Department of
Computer Science
Faculty of Exact Sciences
Bar-Ilan University

RealStats: A Rigorous Real-Only Statistical Framework for Fake Image Detection

AISTATS 2026 spotlight talk

Haim Zisman Uri Shaham
Department of Computer Science, Bar-Ilan University

github.com/shaham-lab/RealStats

Problem: Detection Scores Need Meaning

Problem

Generative models are widespread, their innovation rate is increasing, and their outputs are already hard to distinguish from real images with the human eye.

Goal

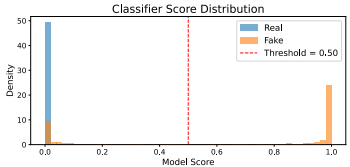
Produce a statistically meaningful method that can adapt to new generators, add new signals, and avoid restrictive assumptions:

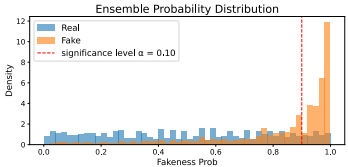
- **Interpretability:** return a null-based p-value under the real-image null.
- **Adaptability:** avoid assumptions about fake-image distributions and be able to incorporate new signals.

Why Existing Methods Are Not Enough

Interpretability Gap

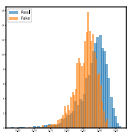
A classifier score may rank images, but its magnitude is not a probability under the real-image population. It does not answer: *how unusual is this image if it were real?*

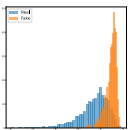




Adaptability Gap

Training-free statistics are more structured, but many still assume a fixed direction of deviation. The same statistic changes behavior when a hyperparameter changes.





Key Idea: Hypothesis Test

We reframe fake-image detection as a hypothesis test:

$$H_0 : x \sim \mathbb{P}_{\text{real}},$$
$$H_1 : x \not\sim \mathbb{P}_{\text{real}}.$$

For each statistic $s(x)$:

$$p(x) = 2\min(\widehat{F}_N(s(x)), 1 - \widehat{F}_N(s(x))).$$

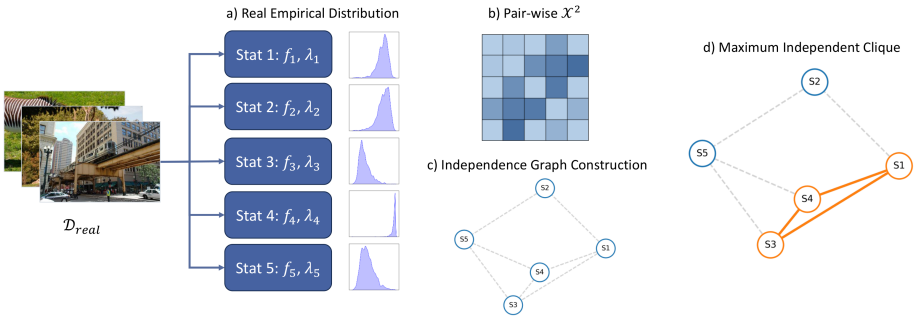
Small p-values indicate statistical deviation from the real-image reference distribution.

RealStats: Real-Only Null Distribution Modeling

Null Distribution Modeling

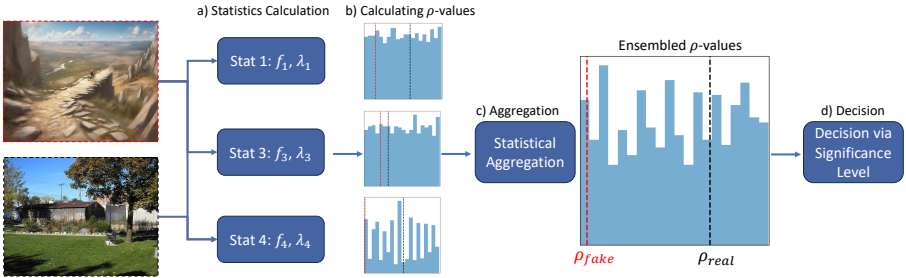
We first model the null distribution using real images only.

1. Compute scalar statistics $s(x)$ from frozen visual encoders under controlled perturbations.
2. Estimate one ECDF $\widehat{F}$ per statistic using the real reference set $\mathcal{D}_{\text{real}}$.
3. Convert each statistic into a two-sided p-value $p(x)$.
4. Select an independent subset via pairwise dependence tests and clique selection.



