

Paper Type: Original Article

Global Sharpness as an Image-Quality Gate for Gas-Pipeline Defect Segmentation: A Retrospective Internal Evaluation

Aiaal Anatolevich Eginov* 

Technical Department, TOIAGAI-2050 LLP, Almaty, Republic of Kazakhstan; eginovaa@gmail.com.

Citation:

Received: 23 May 2026
Revised: 10 August 2026
Accepted: 16 October 2026

Eginov, A. A. (2026). Global sharpness as an image-quality gate for gas-pipeline defect segmentation: A retrospective internal evaluation. *International journal of researches on civil engineering with artificial intelligence*, 3(4), 302-323.

Abstract

Photographs interpreted by machine vision on above-ground gas pipelines vary in sharpness, exposure and compression because of how they are acquired, and that variation degrades network performance. One safeguard is an image-quality gate: Admit photographs passing a criterion, route the rest to manual review. This study asks whether such a criterion can be calibrated to a stated target, what share it admits, and what is withheld. A gate combining a luminance-normalised sharpness statistic with an exposure envelope was calibrated on the validation subset of a family-partitioned corpus of 9,965 pipeline image records, using five instance-segmentation checkpoints as the instrument. The frozen, hashed rule was applied once to a 150-image evaluation subset under fourteen conditions. For deformation, the primary class, no acceptance zone was established: The estimate never reached the target, so the threshold was infinite and the primary endpoint, the change in residual failure among accepted photographs, is not estimable. Two secondary rules accepted 42.7% and 47.2% of image versions, with estimated residual-failure differences of -0.0226 and -0.0248 against the same stream without gating, and intervals including zero; a reduction was not demonstrated. Against nominal performance on unimpaired photographs, a separate comparison point, non-inferiority within the pre-specified margin was also not demonstrated. The quantile-based fallback rule would leave 59.3% of deformation object–seed observations outside automated interpretation, 20.2% of them already correctly matched: Implied workload, not measured cost. The study is a retrospective internal evaluation with a self-documented freeze; deviations found after the evaluation subset had been examined are disclosed.

Keywords: Image-quality gating, Visual inspection of pipelines, Instance segmentation, Acceptance threshold calibration, Coverage and error trade-off, Retrospective evaluation.

1 | Introduction

Visual inspection of above-ground gas distribution infrastructure is increasingly interpreted by machine-vision systems rather than by eye alone. Computer-vision methods for civil infrastructure condition assessment have

 Corresponding Author: eginovaa@gmail.com

 <https://doi.org/10.48314/ijrceai.vi.73>



Licensee System Analytics. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

been reviewed extensively, covering damage characterisation, component recognition and change detection from imagery collected by cameras and unmanned aerial vehicles [1]. For gas distribution networks specifically, machine-vision defect detection has been combined with geographic information systems to support geohazard assessment and maintenance planning in a seismically active urban setting [2], and open software has been published that performs real-time visual recognition of surface anomalies — corrosion, dents and deformation — on inspection photographs through a deployable interface [3]. Instance segmentation provides object masks as well as detections, so it supplies both a detection and an estimate of visible extent. Recent work continues to develop dedicated architectures for this detection task, for example a lightweight distance-content aware detection transformer for gas-pipeline defect detection [4]; that work, like [2] and [3] above, is directed at defect-detection performance on a photograph already provided to the model. Whether a photograph should be admitted to interpretation at all is a separate question, addressed here.

A survey crew photographs pipeline runs under uncontrolled illumination, at variable standoff, with handheld or airborne cameras, and the resulting images are typically compressed and resampled before analysis. The photographs that reach an automated interpreter therefore vary in sharpness, exposure and compression quality, and that variation is a property of the acquisition process rather than of the asset being inspected.

It is well established that such variation degrades network performance. Classification accuracy falls under common quality distortions, most sharply under blur and noise [5]; standardised corruption benchmarks show that the sensitivity persists across architectures [6]; and analogous benchmarks report it for object detection [7]. An operator who admits every photograph to an automated pipeline therefore admits an unknown and variable error rate with it.

One response is to gate the input: Measure observable properties of each photograph, admit those that pass a criterion, and route the remainder to manual review or reacquisition. Gating can draw on two established lines of work. Declining to decide has a long history in pattern recognition, where classification with a reject option trades coverage against error [8], and it has been developed for deep networks as selective prediction [9]; those formulations commonly use the model's own confidence as the deciding quantity. A recent example applies this same reject-option principle — calibrated probabilities and a cost-aware deferral threshold, using the model's own confidence, not the input's properties, as the deciding quantity — to clinical triage rather than industrial inspection [10]. The rule studied here is constrained differently by design: It uses only statistics of the photograph itself, computed without ground truth and without a reference frame, which places it in the setting of no-reference image-quality assessment, where natural-scene-statistics models estimate quality from the image alone [11], [12]. Simple scalar descriptors are also available, among them the variance of the Laplacian, used as a focus measure in autofocus work and as a general sharpness statistic [13].

For an operator the practical question is narrower than the general literature suggests. It is not whether image quality affects a network, which is established, but whether a rule expressible in a few observable image statistics can be calibrated to a stated performance target, what share of photographs such a rule admits, and what is lost on the photographs it does not admit. Answering it requires two things that are easy to omit: Thresholds fitted on data not used to judge them, and an explicit account of the cost of non-acceptance. It also requires two distinct comparisons, because a gate does not alter the model's prediction on any individual photograph and changes only which photographs are interpreted. The effect of gating is therefore measured against interpreting the same stream without a gate. Separately, the level the system attains on unimpaired photographs serves as the reference for asking whether accepted photographs are no worse than that level within a stated margin.

The earlier work cited above addresses adjacent tasks. Reference [2] sets out a monitoring strategy: Detecting defects and combining the detections with GIS hazard layers to prioritise maintenance. Reference [3] documents a software system that performs the visual recognition itself and delivers it to engineers in deployable form. The present study takes the detector as given and asks a different question: Whether a photograph should be admitted to automated interpretation at all, how such an admission rule can be

calibrated, and what non-acceptance costs. The performance figures reported in that work are not compared numerically with the outcome defined in Section 2.4. For reference [2] the reasons are documented: Its dataset, its class scheme and its performance definitions differ from those used here, and its headline figure is a composite over detection, spatial validation and field verification rather than a detector metric. For reference [3] the full text was not available for this study, so comparability was not established either way.

This study addresses the question for one segmentation system and one dataset of above-ground gas-pipeline photographs. A rule based on a luminance-normalised global sharpness statistic, together with an exposure envelope, was calibrated on a validation subset against a pre-specified target derived from the system's nominal performance; the rule and its thresholds were frozen and hashed; and the frozen rule was then applied once to an evaluation subset. Deformation is the primary class, with pipe reported as a secondary descriptive class and corrosion as counts only.

The specific objectives are:

- I. To calibrate an acceptance threshold on the primary class against a stated margin below nominal performance, and to report whether such a threshold exists within the range of the observed feature.
- II. To apply the frozen rule once to a held-aside subset and estimate the change in residual failure among accepted photographs, together with its uncertainty.
- III. To quantify the cost of non-acceptance — the photographs withheld from automated interpretation, the correct automated results withheld with them, and the share of model failures the policy captures.
- IV. To compare the calibrated rule with a raw-sharpness comparator and with an envelope calibrated only on the feature values of unimpaired photographs.
- V. To report the whole procedure, including the deviations found when the implementation was audited after the evaluation stage.

