Bridging the Gap Between Ideal and Real-world Evaluation: Benchmarking AI-Generated Image Detection in Challenging Scenarios

Supplementary Material

Supplementary Overview

Dataset: https://zenodo.org/records/14963880.
Code: https://github.com/ChunXiaostudy/RRDataset.
Weights: https://zenodo.org/records/14991882.
The following provides an overview of the content in each Appendix section:

- **Section A**: Comprehensive Introduction of AI-Generated Image Detectors.
- **Section B**: Real-World Applications of Transmission and Re-digitization.
- **Section C**: Training Details.
- **Section D**: Additional Results for RRBench.
- **Section E**: Comparison of Human Attention Regions.
- **Section F**: Frequency Domain Analysis.
- **Section G**: Visualization of RRDataset.
- **Section H**: Further Discussion and Suggestion.
- **Section I**: Limitation of RRDataset and RRBench.

A. Comprehensive Introduction of AI-Generated Image Detectors

Based on the detection criteria, the methods for detecting AI-generated images can be broadly categorized into the following approaches: detection based on pixel-domain features, detection based on frequency-domain features, detection leveraging pre-trained model features, detection utilizing fused features, as well as rule-based detection methods. Detection based on pixel-domain features of images identifies authentic or synthetic images through common patterns in pixel-level characteristics, which primarily encompasses approaches utilizing steganalysis features [42], texture features [18, 44, 47], color features [5, 58], local features [3, 9, 26], reconstruction features [4, 7, 11, 19, 34, 40, 41, 49, 60, 69], internal correlation features [8, 56, 70] and physical features [50]. Besides, Tan et al. [53] first proposed a gradient-based detection method (LGrad) that employs gradient maps derived from pretrained CNN models. This method trains a binary classifier on gradient map datasets to distinguish AI-generated images from authentic ones. By eliminating the influence of image semantics on the detector, this approach effectively mitigates the model's dependency on training data. However, due to its reliance on specific pretrained architectures, LGrad demonstrates limited generalizability when applied to images generated by non-CNN-based models.

Detection based on frequency domain characteristics of images primarily identifies artificial patterns by analyzing common characteristics exhibited in the frequency domain of generated images. This approach typically converts images into spectrum maps for subsequent analysis through detection networks. Representative methods include: Discrete Fourier Transform (DFT)-based detection [13, 14, 22, 33, 68], Discrete Cosine Transform (DCT)-based detection [2, 16, 46], and Fast Fourier Transform (FFT)-based detection [23]. Detection based on pre-trained model features leverages large-scale models to accomplish detection tasks. Several research efforts directly utilize the powerful feature extraction capabilities of pre-trained models for detection [12, 17, 45, 66]. Notably, Ojha et al. [45] employ the pre-trained CLIP model to extract image features for detector training, enabling the learning of more balanced decision boundaries and thereby enhancing the detector's generalization across different generative models. Koutlis and Papadopoulos [30] propose RINE, which extracts intermediate image representations from the CLIP image encoder, maps them into a learnable forgery-aware vector space, and incorporates a trainable module to predict the importance of each encoder layer. Other approaches focus on fine-tuning vision-language models [6, 27–29, 39, 55, 62, 64] or large language models [15] for detection tasks. Among these, C2P-CLIP [55] introduces class-agnostic prompts to reinforce the conceptual distinction between real and generated images through contrastive learning, while implementing CLIP fine-tuning via Low-Rank Adaptation (LoRA) [20] to develop a universal detector for generated images. Detection methods based on fused features can be further categorized by modality quantity into single-modality fused feature detection methods and multi-modality fused feature detection methods. Single-modality fused feature detection methods refer to approaches that solely integrate different features within images, including pixel-domain feature fusion [1, 24, 25, 32, 48, 63], frequency-domain feature fusion [36, 54], hybrid pixel-frequency domain feature fusion [31, 37, 43, 61], and the integration of spatial features with semantic features in images [65]. Multi-modality fused feature detection methods involve the fusion of features from diverse modalities such as images and texts [35, 51, 67]. Notably, Sha et al. [51] focused on detecting text-to-image model generated content, revealing that generated images closer proximity to their prompt texts compared to real images. Leveraging this observation, they proposed DEFAKE, which concatenates CLIP-extracted textual and visual features to construct discriminative features for detection. Rule-based detection methods typically leverage the differences in sensitivity between two categories of images to the same operations as the detection criterion. These approaches generally require no training process, and since the detection procedure does not rely on classifiers, the results tend to be unbiased [10, 49, 52, 57]. In a distinct approach, Ricker et al. [49] proposed Aeroblade, which utilizes the reconstruction error of autoencoders combined with the Learned Perceptual Image Patch Similarity (LPIPS) metric to detect generated images.

In summary, regarding generative image detection, most studies employ model-based supervised learning, some utilize semi-supervised [21] or unsupervised paradigms [47], few use instance-based approach to perform detection. The features such as noise patterns in the pixel domain and frequency domain, texture, color information, and mapping geometry have turned into detection cues. Some studies aim to enhance the generalization of detection methods when faced with different generative models. Besides, robustness against common image processing also becomes an aspect of evaluating detection methods.

B. Real-World Applications of Transmission and Re-digitization

AI-generated image detection fundamentally differs from traditional image classification tasks. A key distinction is that, for traditional classification tasks such as ImageNet1K, real-world applications typically involve photos captured directly by users or devices, such as autonomous vehicle cameras. However, with the rise of social media and other platforms, an image suspected to be AI-generated may have already undergone multiple rounds of transmission or even appeared in physical form.

In real-life scenarios, the images available for detection are often subjected to lossy internet transmission through social media platforms. Therefore, incorporating such transmitted images into the test set is crucial for a robust evaluation of detection algorithms.

Regarding re-digitization, we first present real-world applications of this process to explain why re-digitization is essential for AI-generated image detection. Re-digitization refers to the process of converting an existing digital image (e.g., stored in formats such as JPG or PNG) into a physical form (e.g., printing or displaying it on a screen) and then converting it back into a digital format through methods such as scanning or photography.

