

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

Chunxiao Li^{1*} Xiaoxiao Wang^{2,*} Meiling Li³ Boming Miao¹
Peng Sun⁴ Yunjian Zhang⁵ Xiangyang Ji⁵ Yao Zhu^{5†}

¹Beijing Normal University, Beijing, China ²University of Chinese Academy of Sciences, Beijing, China

³Fudan University, Shanghai, China ⁴Central University of Finance and Economics, Beijing, China

⁵Tsinghua University, Beijing, China

chunxiaoli@mail.bnu.edu.cn, ee.zhuy@zju.edu.cn

Abstract

With the rapid advancement of generative models, highly realistic image synthesis has posed new challenges to digital security and media credibility. Although AI-generated image detection methods have partially addressed these concerns, a substantial research gap remains in evaluating their performance under complex real-world conditions. This paper introduces the Real-World Robustness Dataset (RRDataset) for comprehensive evaluation of detection models across three dimensions: 1) **Scenario Generalization** – RRDataset encompasses high-quality images from seven major scenarios (War & Conflict, Disasters & Accidents, Political & Social Events, Medical & Public Health, Culture & Religion, Labor & Production, and everyday life), addressing existing dataset gaps from a content perspective. 2) **Internet Transmission Robustness** – examining detector performance on images that have undergone multiple rounds of sharing across various social media platforms. 3) **Re-digitization Robustness** – assessing model effectiveness on images altered through four distinct re-digitization methods.

We benchmarked 17 detectors and 10 vision-language models (VLMs) on RRDataset and conducted a large-scale human study involving 192 participants to investigate human few-shot learning capabilities in detecting AI-generated images. The benchmarking results reveal the limitations of current AI detection methods under real-world conditions and underscore the importance of drawing on human adaptability to develop more robust detection algorithms. Our dataset is publicly available at: <https://zenodo.org/records/14963880>.

*Equal contribution.

†Corresponding author.

1. Introduction

With the emergence of powerful generative models [24, 40, 44, 45], AI-generated images are increasingly difficult to distinguish from real ones. Such indistinguishable AI-generated images pose risks to society, including the spread of misinformation and misleading visual cues. Consequently, the identification of AI-generated images has become a critical task with profound real-world implications.

An increasing number of studies explore various detection methods based on different features, including image texture [33], gradients [48], patch-level features [6, 9, 35], frequency-domain characteristics [12, 13, 15, 59], fine-grained image details [20, 46], and reconstruction loss [43, 54]. Vision-language models (VLMs) have also been adapted for detection tasks [25, 26, 38, 47]. Although conventional benchmarks such as GenImage [62], Fake2M [34], WildFake [18], and Chameleon [56] offer sizeable collections of generative images, they share two significant limitations. First, they primarily consist of everyday-life images, making it hard to assess whether the detection methods can effectively generalize to other scenarios. Second, they fail to account for the influence of internet social media transmissions or re-digitization processes, thus inflating detection performance relative to the complexities of real-world conditions.

To address these gaps, we introduce RRDataset, the first benchmark designed explicitly to evaluate detectors' robustness in practical contexts. RRDataset encompasses images from seven diverse and challenging scenarios—including War & Conflict, Disasters & Accidents, Political & Social Events, Medical & Public Health, Culture & Religion, and Labor & Production—thereby filling crucial content gaps in existing benchmarks. Moreover, it systematically includes internet transmission and four distinct re-digitization processes, enabling a deeper assessment of

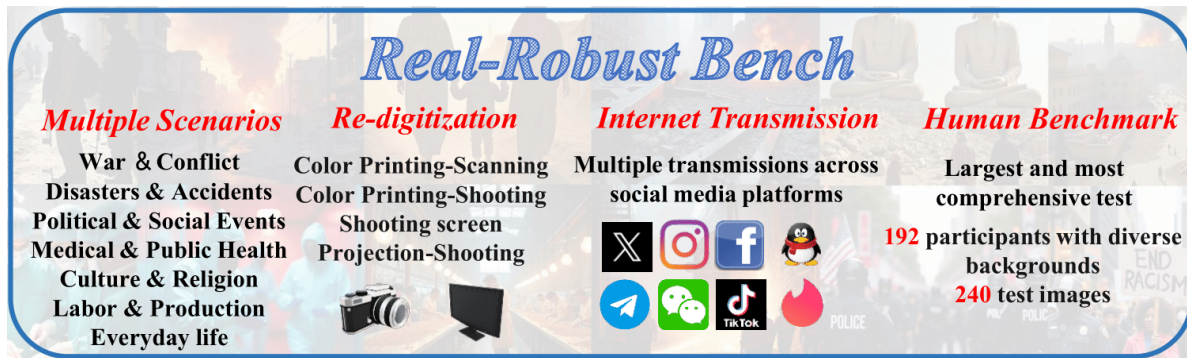


Figure 1. Real-Robust Bench overview, featuring a wide range of scenarios and real-world robustness incorporating four re-digitization methods and internet transmission processes. Additionally, we introduce the largest human benchmark to date, with 192 participants from diverse backgrounds and 240 test images.

