Skepthical review: *A Multi-View Likelihood-Ratio Ensemble of Normalizing Flows for Out-of-Distribution Detection in Weak Lensing Maps*

Summary

The paper proposes a dual-view ensemble for out-of-distribution (OoD) detection in weak-lensing convergence maps using conditional normalizing flows. An OoD proxy is defined by applying a Gaussian blur to evaluation maps to mimic suppression of small-scale non-Gaussian structure while largely preserving large-scale power (Sec. 2.1, Sec. 3.1). Two engineered feature "views" are extracted per map: View 1 uses multi-scale directional gradient statistics plus a radial power spectrum; View 2 uses a compact power-spectrum representation, a bispectrum-like proxy based on adjacent Fourier-mode triplets, and global non-Gaussianity moments (Sec. 2.2). Separate conditional flows model $p(x|\theta)$ for each view (θ: five simulation parameters), and an anomaly score is formed by variance-normalizing per-view NLLs using a small calibration subset and averaging them (Sec. 2.3–2.4, Eqs. (1)–(2)). On the task of distinguishing original maps from blurred counterparts, the method reports strong low-FPR performance (mean TPR $\approx$ **0.8919** over **0.1%–5%** FPR; Sec. 3.2) and qualitative robustness across a grid of cosmologies (Sec. 3.4). The approach is timely and physically interpretable, but key claims currently rest on a narrow OoD setting and limited baselining/ablations; the score's "likelihood-ratio" interpretation and the σ vs σ^2 normalization are unclear; the calibration protocol risks evaluation leakage unless carefully separated; and important implementation/dataset details are missing, limiting reproducibility and the strength of conclusions (Sec. 2–4).

Strengths

- Well-motivated application: OoD detection as simulation validation for weak-lensing analyses, with an emphasis on the low-FPR regime that is operationally relevant (Sec. 1, Sec. 3.2).
- Physically interpretable, complementary feature design targeting both spectral power and non-Gaussian/higher-order information that a blur should affect (Sec. 2.2).
- Conditioning the density models on cosmological/baryonic parameters is well-motivated to reduce spurious "anomalies" at extreme but valid θ, and the robustness analysis is aligned with this goal (Sec. 2.3, Sec. 3.4).
- Clear high-level probabilistic framing (flows modeling $p(x|\theta)$, NLL as base score) and an explicit definition of the ensemble score (Sec. 2.3–2.4; Eqs. (1)–(2)).
- Figures generally support the narrative with helpful qualitative examples and ROC/score distribution visualizations (Sec. 3.1–3.3).

Major issues

1. **Empirical scope is narrow (single dataset and essentially one main OoD mechanism: Gaussian blur), and baselines are limited and not clearly matched. The comparison to the published VCSF baseline appears not to be run under identical conditions (data split, blur strength, metric definition), making the "six-fold improvement" claim hard to interpret (Sec. 2.1, Sec. 3.1–3.2, Conclusions).** This limits evidence that the method generalizes to realistic simulation–data mismatches (noise, masking, PSF/beam effects, baryonic modeling differences, small-scale power changes) rather than detecting one specific blur signature.

   Recommendation: In Sec. 3, broaden the evaluation by (i) sweeping blur strength σ and/or using a distribution of σ values (report performance vs σ), and (ii) adding additional physically motivated OoD scenarios (e.g., shape-noise mismatches, masking/inpainting artifacts, correlated noise, altered small-scale power, alternative baryonic feedback prescriptions if available). Add apples-to-apples baselines under the *same* splits and OoD settings: Flow1-only, Flow2-only, a single conditional flow on concatenated features, and at least one simpler model on the same features (e.g., conditional Gaussian/GMM, one-class SVM, or logistic/regression classifier trained to separate InD vs blurred). For VCSF, either re-run it on your exact dataset/proxy/metric (preferred) or clearly label the comparison as approximate and soften the quantitative "six-fold" claim in Sec. 1 and Conclusions.

