacts” in cross-scene synthetic image detection.
- Propose a patch-based detector, aiming at breaking “semantic artifacts” for generalization detection.
- Validate the effectiveness of our approach in both cross-scene and open-world generalization (including 31 test sets).

Motivation

Existing detectors tend to overfit the specific artifacts of the training data, resulting in substantial **Accuracy** drops in cross-scene generalization.



Training	Metrics	Test	
		Bedroom	Church
Bedroom	Acc. (Real)	99.80	100.0
	Acc. (Fake)	100.0	0.00
	AP	100.0	99.90
Church	Acc. (Real)	100.0	100.0
	Acc. (Fake)	0.00	100.0
	AP	98.20	100.0

Table 1: Cross-scene detection experiments on ResNet-50 models from ForenSynths [4]. We use 2 sets of training images on different scenes, Bedroom and Church, to retrain the detectors. Detection accuracy (Acc.) (at a threshold of 50%) and Average Precision (AP) are reported.

Generator Artifacts

- Synthetic images from different generators exhibit different artifacts.
- Unique artifacts might lead detectors to overfit during training.

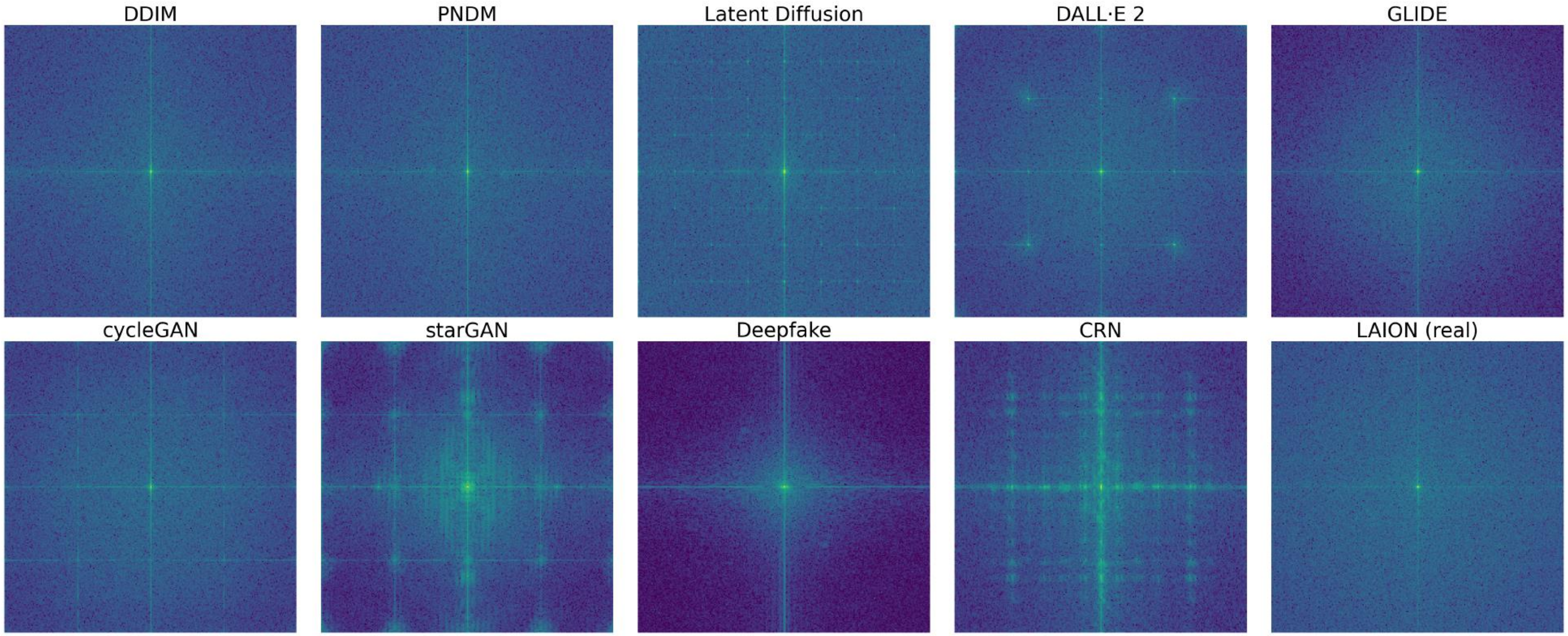


Figure 1: **Generator artifacts:** noise residuals power spectrum of images from 9 generative models and 1 real dataset. Top row: 5 Diffusion Models. Bottom row: 2 GANs, cycleGAN and starGAN, 2 CNN-based generators, Deepfake and CRN, and 1 real dataset, LAION.

