D3: Training-Free AI-Generated Video Detection Using Second-Order Features

Supplementary Material

A. Limitation

In the main text, we have validated the outstanding generalization detection capabilities of D3. However, due to its unique methodology, D3 still exhibits certain limitations.

Unlike deep learning-based approaches, D3 focuses exclusively on temporal artifacts in video content, which means it struggles with video sequences that possess specific temporal characteristics, such as completely static videos composed of a single image or chaotic videos constructed from multiple randomly concatenated images. This explains D3's suboptimal performance on certain datasets, e.g., 45.11% AP on Text2video-zero (T2VZ). T2VZ is based on zero-shot IMAGE generators (for efficiency) with no temporal encoding, which results in the high temporal inconsistency of generated videos and makes them like real videos with large 2nd-order variations (see T2VZ vs. real in Table 1). Figure 1 visualizes a failed example, where large variations in flows confirm the high temporal inconsistency.

Datasets	1st-o std	2nd-o std
MSE	136.201	201.090
OS	152.296	264.395
Pika	138.801	157.827
ST2V	195.296	346.832
T2VZ	**548.477**	**1050.773**
VC2	127.606	174.463
Youku(Real)	338.408	931.518

Table 1. Standard deviation of 1st/2nd-order features.

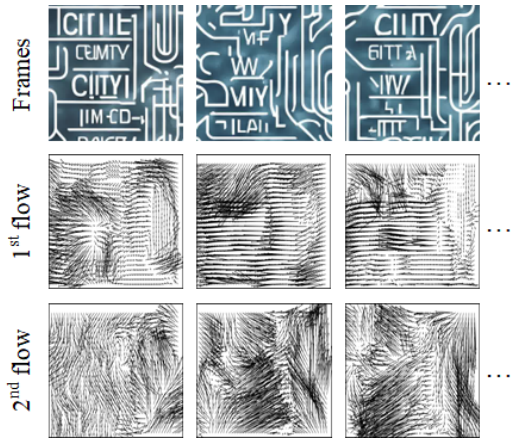

Figure 1. Failed T2VZ example.

B. Classification Threshold Analysis

In the main text, we report AP and AUC, which calculate accumulated results over all possible thresholds (rather than a fixed threshold).

To provide a more comprehensive comparison of D3's performance, we additionally report the accuracy (ACC) results of D3 and the baselines. The classification thresholds for D3 are selected by ensuring the correct classification of the majority (T=85%, 90%, or 95%) of real training videos. Table 2 shows that D3 is not sensitive to the threshold selection, and all three thresholds yield better results than the baselines.

Detection	Datasets (ACC↑)			
Method	GenVideo	EvalCrafter	VideoPhy	VidProM
FID	54.57	63.59	65.01	54.44
NPR	65.41	71.36	57.00	68.04
AIGVDet	49.07	57.62	53.33	47.25
Demamba	54.12	62.45	42.29	42.59
DeCoF	77.59	79.75	67.28	75.47
D3(T=85%)	88.61	89.38	**86.92**	**78.25**
D3(T=90%)	**89.11**	**90.40**	86.52	77.98
D3(T=95%)	87.68	89.41	83.56	75.74

Table 2. Accuracy results on 4 datasets.

C. Impact of Training Datasets

For a more comprehensive comparison, we additionally train the baselines using videos from 10 different generators (ZeroScope, I2VGEN-XL, SVD, VideoCrafter, Pika, DynamiCrafter, SD, SEINE, Latte, and OpenSora from GenVideo), and the test results are shown in Table 3. As seen, the results confirm that D3 remains the best in almost all cases, except on VidProM with T2VZ (see explanations in Section A).

Detection	Datasets (AP↑)			
Method	GenVideo	EvalCrafter	VideoPhy	VidProM
FID	89.18	97.33	97.60	87.79
NPR	88.76	90.07	85.36	**91.94**
Demamba	83.90	76.04	64.91	77.50
DeCoF	73.70	76.41	53.08	85.64
D3	**98.46**	**98.87**	**99.16**	88.46

Table 3. AP results. Multiple training sets are used.