2. **The anomaly score is repeatedly described as a likelihood-ratio (or likelihood-ratio-type) statistic, but the presented construction is an averaged, calibration-normalized NLL and does not correspond to a Neyman–Pearson log-likelihood ratio without additional assumptions. Additionally, Eqs. (1)–(2) divide by σ_k^2 (variance), while surrounding language suggests standardization/z-scoring (division by σ_k). This ambiguity affects interpretation, units/weighting across views, and reproducibility (Sec. 1, Sec. 2.4, Sec. 3.2–3.3; Eqs. (1)–(2)).**

   Recommendation: In Sec. 2.4, either (a) provide a clear derivation/assumptions under which the proposed statistic approximates a likelihood ratio (explicitly defining the alternative hypothesis and how μ_k, σ_k relate), or (b) remove/replace "likelihood-ratio" terminology throughout (Sec. 1, Sec. 2.4, Sec. 3.2–3.3, captions, Conclusions) and describe it precisely as a calibrated/standardized (or precision-weighted) ensemble NLL score. Separately, resolve whether σ_k^2 vs σ_k is intended: if σ_k^2 is deliberate, justify the weighting/units and update terminology accordingly; if not, correct Eqs. (1)–(2) and regenerate results if needed. Include a short comparison of alternative normalizations/fusions (raw NLL average, z-score, learned linear fusion on calibration set) in Sec. 3.3.

3. **Calibration protocol may introduce evaluation leakage or optimistic reporting: Sec. 2.4 states calibration uses 200 maps "drawn from the evaluation set" to compute μ_k and σ_k^2. Unless the calibrated subset is strictly disjoint from all ROC/metric computation and the selection is fixed or repeated over multiple draws, the reported ROC/mean-TPR can be biased. It is also not explicit that calibration uses only InD maps (Sec. 2.4, Sec. 3.2–3.3).**

   Recommendation: Clarify in Sec. 2.4 and Sec. 3.2: (i) calibration uses *only* InD maps; (ii) the 200-map calibration subset is disjoint from both training and the held-out evaluation set used for ROC/TPR; (iii) how the subset is selected (random seed/protocol) and whether calibration is repeated across multiple draws. Report sensitivity to calibration size (e.g., 50/100/200/500) and to subset choice (multiple random draws) and show variability in mean TPR over 0.1%–5% FPR.

4. **Core method claims (need for multi-view modeling, benefit of conditioning on θ, and benefit of calibration/normalization) are not supported by ablation studies. Without ablations, it is unclear whether the gains come from the second view, from conditioning, from the blur being easily captured by one feature subset, or from the calibration/fusion heuristic (Sec. 2.2–2.4, Sec. 3.3).**

   Recommendation: Add an ablation subsection (Sec. 3.3 or new) reporting ROC/mean-TPR@0.1%–5% FPR for: (i) Flow1 alone and Flow2 alone (each calibrated); (ii) ensemble fusion variants (raw-NLL mean, z-score mean, σ^2-weighted mean, max, learned linear fusion); (iii) conditional vs unconditional flows (remove θ); and (iv) a single conditional flow on concatenated features. Use results to refine claims in Sec. 1 and Conclusions.

5. **Feature extraction is under-specified, especially for Fourier-space operations and the bispectrum proxy, limiting reproducibility and interpretability. Missing are: exact feature dimensionality per view; power-spectrum bin edges/spacing and mapping from pixels to ℓ; treatment of windowing/apodization, complex modes, aliasing; definition of "high-frequency region"; precise construction of adjacent mode triplets $(k_1, k_2, k_1 + k_2)$ and the three spectral magnitude moments; and explicit formulas/pooling axes for directional gradient statistics (Sec. 2.2.1–2.2.2).**

   Recommendation: Expand Sec. 2.2 (or add an Appendix) with implementation-level detail: (i) explicit formulas for gradient statistics and pooling (over pixels/orientations/scales); (ii) FFT pipeline details (windowing/apodization, ℓ mapping, handling of complex conjugate modes, any anti-aliasing/deconvolution); (iii) exact definitions/masks for "high-frequency region"; (iv) exact triplet selection procedure and bispectrum-proxy formula with unambiguous indices (e.g., $\phi_{k_1+k_2}$); (v) the final per-view feature vector dimension and ordering (a table is ideal).