Semantic Artifacts

- Real images with different semantics exhibit different artifacts.
- Semantic artifacts can be inherited by generative models.

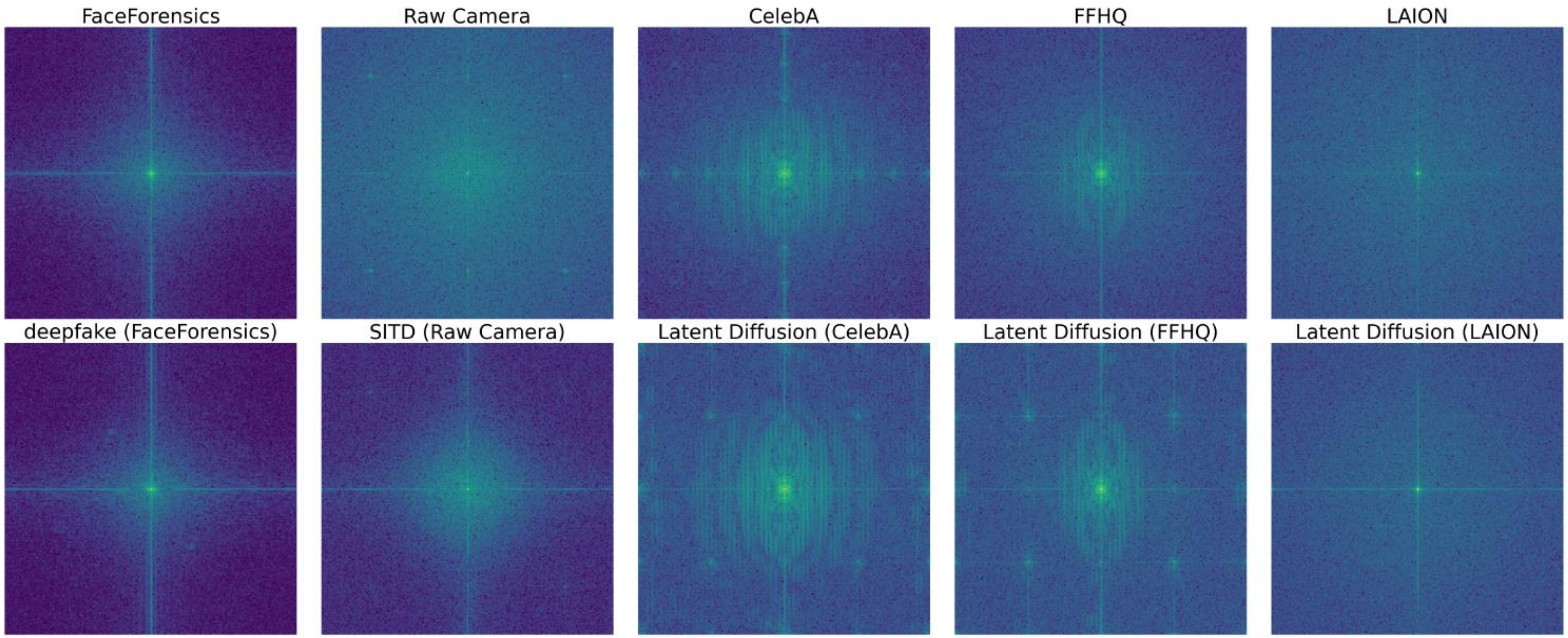


Figure 2: **Semantic artifacts:** noise residuals power spectrum of images from different scenes. Top row: 5 real datasets. Bottom row: 5 generative models in corresponding scenes, deepfake, SITD, and 3 variants of Latent Diffusion on CelebA, FFHQ, and LAION.

