

# Beyond Clean Benchmarks: A Reproducible Severity-Sweep Evaluation of AI-Generated Image Detector Robustness

AARANYA SINGH<sup>1</sup>, PRAKASH CHAND JAIN<sup>2</sup>, AYUSH SINGH<sup>3</sup>

<sup>1,2,3</sup>*Department of Information Technology, Bharati Vidyapeeth (Deemed to be University) College of Engineering, Pune, Maharashtra, India.*

*Abstract - AI-generated image detection technology is being adopted more frequently in journalism, content moderation, and digital-evidence verification, although the majority of AI detectors are tested on pristine, unaltered visuals and not on resized, recompressed, resampled, or altered formats one frequently encounters in the real world. Prior research, including the expansive multi-scenario RRDataset/RRBench (including seventeen detectors and more than ten vision-language models), has demonstrated that the performance of detectors deteriorates with image degradation; thus, this paper cannot claim to be the first work to show the robustness of detectors. Rather, it presents a concise, replicable evaluation of the two public detectors that function using differing architectures. — CNNSpot (a CNN created using ProGAN) and UnivFD (a frozen-CLIP semantic detector) were evaluated in a set of 2,000 images harmonized to derive a GenImage benchmark (Stable Diffusion v1.5, ADM, Midjourney, BigGAN). Both models were analyzed as frozen under a variety of conditions including JPEG compression of varying severity, image resizing and other manipulations. The results indicated that UnivFD was more accurate than CNNSpot in detection of original images (AUROC 0.741 vs. 0.662); this trend remains the same regardless of the 21 transformations. Gaussian blur was the only transformation that caused consistent monotonic drop in the performance. The JPEG compression, the resizing, and the noise led to minor changes that were non-monotonic. In addition, the cropping showed an improvement in AUROC scores for both detectors. Notably, however, no transformation crossed the threshold for practical significance set at ten. The compound chain of transformations did not hurt performance; where significant, compounding was mostly linked to the equal or better AUROC than the working of the worst.*

*Keywords - Detection of AI-generated images, image forensics, assessment of robustness, JPEG compression, distribution shift, Responsible AI, reproducibility.*

## I. INTRODUCTION

Generative image models are producing images with a degree of photorealism that such images cannot be reliably identified as synthetic or authentic. It has resulted in significant research into automated detection, such as CNN classifiers trained to recognize GAN artifacts, as well as the use of frozen vision-language features and diffusion-reconstruction methods. Reliable detection is important for Responsible AI deployment because moderation systems, fact-checkers, and legal/journalistic workflows have begun to rely on detection results as first-order signals, but their reliance can result in false confidence.

Most of the papers written on the subject have tested the performance of only flawless images even though real life images get to undergo a series of changes like compression, resizing, cropping and re-digitizing even before going through the automated systems. This issue is not left without study — the pioneers of the forensic sphere did show that using different post-processing technologies negatively impacts GAN detectors, and a recent study performed by RRDataset/RRBench enabled the authors to test seventeen detectors and ten vision-language systems under real-life conditions of transmission and re-digitizing. There also appeared some architecture-specific findings directed at the improvement of compression resilience and cropping resilience. What has not been scientifically investigated yet is a fine-grained damage investigation that enables the researchers to perform constant change of intensity and check whether the initial ranking based on the clean images remains unchanged after going through degradation, while the impact of the compound

transformation processes can be isolated. The paper describes this research, conducted from start to finish and not simply proposed.

