Akator
Image Matcher

Technical Report

Performance and Robustness

of the Akator Image Matcher

Copy-detection recall and precision on Open Images V7,
at a scale of one million indexed images.

Akator GmbH

Report no.	**AKR-2026-IM-001**
Revision	1.0
Product version	**1.2.18**
Evaluation data	Open Images V7
Index size	1 000 000 images
Date	June 2026

© 2026 Akator GmbH. All rights reserved.

Abstract

The Akator Image Matcher (version 1.2.18) is a perceptual image-matching system for copy- and reuse-detection. Its retrieval performance is characterised on a benchmark derived from Open Images V7 [1]: an index of 1 000 000 images (the “haystack”), and an out-of-index query set manipulated across 85 parameter configurations spanning 21 manipulation families in four classes, with 1 000 queries per configuration. Manipulation families, parameter ranges and step sizes follow established copy-detection methodology [2–4] and are specified in full in Section 3. Recall is reported per configuration, precision through the false-positive rate on 10 000 out-of-index queries.

The system is highly robust to photometric and compression manipulations: the original is retrievable with recall up to 100 % across all tested severities, and is also accepted at the deployed decision threshold for all but the strongest noise, blur and gamma settings, where the threshold trims recall. It retains a false-positive rate of 0.39 % at an index of one million images, while recall declines monotonically under increasing geometric distortion. Its principal limitation is reduced recall under coarse geometric change, which is characterised in full.

Keywords: copy detection; perceptual image matching; robustness evaluation; recall; false-positive rate; Open Images.

1 000 000	0.39 %	100 %	85
indexed images	false-positive rate on 10 000 out-of-index queries	peak recall on colour, compression & watermarks	parameter configurations across 21 families

Contents

1 Introduction

Rights-holders and online platforms increasingly need to detect when a protected image is re-used or re-published elsewhere. Reused images are rarely byte-identical: they are recompressed, recoloured, rescaled, watermarked, or cropped – frequently in order to evade exact-match detection. A useful matching system must therefore recover the original *despite* such manipulations, and must stay precise when the reference database holds millions of images.

This report characterises the Akator Image Matcher in exactly these terms. It measures recall across a broad, standardised battery of manipulations and precision against images outside the database, at an index of one million images, and reports both strengths and limitations. For this use case *recall* is the primary objective: a missed copy is the costly error, whereas a moderate false-positive rate is acceptable because flagged candidates are reviewed before any action is taken. All results are conditional on the data distribution and manipulation parameters defined in Sections 2 and 3; they are not guarantees for inputs outside those conditions.

2 Data

2.1 Image source

All images are drawn from Open Images V7 [1],[1] a publicly available corpus of several million natural photographs. A public, widely-used corpus makes the evaluation independent of any proprietary data and externally verifiable.

2.2 Index and query construction

The corpus is partitioned into two disjoint sets. The first, the *index* (or "haystack"), contains 1 000 000 images and represents a customer database. The second is an *out-of-index* pool of images that are never indexed; it supplies the negative queries below.

Positive queries. Each positive query is a transformed copy of an *indexed* image: one manipulation at one parameter setting is applied to an image that is itself in the index, so a correct system retrieves that original. 1 000 independent source images are used per configuration, yielding 86 000 positive queries in total.

Negative queries. A further 10 000 *unknown* images that are absent from the index are submitted. The out-of-index pool is perceptually de-duplicated against the index, so a near-duplicate of an indexed image cannot leak into it; any retrieval returned for such a query is therefore a false positive, and this set quantifies precision independently of recall. To make the test realistic, one half of the negative queries are themselves manipulated (a 30 % crop followed by JPEG recompression at quality 30): an edited image that is not in the index must still match nothing. The other half are unmodified.

3 Methods

3.1 System under test

The system under test is the Akator Image Matcher version 1.2.18, accessed through its production interface. For a query image it returns a ranked list of candidate matches from the index, each accompanied by a structural similarity score (SSIM) [5]. The system is evaluated as a black box: no internal parameter is modified and no implementation detail is exploited, so the measurements reflect the deployed product. The SSIM score serves as the decision variable, and a match is accepted when it exceeds a fixed operating threshold of 0.50. Queries are issued concurrently (16 in flight) against the 1 000 000-image index.