6. **Normalizing flow architecture and training protocol are insufficiently specified to reproduce results. "Affine coupling" is broad; details on coupling design, number of layers/blocks, hidden widths, activations, (actnorm/batchnorm), conditioning injection (θ embedding/concatenation locations), regularization (AdamW settings), batch size, learning-rate schedule, early stopping/validation, and whether both flows share architecture are unclear (Sec. 2.3).**

   Recommendation: In Sec. 2.3 (or Appendix), provide a complete specification: flow family (e.g., RealNVP/Glow style), number of coupling layers/blocks, permutation strategy, MLP widths/depth, activation functions and where applied, scale-parameter constraints (if any), normalization layers, θ preprocessing/embedding and injection points, optimizer hyperparameters (LR, weight decay), batch size, epochs/steps, validation split and early-stopping criterion, and random-seed handling. Include a small architecture table for both flows.

7. **Robustness and uncertainty are presented mostly qualitatively (e.g., Sec. 3.4 score distributions across cosmologies) without quantitative per-cosmology metrics or confidence intervals, and the headline mean TPR (0.8919) is reported without uncertainty over seeds/training/calibration sampling. For scientific deployment, variability is essential (Sec. 3.2, Sec. 3.4).**

   Recommendation: Add uncertainty quantification: report mean±std (or CI) over multiple training seeds and multiple calibration draws for the main metrics (mean TPR over 0.1%–5% FPR; also TPR@FPR$= 0.1\%, 1\%, 5\%$). In Sec. 3.4, report per-cosmology quantitative summaries (e.g., median InD score, median OoD score, separation, TPR@1% FPR) and summarize their distribution across the 100 cosmologies (table or compact plots).

8. **Inconsistency in the OoD proxy parameter σ (blur strength): methods state $\sigma = 2.0$ pixels, while Figure 1 caption (and related text) uses $\sigma = 1.5$ pixels. Because difficulty depends strongly on σ, this discrepancy directly affects the validity of reported results (Sec. 2.1, Sec. 3.1; Figure 1).**

   Recommendation: Resolve and align σ everywhere (Sec. 2.1, Sec. 3.1, Figure 1 caption/legend). If Figure 1 uses a different σ only for visualization, state that explicitly. If experiments used multiple σ values, label results per σ and report them separately (ideally as part of a σ sweep).

Minor issues

1. Dataset description and splitting are incomplete: provenance of simulations, mapmaking details (resolution, redshift distribution), whether/how shape noise is included (Figure 1 caption mentions noise), and precise composition of training/valida-

tion/evaluation/calibration sets (counts and disjointness) are not fully specified (Sec. 2.1, Sec. 2.4, Sec. 3.2).

Recommendation: In Sec. 2.1 and Sec. 2.4, add a concise dataset/splits table: number of maps per split, whether each InD evaluation map has a paired OoD blurred counterpart, and confirm disjointness between training/calibration/evaluation. Cite simulation suite/code, map resolution, redshift(s), and specify shape-noise level and whether noise is applied before/after blur.

2. Evaluation metric definition ("mean TPR over $\mathbf{0.1\%}$–$\mathbf{5\%}$ FPR") is not fully specified (discrete thresholds vs integral; spacing), and additional standard metrics (AUROC, AUPRC, TPR@fixed FPR) are not consistently reported, limiting comparability (Sec. 2.5, Sec. 3.2).

   Recommendation: In Sec. 2.5, define the averaging procedure precisely (e.g., integral of ROC over $[\mathbf{0.001}, \mathbf{0.05}]$ or average over fixed FPR grid with specified spacing). In Sec. 3.2, report AUROC and TPR@FPR=$\mathbf{0.1\%}, \mathbf{1\%}, \mathbf{5\%}$ (and AUPRC if class imbalance is relevant) alongside mean-TPR.