how these factors degrade detector performance in realistic use cases. We benchmark 17 detection methods and 10 VLMs on RRDataset, finding substantial performance declines under both transmission and re-digitization conditions—shortcomings that have largely been overlooked in previous evaluations. We further establish a large-scale human benchmark, involving 192 participants and 240 test images, to investigate how human observers adapt in similar conditions. Although human accuracy also decreases when confronted with altered images, it improves significantly after a few-shot learning phase, highlighting a capacity for rapid adaptation that may inspire future detection algorithm design. The main contributions are summarized as follows:

- We propose RRDataset, the first dataset incorporating six critical high-risk scenarios and aspects of daily life. To enhance its realism and applicability, the images within this dataset have been transmitted over the internet and re-digitized, thereby simulating real-world image degradation as shown in Fig. 1.
- We conduct RRBench, a comprehensive benchmark evaluating 17 detection methods and 10 VLMs, revealing that current detection strategies suffer notable performance drops under transmission and re-digitization.
- Our RRBench also constructs the largest human AI-Generated image benchmark to date, involving 192 participants and 240 test images. Our analysis uncovers the remarkable few-shot adaptation capabilities of human observers.

2. Related work

AI-Generated Image Datasets: With the rapid evolution of generative models, AI-generated image detection datasets have undergone substantial changes. Early efforts, such as CNNSpot [53], primarily employed ProGAN [22]-produced images to evaluate detection performance. More recent works—including CiFAKE [4] and

DE-FAKE [47]—have shifted toward diffusion-based approaches, yielding more realistic synthetic images.

Large-scale datasets such as GenImage [62], WildFake [18], Fake2M [34], Chameleon [56], and PatchCraft [61] have pushed the boundaries of size and diversity, reaching millions of synthetic samples from multiple model architectures. Some datasets like Chameleon [56], emphasize high-quality human-curated images to expose blind spots in current detectors. However, most of these datasets concentrate on everyday scenarios drawn from sources like ImageNet [11] or COCO [31], neglecting critical contexts such as war, disasters, or public health, where misinformation can be particularly damaging.

Several recent works also explore robustness under data augmentation or online transmission. SEMI-TRUTHS [39] investigates how different augmentations affect detection performance, whereas WildRF [5] and FOSID [21] collect a limited amount of web-transmitted data to examine detector robustness. However, their choice of platforms and overall scale remains narrow, rendering the results less representative of the broader impact of social media transmission.

RRDataset addresses these gaps by covering diverse high-impact scenarios while also modeling multiple rounds of online transmission and various re-digitization processes. This approach enables a more faithful assessment of detection performance in scenarios that closely mirror actual practices and challenges.

Human Benchmark for AI Image Detection: HPBench [34] is the first to collect data from 50 participants on 100 test images, along with insights into why participants identified certain images as AI-generated. However, this study used images exclusively generated by Midjourney, which may introduce bias due to reliance on a single generator, and its limited sample size and question set constrain the scope of its findings. While FakeBench [30] uses images from multiple generators, it does not account for the effects of image transmission or re-digitization and includes



Figure 2. Special-scenario image generation pipeline, illustrating scenario definition, theme expansion, prompt refinement, and final image filtering.

only 34 participants. This paper establish the largest human benchmark to date, addressing these limitations by incorporating 192 participants and 240 test images, encompassing original, transmitted, and re-digitized images. We conducted a detailed analysis of how AIGC and photography backgrounds impact detection accuracy. Additionally, we explored human few-shot learning capabilities in detecting AI-generated images.

AI-Generated Image Detectors: With the continuous development of generative models, numerous detection methods [6, 8–10, 20, 27, 29, 33, 35, 38, 43, 46, 49, 51, 54, 56] have emerged. While these methods achieve impressive accuracy on their respective test sets, robust evaluation remains largely unaddressed. Existing methods typically limit robustness evaluation to resizing, JPEG compression and Gaussian blur, with some studies using only JPEG compression at a quality level of 90, which does not closely simulate real-world conditions. Therefore, it is essential to evaluate these detectors’ performance in realistic scenarios. A more comprehensive overview of AI-generated image detection methods can be found in App. A.

3. Dataset Construction

Sec. 3.1 describes the problem setting and design rationale. Sec. 3.2 outlines our collection and creation process, and Sec. 3.3 explains internet transmission and re-digitization. Additional examples are included in App. G.

3.1. Dataset Rationale

The primary goal of RRDataset is to comprehensively benchmark AI-generated image detectors under real-world conditions, focusing on three critical dimensions: 1) **Multi-scenario detection capability:** We incorporate a broad range of image contexts, including rarely encountered or sensitive scenarios, to ensure that detection models are eval-

uated on content closely mirroring real-world usage. 2) **Internet transmission robustness:** Images routinely undergo repeated sharing and compression on various social media platforms. We expect a reliable detector to maintain accuracy despite these transmission-induced degradations. 3) **Re-digitization robustness:** Re-digitization—via scanning, re-photographing, or similar methods—is a pivotal yet underexplored challenge. Many practical scenarios lack direct digital files (e.g., PNG, JPG), such as verifying images in newspapers, presentation slides, or street advertisements. In RRDataset, an image’s label (Real vs. AI) is determined solely by its original source, underscoring the need for detectors to withstand transformations introduced by re-digitization.