The study is a retrospective internal evaluation with a self-imposed, self-documented freeze. The dataset, the family partition, the five checkpoints and the degradation protocol were not created here: They come from two companion manuscripts, one on source-aware data partitioning and one on segmentation robustness under controlled image degradation, which are identified in Section 2. The evaluation subset had accordingly been used in that earlier work, and there was no external registration, no third-party timestamping and no independent replication. Section 2.8 states what was known before the protocol was frozen, and Section 2.9 states what was corrected afterwards. No claim of novelty is made, and the findings are not extended beyond the feature family, rule form, model and dataset examined here.

2 | Materials and Methods

2.1 | Study Design and Status

A decision rule for accepting inspection photographs into automated screening was calibrated on a validation subset, frozen together with its thresholds and hashed, and then applied once to an evaluation subset. The study is a retrospective internal evaluation. The freeze was self-imposed and self-documented: There was no registration with an external party, no third-party timestamping and no independent validation. The evaluation subset was not naive, and its prior use is set out in Section 2.8.

2.2 | Data and Analysis Units

The image corpus, the grouping of records into source-image families and the family-wise partition were established in a companion manuscript on source-aware data partitioning, referred to below as the partitioning study. The photographs are a field-collected survey of above-ground gas distribution infrastructure in the Almaty region, annotated on a commercial annotation platform and exported in COCO instance-segmentation format.

The corpus comprises 9,965 exported image records derived from 1,029 source images of above-ground gas-pipeline infrastructure, annotated in COCO instance-segmentation format with three classes: Corrosion, deformation and gas-pipelines. Source images were grouped into families and each family was assigned whole to one subset: 729 families for training, 150 for validation and 150 for evaluation, with no family shared between subsets.

Composition of the two subsets used here is given in *Table 1*. One image in each subset was captured at 3024×4032 and is rescaled for analysis; all others are 640×640 .

Table 1. Subsets and analysis units.

Quantity	Validation	Evaluation
Source families	150	150
Images	150	150
Image versions analysed (images $\times$ 14 conditions)	2,100	2,100
Annotated objects, total	310	292
— pipe	228	215
— deformation	80	72
— corrosion	2	5
Images containing deformation	70	62
Images containing corrosion	2	2
Deformation object–seed observations	5,600	5,040
Pipe object–seed observations	15,960	15,050
Corrosion object–seed observations	140	350
Images at native 3024×4032 (rescaled for analysis)	1	1
Nominal valid-match probability, deformation	0.2925	0.2806
Nominal valid-match probability, pipe	0.4553	0.5591

Note (*Table 1*): Source families were assigned whole to one subset; the three family sets are pairwise disjoint. Object–seed observations are annotated objects $\times$ 14 conditions $\times$ 5 checkpoints. The 14 conditions are listed in Supplementary *Table S1*.

Training used 729 further families, which appear in neither subset. Corrosion counts are reported for completeness only; five objects from two families support no inference.

Three units are distinguished throughout. An image version is one image under one condition; each subset yields 2,100 versions. An object–seed observation is one annotated object under one condition as evaluated by one checkpoint. A source family is the resampling unit for all intervals. Acceptance is a property of a version, so acceptance shares over versions and over object–seed observations differ and are both reported.

Validation ground truth was reconstructed in COCO format from the YOLO-format polygon labels, with vertices rounded to 0.001 px to match the export convention. Applying the identical reconstruction to the evaluation labels reproduced all 292 masks of the original evaluation annotation file exactly; this served as the correctness check for the reconstruction.

2.3 | Segmentation System and Degradation Conditions

The task addressed by the segmentation system — real-time visual recognition of corrosion, dents and deformation on gas-pipeline inspection photographs, with YOLO-family instance segmentation and a platform-based annotation pipeline — is the task of the published open software framework [3], and the wider monitoring context in which such a system is deployed is described in [2]. Neither publication establishes which trained weights are used in the present analysis.

The system used here is a specific set of five YOLO11n-seg checkpoints trained on the 729 training families under five random seeds. The training configuration, the software versions and the SHA-256 value of each

checkpoint are documented in the partitioning study and reproduced in Supplementary S5, and all five hashes were verified before any analysis. The checkpoints are therefore identified by hash rather than by name, and their correspondence to the weights distributed with the published software package [3] was not examined. Reference [3] is cited for the task and the deployment context; the five checkpoints are the measurement instrument of this study. No model was retrained, fine-tuned or re-selected here.

Fourteen conditions were analysed: The unmodified image and thirteen synthetic degradations. Nine were inherited unchanged from the earlier degradation study — Gaussian blur at $\sigma = 1.0$ and 2.5 , brightness multipliers 0.75 and 0.50 , JPEG quality 35 and 25 , and downsampling to $1/2$, $1/4$ and $1/6$ with restoration of size. Four intermediate levels drawn from the previously declared severity ladder were added to populate the region between nominal and clearly degraded images: Blur at $\sigma = 0.5$ and 1.5 , and downsampling to $3/4$ and $1/3$. Parameters, implementations and provenance of all fourteen are listed in Supplementary Table S1.

Degraded versions were produced by the frozen perturbation code and saved as JPEG at quality 100 with $4:4:4$ subsampling, so the model and the gate read the same file. Every version was verified by comparing the SHA-256 of its decoded pixel array against the value recorded on the inference host; all $2,700$ host-generated versions matched their local regeneration.

The synthetic mixture of fourteen conditions is a design artefact. Pooled quantities over that mixture describe the constructed stream and are not estimates of the prevalence of degraded photographs in field inspection.

2.4 | Outcomes

Predictions were obtained with the frozen inference settings and evaluated with the frozen matching procedure: Predictions with score ≥ 0.25 retained; class-consistent Hungarian assignment maximising total mask IoU; assignments below mask IoU 0.50 dissolved. An object counts as a valid match when it retains an assignment at that threshold. The assignment was solved a second time at IoU 0.35 to split failures into misses (no assignment, or assignment below 0.35) and localisation failures (assignment between 0.35 and 0.50). Retained but unmatched predictions count as false positives of their class.

Deformation is the primary class, pipe is secondary and descriptive, and corrosion is reported as counts only: Five objects from two families support no inference.

Reproducibility was established before any new analysis. Recomputing the outcomes of the ten inherited conditions from the archived prediction files reproduced all $14,600$ archived object-level results and all 50 false-positive counts exactly; recomputing the archived image-quality features from regenerated images agreed to within $1e-15$ relative difference; and a re-run of the validation baseline inference on the host reproduced the earlier prediction file bit-for-bit.

2.5 | Gate Features

Features are computed by the frozen code on the 640×640 analysis image, which is the image the model receives: Mean luminance L , variance of the Laplacian V , and the fractions of pixels at or below 5 and at or above 250 , whose sum is the clipping fraction C . The gate feature is the luminance-normalised sharpness $S = \log_{10}(V / L^2)$; the raw comparator is $\log_{10}(V)$.

$$S = \log_{10}\left(\frac{V}{L^2}\right). \quad (1)$$

Ground-truth masks, IoU values, condition labels, degradation parameters, prediction outputs and any comparison with an unperturbed reference are excluded as gate inputs. No entropy, contrast or blockiness feature enters any rule.

2.6 | Rules and Decision Procedure

Let p_{nom} be the nominal validation valid-match probability for the calibrated class, δ a policy margin, and G the sorted set of distinct feature values over the calibration versions, that is, all validation versions containing at least one object of that class.