[1] V7 is the latest release of the dataset; the canonical reference paper [1] documents the V4 release, on which later versions build. Open Images V7 photographs are licensed under CC BY 2.0; annotations under CC BY 4.0.

Manipulation	Parameter	Tested values
Colour & exposure		
Brightness	Factor	$\times$ 0.5, 0.7, 1.3 and 1.5
Contrast	Factor	$\times$ 0.5, 0.7, 1.3 and 1.5
Saturation	Factor	$\times$ 0, 0.5, 1.5 and 2
Hue shift	Shift	30°, 60°, 90° and 180°
Gamma	Gamma	γ: 0.5 and 2
Grayscale	–	–
Vignette	–	–
Quality & compression		
JPEG recompression	Quality	q: 75, 50, 30, 20, 15, 10, 8, 5 and 3
Pixelization	Downsample ratio	r: 0.5, 0.25, 0.1 and 0.05
Rescaling	Downscale to	50 %, 25 % and 10 %
Gaussian noise	Std. dev.	σ: 0.05, 0.1 and 0.2
Gaussian blur	Std. dev.	σ: 1, 2, 4, 8 and 16
Geometry		
Crop	Area removed	5 %, 10 %, 15 %, 20 %, 25 %, 30 %, 40 %, 50 %, 70 % and 80 %
Padding	Border added	5 %, 10 %, 15 %, 25 % and 50 %
Rotation	Angle	2°, 5°, 8°, 10°, 12°, 15°, 20°, 30° and 45°
Perspective	Jitter	σ: 10, 25, 40, 50 and 75
Mirror (flip)	Axis	horizontal
Overlays & composition		
Watermark	–	–
Meme frame	–	–
Overlay onto image	Source coverage	75 %, 50 %, 40 %, 25 %, 15 % and 10 %
Collage	Grid	2×2, 3×3, 4×4

Table 1. Manipulation families and the exact parameter values tested. Each listed value defines one configuration; recall is reported per configuration in Section 4.

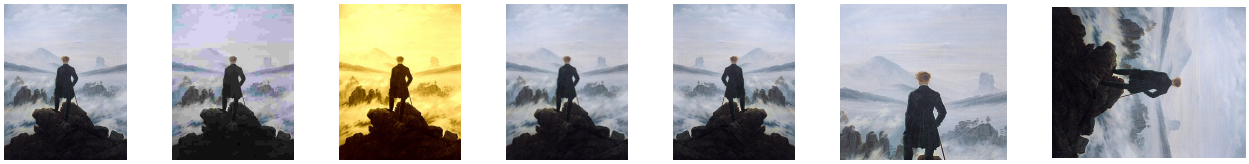

Figure 1. Representative manipulations of a single query image. Difficulty for a perceptual matcher tends to increase from left to right.

3.2 Manipulation suite and parameter grid

The benchmark spans 21 manipulation families grouped into four classes, each evaluated over a ladder of parameter settings (Table 1). Ranges and step sizes follow established copy-detection benchmarks: the Image Similarity Challenge 2021 with its AugLy-based augmentations [2, 3] and the INRIA Copydays protocol [4] (for example the JPEG-quality and crop ladders). Figure 1 illustrates representative manipulations. An unmanipulated *identity* control is included to confirm that exact duplicates are retrieved.