D. Spatial Artifacts Analysis

Intriguingly, many detection methods perform analysis on spatial artifacts [1, 5, 7], while D3 utilizes the temporal ar-

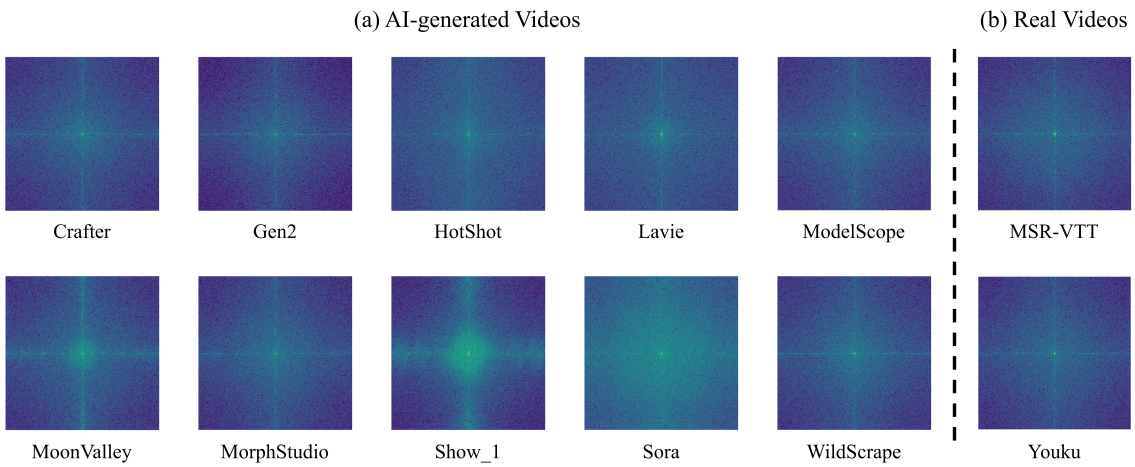

Figure 2. Noise residual power spectra of images extracted from different videos. One frame is extracted from each video. We visualize generated videos from 10 generators (Crafter, Gen2, HotShot, Lavie, ModelScope, MoonValley, MorphStudio, Show_1, Sora, and WildScrape) and real videos from 2 datasets (MSR-VTT and Youku).

tifacts. However, our experiments in the main text indicate that these methods suffer from performance drops for video detection. To address this, we delve into the image-level artifacts (i.e., spatial artifacts) in generated videos, which aims to uncover any unique or intriguing properties of the spatial generation artifacts of existing video generators.

Notably, some recent studies on synthetic image detection have conducted visualization experiments on spatial artifact analysis. Therefore, we employ the same experiments as these studies [2–4], presenting the power spectrum of the noise residuals in images.

Specifically, we randomly extract one frame from each video to create a set of images I for each video subset. Then, we utilize the denoising filter $D(x_i)$ from [8] to extract the noise residuals r_i of an original image x_i:

$$r_i(m,n) = x_i(m,n) - D(x_i(m,n)) \quad i = 1, 2, ..., I \quad (1)$$

Then, we start from the Fourier transform F of the noise residuals of the $M \times N$ image:

$$F_i(k,l) = \sum_{m=1}^{M} \sum_{n=1}^{N} r_i(m,n) e^{-j2\pi(\frac{k}{M}m + \frac{l}{N}n)} \quad (2)$$

where, k and l represent the frequency domain coordinates. To obtain the artifacts of each video subset, we take the average power spectrum S of all single images from it:

$$S_x(k,l) = \frac{1}{I} \sum_{i=1}^{I} |F_i(k,l)|^2 \quad (3)$$

Figure 2 shows the noise residuals power spectrum of images extracted from different videos, including 10 generators (Crafter, Gen2, HotShot, Lavie, ModelScope, MoonValley, MorphStudio, Show_1, Sora, and WildScrape) and 2 real datasets (MSR-VTT and Youku).

