Any-Resolution AI-Generated Image Detection by Spectral Learning

Dimitrios Karageorgiou, Symeon Papadopoulos,
Ioannis Kompatsiaris, Efstratios Gavves

CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS

Information
Technologies
Institute

Too many generative models, too brittle artifacts!

An abundance of image generation approaches is currently available, while more get released daily!

- High-resolution & topic-agnostic generation
- Different architectural families (GANs, Diffusion Models etc.)
- Open-source & commercial approaches
- Low-Rank Adaptation (LoRA) fine-tuning by anyone!

Stable Diffusion 3

GigaGAN

Midjourney v6.1

DALLE3

Any assumptions about the artifacts introduced to the image signal quickly become obsolete!

- Visible artifacts are fixed in newer models.
- Low-level artifacts differ among models with minimal differences.

→ AI-Generated Image Detection approaches fail to generalize on unseen generative approaches!

Spectral domain provides significant discriminative power

- Recent works have established that generative models introduce strong spectral artifacts to the generated images.
- However, they significantly differ even among models with minor differences.

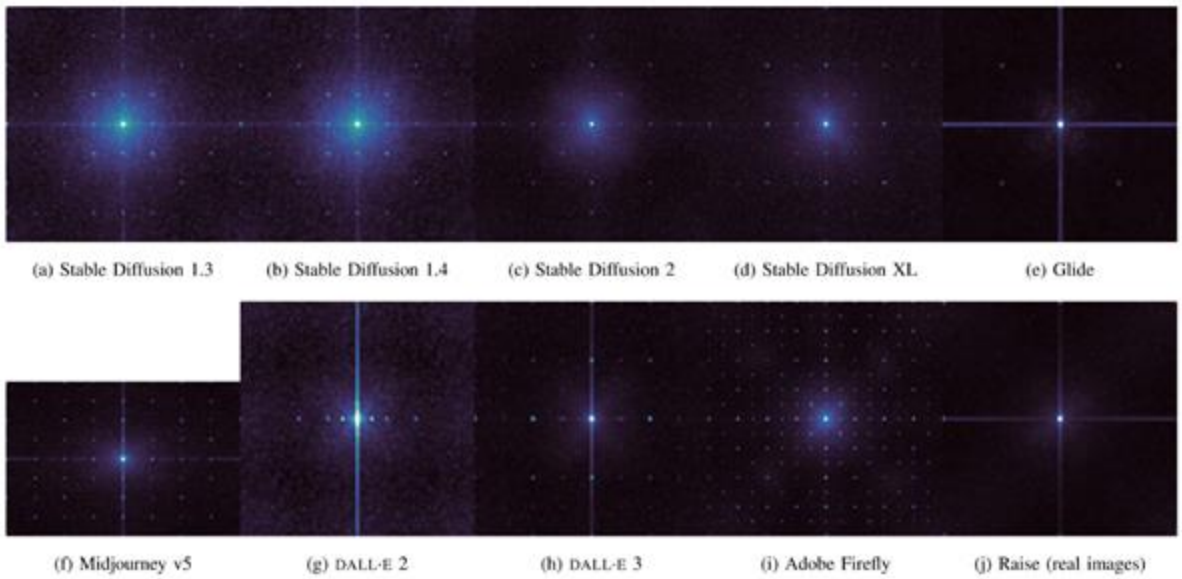
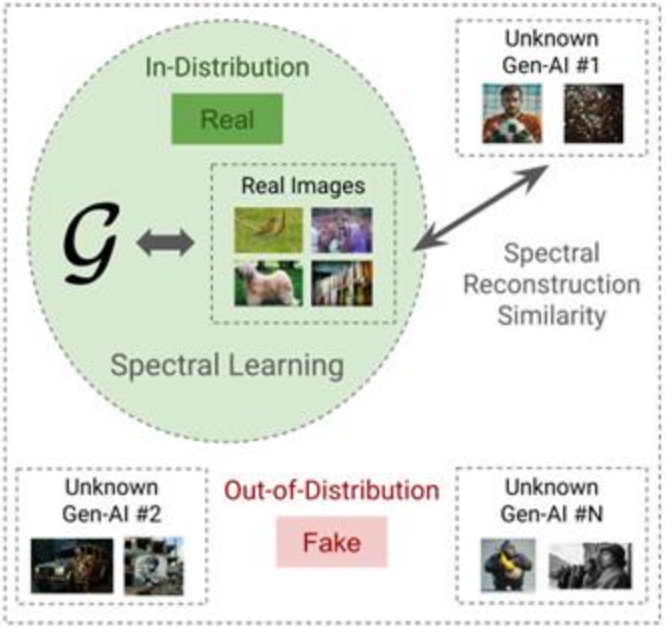


Figure from Bammey, Q. (2023). Synthbuster: Towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing.