3.3 Metrics

For a configuration, let TP and FN denote the counts of correctly retrieved and missed originals. *Recall* is $\mathrm{R} = \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ and is reported per configuration in two complementary forms. *Retrievable recall* (the threshold-free upper bound) counts a query as recovered whenever its true original appears among the returned candidates at all; *operating recall* counts it only when the match also clears the operating threshold of Section 3.1 (SSIM ≥ 0.50). The two coincide for most families—in particular every geometric and overlay configuration,

where a missed original is genuinely absent from the candidates—but diverge where a manipulation depresses the structural similarity between a retrieved copy and its original, so that the copy is found yet scores below the threshold. This is most pronounced for heavy Gaussian noise (retrievable 98 %, operating 41 %) and for gamma at $\gamma = 2.0$ (99 % vs. 84 %); both figures are reported for every configuration in Table 3. For the negative set, the *false-positive rate* is the fraction of unknown queries that yield any accepted match. Ranking quality across the full operating range is summarised by the *micro-averaged precision* (micro-AP), the area under the precision–recall curve pooled over all queries, following the Image Similarity Challenge metric [2]. Because micro-AP and any recall-at-precision figure depend on the benchmark's positive-to-negative ratio rather than a deployment base rate, they serve for within-study comparison; the per-query false-positive rate is the base-rate-independent precision measure.

3.4 Statistics and protocol

Each point estimate carries a 95 % *Wilson score* confidence interval [6], which is well behaved for proportions near 0 and 1 and for the sample sizes used here (1 000 positives per configuration; 10 000 negatives). Images are drawn from the train split of Open Images V7; a fixed random seed (51) and fixed parameter settings are used, so the benchmark construction is deterministic and the evaluation is reproducible against the product interface, up to the negligible fraction of load-dependent transport errors noted in Section 2.

4 Results

We report precision first – it is independent of the manipulation set – and then recall across the full manipulation grid.

4.1 Precision

The system is precise at scale. Of 10 000 out-of-index queries, only 39 produce a match above the decision threshold – a false-positive rate of **0.39 %** (95 % CI: 0.29 – 0.53 %). This is a demanding test: half of these negative queries are themselves manipulated (Section 2), so that an edited image which is not in the database must still match nothing – and almost always does.

4.2 Recall by class

Recall, by contrast, depends strongly on the manipulation. Table 2 summarises it by class; the unmanipulated *identity* control is retrieved at 100 %, confirming that exact duplicates are found reliably and that the measurement pipeline is sound. The manipulations typical of image reuse – recompression, recolouring, exposure changes, rescaling and watermarking – fall in the high-recall classes; recall is lost mainly under coarse geometric change (Section 5). Table 2 reports both the retrievable recall and the operating recall at the decision threshold; the two are close for every class, separating only at the hardest noise, blur and gamma settings, where the structural-similarity gate rejects otherwise-retrieved copies (Section 3.3).

Because recall ranges from near-perfect to near-zero across the grid, a single pooled figure – one that scores every positive query across all manipulations together – is dominated by the hardest configurations and is not representative on its own: pooled over the entire grid – which by design includes near-impossible settings such as 80 % crop, JPEG quality 3 and a 3×3 collage – the system attains 62 % recall at 95.0 % precision (micro-AP 0.635). The per-class medians (Table 2) reflect typical behaviour.

		Retrievable (ceiling)		At threshold 0.50	
Class	**Configs**	**Mean (%)**	**Median (%)**	**Mean (%)**	**Median (%)**
Colour & exposure	20	98.9	99.7	98.1	99.5
Quality & compression	24	87.0	99.5	83.1	99.2
Geometry	30	40.1	23.6	40.1	23.6
Overlays & composition	11	9.8	0.0	9.8	0.0

Table 2. Recall aggregated over the configurations of each class, reported as both retrievable (ceiling) and operating-threshold recall: unweighted mean and median across severities. The mean weights every severity equally and is therefore pulled down by the hardest settings; the median is more representative of typical behaviour. Higher is better.

4.3 Robustness curves

Figure 2 shows recall as a function of severity for the families with a parameter ladder, as both retrievable recall (solid) and operating recall at the decision threshold (dashed). Photometric and compression families remain near 100 % retrievable across their entire range; the two curves track each other closely and separate only at the strongest noise, blur and gamma settings, where retrieved copies score below the threshold. Geometric families degrade monotonically as distortion increases, with retrievable and operating recall essentially identical. JPEG recompression is tolerated down to very low quality, whereas crop and rotation fall off rapidly once a large fraction of the spatial layout is displaced.