G-cal (primary): A weighted isotonic, non-decreasing regression of the valid-match outcome on S is fitted by pool-adjacent-violators, with weights equal to the number of object-seed observations. The fitted function is evaluated as a right-continuous step function. The acceptance threshold is $t_A = \min\{g \in G : \hat{f}(g) - p_{\text{nom}} \geq -\delta\}$; if no grid point qualifies $t_A = +\infty$, and if every grid point qualifies $t_A = -\infty$.

$$t_A = \min\{g \in G : \hat{f}(g) - p_{\text{nom}} \geq -\delta\}. \quad (2)$$

The rejection threshold is $t_R = \min\{g \in G : U(g) \geq -\delta\}$, where $U(g)$ is the upper 95% bound of $\hat{f}(g) - p_{\text{nom}}$ from the calibration bootstrap; if $U(g) \geq -\delta$ at the first grid point $t_R = -\infty$, and if t_R exceeds t_A it is capped at t_A .

The decision rule is evaluated in a fixed order, with Reject taking priority:

- I. Reject if $S < t_R$.
- II. Otherwise Accept if $S \geq t_A$ and L lies within the validated exposure range $[L_{\min}, L_{\max}]$ and $C \leq C_{\max}$, where all three exposure bounds are taken over all 2,100 validation versions.
- III. Otherwise Caution.

The exposure test therefore gates acceptance independently of sharpness, and it remains operative even where the sharpness band between t_R and t_A is empty.

G-raw (comparator): The identical calibration applied to $\log_{10}(V)$, used as a two-state rule: Accept if the feature is at or above t_A , otherwise Reject. Its t_R is computed by the same procedure but is not used in the decision and is reported for completeness only.

G-cal| ϵ (fallback): Specified in advance for the case in which G-cal fails the coverage requirement. The acceptance threshold becomes $\min(t_A, q_\epsilon)$, where q_ϵ is the ϵ -quantile, linear interpolation, of S over the 150 nominal validation frames; t_R is then capped at the new t_A . The procedure begins from the outcome-calibrated G-cal, so the variant is not independent of model outcomes; whether the binding threshold ends up being the isotonic value or the nominal quantile depends on the calibration result. Its states are reported as Accept and Not accepted (quality envelope), with the reason for non-acceptance — feature below threshold, exposure outside the validated range, or both — preserved and reported; the partition of versions is identical to the three-state labelling.

G-env (secondary): A degradation-agnostic envelope calibrated on the 150 nominal validation frames alone: Accept if $S \geq q_{\epsilon/3}(S)$ and $L \geq q_{\epsilon/3}(L)$ and $C \leq q_{1-\epsilon/3}(C)$, all quantiles by linear interpolation; otherwise Caution. It has no Reject state, because no outcome evidence enters its calibration.

Policy parameters: $\delta = 0.026$ is the across-checkpoint standard deviation of the nominal validation deformation valid-match rate, rounded to three decimals; $\delta_{\text{pipe}} = 0.015$ follows the same rule. The false-rejection allowance is $\epsilon = 0.05$, with 0.025 and 0.10 as sensitivity values. Both were fixed before any validation outcome under degradation existed. They are policy choices, not empirically derived requirements, and the coverage requirement is a reporting criterion: thresholds were never adjusted to satisfy it.

Coverage requirement: A rule meets the requirement when the share of nominal validation frames that are not accepted is at most ϵ .

2.7 | Statistical Analysis

Two distinct bootstraps are used, both resampling the 150 source families of the relevant subset with replacement and preserving family multiplicity.

The calibration bootstrap (validation subset, 2,000 replicates, generator seed 20260911) recomputes p_{nom} and the isotonic fit in each replicate and yields $U(g)$, from which t_R is taken. Families carrying no object of the calibrated class contribute zero to both numerator and denominator. A replicate containing no nominal observation of the class is discarded and counted. The distribution of the replicate acceptance thresholds is summarised with $-\infty$ and $+\infty$ counted separately; percentiles of the full distribution are order statistics on the extended real line, taking $-\infty$ when the lower neighbour is $-\infty$ and $+\infty$ when the upper neighbour is $+\infty$, and a summary conditional on a finite threshold excludes both infinities.

The evaluation bootstrap (evaluation subset, 2,000 replicates, generator seed 20260912) recomputes every ratio metric with the thresholds held fixed. Each resampled family carries all of its versions, checkpoints, objects and false positives. The same resample indices are reused for every rule and metric, so contrasts between rules are paired. Intervals are 95% percentile intervals. A replicate in which a denominator is zero leaves the metric undefined; such replicates are counted and the interval is computed over the remainder. Where a rule accepts nothing, the metric has no denominator at all and is reported as not estimable rather than as a value.

A third generator seed, 20260913, was reserved for a matched random-acceptance reference; because the primary rule accepted no version, that reference is degenerate and is not reported.

The five checkpoints enter as a fixed blocking factor: per-checkpoint point estimates and the number of checkpoints agreeing in direction are reported, and no checkpoint variance component is estimated. There is one primary endpoint, so no multiplicity adjustment is applied, and no p-value is used as a decision criterion. The pre-declared criterion for claiming a reduction in residual failure is an upper interval bound below zero; the pre-declared criterion for claiming non-inferiority is a lower interval bound for the margin at or above $-\delta$.

2.8 | Prior Knowledge, Reuse and Pre-Specification

Two companion manuscripts supply the material reused here: The partitioning study, which established the family partition, the training sets and the five checkpoints, and a robustness study of instance segmentation under controlled image degradation, which established the matching protocol, the confidence threshold, nine of the fourteen conditions and the archived object-level outcomes for the evaluation subset.

Earlier published work sets out the monitoring context and the software framework for this class of system [2], [3]. The extent to which the image sets, the trained models and the performance definitions of that work coincide with those used here has not been established: For [2] the dataset, class scheme and reported metrics are described in its Sections II.C, II.D and III.B and differ from the corpus and outcome definition used here, while for [3] the internal evaluation design was not examined for this study. Direct numerical comparison with the valid-match criterion of Section 2.4 is therefore not undertaken in either case. The validation subset had previously informed checkpoint selection, the choice of the confidence threshold and the selection of degradation severities. The evaluation subset had previously been used for a clean-versus-leaky comparison, for the full object-level degradation analysis of the ten inherited conditions, and for a box-versus-mask agreement analysis; its archived outcomes and image-quality features were additionally reproduced by the analyst during the verification stage that preceded the freeze. All of this was recorded in the frozen protocol before calibration began, together with the known condition-level valid-match probabilities and the known cross-sectional associations between sharpness and outcome from the earlier study.

What did not exist at the freeze were validation outcomes under degradation, any outcome at the four intermediate levels, any image-level association between a gate feature and model outcomes, and any gate threshold or decision.

The protocol, the threshold records and each subsequent correction are published with their SHA-256 values. Hashes fix content, not time of creation; the ordering of events rests on self-recorded timestamps in a single writable environment and is not independently attested.

2.9 | Protocol Deviations and Late Corrections

Four implementation deviations from the frozen protocol were identified during a sequential audit of the code and outputs carried out with AI assistance. The audit followed the evaluation stage, so every correction was computed after the evaluation subset had been examined. Full documentation, comparison tables and code are in Supplementary S5.

A weighting error in the isotonic implementation was found and fixed before any threshold was frozen, and the corrected implementation was verified against an independent isotonic solver.

The calibration bootstrap had resampled only families carrying an object of the calibrated class rather than all 150 validation families. The corrected recomputation covers twenty configurations: The ten that the original calibration had produced — the four δ values $\times$ two features for deformation, plus the pipe-calibrated pair — and ten leave-outlier-out configurations that had no earlier counterpart. Three comparisons must be kept apart.