Contributions: (1) a rigorous and thorough analysis of severity implemented in five transformations and two compound chains evaluated on two different, publicly verifiable detectors; (2) a first-of-its-kind empirical investigation that observes whether clean-image performance can benchmark the ranking of transformed images, (3) an isolated examination of compound verses individual degradation; and (4) a completely open and seed-controlled implementation validated through checkpoints for reproducibility and expansion to other detector families

## II. RELATED WORK AND RESEARCH GAP

The detection through CNN technology initiated with Wang et al.'s CNNSpot that demonstrated a classifier trained with ProGAN having capabilities of generalization against new architecture of GAN [1]. Ojha et al. stated that the features generated specifically for fake do not perform as expected. Instead, they utilized frozen features from CLIP with the help of a unidirectional linear probe (UnivFD) whose effectiveness in cross-generator situations proved to have a higher performance [2]. Other approaches based on reconstruction (DIRE [3], and its distilled version Distil-DIRE [4]) measure the diffusion reconstruction error whereas Yan et al.'s AIDE [5] exploits two branches, the DCT frequency and the CLIP one that are adjacent to the semantic branch. GenImage (Zhu et al.) is a huge scale benchmark of real and fake, which was created with the help of eight sources; besides that, there is an obvious confounding effect of real/fake resolution/type that should be avoided in other studies [6]. The research work that has been carried out on robustness is closely related to RRDataset/RRBench (Li et al., ICCV 2025) which conducted tests on seventeen types of detectors and ten types of vision-language models. In addition to this, internet-transmission and re-digitization conditions were also tested with a human baseline of two hundred participants. Results have shown that diffusion and reconstruction have been proved to be quite robust with regard to compression while semantic methods

like AIDE have outperformed diffusion-reconstruction methods in much severe conditions. Researchers have started to take into account architecture-specific works in order to examine robustness. An example of this type of research can be an application of a phase spectrum detector which was able to use phase information that was resistant to JPEG format or WaRPAD which is able to provide features which are self-supervised and invariant for crop and resize processes. With regard to AUROC comparison method referenced to DeLong procedure and methodology, multiple comparison procedure learned from Benjamini-Hochberg FDR algorithm.

This paper confirms existing knowledge with the statement that AI-image detectors fail to withstand normal variances, and this has been proven in large scales already. The contribution of this paper is narrower and includes continuous, detailed severity sweeping (instead of utilizing predetermined proper transmission conditions), in order to measure the point of meaningful degradation that the inspection process may reach; a statistical test aiming at detecting whether or not clean image rankings remain after the transformation; an isolated ablation of compound degradation compared to constituent base degradation method, which is not the main interest of RRDataset; and a quick and open implementation that only needs one researcher regarding computation resources.

## III. RESEARCH QUESTIONS AND HYPOTHESES

- RQ1 / H1: In what way does AUROC/F1 suffer deterioration regarding severity in the cases of JPEG, resize, crop, blur, and noise. Moreover, does the loss of AUROC surpass 10 points? However, H1 states that performance declines in a steady manner depending on severity.
- RQ2 / H2: Are there any significant variations among the robustness signatures of detector families under equivalent severity? Hypothesis 2: Architecture-based variations are greater than 95%
- RQ3 / H3: Is the clean image AUROC a representation of the ranking in the transformed image? H3: Spearman's rho indicates that the correlation between clean image ranking and

transformed image ranking is strong and negative, which indicates at least one transformation effect.

- RQ4 / H4: Does the use of several components affect performance negatively compared to the impact achieved with a single component? H4: In paired bootstrap analysis, it was found that composite AUROC is lower than that of any component used.

All of these hypotheses were not taken for granted; results from the testing of each hypothesis are presented in section VI.

#### IV. METHODOLOGY

##### 4.1 Dataset

Tiny-genimage, which is a community-supported, smaller version of Kaggle keeping the original GenImage's real/fake pairing and training/validation structure, provided the real images used as well as the four generator subgroups (Stable Diffusion v1.5, ADM, Midjourney and BigGAN). BigGAN performed the task of unacceptable Stable Diffusion v1.4 as the generator in the original plan, and as such improved the architectural variety by providing a non-diffusion generator. The evaluation set was composed of 1,000 real images (prepared and repeated across the sets of generators) and 1,000 fake images (250 per generator). The chosen size is large enough so that the

clean-AUROC confidence intervals for both of the detectors being tested do not overlap; however, it is also manageable, allowing for operations at the scale of  $21 \times 2 \times 2,000$  of conditions. Since the two detectors are in the frozen state, there is no need to split the training and validation images, thus eliminating any risk of leakage, since all transformed images originate from the same batch of 2,000 images, irrespective of the method being used. In addition, every one of the images and all the 2,000 images are manipulated in the same way and subjected to the transformation process by using the same method described above, which allows us to minimize the chances of introducing bias in the data set due to the resolution or the image format