Re-digitization has numerous real-world applications, including the following:

- Capturing photos of images displayed on devices, where exporting images is not permitted.
- Photographing screens to save images that are protected against direct downloads or screenshots in web pages or Apps.
- Scanning printed materials, such as newspapers or magazines, to create electronic versions of articles or images.
- Capturing projected content during presentations when direct access to slides is unavailable.
- Photographing artwork in museums or galleries where digital copies cannot be obtained, for documentation or personal use.
- Recording street art, graffiti, posters, or other outdoor artworks for preservation or documentation.

Therefore, detection algorithms must adapt to diverse re-digitization processes, including scanned documents, photographed screens, or printed photographs. By incorporating these scenarios into benchmarks, we ensure that detection algorithms are practical and reliable across domains. Additionally, including re-digitized samples in benchmarks forces algorithms to learn more generalized features rather than relying solely on subtle artifacts in the original digital image. This enhances their performance in unknown environments and against emerging generative models.

C. Training Details

Hardware and Environment: All training and testing were conducted on a single NVIDIA RTX 4090 GPU using PyTorch 2.0.1.

Training Details: For all detectors, we adopted the training setup used in GenImage [71] and AIDE[65]. Specifically, training was conducted on GenImageSD1.4, where a subset of SDv1-4 was used, consisting of 162,000 AI-generated images from SDv1-4 and 162,000 real images from ImageNet, with an additional 6,000 AI-generated and 6,000 real images reserved for validation.

For certain methods, such as CNNSpot [59], GramNet [38], DIRE [60] and DNF [69], only pretrained weights from the ProGAN dataset were available. Therefore, we replicated their respective training settings and data augmentation procedures to ensure thorough training. For methods already trained and validated on SDv1-4, we directly utilized their publicly released pretrained weights.

Fine-Tuning Details To provide a more comprehensive and realistic evaluation of each detector, we fine-tuned all 17 detection models using a subset of RRDataset-Original. Specifically, we selected 1,250 real images and 1,250 AI-generated images for training, with an additional 250 real and 250 AI-generated images for validation.

Fine-tuning was performed using a batch size of 64 for 5 epochs with the AdamW optimizer, starting with an initial learning rate of 5×10^{-4}. We employed a learning rate decay strategy and early stopping to optimize performance.

VLM Prompt Selection: To maximize the detection performance of VLMs, we manually designed 10 different prompt templates and selected the one with the highest average accuracy for use in the main text. Notably, previous studies have rarely explored AI-generated image detection using VLMs, making it impossible to directly adopt existing settings. Therefore, our approach represents a novel effort in optimizing prompt engineering for this task.

D. Additional Results for RRBench

In this section, we present the impact of different training datasets on detector performance and compare the results of the Human-Inspired In-Context Learning Approach with other in-context learning methods.

D.1. Impact of Different Training Sets

To further evaluate the best-performing detector on RRDataset, DRCT-ConvB [7], and the most robust model, AIDE [65], we conduct additional testing on RRBench using the following model weights:

For DRCT-ConvB[7]:

1. Trained on SDv1-4, with RRDataset fine-tuning
2. Trained on SDv1-4, without RRDataset fine-tuning
3. Trained on DRCT-2M, without RRDataset fine-tuning
   - DRCT-2M is a large-scale dataset generated using advanced diffusion models, primarily SD XL. However, it differs from RRDataset, which uses SD 3.5 and FLUX as generators.
4. Trained on DRCT-2M, with RRDataset fine-tuning

For AIDE [65]:

1. Trained on SDv1-4, with RRDataset fine-tuning
2. Trained on SDv1-4, without RRDataset fine-tuning

As shown in Tab. 1, fine-tuning on RRDataset plays a crucial role in improving model performance. For DRCT-ConvB, fine-tuning on RRDataset led to a 33.67% accuracy increase when pretrained on SDv1-4 and a 21.56% increase when pretrained on DRCT-2M. Similarly, for AIDE, fine-tuning resulted in a 26.77% improvement in accuracy.

The choice of pretraining datasets also significantly impacts performance. Without fine-tuning, the accuracy gap between DRCT-ConvB trained on DRCT-2M and SDv1-4 reached 12.71%, indicating that larger and more advanced training datasets exhibit stronger transferability on RRBench. However, after fine-tuning, the accuracy difference between the two models shrank to just 0.69%, suggesting that fine-tuning on a more advanced dataset can compensate for differences in pretraining data.

This finding implies that rather than continually expanding training datasets—which can be computationally expensive in terms of generation and retraining costs—similar performance gains may be achievable through targeted fine-tuning, making it a more efficient and scalable alternative for future AI-generated image detection improvements.

D.2. Comparisons with Other In-context Learning Strategies

Our Approch: We observe a notable shift in human decision-making criteria when evaluating re-digitized and transmitted images. For re-digitization, over 25% of participants attributed their judgments to lighting and reflection-related artifacts, while only 10% considered image layout or realism as key factors. Conversely, for network transmission, factors such as sharpness, edges, and texture quality accounted for more than 52% of responses.

In designing our in-context learning approach, we aim for VLMs to filter out potential confounding factors, such as compression artifacts from transmission and moiré patterns from screens and printed surfaces. To achieve this, we developed the following in-context learning template.

Robustness-Oriented In-Context Learning

Act as a forensic image analyst specializing in origin classification. Analyze images through their intrinsic visual patterns while disregarding transmission/re-digitization artifacts. Focus on fundamental generation traces rather than secondary distortions.