SPAI: Any-Resolution AI-Generated Image Detection by Spectral Learning

Key Idea: The spectral distribution of real images constitutes an invariant and highly-discriminative pattern for the task of AI-Generated Image Detection.

→ **Corollary**: Given a model of the spectral distribution of real images, AI-Generated images can be detected as Out-Of-Distribution (OOD) samples of this model.

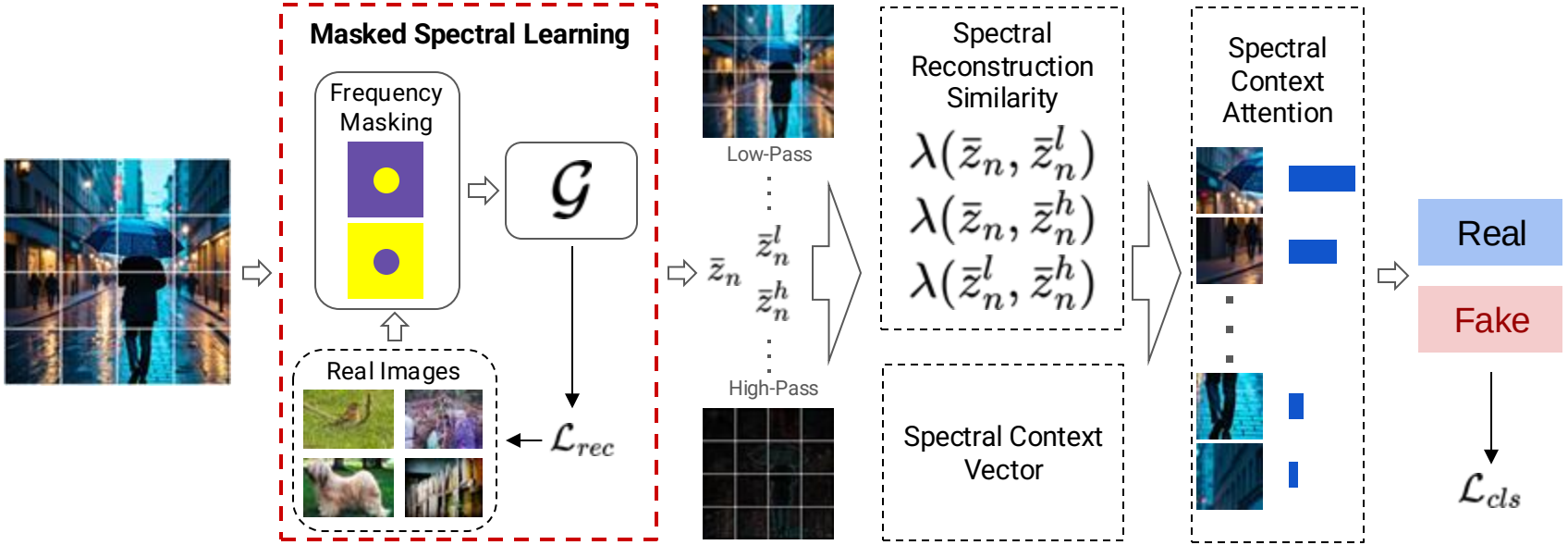


SPAI uses:

Frequency Reconstruction pre-text task to model the spectral distribution of real images.
Spectral Reconstruction Similarity to detect Gen-AI images as OOD samples of this model.
Spectral Context Attention to capture subtle spectral inconsistencies in any-resolution images.

Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E. (2024). Any-Resolution AI-Generated Image Detection by Spectral Learning

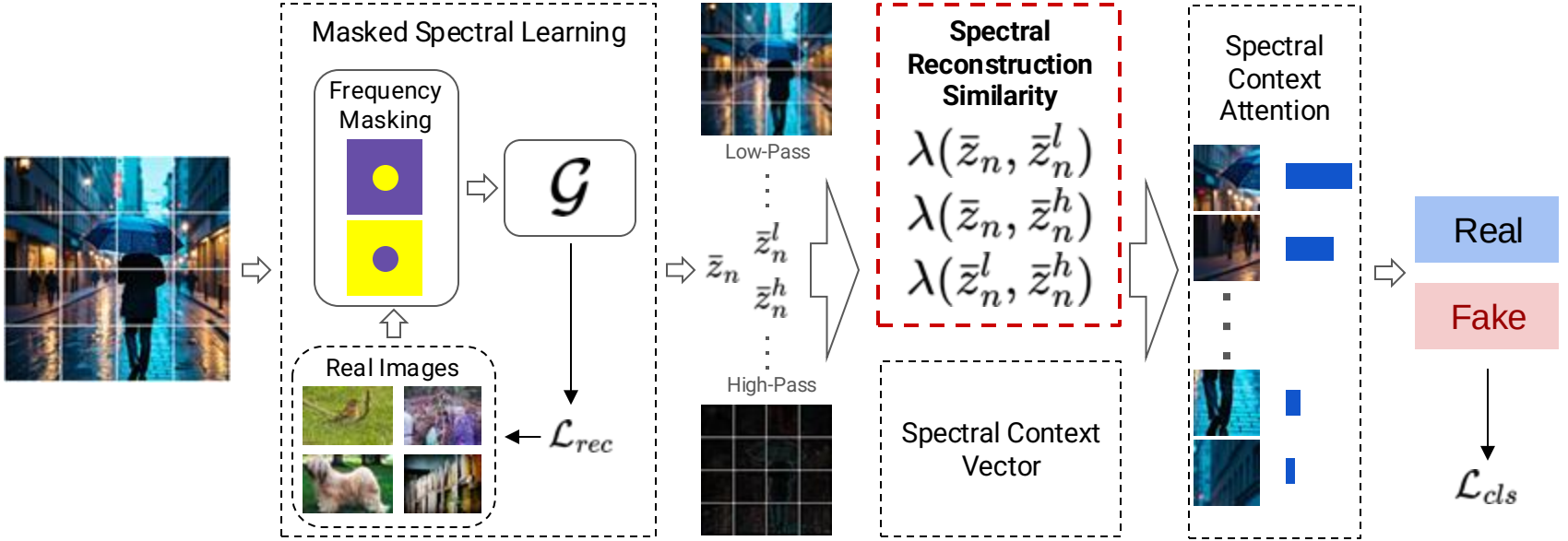
Masked Spectral Learning: Learning the spectral distribution of real images



- Self-supervised training on real images using the pre-text task of frequency reconstruction.
- Inputs are generated by low/high frequency filtering. – Model reconstructs missing frequencies.
- Reconstruction loss is computed on the DFT domain.
- A vision transformer is used for the model $\mathcal{G}$.

Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E. (2024). Any-Resolution AI-Generated Image Detection by Spectral Learning