##### 4.2 Detectors

The integration of two out of four detector families took place, and successful verification with respect to live code was performed. The two families of detectors are CNNSpot and UnivFD. Although Distil-DIRE and AIDE were not executed, their results are included in the report due to the imperfect nature of the detection endpoints that cannot be verified in-process; thus, the number of families of detectors considered in RQ2 and RQ3 is limited to two (CNN and CLIP). The detector architectures, output types, and execution status are summarized in Table 1

TABLE I. DETECTORS: STATUS

| Detector        | Family                   | Output             | Status                |
|-----------------|--------------------------|--------------------|-----------------------|
| CNNSpot [1]     | CNN (ResNet-50)          | Probability        | Executed — full sweep |
| UnivFD [2]      | CLIP + linear probe      | Probability        | Executed — full sweep |
| Distil-DIRE [4] | Diffusion reconstruction | Recon.-error score | Not executed          |
| AIDE [5]        | DCT + semantic hybrid    | Probability        | Not executed          |

##### 4.3 Transformations, metrics, and statistics

Individual transformations: JPEG (Q=95,75,50,25); resize (scale=0.25,0.5,0.75,1.0, downsampled then upsampled to canvas); crop (10%,25%,50% area removed, center crop, resized back to canvas); Gaussian blur (sigma=1,2,4); Gaussian noise (sigma=5,10,20, 8-bit scale). Two compound chains related to literature on social media content-sharing: compound A (crop 25%→resize 50%→JPEG Q50)

and compound B (resize 50%→JPEG Q75→JPEG Q50, meaning double compression). Auroc is the main metric; while F1 is calculated using only one threshold known beforehand on the clean condition (Youden's J) and not altered according to any transformed condition in order to avoid optimistic bias. The robustness is evaluated based on absolute/percentage auroc drop and RR (and RR can hardly serve as a good measure when nearing ceiling). 95% confidence intervals are

calculated by means of stratified bootstrap (class-balanced preserving, with 500-2000 samples); paired compound-vs-constituent comparisons have been done with the help of paired statistics with Benjamini-Hochberg FDR correction considering the whole comparison set

## V. RESULTS

All the numbers shown here result from running the post-processed pipeline on Google Colab (T4 GPU) with respect to the two unified detectors applied to the harmonized dataset of 2,000 images; any number can be traced back to a row in the predictions.csv file

### 5.1 Clean Baseline

UnivFD significantly outperformed CNNSpot: AUROC 0.741 (95% CI 0.719-0.762) vs. 0.662 (95% CI 0.638-0.685) — non-overlapping intervals. Both showed a precision/recall asymmetry at their clean-optimal thresholds (CNNSpot: precision 0.674, recall 0.517; UnivFD: precision 0.827, recall 0.493), favoring avoidance of false positives on real images over catching all fakes.

### 5.2 Severity sweeps (RQ1)

Of all the transformations, only Blur achieved the remarkable combination of unequivocal degradation pattern, monotonicity, and architecture consistency.