Methodology - Patch-based Detection

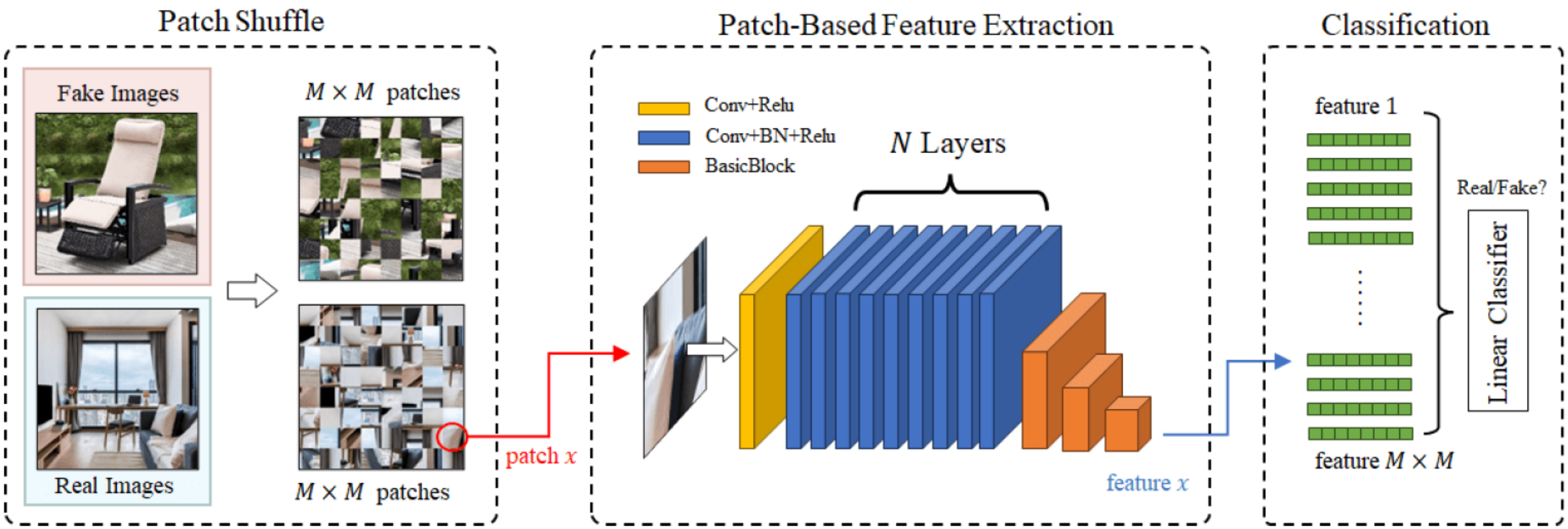