3. Robustness claims in Sec. 3.4 use qualitative language (e.g., "remarkably uniform") without numerical ranges, and the plots do not clearly indicate variability/dispersion beyond medians.

   Recommendation: Tighten wording in Sec. 3.4 and add quantitative ranges (min/median/max or percentiles across cosmologies). Where feasible, add error bars or shaded bands to show dispersion.

4. Figure uncertainty and methodological annotation are limited: ROC curves and spectral plots are shown without confidence bands; some plots omit sample sizes, thresholds, and exact processing parameters, reducing interpretability (Sec. 3.1–3.3; Figures 2, 4, 6 mentioned in the structured report).

   Recommendation: Add uncertainty visualization (e.g., bootstrap bands over maps and/or variation over seeds) to ROC and key summary plots; expand captions to include sample sizes and key preprocessing parameters (binning, normalization, σ, noise level).

5. Baseline comparison to VCSF is presented as a single approximate number without protocol parity details (same OoD proxy σ, same FPR range/metric, same dataset/noise), which can mislead (Sec. 3.2).

   Recommendation: In Sec. 3.2, either re-run VCSF under identical conditions and report matched metrics with uncertainty, or explicitly state it is taken from prior work and list key differences. Adjust claims accordingly.

6. Conditioning mechanism on θ is described briefly; it is unclear whether θ is standardized and how it is injected into coupling layers (Sec. 2.3).

Recommendation: Add 2–3 sentences in Sec. 2.3 describing $\boldsymbol{\theta}$ preprocessing (standardization), any embedding network, and exact injection point(s) into the coupling transforms.

7. Paper positioning relative to broader OoD detection literature is limited and dispersed (Sec. 1, Conclusions), making novelty and connections (likelihood-ratio/complexity-corrected scores, ensemble likelihood methods) harder to evaluate.

   Recommendation: Add a short dedicated related-work subsection (e.g., Sec. 1.1) covering likelihood-based OoD pitfalls and corrections, ensemble/multi-view approaches, and cosmology-specific anomaly detection. Explicitly state what is new vs adopted.

8. Conclusions may overstate general applicability given reliance on engineered features, known $\boldsymbol{\theta}$ at test time, and evaluation restricted to blur-based OoD on one suite (Sec. 4).

   Recommendation: Add a limitations paragraph in Sec. 4 stating these assumptions and outlining concrete next steps (broader OoD types, uncertain $\boldsymbol{\theta}$, learned representations, real survey data).

9. Notation could be made view-specific: $\mathbf{NLL}_k$ is written as a function of $\boldsymbol{x}$ (same symbol across views) even though each flow uses different feature vectors; this can confuse the mathematical meaning (Sec. 2.4; Eqs. (1)–(2)).

   Recommendation: Use $\boldsymbol{x_k}$ (or explicitly define $\boldsymbol{x}$ as the $\boldsymbol{k}$-th view feature vector within the sum) in Eqs. (1)–(2) and surrounding text.

Very minor issues

1. Formatting issues: inconsistent heading markers (e.g., a leading "#" in Sec. 2.5), placeholders like "Figure ??"/"Table ??" and inconsistent section numbering (Sec. 2.5, Sec. 3.1–3.4).

   Recommendation: Standardize headings/numbering and replace all placeholder cross-references with correct figure/table numbers; verify every referenced object exists and is consistently labeled.

2. Minor typography/notation inconsistencies (OoD vs out-of-distribution capitalization; dimension formatting; inconsistent naming of the score as "likelihood-ratio", "standardised", etc.) across text and captions.

   Recommendation: Proofread to standardize terminology and notation; adopt one accurate name for the score (consistent with the final Eq. (1)–(2) definition) and apply it throughout.