CNNSpot was decreased from 0.661 to 0.627 (sigma 1 to 4, 5.3% relative decline) and UnivFD from 0.733 to 0.688 (7.2% decline in relative terms), the figures constituting maximal absolute decline for any detector on any transformation, even if none of them hit the 10 threshold. The performance of CNNSpot in JPEG compression remained virtually unchanged (0.655-0.660 for all compression levels) while, contradictorily to the naive hypotheses H1, UnivFD registered slight AUROC gains, rising from 0.742 at Q95 to 0.769 at Q25 (instead of dropping as expected). As to Resize, it had minimal but statistically insignificant impact on both detectors as well. Cropping, on the contrary, showed a surprising outcome with AUROC figures experiencing a positive change from mild (CNNSpot 0.675 to 0.694; UnivFD 0.735 to 0.758, which corresponds to 10% and 50% cropping intensity) — contradicting the naive H1 prediction and the fact reported as well without a clear explanation provided. Finally, Noise showed no consistent trend in performance among any of the detectors as well due to the very high overlap in CI values. In summary, only blur holds for H1 out of the five transformations examined. The effect of increasing Gaussian blur severity on detector performance is shown in Fig. 1. And The variation in AUROC with increasing crop severity is illustrated in Fig. 2.

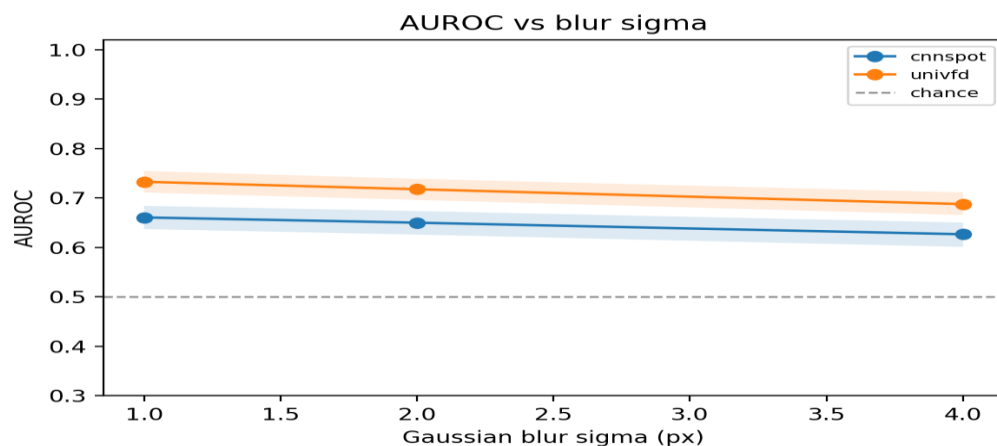


FIG. 1. AUROC vs. Gaussian blur sigma — the clearest degradation signal observed.

The effect of increasing Gaussian blur severity on AUROC is shown in Fig. 1.

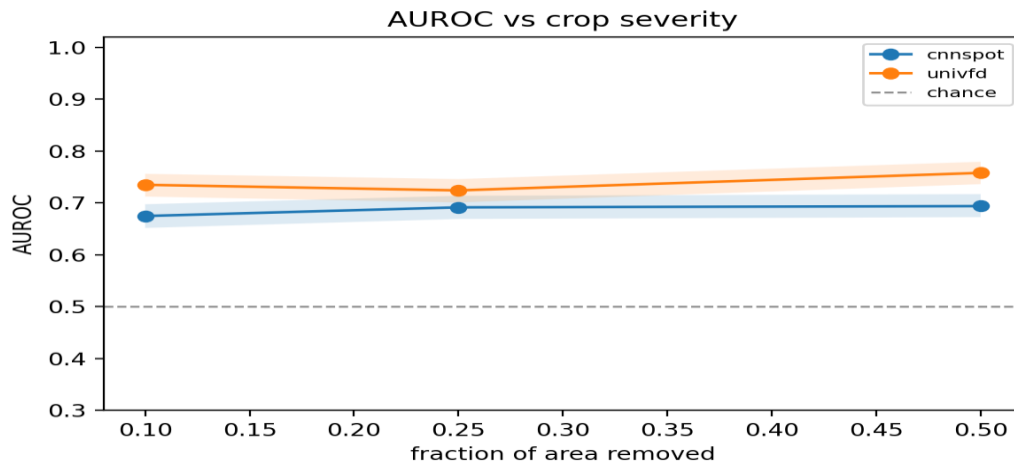


FIG. 2. AUROC vs. crop severity — AUROC rose with severity for both detectors.