By encompassing these three dimensions, RRDataset aims to provide a comprehensive and realistic evaluation framework that addresses the often-overlooked factors limiting current detection methods in existing assessments.

3.2. Data Collection

3.2.1. Special-Scenario Image Collection

To capture diverse and critical contexts, RRDataset includes six specialized scenarios: War & Conflict, Disasters & Accidents, Political & Social Events, Medical & Public Health, Culture & Religion, and Labor & Production. First, each scenario is expanded into 10 manually defined theme (e.g., traffic accidents, floods, earthquakes, wildfires, and plane crashes under Disasters & Accidents). Next, we use Qwen2.57b-instruct [57] to further enrich these sub-topics into prompts, adding location details and additional descriptive elements. This process yields 24,000 prompts, which are then used by SD 3.5 Large [44] and Flux.1 [28] to generate 48,000 images at various resolutions. We filter the generated images using CLIP-score [41] (removing those below 0.27) and an NSFW safety checker, discarding any that

exhibit low text-image alignment or contain unsafe content. After filtering, we retain 6,000 high-quality AI-generated images (1,000 per scenario), the complete image generation workflow is illustrated in Fig. 2. For real images, we collect an additional 6,000 samples from openly licensed news photography websites, including Reuters Pictures, Associated Press Images, BBC News In Pictures, UN Photo, and the International Committee of the Red Cross, ensuring consistent coverage of the same six domains.

3.2.2. Everyday Life-Scenario Image Collection

We then gather 4,000 real-life images from COCO [31], CC3M-val [7], and the publicly licensed photography platform Unsplash. To generate AI counterparts, we use captions from COCO [31] and CC3M-val [7] as prompts for various generative models (SD 3.5 Large [44], Flux.1 [28], DALL-E3 [42], SDv1.4 [44], SDv1.5 [44], and Midjourney [36]). To match the high-resolution photographic works, we also randomly selected a portion of AI-generated images from the Chameleon dataset [56].

Since real-world images include both high-resolution and low-resolution examples [14], we also incorporate a portion of lower-resolution images generated by StyleGAN [23] and ProGAN [22] into RRDataset to better capture this characteristic. This step ensures that RRDataset more accurately reflects the real-world variations in image quality. Ultimately, we obtained a dataset of 4,000 images representing everyday scenes.

3.3. Data transformation

In this section, we describe how internet transmission and re-digitization are applied to our dataset. We also provide a detailed discussion in App. B on why these transformations are critical for AI-generated image detection and how they manifest in real-world scenarios.

3.3.1. Internet Transmission

Social media platforms typically compress images, causing reductions in resolution, compression artifacts, and detail loss. To evaluate detector robustness under realistic transmission conditions, we subjected all 10,000 real images and 10,000 AI-generated images from RRDataset to multiple rounds of sending through popular messaging and social platforms—Telegram, WeChat, Facebook, QQ, WhatsApp, X, Instagram, and Tinder. As summarized in Tab. 1, each image underwent 2 to 6 transmission cycles, covering both cross-platform and single-platform settings. The resulting images were then added to the RRDataset test set, enabling a more faithful assessment of detector performance in real-world scenarios.

3.3.2. Re-digitization

Re-digitization refers to the process of converting a digital image into a physical format—such as printing or display-

Trans-Times	Percentage	Platforms	Self-Trans-Limit
2	10%	1–2	2
3	25%	2–3	2
4	25%	2–4	3
5	25%	3–5	3
6	15%	4–6	3

Table 1. Cross-platform multi-transmission workflow, simulating three levels of information sharing: direct private messaging, limited-scale forwarding, and repeated forwarding of trending news.

ing it on a screen—and then converting it back into digital form, for instance, via scanning or photography. This process inevitably affects image quality, particularly in resolution and color fidelity, and may also introduce geometric distortions, adding another layer of complexity for detection. We employ four common re-digitization methods:

- Scanning a color printout.
- Photographing a color printout.
- Photographing an image displayed on a screen using various camera devices.
- Photographing a projected digital image with different camera devices.

Each method is applied with equal probability to 10,000 real images and 10,000 AI-generated images. The resulting re-digitized images—both real and AI-generated—are then incorporated into RRDataset.

4. RRBenchmark: Benchmarking Model Performance under Real-World Scenarios

4.1. Benchmark Setting

Detector Setup: We evaluated 17 detection methods, including the latest state-of-the-art (SOTA) algorithms introduced at KDD 2025, AAI 2025, and ICLR 2025. Specifically, our benchmarks incorporate CNNSpot [53], F3Net [55], GramNet [33], DIRE [54], UnivFD [38], LNP [32], LGard [48], AIDE [56], SSP [9], Fusing [20], Fredect [15], DNF [60], NPR [51], Freq-Net [50], SAFE [29], DRCT [8], and C2P-clip [49].