First, for the ten original configurations, the corrected calibration left every acceptance threshold t_A unchanged, changed two rejection thresholds — the pipe-calibrated t_R from -2.417648 to -2.416538 and the diagnostic G-raw t_R at $\delta = 0.026$ from 1.527464 to 1.524649 — and changed no evaluation decision; one validation decision changed, in the pipe-calibrated variant. That correction changed the bootstrap distribution, because families carrying no object of the calibrated class were never drawn; the direction and size of its effect on interval width were not quantified.

Second, the validation half of the leave-outlier-out sensitivity analysis had been omitted and was run for the first time as part of the correction, so its ten configurations have no pre-correction values to compare against. It did not produce an acceptance zone for G-cal.

Third, a later audit found that the branch of the acceptance rule returning $-\infty$ when every grid point qualifies had not been implemented. Here all twenty configurations could be compared, and every point threshold was unaffected: The branch was never triggered by a point estimate. It affected 302 of 40,000 bootstrap replicates and changed only summaries of the threshold distribution. Separately, percentiles of the replicate threshold distribution had been computed among finite replicates only; both summaries are now reported, with $-\infty$ and $+\infty$ counted separately.

Fig. 1 is drawn from calibration curves recomputed under the corrected bootstrap. Threshold-distribution summaries come from the final correction. No evaluation metric table required reissue, because no evaluation decision changed.

2.10 | Use of Generative Artificial Intelligence

Anthropic Claude and OpenAI ChatGPT were used in preparing the manuscript and analysis code and in checking the results. The author is responsible for the final content.

3 | Results

3.1 | Calibration on the Validation Subset

The nominal anchor recomputed from the validation baseline was $p_{\text{nom}} = 0.2925$ for deformation, 117 of 400 object-seed observations, matching the value pre-specified in the protocol, so the pre-declared stop rule did not trigger. With $\delta = 0.026$ the target level was 0.2665.

The isotonic estimate of deformation valid-match probability as a function of S ranged from 0.2057 to 0.2339 over the 980 grid points and did not reach the target at any of them. No acceptance zone was established for G-cal: $t_A = +\infty$. The rejection threshold was $t_R = -0.9502$, which placed 77.3% of nominal validation frames in Reject and the remaining 22.7% in Caution, so no nominal frame was accepted and the coverage requirement was not met.

For G-raw the isotonic estimate reached the target only in the extreme upper tail, giving $t_A = 3.150258$; that threshold accepted 60 of 2,100 validation versions and left 135 of 150 nominal frames unaccepted. At $\delta = 0.052$ the same procedure gave $t_A = 1.524649$, accepting 1,219 versions and leaving no nominal frame unaccepted.

Because G-cal failed the coverage requirement, the pre-specified fallback G-cal| ϵ was calibrated. Here the isotonic threshold was infinite, so the minimum operation left the acceptance threshold equal to the nominal quantile: $t_A = -1.9404$ at $\epsilon = 0.05$. What is specific to this configuration is that the binding threshold is a quantile of the nominal S distribution; the procedure that produced the variant nevertheless begins from the outcome-calibrated G-cal, so the variant is not a rule independent of model outcomes. G-env at $\epsilon = 0.05$ used $S \geq -2.1600$, mean luminance ≥ 33.409 and clipping ≤ 0.5782 .

Both G-cal| ϵ and G-env left 8 of 150 nominal validation frames unaccepted, a rate of 0.0533. That rate exceeds the policy allowance $\epsilon = 0.05$, so the coverage requirement was not met by these rules either. The thresholds were not adjusted in response.

Table 2 reports all thresholds at the primary policy setting, with the full δ and ϵ grids and the bootstrap threshold distributions in Supplementary Table S2; Fig. 1 shows the isotonic fits with the target level and the upper bootstrap bound.

Table 2. Thresholds calibrated on validation data at the primary policy setting ($\delta = 0.026$, $\epsilon = 0.05$), and the resulting coverage.

Rule	Acceptance Threshold t_A	Rejection Threshold t_R	Accepted, of 2,100 Validation Versions	Nominal Validation Frames Not Accepted, of 150
G-cal (primary)	$+\infty$ — no acceptance zone established	-0.950151	0 (0.000)	150 (1.0000)
G-raw ($\delta = 0.026$)	3.150258	1.527464 (diagnostic; unused by the two-state rule)	60 (0.029)	135 (0.9000)
G-raw ($\delta = 0.052$)	1.524649	1.006593 (diagnostic; unused)	1,219 (0.580)	0 (0.0000)
G-cal ϵ ($\epsilon = 0.05$)	-1.940450, set by the ϵ -quantile of S on nominal frames	-1.940450 (capped at t_A)	911 (0.434)	8 (0.0533)
G-env ($\epsilon = 0.05$)	$S \geq -2.159977$, $L \geq 33.409249$, clipping ≤ 0.578205	none	969 (0.461)	8 (0.0533)

Note (Table 2): Thresholds act on $S = \log_{10}(V / L^2)$, except for G-raw, which uses $\log_{10}(V)$. Counts are recomputed from the validation features with the full-precision thresholds under the frozen decision rules and agree with the shares stored in the threshold record. Full δ and ϵ grids, the pipe-calibrated variant and the bootstrap threshold distributions are in Supplementary Table S2.

Exposure envelope used by the Accept branch of G-cal and G-cal| ϵ , taken over all 2,100 validation versions: Mean luminance in [8.426995, 187.467029], clipping fraction ≤ 0.710959 . G-cal| ϵ and G-env each leave 8 of 150 nominal validation frames unaccepted, a rate of 0.0533, which exceeds the policy allowance $\epsilon = 0.05$; the thresholds were not adjusted. The G-raw rejection thresholds are listed for completeness: The one at $\delta = 0.026$ moved from 1.527464 (E4) to 1.524649 under the corrected calibration bootstrap and is not used by that two-state rule.

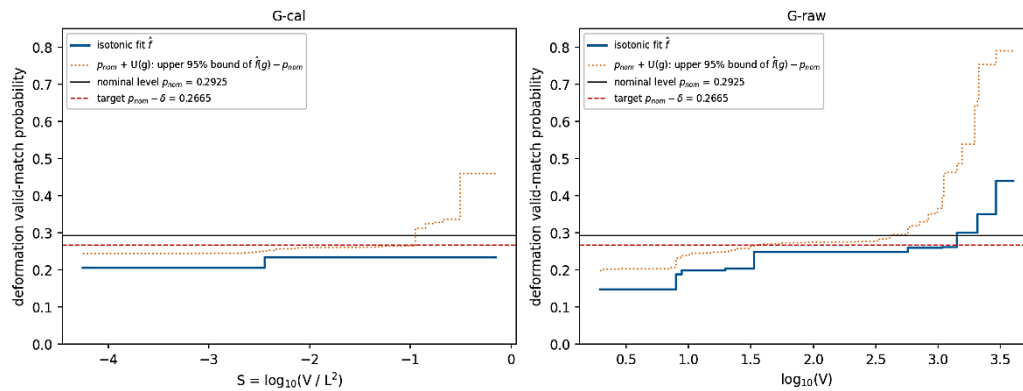


Fig. 1. Validation calibration for the primary feature $S = \log_{10}(V / L^2)$ (left) and the raw comparator $\log_{10}(V)$ (right), deformation class.

Note (Fig. 1): The solid step function is the weighted isotonic fit $\hat{f}(g)$. The dotted curve is $p_{\text{nom}} + U(g)$, where $U(g)$ is the upper 95% cluster-bootstrap bound of the difference $\hat{f}(g) - p_{\text{nom}}$ with p_{nom} recomputed in every replicate; it is a bound on that difference re-centred for display and is not a confidence bound on $\hat{f}(g)$ itself. The solid horizontal line is the nominal level $p_{\text{nom}} = 0.2925$ and the dashed line is the target $p_{\text{nom}} - \delta = 0.2665$. For S the fit does not reach the target at any grid point, so no acceptance threshold exists. Curves are recomputed under the corrected calibration bootstrap over all 150 validation families.