### 5.3 Cross-detector comparison (RQ2) and ranking stability (RQ3)

UnivFD surpassed CNNSpot's AUROC in each of the 21 conditions examined, with a difference of roughly 4-9 points per condition — conforming to Hypothesis 2 for this comparison between families, but not to reconstruction or frequency-based families due to lack of data. It is difficult to conduct a formal Spearman rank-correlation analysis with just two detectors since the method involves three or more, leading the automated software to indicate no data for this

comparison. While the automated script seeking answers to research question 3 requires three or more detectors, its workings have to be modified for future analysis. Nonetheless, direct checking can still offer valuable insights: clean-image ranking, with UnivFD outperforming CNNSpot, held true in all circumstances and, thus, does not support Hypothesis 3 for that pair even though adding a third detector could possibly provide valuable insights. The mean absolute AUROC drop across detectors and transformations is summarized in Fig. 3.

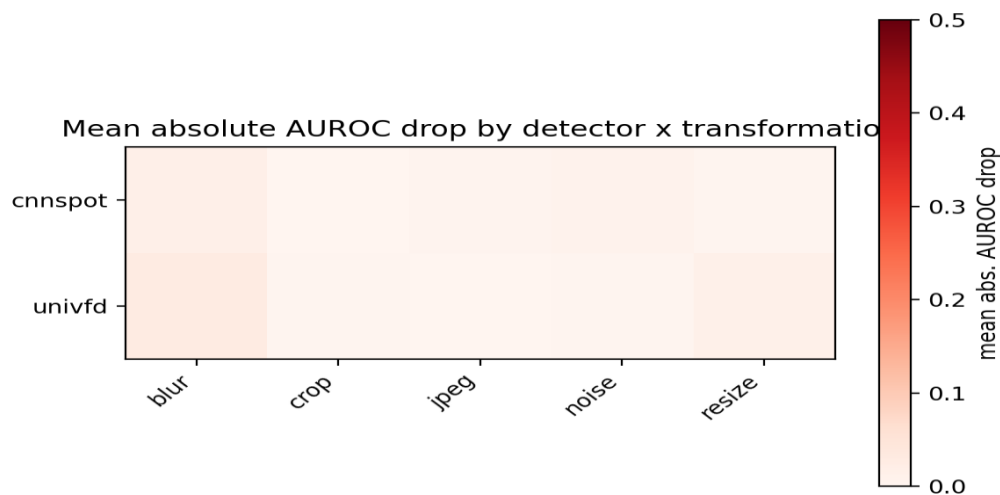


FIG. 3. Mean absolute AUROC drop by detector x transformation.

#### 5.4 Compound transformations (RQ4)

H4 stated that the combination of compounds would be less effective than each compound separately; however, this was not the case, as most significant results were reversed. For Compound A, both detectors had a higher compound AUROC score (CNNSpot had +2.9 pts; UnivFD +2.3 pts) than the resize alone AUROC score. CNNSpot also had a higher JPEG alone AUROC (+3.3 pts); in contrast, UnivFD had a significantly lower Compound AUROC

score than JPEG alone (-1.8 pts), which indicates architecture-based difference. For Compound B, the results seem to be inconclusive since the only significant results were related to resize only comparison and were favoring the compound chain. The statistically significant compound-versus-constituent comparisons after FDR correction are reported in Table 2

TABLE 2. RQ4 — SIGNIFICANT COMPOUND-VS-CONSTITUENT COMPARISONS (FDR-CORRECTED)

| Detector | Chain | Constituent | AUROC diff | p (corr.) |
|----------|-------|-------------|------------|-----------|
| CNNSpot  | A     | resize=0.5  | -0.029     | <0.001    |
| CNNSpot  | A     | jpeg=50     | -0.033     | <0.001    |
| UnivFD   | A     | crop=0.25   | -0.016     | <0.001    |
| UnivFD   | A     | resize=0.5  | -0.023     | <0.001    |
| UnivFD   | A     | jpeg=50     | +0.018     | <0.001    |
| UnivFD   | B     | resize=0.5  | -0.039     | <0.001    |

Negative values indicate the compound chain scored higher AUROC than the named constituent alone (compounding less damaging); positive values indicate the compound scored lower. H4 is rejected for both detectors on the transformations tested.