VLM Setup: To further assess detection capabilities, we tested 10 contemporary vision-language models—gpt-4o-latest [1], Claude-3-7-sonnet [2], Gemini-1-5 pro [52], Gemini-2 flash [52], GLM-4v-plus [16], Grok-2-vision [17], Qwen2.5-VL-72B [3], YI-vision [58], Moonshot-preview-vision-128k[37], and Hunyuan-vision [19], using the prompt in App. D.2:

Each model was prompted to provide a JSON-based prediction—“AI-generated” or “Real”—based on its visual analysis. By including these vision-language systems, we aim to capture cutting-edge approaches to AI-generated im-

age detection, enabling a broader comparison of performance across multiple methodological paradigms.

Metric Setup: We use accuracy on real images and accuracy on AI-generated images as our primary evaluation metrics. Since the dataset is balanced across classes, recall and precision are straightforward to compute.

Training Setup: Following the setup outlined in Chameleon [56] and GenImage [62], we pretrain the detectors on GenImage-SD v1.4 and fine-tune them using RRDataset-subset. Training details and additional results are provided in App. C and App. D.

4.2. Results and Analysis

As shown in Tab. 2, none of the 17 detection methods achieve saturated performance on RRDataset, with the best accuracy reaching only 89.59%. This highlights both the ongoing complexity of AI-generated image detection and the crucial role of RRDataset in examining real-world challenges. Network transmission and re-digitization significantly degrade detection performance, underlining the dataset’s ability to reveal limitations overlooked by conventional evaluations.

Impact of Internet Transmission: For internet-transmitted images, 14 detectors exhibit reduced fake accuracy, suggesting that compression artifacts, lower resolution, and color distortion lead to misclassification of AI-generated images as real. In particular, Freq-Net [50], Fusing [20], and SAFE [29] experience accuracy drops of 71.61%, 79.86%, and 97.41%, respectively, underscoring their lack of robustness. By contrast, DNF [60] and DIRE [54], which rely on diffusion-based denoising features, show only minor performance fluctuations under the same conditions. DRCT-ConvB [8], which employs diffusion for image redrawing, similarly demonstrates strong resistance to transmission artifacts. Methods such as Gram-Net [33] and AIDE [56], which focus on visual artifacts and noise patterns, also maintain relatively stable results when confronted with network-induced degradation.

Impact of Re-digitization: Re-digitization proves even more challenging: 16 of the 17 detection methods suffer decreases in fake accuracy. Notably, the diffusion-based DIRE [54] and DNF [60] drop by 88.30% and 90.57%, respectively. In contrast, AIDE [56] exhibits remarkable robustness, with only 1.89% reduction in fake accuracy and a 4.19% increase in real accuracy—likely owing to its reliance on CLIP-extracted semantic and contextual information, which remains largely intact after re-digitization.

Vision-Language Models Performance: Vision-language models demonstrate strong zero-shot classification capabilities, with GPT-4o [1] even surpassing 16 specialized detectors on the original data. However, their performance declines substantially under network transmission and re-digitization, indicating that reliance on internal model

knowledge becomes a liability when images degrade in quality or exhibit color distortion.

Based on our findings, we can draw the following conclusions:

- Diffusion-based methods are often robust to internet transmission, with models like DRCT-ConvB [8], DNF [60], and DIRE [54] showing nearly unchanged performance against internet transmission.
- A hybrid strategy may contribute to the robustness of detection methods across different scenarios, as demonstrated by AIDE [56], which combines multiple expert models and uses hybrid features for detection, exhibiting the best robustness on RRbench, and by DRCT-ConvB [8], which incorporates diffusion image reconstruction and contrastive training strategies, exhibiting the highest overall performance on RRbench.
- Vision-language models show considerable potential in AI-generated image detection, a promising yet overlooked aspect in existing detection research.

5. Benchmarking Human Performance under Real-World Scenarios

In this section, we conduct a comprehensive analysis of human performance on the RRDataset. Sec. 5.1 introduces the evaluation system and process. Sec. 5.2 presents a detailed analysis of human evaluation results.

5.1. Human Benchmark Evaluation System

Our human benchmark evaluation system consists of four steps, as shown in Fig. 3. A total of 192 participants were randomly assigned to either a special-scenario group or an everyday-scenario group. The special-scenario group viewed images from six high-impact categories: War & Conflict, Disasters & Accidents, Political & Social Events, Medical & Public Health, Culture & Religion, and Labor & Production. To ensure reliability, there was no time limit, the test was conducted in a quiet environment on a standardized 4K display, and participants worked independently without any electronic aids or internet access.

Phase One Testing. To facilitate direct comparisons with RRbench, we selected 20 real and 20 AI-generated images from each of three categories—original, transmitted, and re-digitized—yielding 120 test images per participant. For each image, participants answered two questions:

1. *Is this image AI-generated or real?* If participants classified the image as AI-generated, they selected one of 14 reasons—five low-level criteria (*texture, edge, clarity, distortion, overall hue*), five mid-level criteria (*light & shadow, shape, content deficiency, symmetry, reflection*), and four high-level visual criteria (*layout, perspective, theme, irreal-ity*), consistent with FakeBench [30].
2. *How confident are you in this judgment?* Responses were provided on a five-point Likert scale.