Contextual Examples:

- **[Transmitted Image]** Compressed JPEG with blocking artifacts, but shows consistent micro-textures in hair strands and natural skin pore variation → **Real**
- **[Transmitted Image]** Lossy blurred image with chromatic aberration, yet reveals perfect fractal patterns in the background and asymmetric eyelash duplication → **AI-generated**
- **[Re-digitized Image]** A scanned copy with a moiré pattern, but the content and details of the picture are real → **Real**
- **[Re-digitized Image]** Photographed print showing lens glare, yet contains Diffusion-typical floating specks and impossible light intersections → **AI-generated**

Analysis Protocol:

1. **Primary Focus Areas:**
   - Microscopic texture coherence (brush strokes/sensor noise patterns)
   - Image content detail preservation beyond compression/scanning
   - Biological imperfection consistency (asymmetric irises, skin translucency)
   - Physical light interaction validity (shadow falloff, subsurface scattering, building material)
2. **Artifact Discounting:**
   - Ignore format-specific compression patterns (JPEG blocking, slight color shift)
   - Disregard scanning artifacts (dust particles, Newton rings)
   - Overlook resampling distortions (aliasing, interpolation errors)
3. **Decisive Indicators:**
   - AI-generated artifacts (e.g., diffusion-based oversmoothing, GAN-induced repetition, or hyperrealism)
   - Anatomical plausibility under 3x digital magnification
   - Material property consistency (metallic reflections, cloth draping)

Output JSON:

```
{ "classification": "AI-generated/Real" }
```

CoT prompt for Comparison:

Table 1. Impact of Different Training Datasets on Model Accuracy.

Model	Original		Transmission		Re-digitization		Overall
	Fake(%)	Real(%)	Fake(%)	Real(%)	Fake(%)	Real(%)	ACC(%)
DRCT-ConvB							
SDv14- w finetune	93.52	95.52	92.82	95.09	64.34	96.22	89.59
SDv14- w/o finetune	53.64	58.21	39.91	73.74	51.42	54.99	55.92
DRCT2M- w finetune	95.15	94.09	94.73	92.94	68.35	96.42	**90.28**
DRCT2M- w/o finetune	68.00	73.42	66.73	73.14	49.74	80.74	68.63
AIDE							
SDv14- w finetune	78.95	78.94	74.72	78.75	76.04	83.13	78.42
SDv14- w/o finetune	56.04	61.69	13.40	81.88	28.78	68.10	51.65

Chain-of-Thought (CoT) Prompt for AI-Generated Image Detection

Task: You are an expert in forensic image analysis. To determine whether an image is **AI-generated or real**, follow a structured reasoning process. Do not rely solely on immediate intuition; instead, break down your decision step by step.

Step 1: Content Understanding & Context
- What is the subject matter of the image? Describe its main elements (e.g., people, objects, background).
- Does anything seem out of place or conceptually inconsistent?

Step 2: Fine-Grained Visual Inspection
- Examine details such as **textures, edges, and transitions**. Are there any **abnormal smoothness, inconsistencies, or unnatural repetitions**?
- Evaluate lighting and reflections—do they follow physical laws?
- Consider depth perception and perspective—is everything logically structured?

Step 3: Logical Contradictions & Anomalies
- Are there elements that should logically interact but do not (e.g., missing object shadows, disconnected reflections)?
- If the image contains humans, do facial features, hands, or expressions look **artificial or anatomically incorrect**?
- Does the image resemble known AI-generated artifacts (e.g., **diffusion-based oversmoothing, GAN-induced repetition, or hyperrealism**)?

Step 4: Final Reasoning & Justification
- Summarize the strongest indicators supporting either classification.
- Make a final classification decision based on reasoning.

Output Format (JSON):

```
{ "classification": "AI-generated/Real",
  "reasoning": "Step-by-step breakdown
    of key factors influencing the decision"}
```

Initial prompt for Comparison:

Prompt for Vision-Language Models

Act as an expert in computational photography and generative AI. Analyze the visual characteristics to classify its origin as either real-world captured or AI-generated.
Your analysis should:
- Examine technical artifacts (unnatural textures, perfect symmetry, atypical shadow patterns)
- Check for common GAN/diffusion model fingerprints
- Evaluate biological plausibility (eyes, hair, skin textures)
- Identify hyperrealistic elements vs. physical-world imperfections
Format response as JSON: {"classification": "AI-generated/Real" }

Comparison Results: Tab. 2 presents a comparison of four different prompting strategies. Our human-inspired Robustness-Oriented In-Context Learning achieves the best performance on 3 out of 4 VLMs, demonstrating its effectiveness in enhancing detection robustness, especially in challenging real-world conditions such as internet transmission and re-digitization.

The CoT+In-Context Learning approach achieves notable improvements on original images, suggesting that explicit reasoning chains help VLMs analyze fine details and generative traces more accurately when image quality is intact. However, its performance deteriorates significantly on transmitted and re-digitized images, indicating a strong reliance on pristine image quality. This suggests that CoT-based reasoning alone is insufficient to mitigate the impact of information loss, compression artifacts, and noise distortions introduced in real-world scenarios.

In contrast, our approch not only ensures higher robustness but also achieves the best overall performance across different VLMs, demonstrating that a robustness-oriented learning strategy is more effective in guiding VLMs to maintain high detection accuracy, even under degraded conditions.

E. Comparison of Human Attention Regions

E.1. Human Attention Map Generation

Data Collection: In this experiment, we select a set of images $\{I_1, I_2, \ldots, I_K\}$ for analysis. Each participant views the images on an screen and marks the Regions of Interest (ROIs) by drawing closed contours with a stylus. To ensure consistency, the following conditions should be met: The displayed images should have the same resolution and scaling across all participants.

ROI Data Recording: Suppose there are P participants. For each image I_k, participant p (where $p = 1, \ldots, P$) draws one or more closed polygons that represent their ROIs. We denote these polygons by

$$\{\Omega_{p,k,1}, \Omega_{p,k,2}, \ldots\}.$$

Table 2. Performance Comparison for Human-Inspired Robustness-Oriented In-Context Learning.