4.4 Operating point and threshold sensitivity

The decision threshold trades recall against the false-positive rate. Both are *conditional* rates – recall is measured on copies, the false-positive rate on unknowns – so they transfer across deployments independently of the base rate, unlike precision or any F-score (which depend on the positive-to-negative ratio of the test set). Figure 3 shows the full trade-off pooled over the grid. Rather than assert a single hand-picked threshold, we fix it by a base-rate-independent rule: the lowest threshold whose false-positive rate stays at or below a target of 0.50 %, chosen on a held-out calibration split and evaluated on a disjoint test split so the reported figures are not optimistically biased by selecting and scoring on the same data. This yields a threshold of 0.47, at which pooled recall is 62.7 % and the false-positive rate 0.48 % on the test split – essentially the 0.50 gate used for the per-configuration tables, which confirms that gate as a sound operating point rather than an arbitrary one.

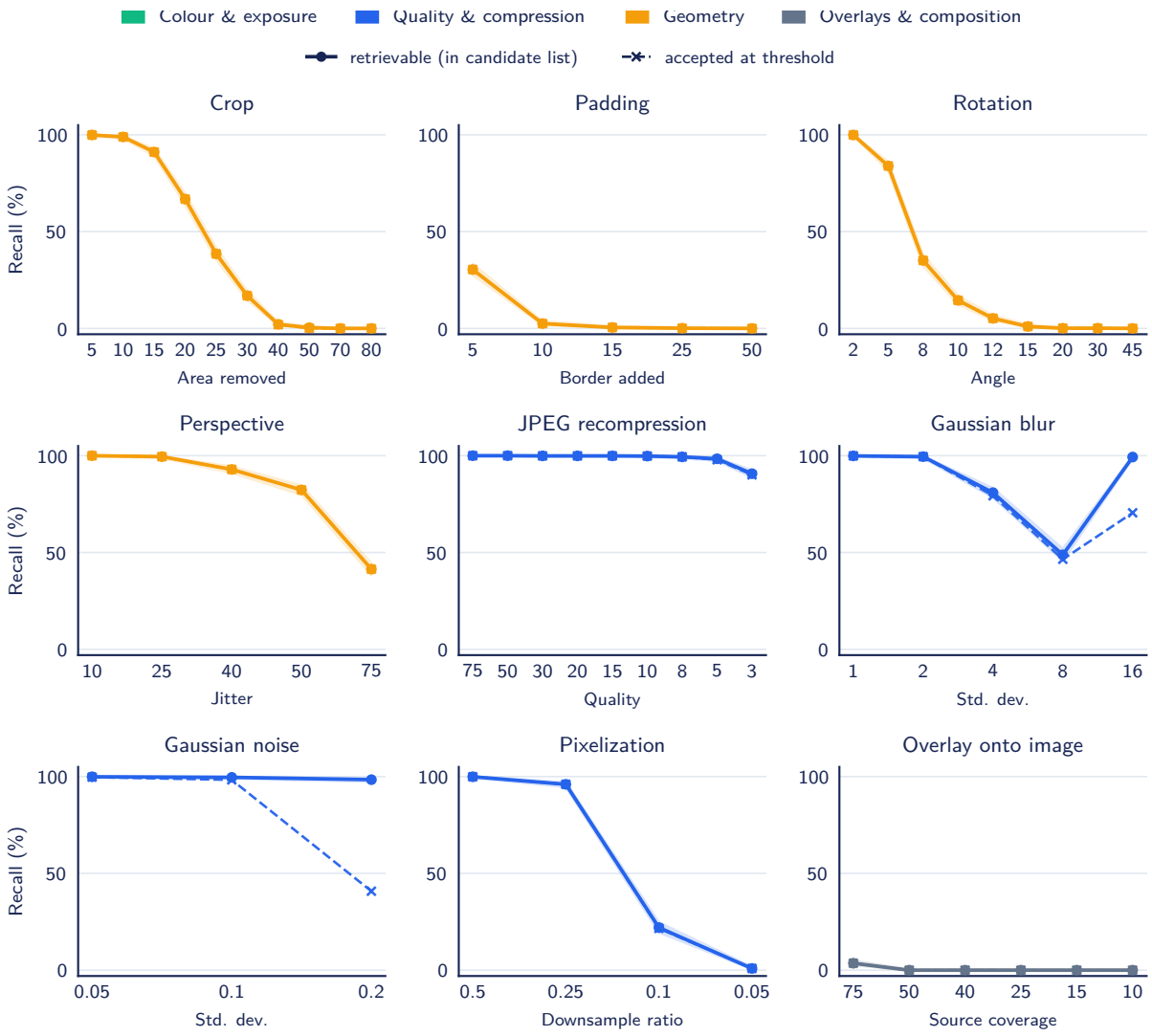