Table 2. Performance Comparison of 17 Detectors and 10 VLMs on RRDataset. All values are presented as percentages and represent the average results from three trials. “Fake” denotes the accuracy on AI-generated images, while “Real” represents the accuracy on real images. Note: Since C2P-clip[49] has not released its training code, we evaluate the model using its pre-trained weights.

Model	Original		Transmission		Re-digitization		Overall ACC(%)
	Fake(%)	Real(%)	Fake(%)	Real(%)	Fake(%)	Real(%)	
Detectors (Train on GenImage-SDv1.4 & fine-tune on RRDataset)							
DRCT-ConvB[8]	93.52	95.52	92.82(-0.70)	95.09(-0.43)	64.34(-29.18)	96.22(+0.70)	89.59
DIRE[54]	89.72	98.25	90.34(+0.62)	97.87(-0.38)	1.42(-88.30)	98.89(+0.64)	79.42
DNF[60]	90.62	99.05	90.98(+0.36)	94.45(-4.60)	0.05(-90.57)	99.98(+0.93)	79.19
AIDE[56]	78.95	78.94	74.72(-4.23)	78.75(-0.19)	76.04(-2.91)	83.13(+4.19)	78.42
GramNet[33]	81.34	74.65	79.49(-1.85)	75.69(+1.04)	79.45(-1.89)	62.02(-12.63)	75.44
CNNspot[53]	72.42	89.09	65.72(-6.70)	88.78(-0.31)	43.12(-29.30)	86.72(-2.37)	74.31
LNP[32]	83.14	89.26	38.23(-44.91)	89.30(+0.04)	31.91(-51.23)	91.05(+1.79)	70.48
Fredect[15]	79.60	75.93	58.13(-21.47)	82.11(+6.18)	46.34(-33.26)	69.95(-5.98)	68.68
NPR[51]	49.21	96.18	28.08(-21.13)	97.13(+0.95)	38.65(-10.56)	92.42(-3.76)	66.95
Fusing[20]	87.24	92.46	7.38(-79.86)	99.04(+6.58)	30.79(-56.45)	73.97(-18.49)	65.15
SAFE[29]	98.29	88.22	0.88(-97.41)	98.85(+10.63)	2.29(-96.00)	98.82(+10.60)	64.56
Freq-Net[50]	76.08	82.18	4.47(-71.61)	98.64(+16.46)	37.48(-38.60)	77.99(-4.19)	62.81
F3Net[55]	65.82	71.35	52.18(-13.64)	75.49(+4.14)	31.16(-34.66)	74.85(+3.50)	61.81
UnivFD[38]	64.79	64.90	44.61(-20.18)	70.80(+5.90)	36.15(-28.64)	75.69(+10.79)	59.49
C2P-CLIP[49]	17.21	97.54	28.82(+11.61)	99.58(+2.04)	18.01(+0.80)	90.29(-7.25)	58.58
SSP[9]	61.29	64.54	40.62(-20.67)	70.33(+5.79)	32.58(-28.71)	79.94(+15.40)	58.22
LGrad[48]	51.00	81.29	18.86(-32.14)	92.54(+11.25)	14.71(-36.29)	88.27(+6.98)	57.78
VLMs (Zero-shot)							
GPT-4o-latest[1]	96.30	92.68	79.41(-16.89)	90.01(-2.67)	69.23(-27.07)	76.92(-15.76)	84.09
Claude-3.7-sonnet[2]	85.12	94.57	71.26(-13.86)	96.17(+1.60)	62.34(-22.78)	85.41(-9.16)	82.48
Gemini-2-flash[52]	72.10	98.43	52.19(-19.91)	97.41(-1.02)	46.11(-25.99)	97.41(-1.02)	77.28
Grok-2-vision[17]	46.15	91.84	52.12(+5.97)	94.03(+2.19)	48.01(+1.86)	81.63(-10.21)	68.96
Gemini-1.5-prof[52]	36.12	97.78	36.36(+0.24)	95.83(-1.95)	22.22(-13.90)	88.64(-9.14)	62.83
Qwen2vl-72B[3]	31.14	88.74	22.87(-8.27)	89.95(+1.21)	26.99(-4.15)	92.55(+3.81)	58.71
GLM4v-plus[16]	22.16	90.57	30.21(+8.05)	86.01(-4.56)	38.16(+16.00)	82.14(-8.43)	58.21
Moonshot-vision[37]	14.68	99.72	16.24(+1.56)	94.75(-4.97)	24.37(+9.69)	96.84(-2.88)	57.77
Hunyuan-vision[19]	28.57	91.84	20.13(-8.44)	86.06(-5.78)	32.14(+3.57)	81.63(-10.21)	56.73
YI-vision[58]	22.73	89.13	20.45(-2.28)	81.81(-7.32)	28.89(+6.16)	78.26(-10.87)	53.55