3. Bispectrum-proxy phase notation is potentially ambiguous (e.g., $\phi_{k_1+k_2}$ vs $\phi_{k_1+k_2}$), which can be misread (Sec. 2.2.2).

Recommendation: Use unambiguous subscripts/braces (e.g., $\phi_{k_1+k_2}$) and define the triplet-selection rule precisely in text.

4. Captions and some Results prose occasionally repeat setup details already described in Methods, reducing concision (Sec. 3.1–3.2).

   Recommendation: Trim repeated exposition by referencing Sec. 2.1–2.4, and use the freed space for key missing details (uncertainty, baselines, ablations).

Key statements and references

- • **Our proposed ensemble method achieves a mean True Positive Rate (TPR) of 0.8919 in the critical False Positive Rate (FPR) range of 0.1% to 5%, representing a roughly six-fold improvement over the published Variational Conditional Scattering Flow (VCSF) baseline score of approximately 0.15 for this task.**
- *Reference(s):* (none)
- • **The Receiver Operating Characteristic (ROC) curve shows that in the low-FPR regime relevant for scientific applications, the method maintains a high TPR, exceeding 0.8 at an FPR of 1% and approaching 0.97 at an FPR of 5%, indicating strong detection performance at very low false alarm rates.**
- *Reference(s):* (none)
- • **Conditioning each flow on the known simulation parameters so that it models $p(x \mid \theta)$ rather than the marginal $p(x)$ ensures that the anomaly score measures how atypical a map is given its own cosmology, which prevents valid but extreme cosmologies from being falsely flagged as out-of-distribution and yields robustness of the detector across the full cosmological parameter space.**
- *Reference(s):* (none)

Mathematical consistency audit

This section audits **symbolic/analytic** mathematical consistency (algebra, derivations, dimensional/unit checks, definition consistency).

Maths relevance: light

The paper contains limited formal mathematics. The main explicit mathematics is the definition of an ensemble anomaly score (Eqs. (1)–(2)) formed by centering each flow's negative log-likelihood using a calibration mean and scaling by a calibration variance, then averaging across views. Other mathematical content is largely descriptive (feature definitions, conditional density modeling) without detailed derivations.

Checked items

1. ⚠ **General K-view fused score definition** (Eq. (1), Introduction, p.2)
   - **Claim:** Defines anomaly score as the average across K views of variance-normalized NLL: $s = (1/K)\sum_{k=1}^{K} \frac{\mathrm{NLL}_k(x|\theta)-\mu_k}{\sigma_k^2}$.
   - **Checks:** symbol/definition consistency, dimensional consistency, sanity check (mean-centering effect)
   - **Verdict:** UNCERTAIN; confidence: medium; impact: critical
   - **Assumptions/inputs:** Each view k has its own conditional density model producing $\mathbf{NLL}_k$, μ_k and σ_k^2 are the mean and variance of $\mathbf{NLL}_k$ over an in-distribution calibration set
   - **Notes:** The expression is algebraically well-formed and dimensionless if NLL is dimensionless. However, the paper uses 'standardize/standardised' language elsewhere while dividing by variance (σ^2) rather than standard deviation (σ), creating ambiguity about intended normalization. Additionally, calling this a 'likelihood-ratio' is not justified by the formula alone.
2. ✔ **Two-view score specialization** (Eq. (2), Sec. 2.4, p.4)
   - **Claim:** Specializes Eq. (1) to $K = 2$: $s = (1/2)\sum_{k=1}^{2} \frac{\mathrm{NLL}_k(x|\theta)-\mu_k}{\sigma_k^2}$.
   - **Checks:** algebra between shown steps, notation consistency
   - **Verdict:** PASS; confidence: high; impact: moderate
   - **Assumptions/inputs:** There are exactly two views/flows ($k = 1, 2$)
   - **Notes:** This is the direct $K = 2$ specialization of Eq. (1) and is internally consistent with the stated ensemble of two flows.
3. ⚠ **View-specific input space vs shared symbol x** (Eqs. (1)–(2), pp.2 and 4; Sec. 2.2, pp.3–4)
   - **Claim:** $\mathbf{NLL}_k$ is evaluated on the feature vector for view k, though written as $\mathbf{NLL}_k(x \mid \theta)$.
   - **Checks:** symbol/definition consistency, type/space consistency
   - **Verdict:** UNCERTAIN; confidence: high; impact: moderate
   - **Assumptions/inputs:** Flow k is trained on features from view k only, Different views generally have different dimensionality/semantics
   - **Notes:** Because each view is a different feature representation, the argument should be view-indexed (x_k). Using a shared x is potentially a notational shortcut, but it is not explicitly defined and can be mathematically misleading (different domains inside a single summation).
4. ⚠ **Claim that the score is a likelihood-ratio test** (Text immediately after Eq. (2), Sec. 2.4, p.4; also Introduction around Eq. (1), p.2)
   - **Claim:** The formulation 'acts as a likelihood-ratio test/mechanism'.