	Original	Transmission	Redigital	ACC
Initial prompt				
GPT-4o-latest	92.94	84.71	73.08	83.58
Claude-3.7-sonnet	89.85	83.72	73.87	82.48
Gemini-2-flash	85.27	74.80	71.76	77.28
Grok-2-vision	68.99	**73.08**	64.82	68.96
Initial prompt + In-Context Learning (2-shot)				
GPT-4o-latest	91.85	84.06	74.55	83.49
Claude-3.7-sonnet	87.10	83.95	75.22	82.09
Gemini-2-flash	85.84	**75.12**	75.27	78.74
Grok-2-vision	67.28	72.86	66.71	68.95
CoT prompt + In-Context Learning (2-shot)				
GPT-4o-latest	**96.27**	83.58	71.45	83.77
Claude-3.7-sonnet	91.28	84.29	73.92	83.16
Gemini-2-flash	**88.79**	73.52	**78.31**	**80.21**
Grok-2-vision	71.02	71.47	65.93	69.47
Robustness-Oriented In-Context Learning (2-shot)				
GPT-4o-latest	95.67	**88.17**	**78.58**	**87.47**
Claude-3.7-sonnet	**92.26**	**84.97**	**77.76**	**85.00**
Gemini-2-flash	88.38	74.99	75.78	79.72
Grok-2-vision	**72.86**	71.52	**71.43**	**71.94**

Each polygon $\Omega_{p,k,j}$ can be represented by a set of vertex coordinates, for example:

$$\Omega_{p,k,j} = \Big\{(x_{p,k,j,1}, y_{p,k,j,1}), \dots, (x_{p,k,j,n_j}, y_{p,k,j,n_j})\Big\}.$$

Constructing the Binary Interest Mask: Each image I_k has a size of $W \times H$. We define a matrix $M_{p,k}$ to represent participant p's interest area on image I_k. Let the final set of aligned ROIs for participant p on I_k be $\{\Omega^*_{p,k,1}, \Omega^*_{p,k,2}, \dots\}$. For any pixel (x, y), we have:

$$M_{p,k}(x,y) = \begin{cases} 1, & \text{if } (x,y) \in \Omega^*_{p,k,1} \cup \Omega^*_{p,k,2} \cup \dots, \\ 0, & \text{otherwise.} \end{cases}$$

$M_{p,k}(x, y) \in \{0, 1\}$ defines a binary mask: 1 indicates that pixel (x, y) lies within the participant's drawn ROI, while 0 indicates no interest in that pixel.

Then, we can compute an averaged mask:

$$H_k(x,y) = \frac{1}{P}\sum_{p=1}^{P} M_{p,k}(x,y).$$

In this case, $H_k(x, y)$ lies within the interval $[0, 1]$ and can be interpreted as the proportion that pixel (x, y) is considered an ROI across all participants.

Generating the Heatmap: To visualize the averaged mask $H_k(x, y)$ we apply a color mapping that maps numerical values to colors and then overlay this heatmap on the original image. Typical steps include:

- Rescale $H_k(x, y)$ from $[0, 1]$ into the required color space.
- Adjust the transparency level to overlay the heatmap on the original image.

The resulting heatmap intuitively reveals which regions of the image draw higher or lower attention among the participant group.

E.2. Result

As shown in Fig. 1 , participants in the original image group primarily focused on the main subject of the image, concentrating around the central area with minimal attention given to the edges and background. In the transmission group, participants also focused on the main subject, indicating that transmission did not significantly alter attention regions; the drop in accuracy here may be attributed to an overall reduction in image quality. However, in the re-digitization group, participants' attention was more dispersed across the background, with less focus on the main subject. This shift suggests that re-digitization may lead to a diffusion of attention, likely contributing to the observed decrease in accuracy.

F. Frequency Domain Analysis of Transmission and Re-digitization

To analyze the frequency characteristics of the images, we applied the Fast Fourier Transform (FFT) to convert the spatial-domain image $f(x, y)$ into its frequency-domain representation $F(u, v)$. The Fourier Transform decomposes the image into its sinusoidal components, with $|F(u, v)|$ representing the amplitude and $\angle F(u, v)$ the phase. The mathematical formulation of the 2D Fourier Transform is expressed as:

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)\cdot e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}\right)}, \quad (1)$$

where (x, y) are pixel coordinates in the spatial domain, and (u, v) are the corresponding frequency coordinates. For better visualization, we applied the FFT shift operation, which repositions the zero-frequency component (DC component) to the center of the spectrum. This is mathematically expressed as:

$$F_{\text{shifted}}(u,v) = F\left(u+\frac{M}{2}, v+\frac{N}{2}\right). \quad (2)$$

The magnitude of the spectrum was computed as:

$$|F(u,v)| = \sqrt{\text{Re}(F(u,v))^2 + \text{Im}(F(u,v))^2}. \quad (3)$$

To improve visualization, the magnitude spectrum was log-transformed to compress the dynamic range of the values. Finally, the spectrum was normalized to the range $[0, 1]$ for better contrast. Fig. 2 and Fig. 3 illustrates the frequency-domain representation of the analyzed images. It

