Beyond Jagged Edges: Content-Controlled Spectral and Edge Forensics for AI-Generated Face Detection

In this WildFake/YuNet face subset, AI-generated images contained less mid-to-high-frequency detail, while direct staircase-like edge evidence was weaker.

Maoyuan Xu · Department of Computer Science and Mathematics · Arcadia University

mxu@arcadia.edu

Faculty advisor: Kathy P. Macropol

2,000 balanced face images

776 content-matched pairs

400 untouched blind-test images

0.7482 fusion blind ROC-AUC

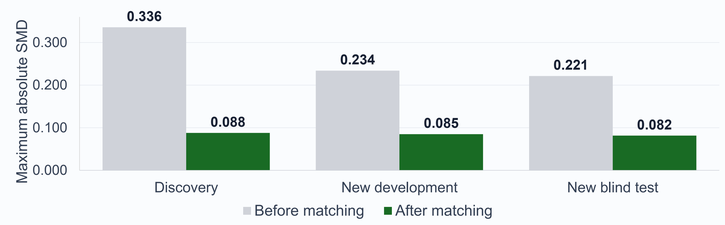
Research question

After matching on measurable content, which spectral and edge features differ between AI-generated and real faces? Do handcrafted features improve a pretrained image model?

Dataset and content control

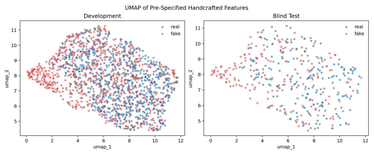
- 2,000 YuNet-filtered 200 × 200 JPEG face crops from WildFake/FaceGuard: 1,000 AI-generated and 1,000 real. WildFake spans several generator families, but this subset lacks image-level generator labels.
- Discovery and development contained 600 and 1,000 images; 400 images formed a locked blind test.
- Matching used low-pass luminance, color, symmetry, center/border composition and a blurred 16 × 16 layout. Edge and spectral outcomes were excluded.
- Hungarian assignment within cohorts, a 95th-percentile distance caliper and balance screening retained 776 pairs. All 2,000 images remained in predictive modeling.

Content balance after matching



Every cohort met the prespecified post-match maximum |SMD| ≤ 0.10.

Feature-space overlap and matched example



Development and blind-test points overlap substantially. The eight handcrafted features yield moderate discrimination and a diffuse class boundary.

Accepted pair example · cost 0.253



Pair F0251–R0200 shows one accepted Hungarian assignment. The cost used low-frequency content variables, including luminance, color, symmetry and composition. Edge and spectral outcomes were excluded. The pair illustrates content control, not authenticity evidence.

Analysis design

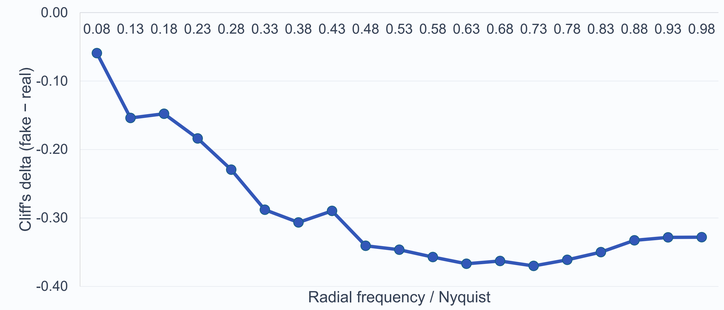
Eight features covered frequency, sharpness and edge shape: spectrum slope, HF70, HF90, Laplacian variance, 1 px high-pass SD, Canny density, orientation roughness and edge-width IQR.

Group differences used Mann–Whitney U, Cliff's delta, bootstrap intervals and FDR correction.

Logistic Regression, XGBoost, frozen ResNet18 and fusion underwent grouped five-fold selection before the locked 400-image blind test.