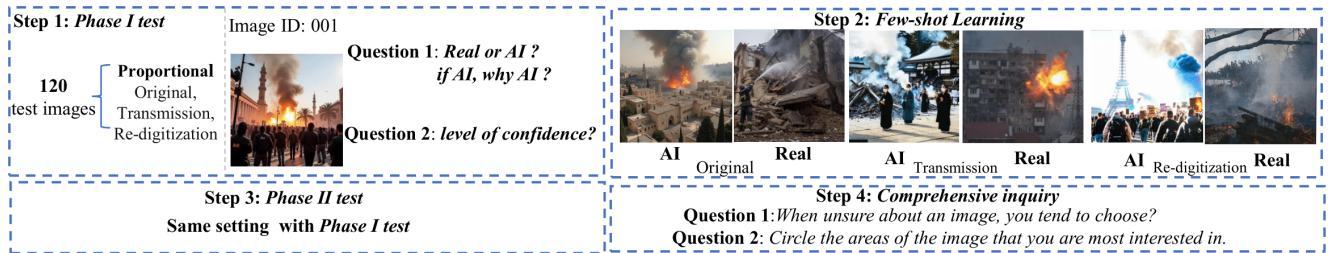


Figure 3. Human Benchmark Evaluation System.

Few-Shot Learning Phase. Each participant viewed two additional images (original, transmitted, and re-digitized) drawn from RRDataset, ensuring no overlap with the main test set.

Phase Two Testing. Phase Two replicated the Phase One

procedure to evaluate whether 2-shot learning affected performance.

Comprehensive Inquiry . Finally, two additional questions were posed: 1. *When uncertain, do you tend to judge an image as AI-generated or real?* 2. *Please highlight the area*

you focus on most. For the second question, participants were split into three subgroups of 32, each viewing six AI-generated images. Subgroup 1 viewed the original images, Subgroup 2 the transmitted versions, and Subgroup 3 the re-digitized versions. Each participant had 10 seconds to mark the region they found most telling, using a freeform drawing tool based on their immediate impression.

5.2. Analysis of Human Evaluation Results

Overall Human Discrimination Ability: As shown in Tab. 3, the everyday life-scenario test group achieved an overall accuracy of 69.17%, while the special-scenario test group reached 59.52%, suggesting that humans are generally more adept at discerning AI-generated images in everyday contexts. For real images, the everyday life-scenario group achieved an average accuracy of 80.05%, while the special-scenario group reached 79.40%, indicating a negligible gap. However, for AI-generated images, the everyday life-scenario group attained 58.29% accuracy, whereas the special-scenario group managed only 39.64%. This difference indicates that in more sensitive, high-stakes contexts, humans are significantly less adept at identifying AI-generated images than they are in everyday life-scenarios. In both groups, internet transmission and re-digitization led to a significant, consistent drop in detection accuracy. During the first testing phase, these two factors reduced accuracy by 14.01% and 14.29%, respectively.

Table 3. Human Benchmark Testing Accuracy Results. The numerical values in the table indicate accuracy (ACC). Values below 50% are highlighted in red.

Group	Original		Trans.		Re-digit.		Overall ACC
	Fake	Real	Fake	Real	Fake	Real	
Pre-learning							
Everyday Life	64.87	81.65	46.93	71.63	41.98	83.5	65.09
Special-Scenario	46.21	84.21	23.26	79.10	29.17	65.15	54.52
Post-learning							
Everyday Life	66.42	85.30	66.31	78.53	63.24	79.70	73.25
Special-Scenario	58.61	79.63	42.15	81.55	38.42	86.76	64.52

Analysis of Human Few-Shot Learning Ability: We compare the results of the first-stage and second-stage tests. Tab. 3 highlights humans’ remarkable few-shot learning ability. For both the everyday-scenario group and the special-scenario group, accuracy for transmitted images increased by 13.14% and 10.67%, respectively, while accuracy for re-digitized images rose by 8.73% and 15.43%. Overall, these improvements show that few-shot learning significantly boosts AI image recognition accuracy—from 42.07% to 55.86%, an increase of 13.79%—and also raises accuracy for real images from 77.54% to 81.92%, a gain of 4.37%.

Confidence Analysis: For the five confidence levels provided—very uncertain, somewhat uncertain, ambiguous,

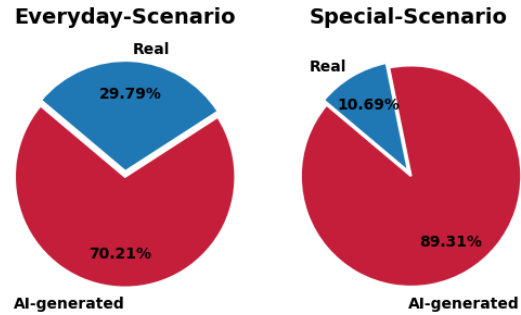


Figure 4. Trust Crisis Across Everyday Scenarios & Special Scenarios.

