Beyond Generation: A Diffusion-based Low-level Feature Extractor for Detecting AI-generated Images

Supplementary Material

1. Network architecture

We employ a ViT-like network as the backbone of the feature extractor. The standard ViT splits a spatial image into multiple patches and then directly flattens them into vectors. However, this operation destroys the spatial pixel correlation of each patch, which is crucial to low-level feature extraction. Therefore, we first apply a set of high-pass filters to each patch and cascade multiple convolution layers. The specific kernels of high-pass filters are shown in Fig. 2. This set of high-pass filters is proposed by SRM [1]. These high-pass filters are widely adopted as a preprocessing module in forensic-related studies. The pipeline of the feature extractor is shown in Fig. 1.

2. Detection performance analysis

Besides accuracy, we employ recall rate and false alarm rate to comprehensively analyze the detection performance of our method. The recall rate is defined as the proportion of actual AI-generated images that are accurately detected by the detector. The false alarm rate is defined as the proportion of real photographs that are mistakenly classified as AI-generated by the detector. A good classifier aims to simultaneously achieve a high recall rate and a low false alarm rate. Table 1 and Table 2 show the experimental results of GenImage and DRCT-2M, respectively. In terms of DRCT-2M set, we find that our detector can identify most AI-generated images with a 100% recall rate. Meanwhile, all subsets of DRCT-2M share the same 5000 real photographic images from MSCOCO, which results in a consistent false alarm rate. Therefore, our detector obtains a consistent 97.13% accuracy.

3. Ablation studies

The number of photographic images serves as a crucial hyperparameter. In previous experiments, we adopt 10000 photographic images to estimate the distribution of their low-level features. We take GenImage dataset as an example to investigate the impact of the number of photographic images. The original GenImage provides 162000 real photographic images from ImageNet. We vary the number of photographic images from 2000 to 18000. As shown in Fig 3, we find that our method only requires 10000 photographic images to achieve approximately 96% detection accuracy. Meanwhile, the accuracy tends to stabilize once the number of photographic images exceeds 10,000.

During the detection phase, we utilize an adaptive threshold to identify AI-generated images. In previous experiments, we set the threshold as the 99.95th percentile of all likelihood scores from the training samples in descending order. The threshold should be set lower than the likelihood scores of most real photographic images to minimize the risk of false alarms. Table 3 illustrates the impact of the threshold on the detection performance. Setting a high threshold increases the recall rate but also raises the false alarm rate.

Furthermore, we also vary the number of Gaussian components w used in the Gaussian Mixture Model, where we set $w = 6$ in previous studies. We still adopt GenImage as an example. As shown in Fig. 4, the final average detection accuracy continues to be greater than 95% when w is less than 6. However, the accuracy decreases when w is too large, resulting in overfitting.

References

[1] Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 7(3):868–882, 2012. 1

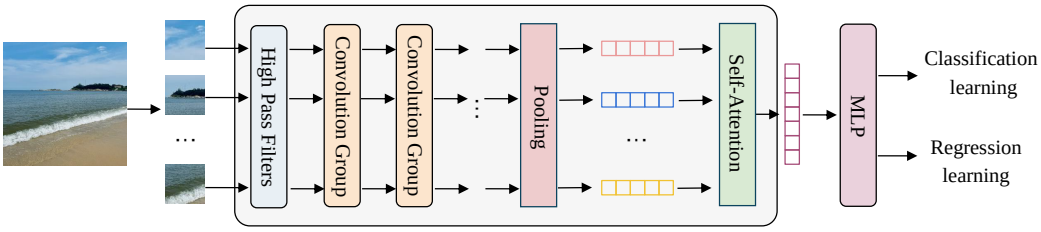

Figure 1. The pipeline of our feature extractor. A spatial image is split into multiple patches. We employ high-pass filters and convolution groups, including a convolution layer, an activation function and a normalization layer, to refine the pixel correlation. Finally, we flatten the feature maps and feed them into the self-attention module to fuse each patch.

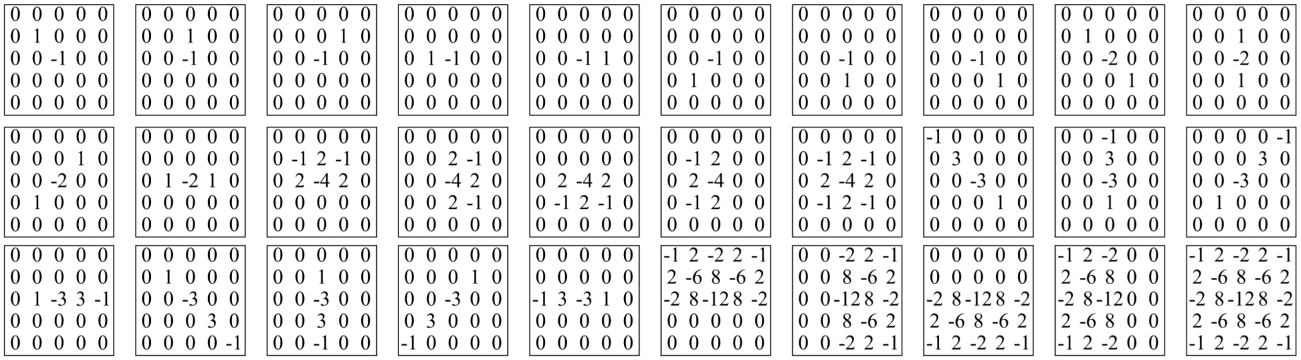

Figure 2. The specific kernel parameters of high-pass filters.