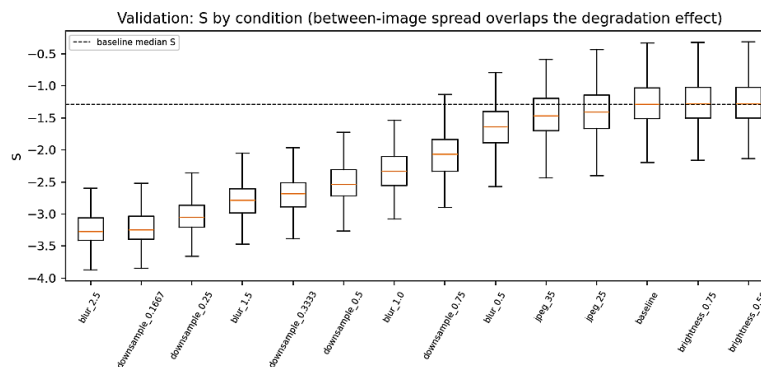


Fig. 2. Distribution of S by condition on the validation subset, showing the spread between source images alongside the shift each degradation produces within an image.

3.2 | Decisions on the Evaluation Subset

The frozen thresholds were applied once to the 2,100 evaluation versions. G-cal accepted no version. G-cal| ϵ at $\epsilon = 0.05$ accepted 897 versions (42.7%), covering 2,050 of 5,040 deformation object–seed observations (40.7%). G-env accepted 991 versions (47.2%) and 2,275 observations (45.1%). G-raw accepted 51 versions (2.4%) at $\delta = 0.026$ and 1,249 versions (59.5%) at $\delta = 0.052$.

On the 150 nominal evaluation frames, 9 were not accepted by G-cal| ϵ and 9 by G-env, a rate of 0.060. The coverage requirement, already unmet on validation at 0.0533, was therefore also unmet on evaluation.

No evaluation version fell outside the validated exposure envelope. The exposure test remains part of the decision rule and was evaluated for every version; it simply removed none here.

3.3 | Primary Endpoint: Deformation

Because G-cal accepted no version, the primary endpoint is not estimable. This is the study's main result. The secondary rules reported below do not substitute for it.

With no gate, residual failure for deformation over the fourteen-condition stream was 0.7821 (95% interval 0.7077 to 0.8482; 5,040 observations), composed of 0.6613 misses and 0.1208 localisation failures. On the nominal condition alone it was 0.7194 (0.6262 to 0.8055; 360 observations).

For every secondary rule the interval for Δ against no gate included zero: -0.0226 (-0.0776 to 0.0359) for G-cal| ϵ , -0.0248 (-0.0763 to 0.0318) for G-env, -0.0260 (-0.0602 to 0.0149) for G-raw at $\delta = 0.052$, and -0.0670 (-0.3315 to 0.1683) for G-raw at $\delta = 0.026$. A reduction in residual deformation failure among accepted versions was therefore not demonstrated for any rule under the pre-declared criterion. Point estimates were negative for all five checkpoints for G-cal| ϵ and G-env, ranging from -0.0418 to -0.0082 and from -0.0451 to -0.0077 .

The G-raw estimate at $\delta = 0.026$ rests on 51 accepted versions and 165 accepted observations. It is an imprecise estimate, as its interval width shows, and it is not treated as the best-performing rule.

Non-inferiority to the nominal level within δ was likewise not demonstrated. Accepted versions under G-cal| ϵ gave a deformation valid-match probability of 0.2405 against the nominal evaluation level of 0.2806, a margin of -0.0401 (-0.0886 to 0.0057); the lower bound lies below $-\delta$. For G-env the margin was -0.0379 (-0.0862 to 0.0049). These intervals include values inside the margin and values well below it, so the data neither establish nor exclude performance within δ .

False positives per accepted version-checkpoint pair were 0.2337 with no gate, 0.2261 under G-cal| ϵ , 0.2367 under G-env and 0.2479 under G-raw at $\delta = 0.052$. These point estimates are close to one another; no equivalence test was performed and none is implied.

Table 3 gives the full set of deformation results.

Table 3. Evaluation subset, deformation failure among accepted versions and its components, the primary contrast against no gate, and the margin to the nominal evaluation level.

Rule	Versions Accepted (of 2,100)	Object-Seed Observations Accepted (of 5,040)	Residual Failure	95% Interval	Misses	Localisation Failures	Image-Level Miss	False Positives per Accepted Version-Seed	Δ vs no Gate	95% Interval	Margin to Nominal	95% Interval
No gate	2,100 (1.000)	5,040 (1.000)	0.7821	0.7077 to 0.8482	0.6613	0.1208	0.7618	0.2337	—	—	−0.0627 to −0.1032	—
G-cal (primary, $\delta = 0.026$)	0 (0.000)	0 (0.000)	not estimable	—	not estimable	not estimable	not estimable	not estimable	not estimable	—	not estimable	—
G-cal ϵ ($\epsilon = 0.05$)	897 (0.427)	2,050 (0.407)	0.7595	0.6679 to 0.8389	0.6029	0.1566	0.7378	0.2261	−0.0226	−0.0776 to 0.0359	−0.0401 to −0.0886	−0.0886 to 0.0057
G-env ($\epsilon = 0.05$)	991 (0.472)	2,275 (0.451)	0.7574	0.6714 to 0.8347	0.6018	0.1556	0.7383	0.2367	−0.0248	−0.0763 to 0.0318	−0.0379 to −0.0862	−0.0862 to 0.0049
G-raw ($\delta = 0.026$)	51 (0.024)	165 (0.033)	0.7152	0.4443 to 0.9539	0.5394	0.1758	0.6867	0.2863	−0.0670	−0.3315 to 0.1683	0.0043 to 0.2645	−0.2273 to 0.2645
G-raw ($\delta = 0.052$)	1,249 (0.595)	2,965 (0.588)	0.7562	0.6740 to 0.8290	0.6047	0.1514	0.7365	0.2479	−0.0260	−0.0602 to 0.0149	−0.0367 to −0.0742	−0.0742 to 0.0004

Note (Table 3): Residual failure = misses (no assignment, or assignment below mask IoU 0.35) + localisation failures (assignment at IoU between 0.35 and 0.50). Image-level miss is the share of accepted version–checkpoint pairs containing at least one deformation object in which at least one object lacks a valid match. False positives are per accepted version–checkpoint pair. Intervals are 95% cluster-bootstrap percentile intervals over 150 source families, 2,000 replicates, paired across rules.

G-cal accepted no version, so every accepted-set quantity and the primary endpoint Δ are not estimable for it. The margin column is the accepted-set valid-match probability minus the nominal evaluation level of 0.2806, and the pre-specified margin is $\delta = 0.026$. In the No-gate row this margin is not an effect of filtering: It is the difference between the valid-match probability over the constructed fourteen-condition stream and the probability on the nominal condition alone, and it therefore measures the composition of that stream. No interval for Δ excludes zero, and no lower interval bound for the margin lies at or above -0.026 . G-raw at $\delta = 0.026$ is a pre-specified comparison retained here; with 51 accepted versions and 165 accepted observations it is an imprecise estimate, as its interval width shows, and it is not put forward as the best-performing rule.