An intriguing conclusion is that, compared to the strong patterns of artifacts exhibited by generated images in the frequency spectrum (see recent studies [2, 3, 6]), spatial artifacts in videos are inconsistent and less pronounced. This can be due to the differences in the generative architectures as well as the post-processing compression applied to videos in cyberspace, which alters spatial artifacts. These findings further underscore the difficulties and challenges of detecting synthetic videos based on spatial artifacts.

E. Additional Detection Results

We provide the detailed detection results of AUC on 4 open-source datasets, represented in Tables 4, 5, 6, and 7.

These results demonstrate the strong detection performance and generalization ability of D3, which is consistent with our findings in the main text. (For the suboptimal performance of Table 7, refer to the explanation in Section A)

Detection Method	Visual Encoder	Datasets (AUC↑)										Avg. AUC
		Crafter	Gen2	HotShot	Lavie	MSE	MV	MSO	Show-1	Sora	WS	
FID	-	92.02	93.16	85.77	82.57	90.94	93.7	91.93	91.19	73.45	81.3	87.60
NPR	-	**99.24**	98.78	26.83	42.32	93.92	**99.92**	**99.38**	18.99	97.56	76.12	75.31
AIGVDet	-	69.77	71.62	74.67	79.55	68.05	59.41	79.42	70.29	60.79	67.82	70.14
Demamba	-	97.99	99.10	59.49	63.25	80.53	99.90	95.52	49.34	91.76	71.15	80.80
D3 (Cosine Sim)	DINOv2-B	90.91	95.78	85.02	83.88	80.75	95.47	83.11	90.68	85.22	80.85	87.17
	DINOv2-L	89.71	94.01	82.92	82.05	77.66	94.18	80.47	87.89	85.18	79.27	85.33
	CLIP-B-16	89.04	96.68	79.16	86.05	82.36	92.93	88.32	89.67	87.75	86.26	87.82
	XCLIP-B-16	94.09	97.07	90.10	89.83	86.37	96.80	89.43	93.37	90.70	85.21	91.30
	CLIP-B-32	87.94	94.98	82.39	86.43	83.37	93.19	87.28	90.43	87.27	85.46	87.87
	XCLIP-B-32	93.06	96.29	89.58	89.65	86.05	96.13	88.95	93.10	87.15	84.36	90.43
	ResNet18	91.85	96.54	88.69	86.18	85.22	96.07	86.50	93.80	88.26	84.23	89.73
	VGG16	92.97	98.19	92.30	90.14	89.02	97.51	90.16	96.00	92.58	87.43	92.63
	EfficientNet-b4	88.87	94.17	84.11	83.95	80.85	92.11	82.12	89.75	81.13	78.08	85.51
	MobileNet-v3	89.11	95.51	86.32	83.45	82.22	94.31	83.05	91.86	84.42	81.17	87.14
D3 (L2 Distance)	DINOv2-B	96.97	98.94	95.76	93.92	92.46	98.79	95.28	97.94	96.36	91.99	95.84
	DINOv2-L	96.23	98.39	94.68	92.99	91.03	98.09	93.95	96.79	95.98	91.12	94.92
	CLIP-B-16	97.06	**99.40**	94.65	96.10	94.92	98.20	98.03	98.03	97.96	**95.67**	97.00
	XCLIP-B/16	98.36	99.27	97.65	**97.13**	**96.02**	99.16	97.78	98.64	**98.35**	94.85	**97.72**
	CLIP-B-32	96.28	98.75	95.21	96.00	94.88	98.10	97.55	97.83	97.74	94.94	96.73
	XCLIP-B-32	97.73	98.93	96.95	96.64	95.22	98.86	97.04	98.21	96.74	93.63	96.99
	ResNet18	96.92	99.05	96.88	94.88	93.74	98.66	95.77	98.58	96.68	92.75	96.39
	VGG16	95.93	99.39	**98.42**	96.71	94.62	98.55	96.02	**99.33**	97.55	93.19	96.97
	EfficientNet-b4	95.59	98.13	94.24	93.51	91.46	97.18	93.46	96.56	93.00	89.67	94.28
	MobileNet-v3	95.90	98.78	95.88	94.02	92.31	98.22	94.61	97.92	95.42	91.59	95.47