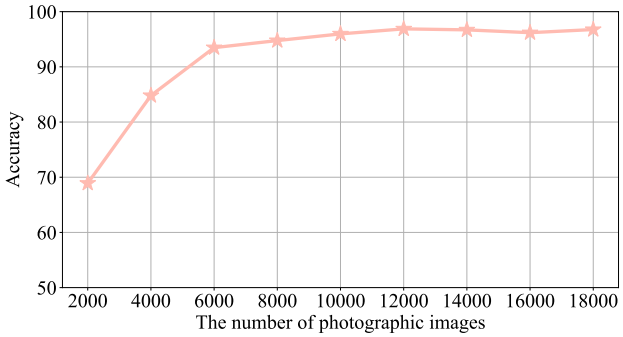

Figure 3. The impact of the number of real photographic images. Accuracy denotes the average detection accuracy across eight generative models from the GenImage set.

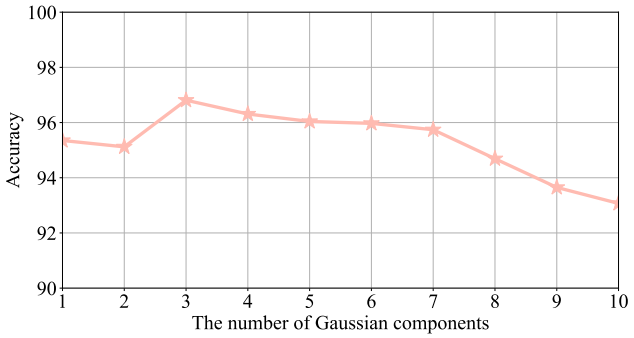

Figure 4. The impact of the number of Gaussian components. Accuracy denotes the average detection accuracy across eight generative models from the GenImage set.

Method	Metric	Midjourney	SDv1.4	SDv1.5	ADM	Glide	Wukong	VQDM	BigGAN	Average
Ours-basic	Acc	95.51	96.06	96.33	95.88	96.00	96.36	95.95	95.71	95.97
	Recall	98.82	100.00	99.99	99.77	99.93	100.00	100.00	99.63	99.77
	False Alarm	7.80	7.88	7.34	8.02	7.93	7.28	8.10	8.22	7.82
Ours-ft	Acc	95.03	95.32	95.53	95.29	95.17	95.58	95.10	95.15	95.27
	Recall	99.63	100.00	100.00	99.97	99.97	100.00	100.00	99.88	99.93
	False Alarm	9.55	9.37	8.94	9.38	9.62	8.85	9.80	9.58	9.39

Table 1. Detection performance analysis on GenImage set.

			SD Variants					Turbo Variants		LCM Variants		ControlNet Variants			DR Variants			Avg.
Method	Metric	LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	SDv1-DR	SDv2-DR	SDXL-DR	
Ours-basic	Acc	97.37	97.60	97.54	97.65	97.64	97.67	97.62	97.42	97.61	97.64	50.33	50.14	52.43	92.75	52.31	50.08	82.74
	Recall	99.4	99.86	99.74	99.96	99.94	100	99.9	99.5	99.88	99.94	5.32	4.94	9.52	90.16	9.28	4.82	70.135
	False Alarm	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66	4.66
Ours-ft	Acc	97.13	97.13	97.13	97.13	97.13	97.13	97.13	97.13	97.13	97.13	97.08	96.81	94.73	92.26	51.79	49.73	90.86
	Recall	100	100	100	100	100	100	100	100	100	100	99.9	99.36	95.2	90.26	9.32	5.2	87.45
	False Alarm	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74	5.74

Table 2. Detection performance analysis on DRCT-2M set.

Threshold	Model	Midjourney	SDv1.4	SDv1.5	ADM	Glide	wukong	VQDM	BigGAN	Average
99.99th	Acc	95.53	96.08	96.34	95.88	96.02	96.37	95.96	95.73	95.99
	Recall	98.82	100.00	99.99	99.77	99.93	100.00	100.00	99.63	99.77
	False Alarm	7.77	7.85	7.30	8.00	7.90	7.27	8.08	8.18	7.79
99.95th	Acc	95.51	96.06	96.33	95.88	96.00	96.36	95.95	95.71	95.97
	Recall	98.82	100.00	99.99	99.77	99.93	100.00	100.00	99.63	99.77
	False Alarm	7.80	7.88	7.34	8.02	7.93	7.28	8.10	8.22	7.82
99.00th	Acc	94.47	94.83	95.02	94.74	94.73	95.18	94.86	94.83	94.83
	Recall	99.33	100.00	100.00	99.85	99.98	100.00	100.00	99.88	99.88
	False Alarm	10.38	10.35	9.96	10.37	10.52	9.63	10.28	10.23	10.22
98.00th	Acc	94.19	94.43	94.60	94.43	94.32	94.85	94.41	94.42	94.46
	Recall	99.38	100.00	100.00	99.90	99.98	100.00	100.00	99.90	99.90
	False Alarm	11.00	11.13	10.80	11.03	11.35	10.30	11.18	11.07	10.98

Table 3. We conduct ablation studies on the threshold, varying it from the top 99.99% likelihood score of the training set down to the top 98.00%.