- **Checks:** derivation logic (missing steps), definition consistency
- **Verdict:** UNCERTAIN; confidence: high; impact: critical
- **Assumptions/inputs:** A likelihood-ratio statistic would compare likelihood under two hypotheses/distributions
- **Notes:** No likelihood ratio is actually written (e.g., $\log p_{\text{in}}(x|\theta) - \log p_{\text{out}}(x|\theta)$), nor is an alternative hypothesis density specified. Without a derivation connecting the presented centered/scaled NLL to a likelihood ratio, the claim is not verifiable from the paper.

5. ✔ **Conditional density / NLL definition consistency** (Sec. 2.3, p.4; references to $p(x|\theta)$ also on p.9)
   - **Claim:** Each flow learns $p(x|\theta)$ and is trained by minimizing the NLL of training data.
   - **Checks:** symbol/definition consistency
   - **Verdict:** PASS; confidence: medium; impact: minor
   - **Assumptions/inputs:** $\text{NLL}_k(x|\theta) = -\log p_k(x|\theta)$, Base distribution is standard multivariate Gaussian and the flow is invertible
   - **Notes:** While the explicit change-of-variables formula is not shown, the stated relationship between conditional density modeling and NLL minimization is internally consistent within the paper's notation.

6. ✖ **Consistency of the OoD proxy blur parameter σ** (Sec. 2.1, p.3; Figure 1 caption, p.5; Figure 3 legend text, p.7 (image text indicates $\sigma = 1.5$))
   - **Claim:** The OoD proxy is a Gaussian blur with a specified σ in pixels.
   - **Checks:** definition consistency
   - **Verdict:** FAIL; confidence: high; impact: moderate
   - **Assumptions/inputs:** σ uniquely specifies the blur strength for the stated OoD proxy task
   - **Notes:** Sec. 2.1 specifies $\sigma = 2.0$ pixels, but Figure 1 caption states $\sigma = 1.5$ pixels (and figure-related text also references 1.5). This is an internal inconsistency in the definition of the OoD proxy used for evaluation.

7. ⚠ **Bispectrum-proxy phase-coupling expression readability** (Sec. 2.2.2, p.4)
   - **Claim:** Defines a phase-coupling proxy as mean of $\cos(\phi_{k_1} + \phi_{k_2} + \phi_{k_1+k_2})$ over adjacent mode triplets.
   - **Checks:** notation clarity/consistency
   - **Verdict:** UNCERTAIN; confidence: medium; impact: minor
   - **Assumptions/inputs:** ϕ_k denotes the phase of Fourier mode k, $k_1 + k_2$ denotes vector addition in Fourier space

- **Notes:** The written form '$\phi_{k_1+k_2}$' is typographically ambiguous without explicit braces/subscripts, and 'adjacent mode triplets' is not defined mathematically, preventing a precise analytic verification of the statistic being computed.