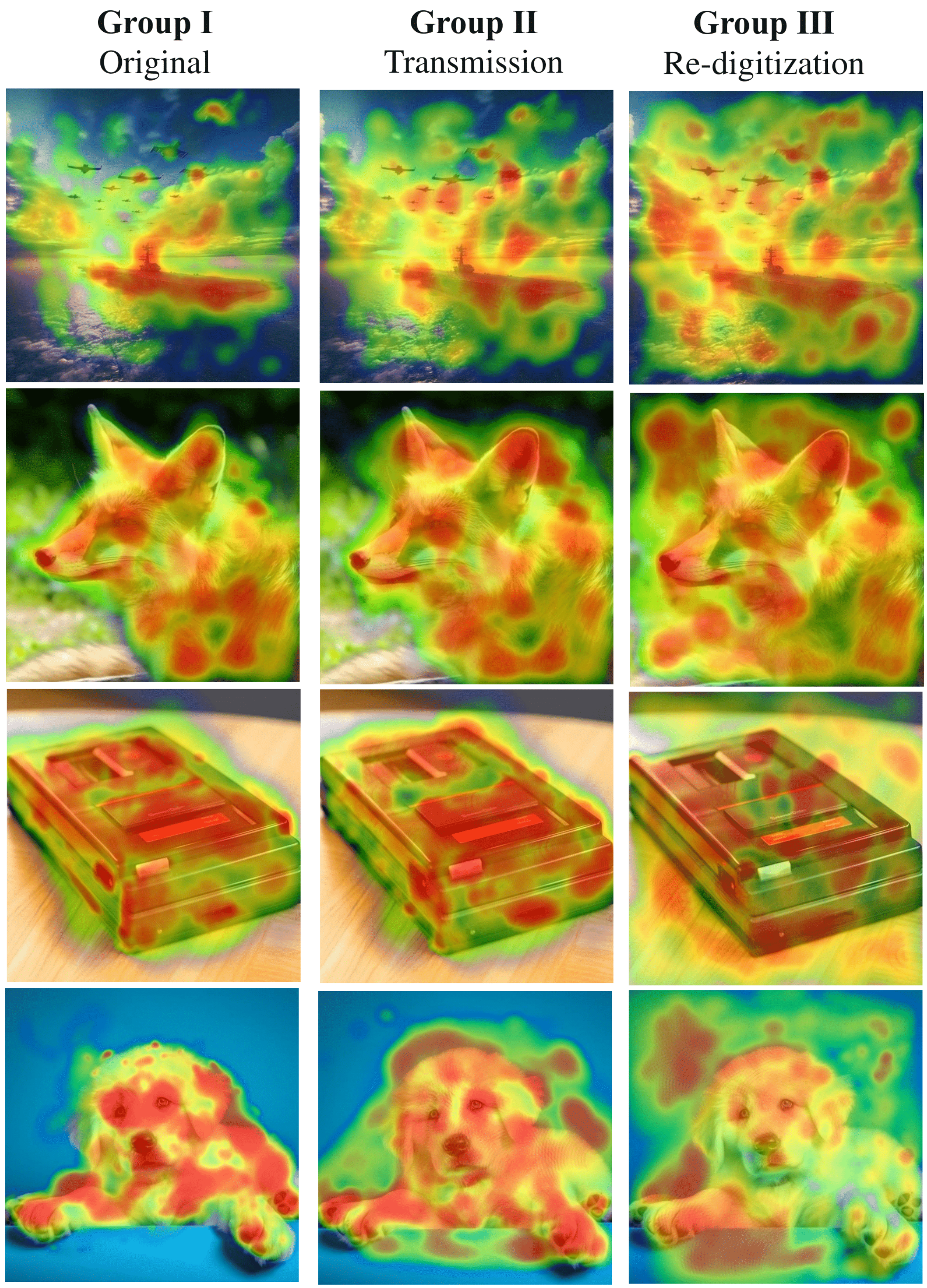

Figure 1. Comparison of Human Attention Regions in Original, Transmitted, and Re-digitized Images.

is evident that both transmission and re-digitization significantly alter the frequency domain characteristics of images, whether real or AI-generated. This alteration may explain the failure of frequency-dependent methods like Freq-Net and Fredect in detecting transmitted and re-digitized images.

G. Visualization of RRDataset.

In this section, we present direct examples from RRDataset to illustrate its key characteristics. Fig. 4 and Fig. 5 showcase AI-generated special-scenario images, while Fig. 6 illustrates a comparison of images in their original, transmitted, and re-digitized states.

H. Further Discussion and Suggestion

The rapid advancements in generative AI technologies have led to profound societal implications, especially in the context of AI-generated image detection. On one hand, these technologies offer creative opportunities, enhancing industries such as media, design, and entertainment. On the other hand, they pose significant challenges, including the spread of misinformation, digital forgeries, and potential misuse in sensitive domains like journalism, legal evidence, and intellectual property.

Our RRDataset provides a critical benchmark for evaluating the robustness of detection algorithms in real-world scenarios, directly addressing these challenges. By incorporating transmission and re-digitization into the dataset, we simulate practical conditions where AI-generated images are often manipulated. This ensures that detection methods are not only technically effective but also reliable in real-world scenarios.

Finally, by revealing robustness gaps in current detection methods, our work calls for the development of more generalized and resilient algorithms. This is essential not only for technical progress but also for fostering trust and accountability in AI-driven societies.

Based on the findings of our study, we propose the following recommendations:

1. **For Researchers**: When developing new AI-generated image detectors, it is crucial to go beyond focusing solely on accuracy and consider their robustness in real-world scenarios. The integration of diverse features and leveraging human-like few-shot learning capabilities could be a promising direction for future advancements.
2. **For Addressing the Trust Crisis**: Given the severe trust crisis currently emerging, especially in highly sensitive special-scenario images, skepticism towards authentic news content may significantly undermine the credibility and effectiveness of information dissemination.
3. **For Generative Model Developers**: Considering the rapid pace of advancements in generative models, we recommend incorporating imperceptible watermarks into AI-generated content. This measure would ensure transparency and uphold people's right to know the origins of the content they encounter.

I. Limitation

Benchmark Limitations: Although we have endeavored to include as many detection methods as possible, including the latest works from KDD 2025, AAAI 2025, and ICLR 2025, some proprietary methods could not be incorporated into RRBench. Additionally, it is possible that newer open-source detectors have not yet been included. We are committed to continuously updating and curating detection methods in future iterations.

Dataset Limitations: RRDataset includes some of the latest and most powerful generative models. However, newer and more advanced models will likely continue to emerge, which may not be represented in RRdataset. To address this, we plan to update the RRDataset continuously in the future.

References

[1] Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024. 1

[2] Nicolo Bonettini, Paolo Bestagini, Simone Milani, and Stefano Tubaro. On the use of benford's law to detect gan-generated images. In *2020 25th international conference on pattern recognition (ICPR)*, pages 5495–5502. IEEE, 2021. 1

[3] Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world, 2024. 1

[4] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10759–10769, 2024. 1

[5] Keshigeyan Chandrasegaran, Ngoc-Trung Tran, Alexander Binder, and Ngai-Man Cheung. Discovering transferable forensic features for cnn-generated images detection. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XV*, pages 671–689. Springer, 2022. 1

[6] You-Ming Chang, Chen Yeh, Wei-Chen Chiu, and Ning Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023. 1