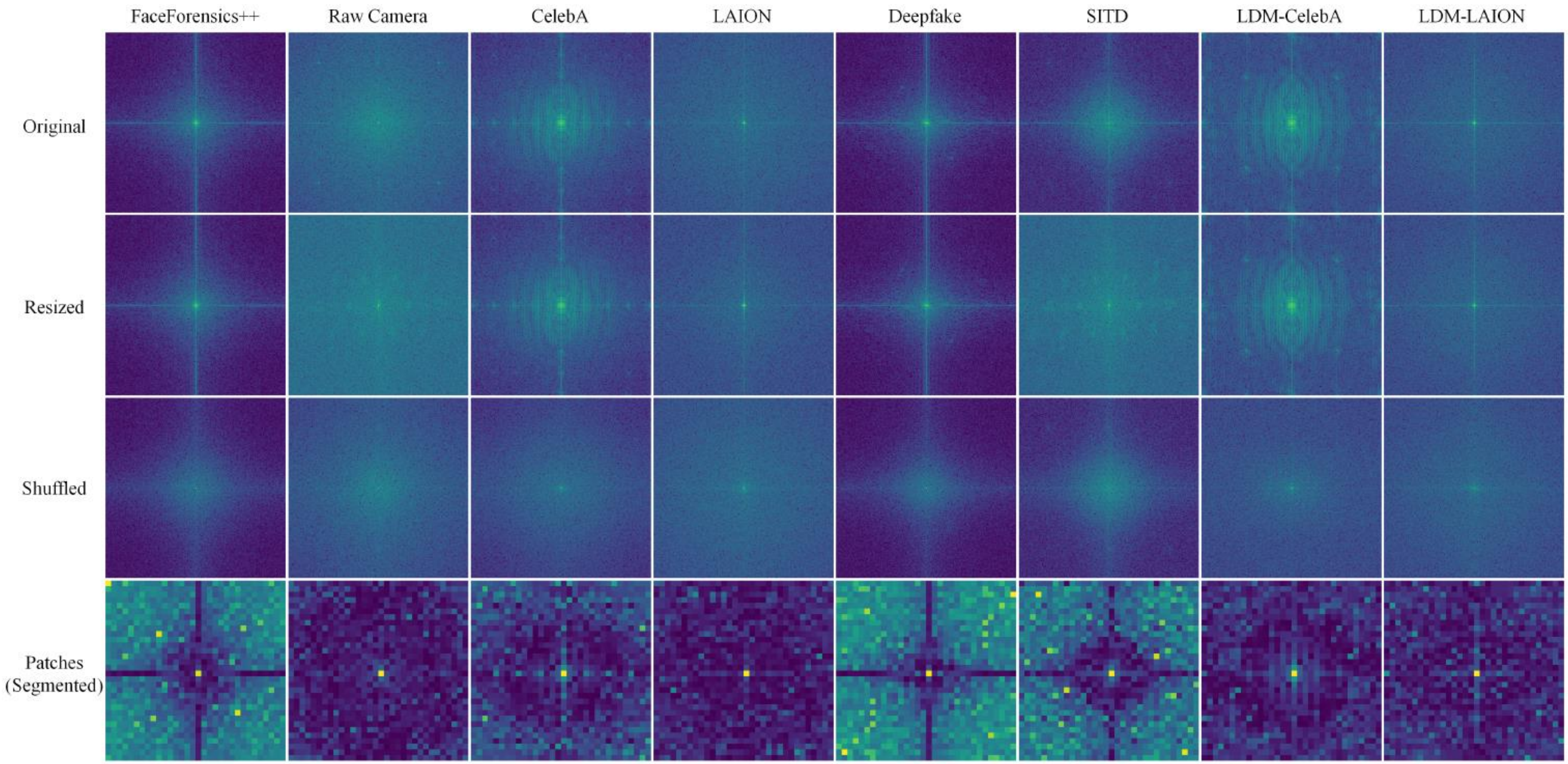