3.4 | Cost of Non-Acceptance

Non-accepted versions are not correct rejections. Under $G\text{-cal}|\epsilon$ at $\epsilon = 0.05$, 2,990 of 5,040 deformation object–seed observations (59.3%) fell on versions that the policy would route to manual review or reacquisition rather than to automated screening. No manual review was observed or evaluated in this study, and no assumption is made about its effectiveness. Among the deferred observations, 20.2% had in fact received a valid match, so the policy would withhold correct automated results at that rate, and the deferred set contained 60.5% of all model failures in the stream. G-env deferred 54.9% of observations, withheld 19.7% correct results among them and captured 56.3% of failures.

Table 4 reports these quantities for all rules; Fig. 3 shows residual failure against coverage, with coverage measured as the share of deformation object–seed observations retained.

Table 4. Cost of non-acceptance on the evaluation subset, deformation class.

Rule	Deformation Object–Seed Observations not Accepted (of 5,040)	Share of Those That Had Received a Valid Match	Share of All Model Failures Falling on Non-Accepted Versions	Nominal Evaluation Frames not Accepted
G-cal (primary, $\delta = 0.026$)	5,040 (1.000)	0.218	1.000	150 of 150 (1.000)
$G\text{-cal} \epsilon$ ($\epsilon = 0.05$)	2,990 (0.593)	0.202	0.605	9 of 150 (0.060)
G-env ($\epsilon = 0.05$)	2,765 (0.549)	0.197	0.563	9 of 150 (0.060)
G-raw ($\delta = 0.026$)	4,875 (0.967)	0.216	0.970	129 of 150 (0.860)
G-raw ($\delta = 0.052$)	2,075 (0.412)	0.181	0.431	1 of 150 (0.007)

Note (Table 4): Non-accepted versions are not correct rejections: under the policy they would be routed to manual review or reacquisition. No manual review was observed or evaluated in this study, and no assumption is made about its effectiveness.

The column "Share of those that had received a valid match" quantifies correct automated results that the policy would withhold. The policy allowance is $\epsilon = 0.05$; both $G\text{-cal}|\epsilon$ and G-env exceed it on the evaluation subset, as they already did on validation (Table 2).

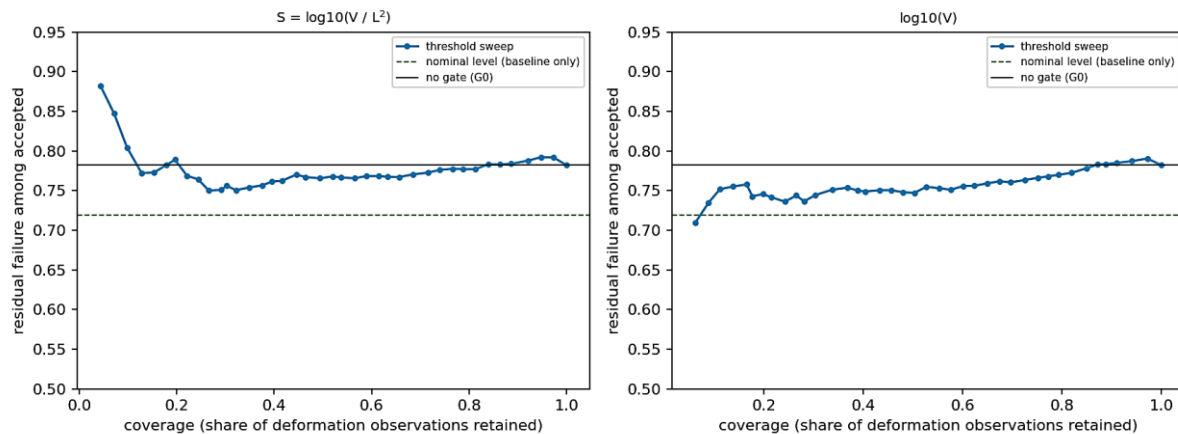


Fig. 3. Evaluation subset, deformation class: Residual failure among retained observations against coverage, for a sweep of absolute thresholds on S and on $\log_{10}(V)$.

Nore (Fig. 3): Coverage is the share of deformation object–seed observations retained, not the share of images. Horizontal lines mark the no-gate value over the fourteen-condition stream and the nominal-condition value.

3.5 | Behaviour across Conditions

Acceptance under $G\text{-cal}|\epsilon$ at $\epsilon = 0.05$ tracked the degradations that shift global sharpness. Of 150 versions per condition it accepted 141 nominal versions, 141 at brightness $\times 0.50$, 137 at JPEG quality 25, 131 at JPEG 35, 109 at blur $\sigma = 0.5$, 50 at downsampling to $3/4$, 18 at blur $\sigma = 1.0$, 15 at downsampling to $1/2$, and one version each at blur $\sigma = 2.5$ and downsampling to $1/6$.

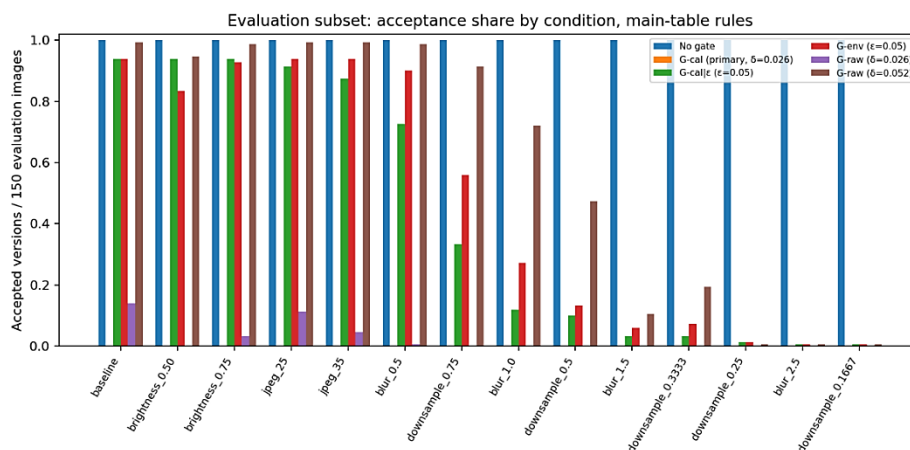


Fig. 4. Acceptance shares by condition for the evaluated rules, evaluation subset.

Among accepted versions, residual deformation failure was 0.755 at baseline (330 observations), 0.729 at blur $\sigma = 0.5$ (255), 0.751 at JPEG 35 (305), 0.778 at JPEG 25 (320) and 0.767 at downsampling to $3/4$ (120), against the nominal no-gate value of 0.7194. This is a descriptive comparison of point estimates across conditions, not a demonstration that performance on accepted versions is no worse than nominal; the formal margin is reported in Section 3.3. In the strongly degraded conditions the accepted denominators fall to 25 observations or fewer, so those cells are counts rather than estimates. Supplementary Table S3 gives per-condition results for every rule and metric with denominators and intervals.

3.6 | Pipe Class: Secondary Descriptive Results

Pipe results are secondary and descriptive. They are reported because the calibration and decision machinery is shared, and they do not make any rule a validated gate.

With no gate, residual failure for pipe was 0.5015 (0.4348 to 0.5660; 15,050 observations). Under $G\text{-cal}|\epsilon$ it was 0.4539 (0.3797 to 0.5278; 6,455 observations), a difference of -0.0476 (-0.0784 to -0.0179); under $G\text{-env}$, -0.0485 (-0.0760 to -0.0234). Both point estimates were negative for all five checkpoints. Restricting the stream to the ten inherited conditions gave -0.0570 and -0.0563 , and excluding the large evaluation image changed the values in the fourth decimal.