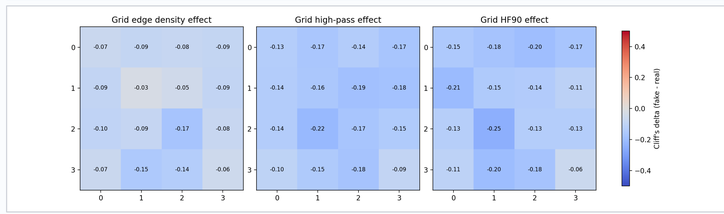
Spectral result

Strongest shift: 0.575–0.775 Nyquist
Cliff's delta ≈ −0.36 to −0.37



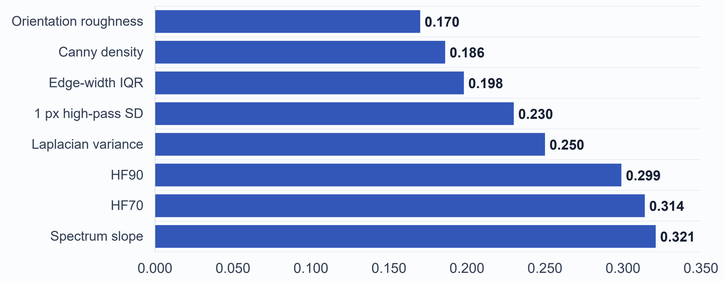
Negative Cliff's delta means real images had stronger band energy. The signal spans a continuous mid-to-high-frequency interval.

Where the signal appears



Real faces had higher high-frequency responses across much of the image. No single contour explains the difference.

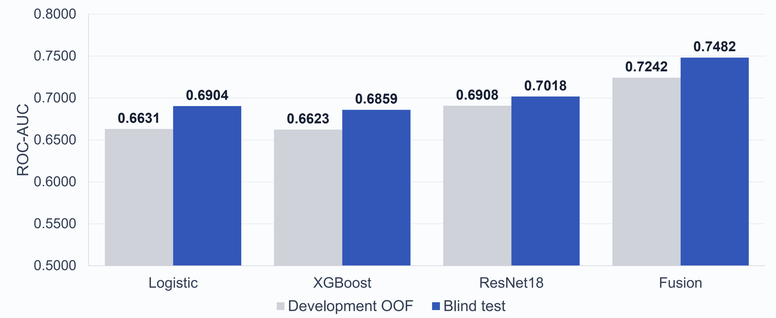
Primary handcrafted effects



Direct jagged-edge evidence was weaker

Only 3 of 9 staircase measures survived FDR. Overshoot (δ = −0.290) and monotonicity violations (δ = −0.239) were higher in real faces. Staircase-turn fraction showed no difference. The stronger finding is reduced fine detail in fake faces, not stronger jagged-edge aliasing.

Blind-test model performance



Fusion ROC-AUC 0.7482 · Balanced accuracy 0.6875
95% bootstrap CI 0.7001–0.7952

Blind-test confusion matrix

Actual / predicted	Real	AI-generated
Real	140	60
AI-generated	65	135

PR-AUC 0.7846 **Brier score 0.2018**
Fake sensitivity 0.675 **Real specificity 0.700**

Blind-test interpretation

ROC-AUC 0.7482 shows moderate discrimination; balanced accuracy 0.6875 weights both classes equally.

At the chosen threshold, 60 real and 65 fake faces were mislabeled. Brier score 0.2018 measures probability error (lower is better).

Fusion combines ResNet18 and handcrafted features; no prediction proves authenticity.

Limits of the evidence

Image-level generator labels were unavailable, so held-out-generator performance remains untested.

Independent datasets and demographic groups need separate evaluation.

A prediction cannot prove the authenticity of one image.

Next validation steps

Validate staircase measures on synthetic aliased and antialiased edges.

Compare aligned jaw and hairline regions at fixed edge orientations.

Test unseen generators and external images after compression and resizing.

References

1. Hong et al. WildFake: A Large-Scale and Hierarchical Dataset for AI-Generated Images Detection. AAAI, 2025.
2. He et al. Deep Residual Learning for Image Recognition. CVPR, 2016.
3. Kuhn. The Hungarian Method for the Assignment Problem. Naval Research Logistics Quarterly, 1955.
4. Benjamini and Hochberg. Controlling the False Discovery Rate. JRSS B, 1995.

Acknowledgments and reproducibility

Project author: Maoyuan Xu. Academic guidance: Kathy P. Macropol. Source face images are excluded from the public repository because redistribution rights are unclear.

github.com/spiderman-in-philly/AI-Generated-vs-Real-Face-Forensics

AI assistance disclosure: OpenAI Codex supported code and document editing. The author directed and validated the analyses.