Table 4. Detection results (AUC) on 10 GenVideo datasets.

Detection Method	Visual Encoder	Datasets (AUC↑)														Avg. AUC
		MV	Floor32	Gen2	Gen2-D	HotShot	LaVie-V	LaVie-I	Mix-SR	MSE	Pika	Pika-v1	Show-1	VC	ZS	
FID	-	98.28	96.26	97.46	98.58	90.05	92.32	83.42	98.23	95.27	99.48	99.24	96.79	95.26	95.29	95.42
NPR	-	**99.92**	**99.38**	97.74	**99.80**	26.83	50.23	36.00	**99.22**	93.92	**99.92**	**99.71**	18.99	**99.25**	92.60	79.54
AIGVDet	-	59.90	79.52	70.10	73.03	73.92	85.34	74.73	64.44	68.17	92.78	90.60	70.23	74.92	67.75	74.67
Demamba	-	99.87	95.41	98.60	99.70	59.10	72.87	55.39	98.91	79.11	99.51	99.70	48.89	97.08	76.98	84.37
D3 (Cosine Sim)	DINOv2-L	94.18	80.47	91.56	96.75	82.92	79.97	84.64	91.52	77.66	94.98	94.87	87.89	87.39	77.56	87.31
	CLIP-B-16	92.93	88.32	95.52	97.76	79.16	84.45	88.54	88.89	82.36	96.43	97.38	89.67	89.73	86.31	89.82
	XCLIP-B-16	96.80	89.43	96.25	98.19	90.10	88.73	91.41	95.29	86.37	96.01	96.77	93.37	92.91	87.77	92.81
	CLIP-B-32	93.19	87.28	93.83	96.17	82.39	84.71	89.21	87.74	83.37	95.21	95.30	90.43	88.81	85.75	89.53
	XCLIP-B-32	96.13	88.95	95.34	97.61	89.58	88.60	91.21	93.60	86.05	95.41	96.10	93.10	92.27	88.37	92.31
	ResNet18	96.07	86.50	95.28	97.84	88.69	83.76	89.53	91.84	85.22	97.31	96.72	93.79	91.27	89.20	91.64
	VGG16	97.51	90.16	97.24	99.16	92.30	88.26	92.87	92.53	89.02	98.66	98.16	96.01	93.07	93.34	94.16
	EfficientNet-b4	92.11	82.12	92.58	96.06	84.11	82.59	86.19	89.40	80.85	93.88	93.87	89.74	87.93	81.62	88.08
	MobileNet-v3	94.31	83.05	94.02	97.10	86.32	81.18	86.68	89.14	82.22	95.92	95.31	91.85	88.52	87.41	89.50
D3 (L2 Distance)	DINOv2-B	98.79	95.28	98.38	99.49	95.76	92.99	95.64	97.60	92.46	99.19	99.17	97.94	96.33	95.62	96.76
	DINOv2-L	98.09	93.95	97.58	99.22	94.68	91.87	94.74	97.03	91.03	98.88	98.63	96.79	95.39	93.95	95.84
	CLIP-B-16	98.20	98.03	**99.20**	99.58	94.65	95.46	97.21	96.90	94.92	99.34	99.54	98.03	97.58	98.19	97.63
	XCLIP-B/16	99.16	97.78	99.10	99.56	97.65	**96.82**	97.78	98.67	**96.03**	99.08	99.28	98.64	98.15	97.66	**98.24**
	CLIP-B-32	98.10	97.55	98.47	99.02	95.21	95.39	97.11	96.23	94.88	98.77	98.85	97.83	96.79	97.46	97.26
	XCLIP-B-32	98.86	97.04	98.67	99.33	96.95	96.28	97.33	97.81	95.22	98.68	98.90	98.22	97.61	97.12	97.72
	ResNet18	98.66	95.77	98.65	99.43	96.88	93.84	96.51	96.76	93.74	99.31	99.06	98.58	96.90	97.52	97.26
	VGG16	98.55	96.02	99.12	99.61	**98.42**	95.94	**97.95**	95.29	94.62	99.63	99.46	**99.33**	96.64	**99.23**	97.84
	EfficientNet-b4	97.18	93.46	97.55	98.82	94.24	92.78	94.77	95.79	91.47	97.98	98.01	96.56	95.31	93.02	95.49
	MobileNet-v3	98.22	94.61	98.32	99.26	95.88	93.06	95.62	96.02	92.32	98.87	98.76	97.92	95.76	96.18	96.48

