Feature Matching Benchmark: Component Evidence for Robustness Under Distribution Shift

Shlok Jadhav

August 2026

Abstract

This report studies where sparse visual correspondence systems lose robustness under controlled and natural variation. The benchmark keeps feature extraction, matching, geometric verification, metrics, storage, and analysis separate so each result can be traced to a fixed pair, configuration, seed, method version, and source revision. The central held-out study compares ALIKED N16 at 512 features with XFeat at 2,048 features on HPatches scenes. XFeat is substantially faster on CPU, while ALIKED is more precise at a strict correspondence threshold. Paired stage signals and selected match-level replay support an association and repeatability mechanism hypothesis for the viewpoint gap. The evidence is a feature-family operating-point comparison, not a detector-versus-descriptor causal decomposition.

1 Question and scope

The benchmark asks which stage of a visual correspondence pipeline controls robustness: detection, description, matching, or geometric verification. A primary observation is an image pair, pipeline, condition, severity, and seed. Sparse pipelines expose detector, descriptor, matcher, and homography verifier boundaries. Complete systems remain separate when their internal components cannot be exchanged without changing the method.

The initial benchmark does not train or fine-tune models. Learned methods use explicitly pinned pretrained checkpoints. The public repository contains code, configurations, licenses, and this manuscript source. Datasets, weights, raw results, figures, and local audit outputs remain ignored artifacts.

2 Evaluation protocol

The main evidence uses the pinned full HPatches sequence release with 120 pairs: 60 illumination pairs and 60 viewpoint pairs. Both pipelines use mutual brute-force matching, a 1,024-pixel maximum side, CPU execution, one warmup repetition, four counterbalanced measured repetitions, and shared RANSAC settings. The canonical held-out run contains 240 observations and 960 retained timing samples. Accuracy is reported as the pair-level mean of precision at 5 pixels. Latency is summarized by the pair-level median of total measured time.

All raw observations are immutable Parquet rows with configuration and provenance. Analyses retain source run IDs and observation hashes. Paired uncertainty is computed at the scene level where a comparison requires resampling. Selected match diagnostics follow a preregistered loss-candidate rule and are diagnostic rather than population estimates.

3 Results

3.1 Practical CPU frontier

Table 1 shows the practical operating-point comparison on disjoint held-out scenes. ALIKED retains higher strict correspondence accuracy. XFeat is substantially lower latency in both conditions.

The accuracy difference is -0.0345 for illumination and -0.1982 for viewpoint when computed as XFeat minus ALIKED. These values describe the registered feature-family operating points. They do not establish a universal ranking across hardware, image sizes, feature counts, or XFeat semi-dense modes.

Table 1: Held-out accuracy and CPU latency.

Condition	ALIKED accuracy	ALIKED ms	XFeat accuracy	XFeat ms	Speedup
Illumination	0.8217	456.3	0.7872	54.8	8.3×
Viewpoint	0.7452	1129.2	0.5470	139.4	8.1×

3.2 Viewpoint mechanism signals

The paired viewpoint analysis compares 60 exact pair identities across the two pipelines. XFeat has lower repeatability, recall, and inlier concentration. Grid coverage, convex-hull coverage, and spatial entropy are slightly higher for XFeat, with confidence intervals that include zero. This separates the observed gap from a simple spatial-coverage deficit.

Table 2: Paired viewpoint stage signals. Differences are XFeat minus ALIKED.

Signal	ALIKED	XFeat	Difference	95% interval
Repeatability	0.4798	0.3131	-0.1666	[-0.2164,-0.1091]
Precision at 5 px	0.7452	0.5470	-0.1982	[-0.2314,-0.1610]
Recall at 5 px	0.8507	0.5636	-0.2871	[-0.3409,-0.2297]
Inlier ratio	0.7019	0.4075	-0.2945	[-0.3328,-0.2527]
Grid coverage	0.7479	0.7677	+0.0198	[-0.0167,+0.0542]

These signals support a detector or descriptor association hypothesis for the selected operating points. They do not isolate those stages because ALIKED and XFeat differ in architecture, detector, descriptor, preprocessing, and feature count.