Limitations

- The paper provides very few explicit derivations; several key claims (notably the 'likelihood-ratio' interpretation) cannot be verified analytically without additional omitted steps/definitions.
- No explicit formula for the conditional flow likelihood (change-of-variables with Jacobian determinant) is shown, so only high-level consistency (not step-by-step correctness) can be assessed.
- This audit is restricted to the content present in the provided PDF text/images; it does not assess whether the chosen statistics are optimal or empirically valid.

Numerical results audit

This section audits **numerical/empirical** consistency: reported metrics, experimental design, baseline comparisons, statistical evidence, leakage risks, and reproducibility.

17 numeric consistency checks were executed: 16 PASS and 1 FAIL. The only detected inconsistency is a mismatch in the Gaussian blur σ value between the methods description and a figure caption; other arithmetic/logical checks (dataset counts, equation specialization, table range/mean/std sanity, percent-to-fraction conversion, and an approximate six-fold ratio claim) are internally consistent.

Checked items

1. ✔ **C1_dataset_total_maps** (Page 3, Section 2.1 "Dataset and out-of-distribution proxy")
    - **Claim:** "The full dataset contains $\mathbf{20,507}$ maps for training and $\mathbf{10,203}$ maps for evaluation."
    - **Checks:** parts_to_total (compute total size from parts)
    - **Verdict:** PASS
    - **Notes:** Checked training_maps + evaluation_maps equals **30710**.
2. ✔ **C2_view1_gradient_feature_count** (Page 3, Section 2.2.1 "View 1: Directional gradient and spectral features")
    - **Claim:** Directional gradient statistics: **6** orientations, repeated at **4** smoothing scales; retain mean and standard deviation for each scale and orientation.
    - **Checks:** dimensionality_from_counts
    - **Verdict:** PASS
    - **Notes:** Computed implied gradient feature count.

3. ✔ **C3_view1_total_feature_count_minimum** (Pages 3-4, Section 2.2.1 (View 1 includes gradient stats + 128-bin radial power spectrum))
    - **Claim:** View 1 comprises directional gradient statistics (implied 48 features) plus a radial power spectrum averaged into 128 bins.
    - **Checks:** dimensionality_from_counts
    - **Verdict:** PASS
    - **Notes:** Minimum implied View 1 dimensionality from stated components.
4. ✔ **C4_view2_total_feature_count_minimum** (Page 4, Section 2.2.2 "View 2: Compact power-spectrum and bispectrum vector")
    - **Claim:** View 2 includes a 128-bin radial power spectrum, three spectral magnitude moments, a phase-coupling cosine mean, and pixel-level skewness and kurtosis.
    - **Checks:** dimensionality_from_counts
    - **Verdict:** PASS
    - **Notes:** Minimum implied View 2 dimensionality from stated components.
5. ✔ **C5_equation1_vs_equation2_K_value** (Page 2 Eq. (1) and Page 4 Eq. (2))
    - **Claim:** Eq. (1) defines score $s = (1/K) \sum_{k=1..K} \frac{\text{NLL}*k - \mu_k}{\sigma_k^2}.**Eq.(2)definess = (1/2) \sum**$.} \frac{\text{NLL}_k - \mu_k}{\sigma_k^2
    - **Checks:** symbolic_specialization_consistency
    - **Verdict:** PASS
    - **Notes:** Checked Eq.(1) specialized to $K = 2$ matches Eq.(2) prefactor and summation limit.
6. ✖ **C6_blur_sigma_inconsistency_method_vs_figure** (Page 3 Section 2.1 vs Page 5 Figure 1 caption)
    - **Claim:** Section 2.1: Gaussian blur with $\sigma = 2.0$ pixels. Figure 1 caption: Gaussian blur ($\sigma = 1.5$ pixels).
    - **Checks:** repeated_constant_mismatch
    - **Verdict:** FAIL
    - **Notes:** Expected exact match; mismatch indicates inconsistency unless explained.
7. ✔ **C7_image_resolution_consistency** (Page 3 Section 2.2 and Page 5 Figure 1 caption)
    - **Claim:** Maps are reconstructed to 176×176; Figure 1 caption also references 176×176 pixel map.
    - **Checks:** repeated_constant_match
    - **Verdict:** PASS

- **Notes:** Checked repeated resolution constant matches.