[7] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. 1, 3

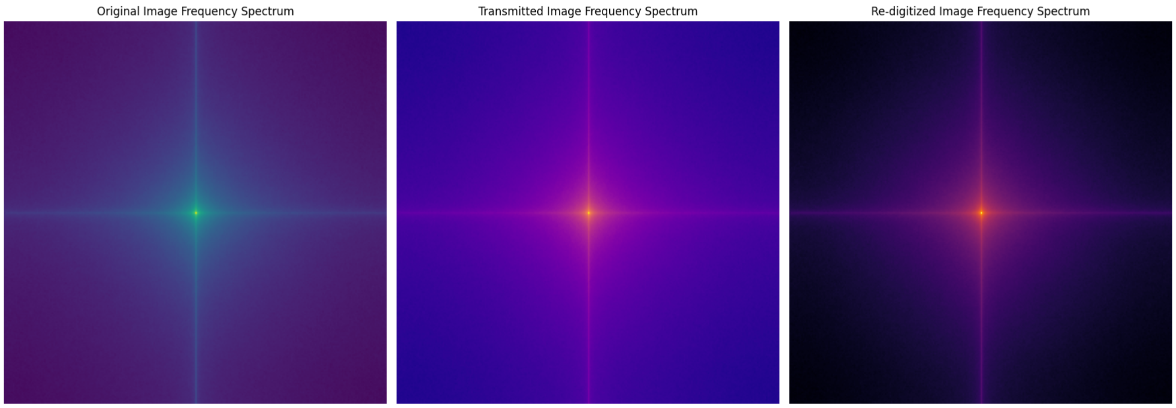

Figure 2. Real Image Frequency Spectrum.

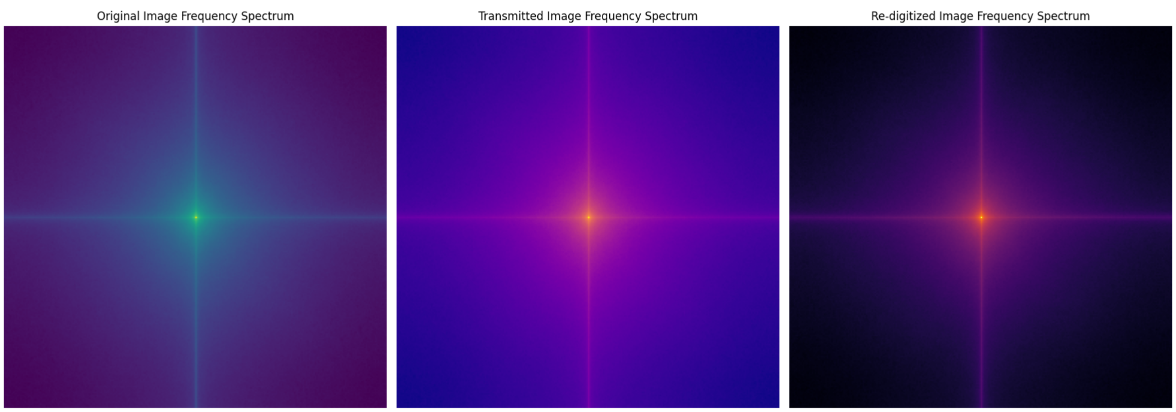

Figure 3. AI Generated Image Frequency Spectrum.

[8] Jiahan Chen, Mengtin Lo, Hailiang Liao, and Tianlin Huang. Ipd-net: Detecting ai-generated images via inter-patch dependencies. *International Journal of Advanced Computer Science and Applications*, 15(7), 2024. 1

[9] Jiaxuan Chen, Jieteng Yao, and Li Niu. A single simple patch is all you need for ai-generated image detection. *arXiv preprint arXiv:2402.01123*, 2024. 1

[10] Sungik Choi, Sungwoo Park, Jaehoon Lee, Seunghyun Kim, Stanley Jungkyu Choi, and Moontae Lee. Hfi: A unified framework for training-free detection and implicit watermarking of latent diffusion model generated images. *arXiv preprint arXiv:2412.20704*, 2024. 2

[11] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error, 2025. 1

[12] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024. 1

[13] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 7887–7896. Computer Vision Foundation / IEEE, 2020. 1

[14] Tarik Dzanic, Karan Shah, and Freddie D. Witherden. Fourier spectrum discrepancies in deep network generated images. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 1

[15] Yuming Fan, Dongming Yang, Jiguang Zhang, Bang Yang, and Yuexian Zou. Fake-gpt: Detecting fake image via large language model. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 122–136. Springer,

Visualization of AI- Generated Special-Scenario Images-1

Figure 4. Visualization of AI- Generated Special-Scenario Images.

Visualization of AI- Generated Special-Scenario Images-2

Figure 5. Visualization of AI- Generated Special-Scenario Images.