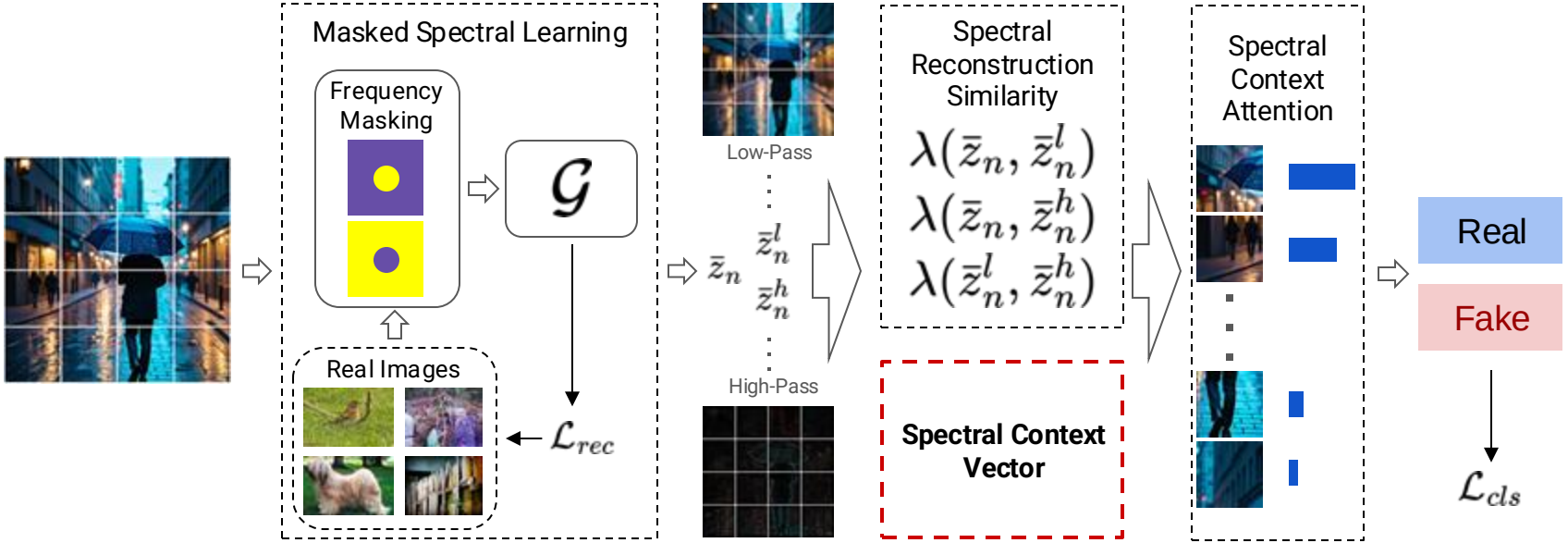
Spectral Reconstruction Similarity (SRS): Detecting OOD images



- Low- & high-pass filtered images are embedded using the learned spectral model.
- Cosine similarity among the three pairs of original, low-pass and high-pass filtered images.
- Spectral reconstruction similarity is computed for the features of each transformer block of $\mathcal{G}$.

Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E. (2024). Any-Resolution AI-Generated Image Detection by Spectral Learning

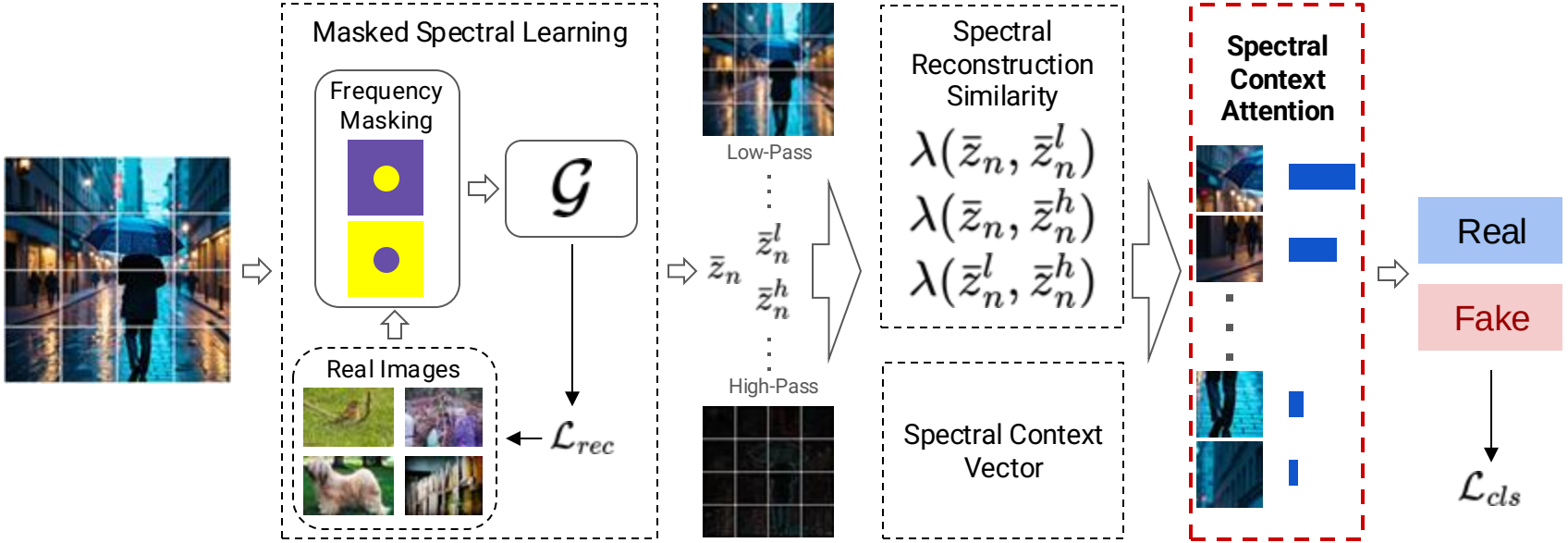
Spectral Context Vector: Understanding which values of SRS are useful