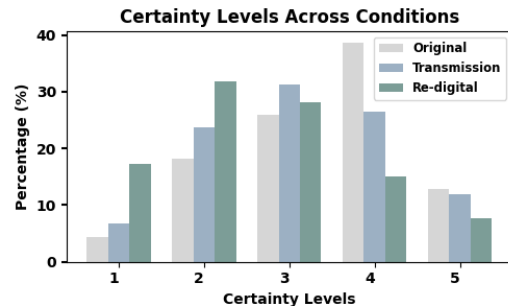


Figure 5. Uncertainty Levels across original images, transmission images and re-digital images where 1, 2, 3, 4, and 5 correspond to “Very Uncertain,” “Somewhat Uncertain,” “Ambiguous,” “Somewhat Certain,” and “Very Certain,” respectively.

somewhat certain, and very certain—we found that transmission and re-digitization significantly decreased participants’ confidence in their judgments as shown in Fig. 5. For original images, the combined proportion of “somewhat certain” and “very certain” responses was 51.8%, which dropped to 38.3% for transmitted images and 22.7% for re-digitized images. This indicates that transmission and re-digitization greatly impacted participants’ confidence in their judgments.

Trust Crisis Emerged in AIGC Era: In the final part of the testing system, we asked all participants, “When you cannot decide whether an image is AI-generated or real, which type are you more inclined to choose?” Surprisingly, in Fig. 4, 70.21% of the everyday-scenario group assumed the image was AI-generated, while this figure rose to 89.31% in the special-scenario group. These findings indicate a significant crisis of trust in images’ origins, driven by the rapid advancement of AIGC technologies. The impact is especially severe for highly sensitive topics—such as War & Conflict, Disasters & Accidents, Political & Social Events, Medical & Public Health, Culture & Religion, and Labor & Production—where genuine news may be met with unwarranted skepticism, negatively influencing the broader information ecosystem. We therefore urge the community to de-

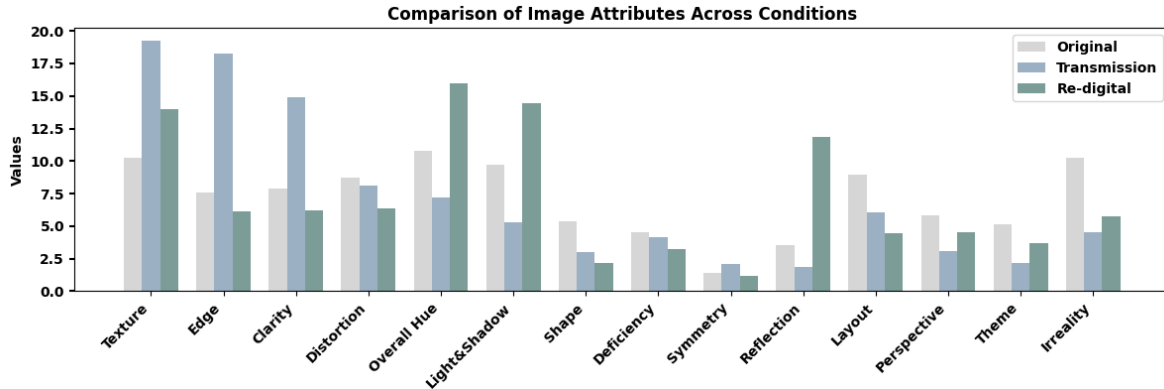


Figure 6. Comparison of Image Attributes Across Original, Transmission and Re-digital.

velop more accurate, real-world-aligned detection methods to help mitigate this growing crisis of trust.

Analysis of Judgemental Attribution: As shown in Fig. 6, for the 14 possible reasons provided, transmission significantly increased the proportion of responses citing texture, edge, and clarity, rising from 25.78% to 52.42%. Re-digitization notably increased the proportion citing light&shadow, reflection, and overall hue, from 24.08% to 42.33%. This suggests that the effects of transmission and re-digitization altered participants’ judgment criteria: transmission impacts are more related to reduced image quality, such as compression artifacts and detail loss, while re-digitization effects are more associated with color changes and loss of light and shadow details.

5.3. Human-Inspired In-Context Learning Approach for VLM Detection

Inspired by humans’ remarkable few-shot learning abilities—where participants rapidly improved their robustness against network transmission and re-digitization with only a small number of samples—as well as the factors influencing human decision-making under these conditions, we aim to harness the strong zero-shot capability of VLMs to further enhance detection resilience.

Our in-context learning approach explicitly emphasizes the impact of transmission and re-digitization artifacts, guiding the VLM to focus on core image content while disregarding irrelevant distortions. Additional comparisons with other in-context learning strategies, as well as our method, are presented in App.D.

As shown in Tab. 4, our approach significantly enhances VLM robustness, particularly in re-digitization scenarios. Notably, on GPT-4o, our method achieves a **5.50%** improvement in re-digitization robustness and a **3.90%** increase in overall accuracy, reaching an average accuracy of **87.47%**. This performance is approaching that of DCRT-ConvB (89.59%), the strongest detector on RRbench to

Table 4. Performance Comparison for Human-Inspired Robustness-Oriented In-Context Learning.