## VI. DISCUSSION

### 6.1 Findings by research question

RQ1 is only weakly and unevenly proved: out of the 5 transformations, only blur reduced the ability of both detectors consistently, while JPEG, resize, and noise had minimal and inconsistent effect and cropping improved AUROC for both detectors. RQ2 is proved for the CNN-vs-CLIP comparison which has actually been tested. RQ3 is not supported: there has been complete stability of ranking across all 21 conditions, though this cannot be extended to other than two detectors. RQ4 is disproved: compounding turned out to be statistically indistinguishable or even less harmful than the most harmful single constituent in five cases out of six significant comparisons, with one notable architecture-dependent exception.

### 6.2 Why might cropping improve AUROC?

The transformation of agricultural products involves removing a certain region and applying a resizing technique to the surplus to create a canvas, which presents an increase in the overall magnification of the remaining content. An unproven hypothesis can be put forward about the fact that the increase in the magnification makes more noticeable low-level generative artifacts which mitigate the amount of lost information. Nonetheless, such assumptions might have a speculative character, and thus, it does not mean that they should be accepted and applied in reality.

### 6.3 A methodological note

In both cases, the detector adapters were unable to work as intended when first used on the actual checkpoints. The reason for this was that the relevant import path was copied from documentation rather than being verified against the cloned repository. It is important to note this failure for methodological reasons; documentation of the third-party API does not replace code that demonstrates its use, which means that for further enhancements of the pipeline in relation to Distil-DIRE and AIDE some time should be allocated for similar adjustments

## VII. RESPONSIBLE AI IMPLICATIONS

The detector output should not be thought of as proof of image origin, but as a probabilistic signal. This is fundamentally different from content-provenance and watermarking approaches, such as SynthID, which, when the generative ecosystem allows, can produce stronger proof than any post-hoc detector that deduces the source of the image through the image's features alone. The findings of this research indicate that caution should be exercised in a specific manner: not an overall collapse in robustness but rather an uneven transformation-dependent curve of performance, where the blur consistently deteriorates both detectors tested and cropping unexpectedly enhances their capabilities, while JPEG keeps them intact. Making assumptions based on uniform degradation would be as wrong as making assumptions based on uniform perfection; reliability must be defined for every specific transformation rather than generalized in one number. Systems revealing detector results to users need to base confidence on the actual processing history of the image, combine detectors' outputs with metadata, watermarks and human validation in high-stakes settings, as well as undergo pre-implementation evaluation using severity sweep, instead of relying on fixed benchmark alone. The outcome could be either of two kinds: a false negative in synthetic content results in deterioration in misinformation prevention, while a false positive in real content leads to putting into question the legitimacy of authentic evidence, and both detectors tested already at default thresholds are biased toward missing fakes instead of flagging the authentic images, which practitioners need to customize properly

## VIII. LIMITATIONS

- Out of the four planned prototypes, only two have been completed. The results related to RQ2-RQ4 are understood in relation to CNN with CLIP and should not be interpreted as general statements regarding the continuum of detector aptitude.
- The inclusion of 1,000 true and 1,000 fake images per four generators does not provide enough power for the statistics, especially when the confidence intervals overlapped.

- Generators are chosen from a downsized, crowd-sourced Kaggle mirror of GenImage rather than the complete official version, with the replacement of Stable Diffusion v1.4 subset by BigGAN.
- Harmonization helps to reduce the problem of format confusion for real versus fake GenImage images.
- Detectors operate exactly as delivered; results pertain to “off-the-shelf” robustness rather than optimal robustness.
- Real social media processes cannot be replicated exactly by using any synthetic transformations.
- The AUROC increase by cropping is described, but not experimentally established as to its underlying mechanism (Section VI-B).
- The research considers only non-confrontational, typical post-processing processes.