Inference: Aggregate Evidence Into One Interpretable Score

At test time, only the selected statistics are evaluated. Their p-values are aggregated into a single interpretable score.



Stouffer

$$z_i = \Phi^{-1}(p_i), \quad P_{\text{Stouffer}} = \Phi\left(\frac{\sum_i z_i}{\sqrt{K}}\right)$$

Combines moderate evidence across several statistics.

Min-p

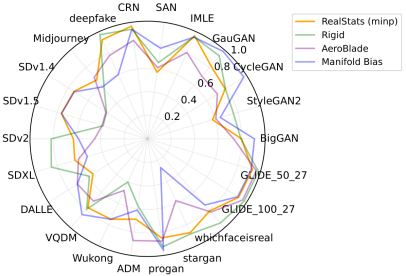
$$P_{\min} = \min_i p_i, \quad F_{P_{\min}}(t) = 1 - (1-t)^K$$

Emphasizes the strongest individual evidence.

Interpretability Does Not Come at the Price of Performance

Evaluation uses 187K images from CNNSpot, Universal Fake Detect, Stable Diffusion Face, Synthbuster, and GenImage, spanning GAN, diffusion, and real-world generator families. Real images are split into disjoint reference and evaluation sets.

Method	AUC	AP
Manifold Bias	0.761 ± 0.179	0.753 ± 0.169
RIGID	0.769 ± 0.194	0.765 ± 0.189
AEROBLADE	0.697 ± 0.161	0.697 ± 0.163
Ours (Stouffer)	0.756 ± 0.135	0.743 ± 0.133
Ours (Min-p)	$\mathbf{0.775 \pm 0.126}$	$\mathbf{0.756 \pm 0.119}$

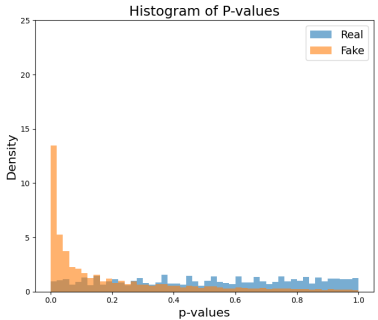


- Metrics are threshold-free: AUC and AP are computed per generator split.
- The same held-out real/fake subsets are used for all baselines.
- Lower variance indicates more consistent behavior across generator families.

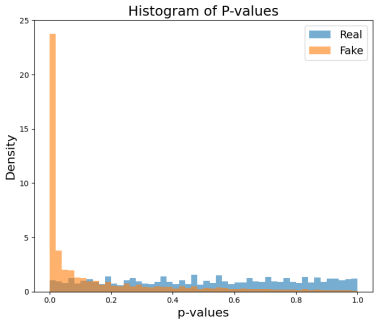
Comparability, lower variance, and interpretable outputs.

Adaptability in Action: Improving Performance on Challenging Generators

When a useful independent detector appears, we can incorporate it into the original null-modeling setup and adapt to challenging new generators.



Before



After

- Adding *ManifoldBias* improves p-value separation: fake samples shift toward zero while real samples remain close to the expected null behavior.
- Under H_0, the added signal follows the null, with p-values approximately $\mathrm{Uniform}(0, 1)$.

Practical Robustness and Limitations

Practical Robustness

- The aggregated score remains strongly predictive for fake samples under distributional shift.
- Robust to blur; JPEG gives a moderate drop.
- GPU parallelism improves throughput with moderate memory use.

Limitations

- Valid p-values need a representative real reference set.
- Distributional shift can violate probabilistic meaning.
- The selected statistic clique affects separability.

Bottom line: distributional shift can hurt interpretability, yet the method remains scalable, practical and comparable.

Takeaway

RealStats frames fake-image detection as statistical inference under the real-image null.

Our Contribution

A complete real-only framework: it formulates AI-image detection as a hypothesis test and integrates existing training-free statistics into a null-based p-value.

- Real-image null hypothesis testing.
- Statistic-level p-values from real-only reference ECDFs.
- Independent statistic selection and evidence aggregation.
- A unified, interpretable, adaptable, training-free framework.

Bottom line: detection can be interpretable, adaptable, and competitive without modeling fake-image distributions.

Thank you

Questions?

github.com/shaham-lab/RealStats