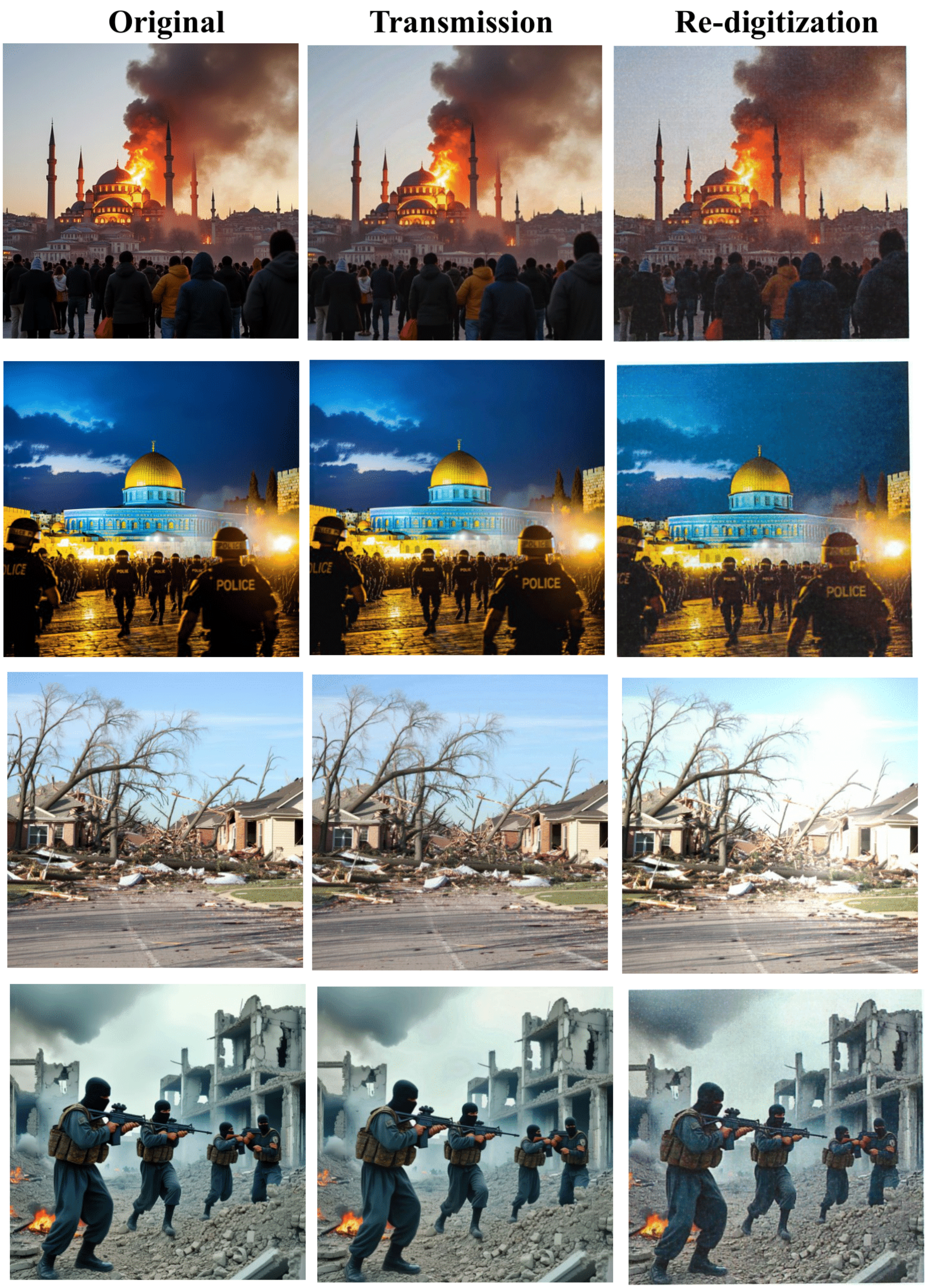

Figure 6. Comparison of Original, Transmitted, and Re-digitized Images.

2024. 1

[16] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 3247–3258. PMLR, 2020. 1

[17] Tatiana Gaintseva, Laida Kushnareva, German Magai, Irina Piontkovskaya, Sergey Nikolenko, Martin Benning, Serguei Barannikov, and Gregory Slabaugh. Improving interpretability and robustness for the detection of ai-generated images. *arXiv preprint arXiv:2406.15035*, 2024. 1

[18] Michael Goebel, Lakshmanan Nataraj, Tejaswi Nanjundaswamy, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, and B. S. Manjunath. Detection, attribution and localization of gan generated images, 2020. 1

[19] Yang He, Ning Yu, Margret Keuper, and Mario Fritz. Beyond the spectrum: Detecting deepfakes via re-synthesis. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 2534–2541. International Joint Conferences on Artificial Intelligence Organization, 2021. 1

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[21] Hyeonseong Jeon, Youngoh Bang, Junyaup Kim, and Simon S. Woo. T-GD: transferable gan-generated images detection framework. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, pages 4746–4761. PMLR, 2020. 2

[22] Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 48–57, 2022. 1

[23] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, Pyounggeon Kim, and Jongwon Choi. Fingerprintnet: Synthesized fingerprints for generated image detection. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, pages 76–94. Springer, 2022. 1

[24] Yan Ju, Shan Jia, Lipeng Ke, Hongfei Xue, Koki Nagano, and Siwei Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *2022 IEEE International Conference on Image Processing, ICIP 2022, Bordeaux, France, 16-19 October 2022*, pages 3465–3469. IEEE, 2022. 1

[25] Yan Ju, Shan Jia, Jialing Cai, Haiying Guan, and Siwei Lyu. GLFF: global and local feature fusion for ai-synthesized image detection. *IEEE Trans. Multim.*, 26:4073–4085, 2024. 1

[26] Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning, 2024. 1

[27] Mamadou Keita, Wassim Hamidouche, Hassen Bougueffa, Abdenour Hadid, and Abdelmalik Taleb-Ahmed. Harnessing the power of large vision language models for synthetic image detection. *arXiv preprint arXiv:2404.02726*, 2024. 1

[28] Mamadou Keita, Wassim Hamidouche, Hessen Bougueffa Eutamene, Abdelmalik Taleb-Ahmed, David Camacho, and Abdenour Hadid. Bi-lora: A vision-language approach for synthetic image detection. *Expert Systems*, 42(2):e13829, 2025.

[29] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval, ICMR 2024, Phuket, Thailand, June 10-14, 2024*, pages 1006–1015. ACM, 2024. 1

[30] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on Computer Vision*, pages 394–411. Springer, 2024. 1

[31] Romeo Lanzino, Federico Fontana, Anxhelo Diko, Marco Raoul Marini, and Luigi Cinque. Faster than lies: Real-time deepfake detection using binary neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3771–3780, 2024. 1

[32] Giorgio Leporoni, Luca Maiano, Lorenzo Papa, and Irene Amerini. A guided-based approach for deepfake detection: Rgb-depth integration via features fusion. *Pattern Recognition Letters*, 181:99–105, 2024. 1

[33] Yanhao Li, Quentin Bammey, Marina Gardella, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone Von Gioi. Masksim: Detection of synthetic images by masked spectrum similarity analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3855–3865, 2024. 1

[34] Yewon Lim, Changyeon Lee, Aerin Kim, and Oren Etzioni. Distildire: A small, fast, cheap and lightweight diffusion synthesized deepfake detection. *arXiv preprint arXiv:2406.00856*, 2024. 1

[35] Li Lin, Irene Amerini, Xin Wang, Shu Hu, et al. Robust clip-based detector for exposing diffusion model-generated images. In *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–7. IEEE, 2024. 1