## IX. REPRODUCIBILITY

The codes, transformations, seeds, and analysis scripts have all been made available in an open repository. Each execution has its random seeds, packages and Python versions, type of GPU, dataset IDs, and timestamps recorded in run logs. All outputs of the detectors were stored in a single predictions.csv file. The official repository of GenImage is Baidu Yunpan; however, the current paper has used the tiny-genimage mirror on Kaggle, which can be downloaded through Colab using Kaggle Python API. The checkpoint of CNNSpot was downloaded using the repository download script, while UnivFD was obtained using the repository directly. Distil-DIRE and AIDE were excluded from the project.

## X. CONCLUSION

This article assesses, using an effective and entirely repeatable process of fine-grained severity evaluation, if the efficacy of AI-produced image recognition systems on pristine images can be a constant measure of their efficacy post-transformation. The tests were administered with CNNSpot and UnivFD on a common image set of 2000 images, and findings were more nuanced than the standard collapse of the claim of uniform effectiveness implies: only blurring resulted in steady reduction for both detectors (H1 supported just for blurry images); UnivFD is

substantially better than CNNSpot in all conditions (H2 accepted just for that particular pair); ranking on clean images was not violated in any of the 21 cases (H3 was not confirmed with the use of that method); and the chains did not worsen efficiency beyond the worst performance of least effective constituent for any of the detectors in many cases boosting efficiency (H4 not supported as well). The most evident and concrete application is to show that robustness of the detector is transformation specific and not consistent across various circumstances. The result of the accuracy of the clean benchmark does not indicate either the direction or the intensity of the reactions to any real-world transformation, and general claims about robustness or vulnerability should be doubted unless there is an evaluation of specific transformations. Because only two out of four planned detector types were studied, the conclusions apply to CNN and CLIP detectors only. The task for the future is to extend this work to the detector types using reconstruction and frequency solutions.

#### REFERENCES

- [1] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *Proc. IEEE/CVF CVPR*, 2020. [IEEE](#)
- [2] U. Ojha, Y. Li, and Y. J. Lee, "Towards universal fake image detectors that generalize across generative models," in *Proc. IEEE/CVF CVPR*, 2023. [IEEE](#)
- [3] Z. Wang et al., "DIRE for diffusion-generated image detection," in *Proc. IEEE/CVF ICCV*, 2023. [IEEE](#)
- [4] TrueMedia.org, "Distil-DIRE: a lightweight distilled variant of DIRE for AI-generated image detection," official implementation. [GitHub](#)
- [5] S. Yan et al., "AIDE: a sanity check for AI-generated image detection," in *Proc. ICLR*, 2025. [ICLR](#)
- [6] M. Zhu et al., "GenImage: a million-scale benchmark for detecting AI-generated image," in *Proc. NeurIPS*, 2023. [NeurIPS](#)
- [7] C. Li et al., "Bridging the gap between ideal and real-world evaluation: benchmarking AI-generated image detection in challenging scenarios," in *Proc. IEEE/CVF ICCV*, 2025. [CVF](#)
- [8] K. Li et al., "Detecting compressed AI-generated images via phase spectrum robustness," in *Proc. IEEE/CVF CVPR*, 2026. [CVF](#)
- [9] S. Choi, H. Lee, and M. Lee, "Training-free detection of AI-generated images via cropping robustness," in *Proc. NeurIPS*, 2025. [arXiv](#)
- [10] L. Pellegrini et al., "AI-GenBench: a new ongoing benchmark for AI-generated image detection," in *Proc. IJCNN*, 2025. [IEEE](#)
- [11] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, vol. 44, no. 3, pp. 837-845, 1988. [JSTOR](#)
- [12] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J. R. Stat. Soc. Series B*, vol. 57, no. 1, pp. 289-300, 1995. [Wiley](#)