Figure 2. Recall versus manipulation severity (1 000 queries per point). Solid lines show retrievable recall (the true original among the returned candidates); dashed lines show operating recall at the decision threshold (SSIM $\geq$ 0.50). Markers are the tested parameter steps; shaded bands are 95 % Wilson score confidence intervals on the retrievable recall – narrow at this sample size and often no wider than the markers. Colour encodes the manipulation class.

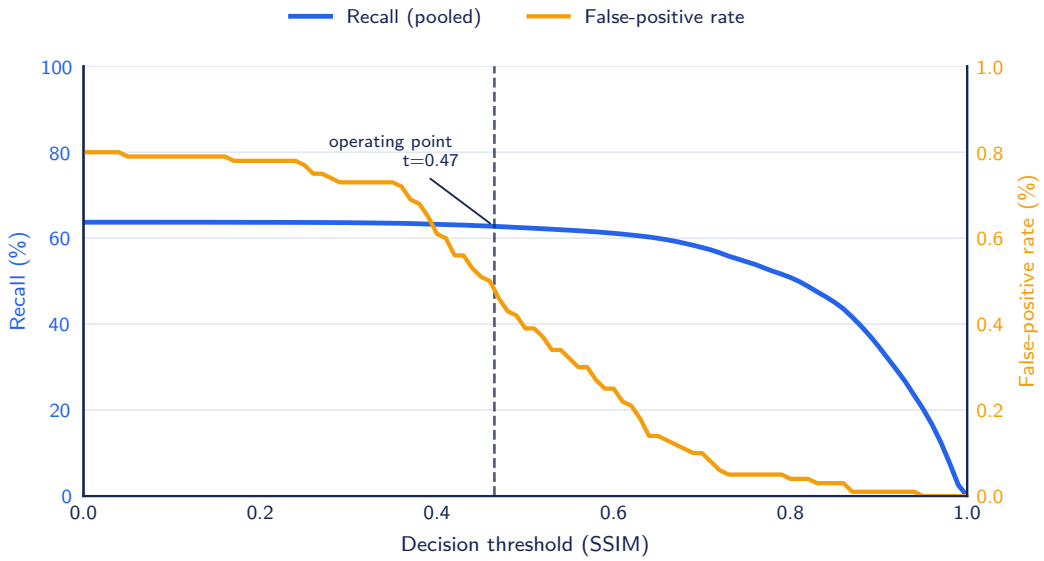

Figure 3. Pooled recall and false-positive rate as a function of the decision threshold; both are base-rate-independent conditional rates. The dashed line marks the operating point selected on held-out data to keep the false-positive rate at or below 0.50 %.

4.5 Complete measurements

Table 3 lists recall with confidence intervals for every one of the 85 configurations, grouped by class.

Table 3. Per-configuration recall on the 1 000 000-image index across the full severity grid (1 000 queries per configuration). *Retrievable* recall counts the original appearing among the returned candidates; *operating* recall counts only matches that also clear the decision threshold (SSIM ≥ 0.50). Confidence intervals are 95 % Wilson score.