3.3 Selected match diagnostics

The diagnostic replay covers five deduplicated viewpoint-loss pairs and both pipelines, retaining 5,345 proposed matches. Match counts, correct-at-3-pixels counts, and RANSAC inlier counts exactly reproduce the source observations. Across the selected pairs, XFeat correct-match fractions range from 10.3% to 68.6%, while ALIKED ranges from 58.2% to 97.9%. In the lowest descriptor-distance quartile, incorrect-match fractions range from 16.3% to 80.7% for XFeat and 0% to 3.8% for ALIKED.

Descriptor distances are interpreted within each method only. Both pipelines use brute-force matching in this control, so matcher confidence and confidence calibration are unavailable.

3.4 Second-process replication

A new process invocation reused the registered configuration, pairs, checkpoints, preprocessing, feature caps, matcher, verifier, seed, and timing protocol. It completed 240/240 observations and 960/960 timing samples. All 28 comparison cells for precision, recall, repeatability, inlier ratio, match counts, correct counts, and inlier counts were exact at zero tolerance. ALIKED remained more accurate and XFeat remained lower latency in both conditions. Timing magnitudes moved across processes and source revisions, so the result is classified as output reproducibility with source-revision drift rather than strict same-revision repeatability.

4 Limitations

1. HPatches is planar and does not cover non-rigid motion, severe occlusion, or all deployment domains.
2. Efficiency measurements are CPU-only and descriptive. They are not hardware-normalized systems measurements.
3. ALIKED and XFeat form a complete sparse feature-family comparison. Fixed-keypoint descriptor swaps are required for causal stage attribution.

4. The match-level sample is selected from registered losses and is not an unbiased estimate over all viewpoint pairs.
5. Confidence-bearing matching is not present in the diagnostic control, so calibration and high-confidence catastrophic-failure claims remain deferred.
6. Descriptor distances are method-specific and are not compared across descriptor families.

5 Reproducibility

The release descriptor is `configs/releases/phase6-heldout-efficiency-v1.json`. It records the configuration digest, immutable source and replica run manifests, repeatability manifest, report registry, audit digest, expected counts, and reproduction commands. A local handoff is checked with:

```
python -m correspondence.analysis.release_bundle \
  configs/releases/phase6-heldout-efficiency-v1.json --repository .
```

The canonical source run is `hpatches-xfeat-aliked-efficiency-heldout-v1-local`. The report registry contains 28 finding rows and an indexed accuracy-latency figure. The release audit validates the source run, shared observation hashes, figure digest, and the repository rule that only the root README is tracked as Markdown.

6 Plain-language summary

ALIKED is slower, but it usually produces cleaner correspondences. XFeat is much faster, but loses more matches when the viewpoint changes. The evidence points toward difficulty finding and associating the same visual points, rather than a simple shortage of spatial coverage. This conclusion applies to the tested settings and does not prove that one internal component is solely responsible.

7 Conclusion

At the selected CPU operating points, XFeat trades strict correspondence accuracy for lower latency. The viewpoint gap is accompanied by lower repeatability, association quality, and inlier concentration, while coverage remains tied. The evidence supports an association and repeatability mechanism hypothesis for the evaluated feature families, not a causal detector-versus-descriptor claim. The next discriminating study is a fixed-keypoint descriptor boundary with the same matcher and geometric verifier.

References

[1] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. HPatches: A benchmark and evaluation of handcrafted and learned local descriptors. CVPR, 2017.

[2] P.-E. Sarlin, P. Lindenberger, and M. Pollefeys. LightGlue: Local Feature Matching at Light Speed. ICCV, 2023.

[3] G. Potje et al. XFeat: Accelerated Features for Lightweight Image Matching. CVPR, 2024.

[4] M. Tyszkiewicz, P. Fua, and E. Trulls. DISK: Learning local features with policy gradient. NeurIPS, 2020.