Table 5. Detection results (AUC) on 14 EvalCrafter datasets.

Detection Method	Visual Encoder	Datasets (AUC↑)										Avg. AUC
		LaVie	OpenSora	CogVideoX-5B	CogVideoX	Dream-Machine	Gen-2	Pika	SVD	VC2	ZeroScope	
FID	-	96.03	86.03	91.72	93.37	97.62	98.11	99.59	95.81	95.68	89.38	94.33
NPR	-	42.90	83.50	73.60	72.10	**99.70**	**99.80**	**99.80**	**99.50**	47.20	52.90	77.10
AIGVDet	-	71.25	69.57	67.05	78.13	68.40	75.87	97.83	72.87	81.36	77.15	75.95
Demamba	-	63.66	51.32	61.44	59.65	98.42	99.64	99.19	97.72	63.57	46.56	74.12
D3 (Cosine Sim)	DINOv2-B	71.96	61.06	79.86	71.10	90.29	97.37	94.27	84.48	88.12	82.91	82.1
	DINOv2-L	67.98	55.12	77.39	66.43	88.32	95.88	92.41	82.80	86.07	78.78	79.12
	CLIP-B-16	77.60	70.97	85.63	82.18	92.92	97.30	94.57	89.64	89.81	81.73	86.24
	XCLIP-B-16	82.27	79.23	87.18	84.76	93.09	98.01	95.18	88.82	92.09	90.37	89.10
	CLIP-B-32	78.86	78.44	85.69	84.49	91.61	95.87	93.10	88.37	88.23	85.78	87.04
	XCLIP-B-32	83.04	80.27	86.74	84.48	91.77	97.43	94.94	87.88	90.61	90.19	88.74
	ResNet18	78.19	75.19	82.87	81.16	92.45	97.40	96.12	84.67	89.95	90.35	86.83
	VGG16	85.49	83.49	88.07	88.32	96.02	98.80	98.01	86.93	93.39	93.55	91.21
	EfficientNet-b4	75.65	73.72	77.60	77.76	86.70	94.19	90.82	80.28	84.03	83.25	82.40
	MobileNet-v3	75.40	73.83	80.20	79.76	89.58	96.34	94.58	82.24	86.59	88.54	84.71
D3 (L2 Distance)	DINOv2-B	90.32	87.61	93.79	90.39	97.56	99.48	98.76	89.72	96.78	95.41	93.98
	DINOv2-L	88.08	84.64	92.40	87.88	96.58	99.13	98.03	88.60	95.80	93.80	92.49
	CLIP-B-16	95.19	93.97	97.41	96.18	99.00	99.45	98.93	95.75	98.14	96.08	97.01
	XCLIP-B/16	95.51	95.45	97.01	96.48	98.54	99.54	98.93	93.65	98.31	98.00	97.14
	CLIP-B-32	94.89	95.12	96.79	95.98	98.29	98.80	98.13	94.22	97.37	96.53	96.61
	XCLIP-B-32	95.04	94.65	96.18	95.47	97.76	99.31	98.60	91.67	97.52	97.27	96.35
	ResNet18	93.08	94.58	95.01	95.35	97.99	99.32	98.86	87.33	97.23	97.93	95.67
	VGG16	**97.07**	**98.61**	**97.75**	**98.76**	99.28	99.62	99.52	86.27	**98.65**	**99.48**	**97.50**
	EfficientNet-b4	90.17	89.23	91.48	91.00	95.33	98.02	96.52	85.51	94.50	92.83	92.46
	MobileNet-v3	92.15	93.09	94.05	94.38	96.97	98.95	98.27	86.51	96.26	96.40	94.70