Figure 4: **Pipeline of our approach.** First, for pre-processing, we divide the input image into patches and shuffle these patches to obtain a randomized sequence. Then, we train a patch-based convolutional network for feature extraction. Finally, we flatten these features into a one-dimensional vector and then apply a linear classifier for classification.

- Patch Shuffling and patch-based feature extraction is able to break the "semantic artifacts" for generalization detection.
- By extracting local features, our detector is able to reduce the impact of global semantics in images.

Effect of pre-processing



- Most visible artifacts are reduced during the patch shuffling.
- Low-frequency features are weakened but high-frequency features (corresponding to artifacts) are enhanced in image patches.

Cross-Diffusion Evaluation

- Extensive experiments on 18 variants of Diffusion Models are performed.
- Significant gap between Acc. and AP exists in several detectors.

Methods	Variants	Bedroom				Church			ImageNet		CelebA		FFHQ	LAION						Average	
		DDIM	iDDPM	PNDM	LDM	DDIM	PNDM	LDM	LDM	ADM	LDM	RDM	LDM	DALLE2	GLIDE	LDM	SDv1	SDv2	SDv2-HR	Acc.	AP
		Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.		
CNN	Blur+JEPG(0.1)	52.80	50.70	50.85	55.60	50.85	51.10	50.00	52.10	50.65	52.95	47.40	49.60	49.40	50.40	61.68	87.30	72.65	60.70	55.37	65.20
	Blur+JEPG(0.5)	50.75	52.00	50.20	52.75	50.75	50.35	50.35	52.60	49.95	57.40	48.05	50.90	50.00	50.15	60.93	84.40	75.50	66.40	55.75	66.91
PatchFor	ResNet18	69.85	46.35	68.00	73.55	77.45	72.40	57.20	56.15	40.65	77.10	63.65	64.45	47.80	55.05	68.33	67.35	49.80	44.80	61.11	68.83
	Xception	50.10	51.35	50.15	51.95	51.30	50.15	50.75	66.80	49.95	95.90	50.00	54.80	50.70	60.00	99.25	99.35	68.40	67.95	62.16	69.14
F3Net	F3Net	50.15	50.00	50.00	50.15	50.15	50.00	50.00	60.35	42.55	89.05	52.85	76.85	50.45	49.95	64.88	86.20	76.70	84.20	60.25	77.31
	LFS	50.10	50.20	50.10	50.80	50.05	50.10	51.25	54.05	31.20	76.15	50.30	52.10	52.85	56.50	60.83	67.10	58.70	85.95	55.46	82.89
	Both	50.00	50.05	50.00	50.00	50.00	50.00	50.00	51.75	47.25	65.35	50.00	67.45	50.00	52.60	56.38	71.10	63.75	80.65	55.91	83.09
Durall	SVM	65.10	54.40	57.90	66.90	57.60	62.00	60.00	49.20	44.60	78.00	40.80	68.40	43.10	50.40	62.90	64.10	59.10	57.80	57.91	55.30
	LR	55.40	44.90	50.50	62.30	53.80	49.70	69.60	49.20	47.40	89.30	44.50	82.70	39.30	53.80	56.60	51.10	50.90	42.50	55.19	54.62
DIRE	CelebA-SDv2	58.40	61.20	55.60	63.55	64.15	82.15	71.80	72.80	45.40	83.25	81.55	37.30	45.63	57.85	50.45	54.55	53.50	59.36	61.03	70.01
	ImageNet-ADM	50.00	51.70	49.70	47.90	51.85	51.90	49.25	73.95	42.70	81.80	81.45	50.05	54.22	63.95	47.14	43.85	45.95	52.77	55.01	60.90
	LSUN-ADM	50.00	49.95	50.40	50.20	50.60	51.00	50.45	96.65	46.30	99.90	**100.0**	49.65	52.01	53.25	53.01	53.30	53.85	75.67	60.34	61.92
Dogoulis	Top 10k	50.20	56.25	50.35	50.35	52.50	47.05	50.90	57.45	**53.85**	46.90	44.65	48.15	50.40	50.40	62.43	85.40	81.95	60.85	55.56	60.77
	Top 24k	50.35	51.30	49.90	50.10	51.30	49.45	49.95	53.15	52.20	47.25	46.15	49.55	51.40	51.30	61.48	89.10	82.20	59.90	55.34	63.49
Ojha	CLIP:ViT-L/14+FC	58.05	82.15	55.30	54.10	67.15	54.35	59.65	68.10	48.80	81.15	60.90	66.40	66.18	64.60	68.87	86.10	77.60	66.10	65.86	79.51
LGrad	-	56.90	59.90	54.20	51.15	51.60	54.25	50.35	58.10	36.65	96.90	64.25	57.95	53.75	58.10	63.03	77.65	68.60	68.71	60.11	85.78
NPR	-	52.80	56.90	54.60	**99.75**	59.20	54.60	83.65	92.30	44.15	99.90	97.55	68.45	77.28	**90.15**	98.60	96.20	94.85	89.30	78.35	**94.08**
Ours	Resizing	99.30	**83.00**	**99.00**	98.95	**99.65**	**99.55**	**99.30**	79.50	13.55	**99.95**	95.95	**89.25**	79.95	84.15	98.57	98.65	92.50	91.45	89.01	93.58
	Zero padding	**99.40**	80.40	98.65	93.45	96.70	98.20	82.70	**97.40**	26.70	**99.95**	93.15	81.95	**88.83**	89.90	**99.69**	**99.70**	**97.60**	**98.90**	**90.18**	93.64

Table 4: Cross-Diffusion generalization results. We evaluate the detectors on all 18 variants of Diffusion Models.

Cross-GAN/CNN Evaluation

- Extensive experiments on 13 variants of GANs or CNN-base models are performed.
- Significant gap between Acc. and AP exists in several detectors.