8. ✔ **C8__table1__range__ordering** (Page 7, Table 1 “Score distribution statistics on the evaluation set.”)
   - **Claim:** Score range is $[-1.432, 9.368]$.
   - **Checks:** range_validity
   - **Verdict:** PASS
   - **Notes:** Checked range_min < range_max.
9. ✔ **C9__table1__mean__within__range** (Page 7, Table 1)
   - **Claim:** Mean score (all maps) is 2.641 and score range is $[-1.432, 9.368]$.
   - **Checks:** mean_within_range
   - **Verdict:** PASS
   - **Notes:** Checked mean within stated range.
10. ✔ **C10__table1__std__nonnegative** (Page 7, Table 1)
    - **Claim:** Score standard deviation is 2.618.
    - **Checks:** nonnegativity
    - **Verdict:** PASS
    - **Notes:** Checked nonnegativity.
11. ✔ **C11__table1__range__width__vs__std__sanity** (Page 7, Table 1)
    - **Claim:** Score std is 2.618 and score range is $[-1.432, 9.368]$.
    - **Checks:** cheap_sanity_check_range_vs_std
    - **Verdict:** PASS
    - **Notes:** Computed width and width/std for sanity; no strict violation.
12. ✔ **C12__training__hyperparams__lr__scientific__notation** (Page 4, Section 2.3)
    - **Claim:** Learning rate is 5×10^{-4}.
    - **Checks:** scientific_notation_parse
    - **Verdict:** PASS
    - **Notes:** Checked raw parsing and provided numeric value equal 5×10^{-4}.
13. ✔ **C13__epochs__integer** (Page 4, Section 2.3)
    - **Claim:** “Each flow is trained for 6 epochs.”
    - **Checks:** integer_validity
    - **Verdict:** PASS
    - **Notes:** Checked value is a positive integer.
14. ✔ **C14__coupling__layers__integer** (Page 4, Section 2.3)
    - **Claim:** “The architecture of each flow consists of 8 affine coupling layers.”
    - **Checks:** integer_validity

- **Verdict:** PASS
- **Notes:** Checked value is a positive integer.

15. ✔ **C15_calibration_subset_vs_evaluation_size** (Page 4 Section 2.4 and Page 3 Section 2.1)
    - **Claim:** Calibration subset is 200 maps drawn from evaluation set of $10,203$ maps.
    - **Checks:** subset_size_feasibility
    - **Verdict:** PASS
    - **Notes:** Checked calibration subset size does not exceed evaluation size; computed fraction.
16. ✔ **C16_low_fpr_range_conversion** (Page 5 Section 2.5; Page 6-7 Section 3.2; Page 8 Figure 4 caption)
    - **Claim:** Low-FPR metric range is from 0.1% to 5% FPR.
    - **Checks:** percent_to_fraction_conversion
    - **Verdict:** PASS
    - **Notes:** Converted percents to fractions and checked ordering.
17. ✔ **C17_six_fold_improvement_claim** (Page 6, Section 3.2 "Overall detection performance")
    - **Claim:** Mean TPR is 0.8919; baseline approximately 0.15; claim "roughly six-fold improvement".
    - **Checks:** ratio_check
    - **Verdict:** PASS
    - **Notes:** Checked ratio against $6\times$ with provided relative tolerance (baseline is approximate).

Limitations

- Checks are restricted to arithmetic/logical consistency using only explicit numeric values in the provided PDF text; no access to underlying data, code, or supplementary materials.
- Figure-based quantitative claims (e.g., ROC values at specific FPRs, power-spectrum ratios) cannot be verified without extracting plotted data; plot-pixel/value extraction is excluded by scope.
- Several feature-vector dimensionalities are only partially specified; dimensionality checks provided are minimum implied counts based on stated components and may not equal the implementation if additional features were included but not described.