Table 6. Detection results (AUC) on 10 VideoPhy datasets.

Detection Method	Visual Encoder	Datasets (AUC↑)						Avg. AUC
		MSE	OS	Pika	ST2V	T2VZ	VC2	
FID	-	89.86	87.36	99.62	98.10	67.01	85.65	87.93
NPR	-	82.61	**98.56**	**99.84**	**98.92**	**93.32**	56.70	**88.33**
AIGVDet	-	60.11	45.27	48.75	34.14	59.96	46.21	49.07
Demamba	-	54.44	84.02	99.29	84.94	76.34	78.61	79.61
D3 (Cosine Sim)	DINOv2-B	73.08	82.89	93.50	59.99	38.77	91.14	73.23
	DINOv2-L	68.07	80.05	91.90	58.23	36.96	89.76	70.83
	CLIP-B-16	78.88	89.30	93.97	68.69	37.68	86.11	75.77
	XCLIP-B-16	80.72	89.26	94.80	73.38	46.98	94.06	79.87
	CLIP-B-32	81.55	89.32	93.19	67.59	36.30	85.16	75.52
	XCLIP-B-32	81.93	89.65	94.21	73.02	46.65	92.26	79.62
	ResNet18	80.85	87.94	95.94	65.30	33.51	90.57	75.68
	VGG16	87.49	92.29	97.81	73.10	19.35	92.06	77.02
	EfficientNet-b4	76.28	84.25	91.40	62.04	37.07	86.93	73.00
	MobileNet-v3	78.70	85.94	94.21	60.81	34.94	87.87	73.74
D3 (L2 Distance)	DINOv2-B	92.09	94.60	98.54	81.52	28.89	97.39	82.17
	DINOv2-L	89.74	93.22	97.98	79.67	28.26	96.51	80.90
	CLIP-B-16	95.98	97.67	98.86	89.06	31.73	95.47	84.79
	XCLIP-B/16	96.04	97.41	98.93	91.91	39.73	**98.45**	87.08
	CLIP-B-32	95.91	97.33	98.35	88.82	28.26	95.13	83.97
	XCLIP-B-32	95.30	96.92	98.39	90.64	34.81	97.36	85.57
	ResNet18	94.22	96.21	98.99	85.02	18.61	96.52	81.59
	VGG16	**97.78**	97.93	99.45	88.78	8.39	96.93	81.54
	EfficientNet-b4	90.48	93.64	97.05	81.27	27.40	94.53	80.73
	MobileNet-v3	92.73	94.95	98.35	80.49	22.70	95.32	80.76

Table 7. Detection results (AUC) on 6 VidProm datasets.

References

[1] Davide Alessandro Coccomini, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. Mintime: Multi-identity size-invariant video deepfake detection. *IEEE Transactions on Information Forensics and Security*, 2024. 1

[2] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 973–982, 2023. 2

[3] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[4] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019. 2

[5] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection. *arXiv preprint arXiv:2312.10461*, 2023. 1

[6] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 2

[7] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023. 1

[8] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017. 2