Non-inferiority within $\delta_{\text{pipe}} = 0.015$ was not demonstrated: accepted versions under $G\text{-cal}|\epsilon$ gave 0.5461 against the nominal evaluation level of 0.5591, a margin of -0.0130 (-0.0393 to 0.0141).

The pipe-calibrated variant of $G\text{-cal}$ specified as sensitivity S5 yielded a finite acceptance threshold, $t_A = -1.3388$, but left 43.3% of nominal validation frames unaccepted. Its raw-sharpness counterpart was not specified in the frozen protocol and is reported as exploratory only. *Table 5* and Supplementary *Table S4* give the details.

Table 5. Evaluation subset, pipe class: Secondary descriptive estimates.

Rule	Object–Seed Observations Accepted (of 15,050)	Residual Failure	95% Interval	Δ vs no Gate	95% Interval	Margin to Nominal	95% Interval
No gate	15,050	0.5015	0.4348 to 0.5660	—	—	-0.0605	-0.0806 to -0.0399
$G\text{-cal}$ (primary, $\delta = 0.026$)	0	not estimable	—	not estimable	—	not estimable	—
$G\text{-cal} \epsilon$ ($\epsilon = 0.05$)	6,455	0.4539	0.3797 to 0.5278	-0.0476	-0.0784 to -0.0179	-0.0130	-0.0393 to 0.0141
$G\text{-env}$ ($\epsilon = 0.05$)	7,120	0.4529	0.3789 to 0.5244	-0.0485	-0.0760 to -0.0234	-0.0120	-0.0347 to 0.0155
$G\text{-raw}$ ($\delta = 0.026$)	460	0.3848	0.2571 to 0.4909	-0.1167	-0.2442 to -0.0106	0.0561	-0.0498 to 0.1863
$G\text{-raw}$ ($\delta = 0.052$)	8,950	0.4517	0.3816 to 0.5209	-0.0497	-0.0667 to -0.0343	-0.0108	-0.0288 to 0.0080

Note (*Table 5*): The pipe class shares the calibration and decision machinery but is not the class on which the primary rule was calibrated; these estimates do not make any rule a validated gate. The pre-specified margin for this class is $\delta_{\text{pipe}} = 0.015$ and the nominal evaluation level is 0.5591.

As in *Table 3*, the margin in the No-gate row is not an effect of filtering: It is the difference between the fourteen-condition stream and the nominal condition. For $G\text{-cal}|\epsilon$ and $G\text{-env}$ the interval for Δ excludes zero and the point estimate is negative for all five checkpoints; no lower interval bound for the margin reaches -0.015 , so non-inferiority within the margin is not demonstrated for this class either. The pipe-calibrated variant of $G\text{-cal}$ (sensitivity S5) and its raw-sharpness counterpart are in Supplementary *Table S4*; the latter was not pre-specified and is exploratory.

3.7 | Corrosion

The evaluation subset contains five corrosion objects from two source families, giving 350 object–seed observations, of which 41 received a valid match with no gate. Under $G\text{-cal}|\epsilon$ at $\epsilon = 0.05$, 145 observations fell on accepted versions and 26 of these received a valid match. No threshold was calibrated for this class and no inference is drawn.

3.8 | Sensitivity

The absence of an acceptance zone for the primary deformation-calibrated G-cal held across the examined δ grid of 0.013, 0.026, 0.039 and 0.052, and when the single large validation image was excluded from calibration. This statement is not extended to the secondary rules or to parameter values outside the examined grids. Results for the secondary rules across the ϵ grid, across the reduced ten-condition stream and with the large evaluation image excluded are in Supplementary *Table S4*.

4 | Discussion

4.1 | What the Study Found

Under a calibration procedure fixed in advance, the primary rule did not acquire an acceptance zone. The isotonic estimate of deformation valid-match probability as a function of the luminance-normalised sharpness S rose from 0.2057 to 0.2339 across the calibration grid and did not reach the target level of 0.2665 at any point, so the acceptance threshold was $+\infty$ and no validation version, and consequently no evaluation version, could be accepted. The primary endpoint is therefore not estimable, and that is the study's principal outcome.

The two secondary rules that accepted 42.7% and 47.2% of evaluation versions produced small negative point estimates for the change in residual deformation failure among accepted versions, -0.0226 for G-cal| ϵ and -0.0248 for G-env, with intervals from -0.0776 to 0.0359 and from -0.0763 to 0.0318 . Neither interval excludes zero, so under the pre-declared criterion a reduction was not demonstrated. Non-inferiority to the nominal level within the pre-specified margin was not demonstrated either: the margin for G-cal| ϵ was -0.0401 with an interval from -0.0886 to 0.0057 , which contains values inside the margin and values well below it. The data are compatible with a modest improvement, with no change, and with a meaningful shortfall relative to nominal operation.

These results are compatible with two different readings, and the present design cannot separate them. One is that a global sharpness statistic carries too little information about deformation detection at the level of an individual photograph. The other is that the available sample — 72 deformation objects in 62 evaluation images, and 80 objects in 70 validation images — is too small to resolve an effect of the size that a useful gate would need to produce. Nothing in these data selects between them, and the study does not claim that image-quality gating in general cannot work.

4.2 | The Feature Behaves as Intended; the Decision does not Follow

The gate feature responded to degradation in the expected direction. On validation, the median S fell by 1.06 log units under Gaussian blur at $\sigma = 1.0$ and by 0.79 under downsampling to $3/4$, while brightness scaling left it essentially unchanged, which is the behaviour the luminance normalisation was introduced to produce. The condition-level acceptance pattern of G-cal| ϵ reflects this: of 150 versions per condition it accepted 141 nominal versions, 141 at brightness $\times 0.50$, 137 at JPEG quality 25, 109 at blur $\sigma = 0.5$, 18 at blur $\sigma = 1.0$, and one version each at blur $\sigma = 2.5$ and downsampling to $1/6$.

Detecting degradation and improving decisions are, however, different things. Between-image variation in S is substantial on nominal frames alone, where the standard deviation is 0.433 and the observed range spans 2.18 log units, which is of the same order as the shift that mild blur produces within an image. An absolute threshold accordingly separates conditions only partially, and the threshold sweep in *Fig. 3* shows residual failure varying little across a wide band of coverage. This is a description of the observed data, not an established causal account: the study did not test a mechanism, and the relative contributions of feature informativeness, sample size and annotation quality remain unresolved.

The direction of the earlier cross-sectional association is consistent with this picture. In the preceding degradation study the odds ratio per standard deviation of log Laplacian variance was 0.50 for deformation,

with an interval from 0.14 to 1.04 that includes unity, and 1.71 for the pipe class, with an interval from 1.19 to 2.55 that does not. The present result adds a calibrated, decision-level counterpart to that observation.

4.3 | The Cost Side of a Gate is Measurable and Substantial

A gate that is not shown to help still removes images from automated processing. Under $G\text{-cal}|\epsilon$ at $\epsilon = 0.05$, 2,990 of 5,040 deformation object–seed observations, 59.3%, fell on versions that the policy would route to manual review or reacquisition; 20.2% of those observations had in fact received a valid match from the model, so that share of correct automated results would be withheld. The deferred set contained 60.5% of all model failures in the stream. $G\text{-env}$ behaves similarly, deferring 54.9% of observations and capturing 56.3% of failures.

This trade is the practical content of the result. Reporting only an improvement among retained images would omit it, and the omission matters: the deferred images are not correct rejections and do not resolve themselves. Whether routing three-fifths of observations to human review is worthwhile depends on inspection capacity and on how effective that review is, neither of which this study observed or evaluated.