[36] Bo Liu, Fan Yang, Xiuli Bi, Bin Xiao, Weisheng Li, and Xinbo Gao. Detecting generated images by real images. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XIV*, pages 95–110. Springer, 2022. 1

[37] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[38] Zhengzhe Liu, Xiaojuan Qi, and Philip H. S. Torr. Global texture enhancement for fake face detection in the wild. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8057–8066. Computer Vision Foundation / IEEE, 2020. 2

[39] Zihan Liu, Hanyi Wang, Yaoyu Kang, and Shilin Wang. Mixture of low-rank experts for transferable ai-generated image detection. *arXiv preprint arXiv:2404.04883*, 2024. 1

[40] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024. 1

[41] RuiPeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. In *The Second Workshop on New Frontiers in Adversarial Machine Learning*, 2024. 1

[42] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *IEEE 1st Conference on Multimedia Information Processing and Retrieval, MIPR 2018, Miami, FL, USA, April 10-12, 2018*, pages 384–389. IEEE, 2018. 1

[43] Zheling Meng, Bo Peng, Jing Dong, Tieniu Tan, and Haonan Cheng. Artifact feature purification for cross-domain detection of ai-generated images. *Computer Vision and Image Understanding*, 247:104078, 2024. 1

[44] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, B. S. Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H. Bappy, and Amit K. Roy-Chowdhury. Detecting GAN generated fake images using co-occurrence matrices. In *Media Watermarking, Security, and Forensics 2019, Burlingame, CA, USA, 13-17 January 2019*. Society for Imaging Science and Technology, 2019. 1

[45] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 24480–24489. IEEE, 2023. 1

[46] Orazio Pontorno, Luca Guarnera, and Sebastiano Battiato. On the exploitation of dct-traces in the generative-ai domain. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3806–3812. IEEE, 2024. 1

[47] Tong Qiao, Yuxing Chen, Xiaofei Zhou, Ran Shi, Hang Shao, Kunye Shen, and Xiangyang Luo. Csc-net: Cross-color spatial co-occurrence matrix network for detecting synthesized fake images. *IEEE Transactions on Cognitive and Developmental Systems*, 16(1):369–379, 2023. 1, 2

[48] Syed Ali Raza, Usman Habib, Muhammad Usman, Adeel Ashraf Cheema, and Muhammad Sajid Khan. Mmganguard: a robust approach for detecting fake images generated by gans using multi-model techniques. *IEEE Access*, 2024. 1

[49] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9130–9140, 2024. 1, 2

[50] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, David A. Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry...for now. *CoRR*, abs/2311.17138, 2023. 1

[51] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. DE-FAKE: detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, CCS 2023, Copenhagen, Denmark, November 26-30, 2023*, pages 3418–3432. ACM, 2023. 1

[52] Zeyang Sha, Yicong Tan, Mingjie Li, Michael Backes, and Yang Zhang. Zerofake: Zero-shot detection of fake images generated and edited by text-to-image generation models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 4852–4866, 2024. 2

[53] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 12105–12114. IEEE, 2023. 1

[54] Chuangchuang Tan, Ping Liu, RenShuai Tao, Huan Liu, Yao Zhao, Baoyuan Wu, and Yunchao Wei. Data-independent operator: A training-free artifact representation extractor for generalizable deepfake detection. *arXiv preprint arXiv:2403.06803*, 2024. 1

[55] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. *arXiv preprint arXiv:2408.09647*, 2024. 1

[56] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 1

[57] Chung-Ting Tsai, Ching-Yun Ko, I Chung, Yu-Chiang Frank Wang, Pin-Yu Chen, et al. Understanding and improving training-free ai-generated image detections with vision foundation models. *arXiv preprint arXiv:2411.19117*, 2024. 2

[58] Lea Uhlenbrock, Davide Cozzolino, Denise Moussa, Luisa Verdoliva, and Christian Riess. Did you note my palette? unveiling synthetic images through color statistics. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, pages 47–52, 2024. 1

[59] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8692–8701. Computer Vision Foundation / IEEE, 2020. 2

[60] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22445–22455, 2023. 1, 2

[61] Alexander Wißmann, Steffen Zeiler, Robert M Nickel, and Dorothea Kolossa. Whodunit: Detection and attribution of synthetic images by leveraging model-specific fingerprints.

In *Proceedings of the 3rd ACM International Workshop on Multimedia AI against Disinformation*, pages 65–72, 2024. 1

[62] Haiwei Wu, Jiantao Zhou, and Shile Zhang. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv preprint arXiv:2305.13800*, 2023. 1

[63] Ziyi Xi, Wenmin Huang, Kangkang Wei, Weiqi Luo, and Peijia Zheng. Ai-generated image detection using a cross-attention enhanced dual-stream network. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1463–1470, 2023. 1

[64] Juncong Xu, Yang Yang, Han Fang, Honggu Liu, and Weiming Zhang. Famsec: A few-shot-sample-based general ai-generated image detection method. *IEEE Signal Processing Letters*, 2024. 1

[65] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024. 1, 2, 3

[66] Di Yang, Yihao Huang, Qing Guo, Felix Juefei-Xu, Xiaojun Jia, Run Wang, Geguang Pu, and Yang Liu. Text modality oriented image feature extraction for detecting diffusion-based deepfake. *arXiv preprint arXiv:2405.18071*, 2024. 1

[67] Xiao Yu, Kejiang Chen, Kai Zeng, Han Fang, Zijin Yang, Xiuwei Shang, Yuang Qi, Weiming Zhang, and Nenghai Yu. Semgir: Semantic-guided image regeneration based method for ai-generated image detection and attribution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8480–8488, 2024. 1

[68] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in GAN fake images. In *IEEE International Workshop on Information Forensics and Security, WIFS 2019, Delft, The Netherlands, December 9-12, 2019*, pages 1–6. IEEE, 2019. 1

[69] Yichi Zhang and Xiaogang Xu. Diffusion noise feature: Accurate and fast generated image detection. *arXiv preprint arXiv:2312.02625*, 2023. 1, 2

[70] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2024. 1

[71] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *arXiv preprint arXiv:2306.08571*, 2023. 2