- All the values of Spectral Reconstruction Similarity (SRS) are not always equally useful.
- E.g. reconstructing high frequencies of images without high-frequency content yield limited info.
- Spectral Context Vector uses an attention mechanism to learn to encode the context of the image.

Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E. (2024). Any-Resolution AI-Generated Image Detection by Spectral Learning

Spectral Context Attention: Embedding arbitrary resolution images



- Image is split into patches (ViT resolution – 224x224).
- The most discriminative SRS values according to the spectral context of each patch are considered.
- Subtle details are captured, as images are processed in their native resolution, with linear complexity.

Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E. (2024). Any-Resolution AI-Generated Image Detection by Spectral Learning

Comparison against state-of-the-art

Generalization on generators of different architectures, resolution, image quality, open-source & commercial.

Image Size	< 0.5 MPixels			0.5 - 1.0 MPixels						> 1.0 MPixels				AVG
Approach	Glide	SD1.3	SD1.4	Flux	DALLE2	SD2	SDXL	SD3	GigaGAN	MJv5	MJv6.1	DALLE3	Firefly	
NPR [66]	72.2	89.6	60.5	19.8	3.9	12.5	18.1	60.6	83.2	15.3	19.8	97.1	38.0	45.4
Dire [72]	33.3	59.9	61.3	45.7	52.2	68.5	46.9	49.2	36.3	41.9	50.3	65.2	49.9	50.8
CNNDet. [71]	59.2	59.0	61.2	39.8	71.5	57.5	67.4	30.2	73.4	48.8	56.7	23.5	73.4	55.5
FreqDet. [23]	43.6	92.3	92.7	36.5	47.4	42.5	66.5	69.8	63.2	36.9	27.5	42.2	80.9	57.1
Fusing [34]	63.0	62.8	62.2	57.5	76.7	66.9	62.1	38.8	80.4	64.0	74.0	25.2	76.3	62.3
LGrad [65]	76.5	82.4	83.4	74.9	85.7	60.7	70.2	12.7	89.9	69.2	79.6	30.0	42.0	65.9
UnivFD [52]	63.3	80.8	81.2	36.3	91.4	84.3	78.3	28.6	86.2	57.1	60.5	31.0	95.5	67.3
GramNet [48]	78.2	83.9	84.3	78.6	85.2	66.7	77.8	19.2	85.0	63.8	84.9	42.9	38.0	68.4
DeFake [63]	86.1	64.2	63.6	90.5	41.4	66.2	52.3	87.7	71.7	67.0	87.5	93.3	39.4	70.1
PatchCr. [77]	78.4	95.7	96.2	86.9	81.8	95.7	96.7	33.8	98.0	79.0	96.1	28.1	79.1	80.4
DMID [7]	73.1	100.0	100.0	97.2	54.3	99.7	99.6	67.9	67.9	99.9	94.4	41.3	90.2	83.5
RINE [39]	95.6	99.9	99.9	93.0	93.0	96.6	99.3	39.1	92.9	96.4	81.2	41.8	82.9	85.5
SPAI (Ours)	90.2	99.6	99.6	83.0	91.1	96.5	97.4	75.9	85.4	94.5	84.0	90.2	96.0	**91.0**

Table 1. Comparison against state-of-the-art. Average AUC over 5 sources of real images is reported. Lower values are highlighted in red, while higher values are highlighted in green. Best overall average value is highlighted in bold, while second best is underlined. Our approach generalizes across all the considered generative approaches, even on ones producing imagery of extreme fidelity, such as SD3, where the single method [63] that scores better was required to explicitly train on relevant data.

Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E. (2024). Any-Resolution AI-Generated Image Detection by Spectral Learning

Ablations Studies & Robustness Against Online Perturbations

Ablation		AUC
SPAI (Ours)		**91.0**
Components	w/o SRS	71.0
	w/o SCV	84.9
	w/o SCA	83.2
	w/o SCA + TenCrop (mean)	85.3
	w/o SCA + TenCrop (max)	84.2
Augm.	w/o JPEG compression	89.1
	w/o distortions	84.2
	with WebP compression	89.3
	with chromatic augm.	80.5

Table 3. Ablation studies of the key components. Average AUC over 5 sources of real images and 13 generative models is reported. Best value is highlighted in bold.

The key design choices of SPAI are crucial for reaching the reported performance. In particular, when removing the Spectral Reconstruction Similarity performance plunges, highlighting the importance of this key idea!

The frequency reconstruction pre-text task greatly enhances the detection performance, using only a fraction of the data of popular encoders trained on spatial pre-text tasks.

Backbone	# Training Data	AUC
CLIP ViT-B/16 [53]	400 million	87.6
DINOv2 ViT-B/14 [56]	142 million	87.5
MFM ViT-B/16 (Ours) [73]	1.2 million	**91.0**

Table 2. Evaluation of different backbones. Average AUC over 5 sources of real images and 13 generative models is reported. Best value is highlighted in bold.

SPAI achieves superior robustness to online perturbations.

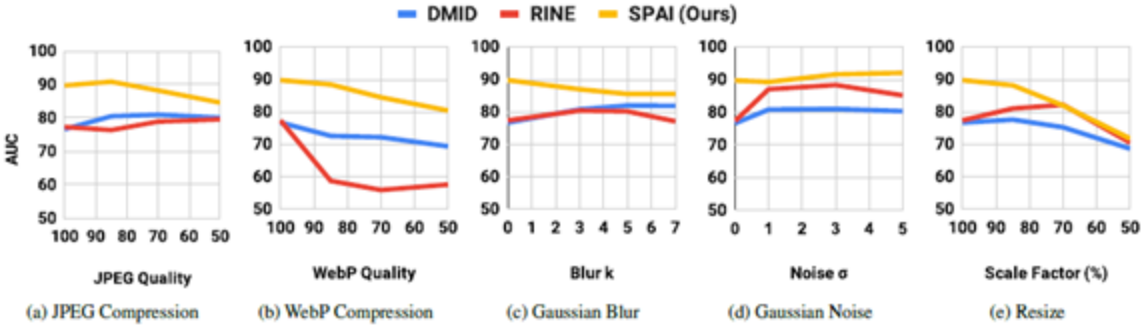


Figure 3. Robustness evaluation on common perturbations. Average AUC is presented over the perturbed versions of two sources of authentic images from smartphones and DSLR cameras respectively and 13 generative models.

Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E. (2024). Any-Resolution AI-Generated Image Detection by Spectral Learning

Beyond binary detection and Open Challenges

Spectral context attention natively provides a mechanism to understand which regions of the image were more important for the final decision.

6-fingers case correctly spotted

Attending texture-rich regions.

Figure 4. Qualitative analysis of spectral context attention. A cool-warm overlay has been applied on each patch. Red color indicates significant patches for deciding whether the image is AI-generated (high attention values), while blue color indicates irrelevant patches (low attention values). The attention values have been normalized in $[0, 1]$.

AI-Generated content commonly appears online in the form of derivative images, i.e. screenshots of posts, photos of a screen etc.

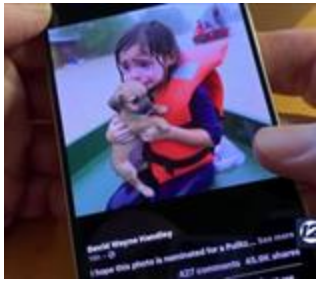
The intermediate medium (digital or analog) heavily distorts the spectral distribution of the AI-generated images.

Detecting such images remains an open issue for any detector that relies on the image signal.

Detection: 86%

Detection: 79%

Detection: 0%



Detection: 2%