		Retrievable (ceiling)		At threshold 0.50	
Manipulation	**Severity**	**Recall (%)**	**95 % CI**	**Recall (%)**	**95 % CI**
Control					
No manipulation	–	100.0	99.6 – 100.0	100.0	99.6 – 100.0
Colour & exposure					
Brightness	$\times 0.5$	100.0	99.6 – 100.0	100.0	99.6 – 100.0
Brightness	$\times 0.7$	100.0	99.6 – 100.0	100.0	99.6 – 100.0
Brightness	$\times 1.3$	99.7	99.1 – 99.9	99.7	99.1 – 99.9
Brightness	$\times 1.5$	98.3	97.3 – 98.9	98.3	97.3 – 98.9
Contrast	$\times 0.5$	100.0	99.6 – 100.0	99.5	98.8 – 99.8
Contrast	$\times 0.7$	100.0	99.6 – 100.0	99.9	99.4 – 100.0
Contrast	$\times 1.3$	99.8	99.3 – 99.9	99.8	99.3 – 99.9
Contrast	$\times 1.5$	99.7	99.1 – 99.9	99.6	99.0 – 99.8
Saturation	$\times 0$	95.8	94.4 – 96.9	95.7	94.3 – 96.8
Saturation	$\times 0.5$	99.9	99.4 – 100.0	99.9	99.4 – 100.0
Saturation	$\times 1.5$	99.9	99.4 – 100.0	99.9	99.4 – 100.0
Saturation	$\times 2$	99.1	98.3 – 99.5	99.1	98.3 – 99.5
Hue shift	30°	99.7	99.1 – 99.9	99.7	99.1 – 99.9
Hue shift	60°	97.6	96.5 – 98.4	97.6	96.5 – 98.4
Hue shift	90°	98.4	97.4 – 99.0	98.4	97.4 – 99.0
Hue shift	180°	93.4	91.7 – 94.8	93.3	91.6 – 94.7
Gamma	$\gamma = 0.5$	99.8	99.3 – 99.9	98.2	97.2 – 98.9
Gamma	$\gamma = 2$	98.9	98.0 – 99.4	84.5	82.1 – 86.6
Grayscale	–	100.0	99.6 – 100.0	100.0	99.6 – 100.0
Vignette	–	98.9	98.0 – 99.4	98.9	98.0 – 99.4
Quality & compression					
JPEG recompression	$q = 75$	100.0	99.6 – 100.0	100.0	99.6 – 100.0
JPEG recompression	$q = 50$	100.0	99.6 – 100.0	100.0	99.6 – 100.0
JPEG recompression	$q = 30$	99.9	99.4 – 100.0	99.9	99.4 – 100.0
JPEG recompression	$q = 20$	99.9	99.4 – 100.0	99.9	99.4 – 100.0
JPEG recompression	$q = 15$	99.9	99.4 – 100.0	99.9	99.4 – 100.0
JPEG recompression	$q = 10$	99.8	99.3 – 99.9	99.8	99.3 – 99.9
JPEG recompression	$q = 8$	99.4	98.7 – 99.7	99.4	98.7 – 99.7
JPEG recompression	$q = 5$	98.4	97.4 – 99.0	97.9	96.8 – 98.6
JPEG recompression	$q = 3$	90.7	88.7 – 92.3	90.1	88.1 – 91.8
Pixelization	$r = 0.5$	99.9	99.4 – 100.0	99.9	99.4 – 100.0
Pixelization	$r = 0.25$	96.0	94.6 – 97.0	96.0	94.6 – 97.0
Pixelization	$r = 0.1$	22.0	19.5 – 24.7	21.6	19.2 – 24.3
Pixelization	$r = 0.05$	0.9	0.5 – 1.7	0.9	0.5 – 1.7
Rescaling	50 %	99.9	99.4 – 100.0	99.9	99.4 – 100.0
Rescaling	25 %	99.1	98.3 – 99.5	99.0	98.2 – 99.5
Rescaling	10 %	57.0	53.9 – 60.0	55.8	52.7 – 58.9
Gaussian noise	$\sigma = 0.05$	99.9	99.4 – 100.0	99.7	99.1 – 99.9
Gaussian noise	$\sigma = 0.1$	99.6	99.0 – 99.8	98.3	97.3 – 98.9
Gaussian noise	$\sigma = 0.2$	98.4	97.4 – 99.0	40.8	37.8 – 43.9
Gaussian blur	$\sigma = 1$	99.9	99.4 – 100.0	99.9	99.4 – 100.0
Gaussian blur	$\sigma = 2$	99.5	98.8 – 99.8	99.5	98.8 – 99.8
Gaussian blur	$\sigma = 4$	80.9	78.3 – 83.2	79.3	76.7 – 81.7
Gaussian blur	$\sigma = 8$	48.8	45.7 – 51.9	46.5	43.4 – 49.6
Gaussian blur	$\sigma = 16$	99.3	98.6 – 99.7	70.5	67.6 – 73.2
Geometry					
Crop	5 %	99.9	99.4 – 100.0	99.9	99.4 – 100.0
Crop	10 %	98.9	98.0 – 99.4	98.9	98.0 – 99.4
Crop	15 %	91.1	89.2 – 92.7	91.1	89.2 – 92.7
Crop	20 %	66.8	63.8 – 69.6	66.8	63.8 – 69.6
Crop	25 %	38.5	35.5 – 41.6	38.5	35.5 – 41.6
Crop	30 %	16.9	14.7 – 19.3	16.9	14.7 – 19.3