	Original	Transmission	Redigital
Zero-shot			
GPT-4o-latest	94.49	84.71	73.08
Claude-3.7-sonnet	89.85	83.72	73.87
Gemini-2-flash	85.27	74.80	71.76
Grok-2-vision	68.99	73.08	64.82
Robustness-Oriented In-Context Learning			
GPT-4o-latest	95.67(+1.18)	88.17(+3.46)	78.58(+5.50)
Claude-3.7-sonnet	92.26(+2.41)	84.97(+1.25)	77.76(+3.79)
Gemini-2-flash	88.38(+3.11)	74.99(+0.19)	75.78(+4.02)
Grok-2-vision	72.86(+3.87)	71.52(-1.56)	71.43(+6.61)

date. These findings underscore the potential of VLMs in AI-generated image detection.

6. Conclusion

In this paper, we introduced the RRDataset, rethinking the evaluation of AI-generated image detection from the perspective of real-world robustness. Our RRbench includes 17 detectors as well as comparisons with 10 VLMs, revealing a significant drop in accuracy for current detection methods under internet transmission and re-digitization conditions. Additionally, we developed the largest human benchmark to date, with 192 participants and 240 test images. We found that human accuracy dropped dramatically when faced with transmitted and re-digitized images; however, after few-shot learning, this effect of transmission and re-digitization was effectively mitigated. We hope this work encourages researchers to focus on the robustness of AI-generated image detection in real-world scenarios and to draw inspiration from humans’ exceptional few-shot learning abilities in developing more robust and effective detection methods.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 4, 5, 6
- [2] Anthropic. Introducing the next generation of claude. <https://www.anthropic.com/news/claude-3-family>, 2024. 4, 6
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 4, 6
- [4] Jordan J Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images. *IEEE Access*, 2024. 2
- [5] Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world, 2024. 2
- [6] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 103–120. Springer, 2020. 1, 3
- [7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 4
- [8] 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. 3, 4, 5, 6
- [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, 4, 6
- [10] 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. 3
- [11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2
- [12] 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
- [13] 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
- [14] 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, 2024. 4
- [15] 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, 4, 6
- [16] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. 4, 6
- [17] grok. grok-2. <https://x.ai/blog/grok-2>, 2025. 4, 6
- [18] Yan Hong and Jianfu Zhang. Wildfake: A large-scale challenging dataset for ai-generated images detection. *arXiv preprint arXiv:2402.11843*, 2024. 1, 2
- [19] hunyuan. hunyuan-vision. <https://hunyuan.tencent.com/>, 2025. 4, 6
- [20] 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, 3, 4, 5, 6
- [21] Dimitrios Karageorgiou, Quentin Bammey, Valentin Porcellini, Bertrand Goupil, Denis Teyssou, and Symeon Papadopoulos. Evolution of detection performance throughout the online lifespan of synthetic images. *arXiv preprint arXiv:2408.11541*, 2024. 2
- [22] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 4
- [23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 4
- [24] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. 1
- [25] 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
- [26] 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
- [27] 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. 3
- [28] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. 3, 4
- [29] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. *arXiv preprint arXiv:2408.06741*, 2024. 3, 4, 5, 6
- [30] Yixuan Li, Xuelin Liu, Xiaoyang Wang, Shiqi Wang, and Weisi Lin. Fakebench: Uncover the achilles’ heels of fake images with large multimodal models. *arXiv preprint arXiv:2404.13306*, 2024. 2, 5
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 4
- [32] 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. 4, 6
- [33] 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. 1, 3, 4, 5, 6
- [34] Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xihui Liu, and Wanli Ouyang. Seeing is not always believing: benchmarking human and model perception of ai-generated images. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2
- [35] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *2022 IEEE International Conference on Image Processing, ICIP 2022, Bordeaux, France, 16-19 October 2022*, pages 3091–3095. IEEE, 2022. 1, 3
- [36] Midjourney. Midjourney, 2024. 4
- [37] moonshot. moonshot-preview-vision. <https://www.moonshot.cn/>, 2025. 4, 6
- [38] 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, 3, 4, 6
- [39] Anisha Pal, Julia Kruk, Mansi Phute, Manogna Bhataram, Diyi Yang, Duen Horng Chau, and Judy Hoffman. Semi-truths: A large-scale dataset of ai-augmented images for evaluating robustness of ai-generated image detectors. *Advances in Neural Information Processing Systems*, 37: 118025–118051, 2025. 2
- [40] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. Pmlr, 2021. 3
- [42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 4
- [43] 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, 3
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 3, 4
- [45] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1
- [46] 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, 3
- [47] 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*, pages 3418–3432, 2023. 1, 2
- [48] 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, 4, 6
- [49] 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. 3, 4, 6
- [50] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space

- domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5052–5060, 2024. 4, 5, 6
- [51] 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. 3, 4, 6
- [52] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 4, 6
- [53] 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, 4, 6
- [54] 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, 3, 4, 5, 6
- [55] Jun Wei, Shuhui Wang, and Qingming Huang. F³net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12321–12328, 2020. 4, 6
- [56] 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, 4, 5, 6
- [57] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. 3
- [58] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*, 2024. 4, 6
- [59] 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
- [60] Yichi Zhang and Xiaogang Xu. Diffusion noise feature: Accurate and fast generated image detection. *arXiv preprint arXiv:2312.02625*, 2023. 4, 5, 6
- [61] Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2024. 2
- [62] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5