Methods	Variants	proGAN	cycleGAN	bigGAN	styleGAN	styleGAN2	gauGAN	starGAN	deepfake	SITD	SAN	CRN	IMLE	WFIR	Average	
		Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	AP
CNN	Blur+JEPG(0.1)	51.00	50.00	49.60	49.80	51.16	65.95	50.20	62.25	51.11	56.62	74.45	**85.35**	48.80	57.41	65.38
	Blur+JEPG(0.5)	52.25	49.28	49.85	51.25	52.87	66.51	50.00	60.90	53.06	49.09	60.55	69.85	46.60	54.77	60.28
PatchFor	ResNet18	54.25	52.73	53.35	52.65	59.26	62.92	58.70	56.95	46.67	50.23	43.50	46.90	49.95	52.93	56.73
	Xception	52.00	51.21	51.70	50.75	50.44	65.68	50.15	50.05	49.72	50.23	50.00	50.00	51.70	51.82	57.53
F3Net	F3Net	50.00	46.89	52.15	49.35	51.51	65.88	51.10	70.00	46.11	47.72	49.95	50.15	50.65	52.42	57.18
	LFS	64.25	50.45	50.75	50.20	53.06	65.06	50.15	50.00	50.00	39.95	57.25	68.65	59.15	54.53	68.04
	Both	50.25	50.00	49.85	49.95	47.66	65.51	50.00	55.40	41.94	47.95	50.10	51.05	55.90	51.20	62.70
Durall	SVM	21.20	58.20	55.40	57.70	73.80	49.40	94.70	50.00	19.50	16.20	45.20	52.20	**96.30**	53.06	51.63
	LR	80.50	56.70	54.40	57.20	74.50	48.90	79.00	55.90	**82.00**	76.40	57.30	65.00	52.40	64.63	52.51
DIRE	CelebA-SDv2	48.44	53.80	49.80	52.55	54.65	19.65	50.95	50.05	49.43	65.50	64.15	63.50	47.50	51.54	55.24
	ImageNet-ADM	53.13	52.70	48.25	53.65	57.65	66.60	49.00	49.90	50.57	51.00	51.10	52.45	57.95	53.38	56.53
	LSUN-ADM	52.34	50.45	50.65	51.65	51.10	50.90	49.85	50.00	50.00	51.25	50.00	50.00	50.50	50.67	49.46
Dogoulis	Top 10k	51.00	49.55	49.50	48.05	48.93	65.99	50.00	50.00	43.06	50.91	48.40	56.05	49.75	50.86	51.68
	Top 24k	50.25	48.94	49.25	49.10	48.70	66.02	50.05	49.95	36.67	56.39	48.95	51.40	49.05	50.36	56.01
Ojha	CLIP:ViT-L/14+FC	91.25	74.90	79.05	84.75	71.25	73.05	72.30	62.05	49.72	64.38	50.50	53.15	68.90	68.87	83.74
LGrad	-	59.75	54.62	49.10	55.40	55.57	65.99	52.75	51.80	35.56	51.14	52.85	62.20	59.60	54.33	64.63
NPR	-	**94.75**	**93.14**	62.65	61.05	85.82	85.79	**99.55**	51.70	58.33	56.62	58.05	58.05	61.00	71.27	81.38
Ours	Resizing	86.50	84.77	**89.85**	90.25	**88.85**	**91.46**	96.50	66.80	10.28	60.00	47.90	58.55	71.85	72.58	81.12
	Zero padding	79.75	88.37	85.20	**95.20**	71.54	60.66	**99.55**	**70.65**	61.94	**86.25**	**74.80**	80.05	87.70	**80.13**	**84.97**

Table 5: Cross-GAN/CNN generalization results. We evaluate the detectors on all 7 Generative Adversarial Networks and 6 CNN-based generative models.

Open-world Evaluation

Methods	Variants	Cross-Scene		Open-World	
		Avg. Acc.	mAP	Avg. Acc.	mAP
CNN (CVPR'20)	Blur+JEPG (0.1)	53.66	63.40	56.23	65.27
	Blur+JEPG (0.5)	54.16	67.02	55.34	64.13
PatchFor (ECCV'20)	ResNet18	66.13	74.67	57.68	63.76
	Xception	69.91	75.26	57.82	64.27
F3Net (ECCV'20)	F3Net	65.21	81.49	56.97	68.87
	LFS	57.53	79.36	55.07	76.66
	Both	56.82	85.39	53.93	74.54
Durall (CVPR'20)	SVM	64.23	59.80	55.87	53.76
	LR	68.28	64.81	59.15	53.74
DIRE (ICCV'20)	CelebA-SDv2	63.19	69.93	57.05	63.81
	ImageNet-ADM	58.35	66.73	54.32	59.07
	LSUN-ADM	66.64	65.66	56.29	56.70
Dogoulis (MAD'23)	Top 10k	52.70	55.75	53.59	56.96
	Top 24k	51.91	57.76	53.25	60.35
Ojha (CVPR'23)	CLIP:ViT-L/14+FC	66.38	79.36	67.12	81.28
LGrad (CVPR'23)	-	62.91	80.99	57.69	76.91
NPR (CVPR'24)	-	90.44	94.84	75.38	<u>88.76</u>
Ours	Resizing	**94.25**	**96.28**	<u>82.12</u>	88.36
	Zero padding	<u>92.52</u>	<u>95.58</u>	**85.97**	**90.00**

Table 3: Results of cross-scene generalization and open-world generalization. For cross-scene generalization, we average the results on 6 variants of Latent Diffusion (LSUN-Bedroom, LSUN-Church, ImageNet, CelebA, FFHQ, LAION). For open-world generalization, we average the results on all 31 test sets (including 18 DMs, 7 GANs, and 6 CNN-based generators). **Bold** represents the best and <u>underline</u> represents the second best. More Detailed results are shown in Table 4 and Table 5.

Conclusion

- Existing detectors suffer from performance drops on cross-scene images, even though the images are generated by models with the same structure (LDM).
- Extensive experiments on 31 test sets validate the generalization performance of our approach, demonstrating our contributions to the universal detection of AI-generated images.

Ending

Thank you for listening

Contact

Code