continued on next page

		Retrievable (ceiling)		At threshold 0.50	
Manipulation	**Severity**	**Recall (%)**	**95 % CI**	**Recall (%)**	**95 % CI**
Crop	40 %	2.1	1.4–3.2	1.9	1.2–2.9
Crop	50 %	0.4	0.2–1.0	0.1	0.0–0.6
Crop	70 %	0.0	0.0–0.4	0.0	0.0–0.4
Crop	80 %	0.0	0.0–0.4	0.0	0.0–0.4
Padding	5 %	30.4	27.6–33.3	30.4	27.6–33.3
Padding	10 %	2.5	1.7–3.7	2.5	1.7–3.7
Padding	15 %	0.5	0.2–1.2	0.5	0.2–1.2
Padding	25 %	0.1	0.0–0.6	0.1	0.0–0.6
Padding	50 %	0.0	0.0–0.4	0.0	0.0–0.4
Rotation	2°	99.9	99.4–100.0	99.9	99.4–100.0
Rotation	5°	83.9	81.5–86.0	83.9	81.5–86.0
Rotation	8°	35.1	32.2–38.1	35.1	32.2–38.1
Rotation	10°	14.5	12.5–16.8	14.5	12.5–16.8
Rotation	12°	5.2	4.0–6.8	5.2	4.0–6.8
Rotation	15°	1.0	0.5–1.8	1.0	0.5–1.8
Rotation	20°	0.1	0.0–0.6	0.1	0.0–0.6
Rotation	30°	0.1	0.0–0.6	0.1	0.0–0.6
Rotation	45°	0.0	0.0–0.4	0.0	0.0–0.4
Perspective	$\sigma = 10$	100.0	99.6–100.0	100.0	99.6–100.0
Perspective	$\sigma = 25$	99.5	98.8–99.8	99.5	98.8–99.8
Perspective	$\sigma = 40$	92.9	91.1–94.3	92.9	91.1–94.3
Perspective	$\sigma = 50$	82.3	79.8–84.5	82.3	79.8–84.5
Perspective	$\sigma = 75$	41.4	38.4–44.5	41.4	38.4–44.5
Mirror (flip)	horizontal	99.3	98.6–99.7	99.3	98.6–99.7
Overlays & composition					
Watermark	–	99.4	98.7–99.7	99.4	98.7–99.7
Meme frame	–	5.2	4.0–6.8	5.2	4.0–6.8
Overlay onto image	75 %	3.6	2.6–4.9	3.6	2.6–4.9
Overlay onto image	50 %	0.0	0.0–0.4	0.0	0.0–0.4
Overlay onto image	40 %	0.0	0.0–0.4	0.0	0.0–0.4
Overlay onto image	25 %	0.0	0.0–0.4	0.0	0.0–0.4
Overlay onto image	15 %	0.0	0.0–0.4	0.0	0.0–0.4
Overlay onto image	10 %	0.0	0.0–0.4	0.0	0.0–0.4
Collage	2×2	0.0	0.0–0.4	0.0	0.0–0.4
Collage	3×3	0.0	0.0–0.4	0.0	0.0–0.4
Collage	4×4	0.0	0.0–0.4	0.0	0.0–0.4

5 Limitations and Intended Use

The complete measurements (Table 3) delimit the operating range. Retrievable recall is high and stable for colour, exposure, compression, noise, rescaling and watermarking across all tested severities. It declines for **coarse geometric change** that substantially displaces the spatial layout: large crop, rotation, padding/framing, and embedding the source into a host image (overlay, collage) reduce recall, in the strongest settings to near zero. For the core use case – recovering recompressed, recoloured, or watermarked/logo-stamped copies – such edits are rarely the relevant transformation, and the system is strong. A deployment whose primary goal is to match heavily cropped or rotated sub-regions should treat this as a material limitation. Geometric robustness is the natural target of future product versions, against which the same benchmark measures any progress directly.

A second, milder limitation concerns the decision threshold rather than retrieval. The threshold is set conservatively for precision (SSIM ≥ 0.50), so a copy that is correctly retrieved but whose structural similarity to its original has been lowered – by heavy noise, strong blur, or a large gamma shift – can score below it and be rejected. Operating recall then falls short of the retrievable upper bound for those settings (Table 3); a lower threshold would recover most of them at some cost in the false-positive rate, a trade-off the operator can tune. For the typical reuse manipulations the two coincide.

The conditions under which these figures hold are stated in the Introduction and the Legal Notice.

6 Conclusion

This report characterised version 1.2.18of the Akator Image Matcher on a one-million-image benchmark derived from Open Images V7. Two findings define its operating envelope: near-perfect, severity-independent recall against photometric, compression and watermark manipulations at a false-positive rate of 0.39 %, set against a monotonic decline in recall under strong geometric edits. The complete per-configuration measurements are in Table 3.

For its intended use – detecting re-used images that have been recompressed, recoloured, rescaled or watermarked – the system is therefore both accurate and precise at the scale of real customer databases. The manipulations it does not yet handle well are coarse geometric edits, which are rarely the means of image reuse; closing that gap is the clear next step, and this benchmark provides the yardstick to measure it.

7 Legal Notice

This document reports empirical measurements of the Akator Image Matcher (version 1.2.18) obtained under the conditions defined in Sections 2 and 3. The reported figures describe the observed behaviour of the system on the Open Images V7 image distribution and the stated manipulation parameters. They are statistical estimates, not deterministic guarantees, and do not characterise behaviour on inputs, data distributions, or manipulations outside those tested. Performance may differ for other datasets, content types, manipulations, or future software versions. Nothing in this document constitutes a warranty, a guarantee of fitness for a particular purpose, or a contractual assurance of performance. Akator GmbH accepts no liability for decisions made in reliance on these figures beyond what is expressly agreed in writing. The evaluation data originate from Open Images V7 and remain subject to their respective licences.

Revision	Date	Description
1.0	June 2026	Initial release (product version 1.2.18).

References

[1] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: *The Open Images Dataset V4.* International Journal of Computer Vision **128**, 1956–1981 (2020).

[2] Douze, M., Tolias, G., Pizzi, E., Papakipos, Z., Chanussot, L., Radenovic, F., Jenicek, T., Maximov, M., Leal-Taixé, L., Elezi, I., Chum, O., Ferrer, C.C.: *The 2021 Image Similarity Dataset and Challenge.* arXiv:2106.09672 (2021).

[3] Papakipos, Z., Bitton, J.: *AugLy: Data Augmentations for Robustness.* arXiv:2201.06494 (2022).

[4] Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., Schmid, C.: *Evaluation of GIST descriptors for web-scale image search.* In: Proc. ACM International Conference on Image and Video Retrieval (CIVR) (2009).

[5] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: *Image quality assessment: from error visibility to structural similarity.* IEEE Transactions on Image Processing **13**(4), 600–612 (2004).

[6] Wilson, E.B.: *Probable inference, the law of succession, and statistical inference.* Journal of the American Statistical Association **22**(158), 209–212 (1927).