Both rules also exceeded the policy allowance for rejecting nominal frames. Eight of 150 nominal validation frames were not accepted, a rate of 0.0533 against $\epsilon = 0.05$, and nine of 150 on evaluation, a rate of 0.060. The thresholds were not adjusted in response, because adjusting them after the freeze would have converted a pre-specified criterion into a fitted one.

4.4 | The Pipe Class

Pipe results are secondary and descriptive. They are reported because the same calibration and decision machinery applies to them, not because the rules were calibrated for that class. Under $G\text{-cal}|\epsilon$ the change in residual failure was -0.0476 with an interval from -0.0784 to -0.0179 , and under $G\text{-env}$ -0.0485 from -0.0760 to -0.0234 ; both point estimates were negative for all five checkpoints, and restricting the stream to the ten inherited conditions gave -0.0570 and -0.0563 .

Interpretation must stay narrow, and the two rules reached their thresholds by different routes. $G\text{-cal}|\epsilon$ follows the outcome-calibrated procedure of the primary rule, which is calibrated on deformation outcomes; in the configuration obtained here its binding threshold ended up being the ϵ -quantile of S on nominal validation frames. $G\text{-env}$ is calibrated on the feature values of nominal validation frames alone and uses no outcome data of any class. Neither rule was calibrated on pipe outcomes. Non-inferiority within the pipe margin of 0.015 was not demonstrated, the margin under $G\text{-cal}|\epsilon$ being -0.0130 with an interval from -0.0393 to 0.0141 . Pipe masks are large and occupy much of the frame, which makes a sharpness-related difference between the classes plausible, but this study did not test that explanation and does not assert it. The pipe-calibrated variant specified as sensitivity $S5$ did produce a finite acceptance threshold at -1.3388 , yet left 43.3% of nominal validation frames unaccepted, which is far outside any policy allowance considered here.

4.5 | Interpreting a Self-Documented Freeze

The calibration was frozen and hashed before the evaluation stage of this study: before the four new intermediate evaluation conditions were processed, before the gate features were derived for any evaluation version, and before any gate decision or metric was computed on the evaluation subset. It was not frozen before the analyst had any contact with evaluation data: Archived object-level outcomes and image-quality features for ten of the fourteen conditions already existed from earlier work and had been reproduced during the verification stage that preceded the protocol.

After the evaluation stage the acceptance thresholds remained unchanged in every configuration that had a pre-correction counterpart; some rejection thresholds were corrected, and those corrections changed no evaluation decision. According to the records presented here, no parameter was chosen in response to the evaluation results, and the published thresholds reproduce every evaluation decision exactly, which a reader

can recompute. What the procedure does not provide is independent confirmation of that chronology: the freeze was self-imposed, timestamps are self-recorded, and the arrangement does not remove the influence of what was already known about the evaluation subset when the protocol was written.

It does not, however, make the study a confirmatory one. The evaluation subset had been used in the companion studies described in Section 2.8, the freeze was self-imposed with no external registration and no third-party timestamping, and the analyst had already seen archived evaluation outcomes for ten of the fourteen conditions before the protocol was written. Hashes fix content, not the time of creation. The correct description is a retrospective internal evaluation with a documented, self-imposed freeze.

Four implementation deviations from the frozen procedure were found in a sequential audit conducted with AI assistance after the evaluation stage, so all corrections were computed with the evaluation results already known. Their effect was bounded and is fully reported: Acceptance thresholds were unchanged in all configurations that had pre-correction values, no evaluation decision changed, and no metric table was reissued. Reporting them is part of the result rather than an appendix to it, because a freeze is only as informative as the audit that tests whether it was implemented as written.

4.6 | Engineering Reading

For an operator, the practical content is narrow and negative, and should be stated as such. On this dataset, with this segmentation system, a threshold on a single global sharpness statistic was not shown to raise the quality of automated deformation findings on the images it admits, while it would remove a large share of images from automated processing. An acceptance rule of this form was not established here; that is a statement about this rule, this feature and this dataset, not a general conclusion about input-quality control.

The nominal operating level provides the necessary context. With no gate, residual deformation failure on unmodified images was 0.7194, so roughly seven in ten annotated deformations were missed or poorly localised even before any degradation was applied. A gate does not alter the model's prediction on any individual image; it can in principle lower the error rate among the images it admits, by selecting which images are processed, but for deformation no such reduction was demonstrated here. No acceptance decision in this study should be read as a statement about the condition of a pipeline.

5 | Limitations

Sample size for the primary class: The evaluation subset contains 72 deformation objects in 62 images, and calibration rests on 80 objects in 70 images. Intervals are correspondingly wide, and the study is not able to resolve small effects. A negative finding at this size does not establish that no rule of this family could work.

Corrosion: Five objects from two source families support no inference whatever. Counts are reported for completeness only, no threshold was calibrated for the class, and no statement about corrosion gating follows from this work.

Synthetic degradations: The fourteen conditions are constructed, and the equal weighting of conditions in the pooled stream is a design choice. Pooled quantities describe that constructed stream and do not estimate the prevalence or the mixture of image defects in field inspection. Gaussian blur stands in for loss of sharpness in general; physically realistic optical defocus was not modelled as such, and motion blur, rain, haze, oblique viewing angles, occlusion and sensor noise were not modelled separately.

A single feature family and a simple rule form: Only global sharpness and exposure statistics entered the rules, evaluated as absolute thresholds. Simple joint thresholds were used: G-env combines conditions on S, mean luminance and the clipping fraction, and the acceptance branch of G-cal and G-cal| ϵ likewise combines a sharpness threshold with the exposure envelope. What was not examined is learned multivariate quality models, feature interactions beyond such conjunctions, local sharpness maps, region-restricted statistics and per-image reference comparisons. The findings apply to the rule form tested.

Resolution and capture geometry: Nearly all images were already at 640×640 when analysed. Nothing here supports requirements on sensor resolution, standoff distance or ground sampling distance.

One model family and one matching criterion: Results were obtained with five YOLO11n-seg checkpoints under a fixed confidence threshold of 0.25 and a valid-match criterion of mask IoU ≥ 0.50 . Other architectures, thresholds or matching criteria could shift both the nominal level and the behaviour of a gate.

Annotation quality: Masks are human-drawn and their accuracy bounds the outcome definition. Localisation failures at IoU between 0.35 and 0.50 accounted for 0.1208 of the no-gate outcome, and how much of that reflects annotation variation rather than model behaviour was not determined.

Not a naive evaluation subset: The evaluation images had been used in the companion studies described in Section 2.8, including the robustness study that supplied ten of the fourteen conditions. The freeze limits threshold tuning; it does not restore naivety.

Self-documented freeze: There was no external registration, no third-party timestamping and no independent replication. Event ordering rests on self-recorded timestamps in a single writable environment.

Late corrections: Four implementation deviations were identified after the evaluation stage and corrected with the evaluation results already known. Their effects are documented and bounded — no evaluation decision changed — but they were not corrections made blind to the outcome.

Policy parameters, not requirements: The margin δ and the false-rejection allowance ϵ are stipulated choices. δ was set from the across-checkpoint variability of the nominal validation rate and ϵ from operational judgement; neither is derived from a safety requirement, and different values would change which rules pass which criterion.

No observation of the human path: Every quantity concerning non-accepted images describes what the policy would do, not what happened. Manual review and reacquisition were neither carried out nor evaluated, and no assumption about their effectiveness enters the analysis.

No field validation: The study is retrospective and offline. Nothing here has been tested in an operational inspection workflow, and field validation would be required before any rule of this kind is used.

6 | Conclusion

This study calibrated an image-quality gate on a validation subset of above-ground gas-pipeline photographs, froze the rule and its thresholds, and applied the frozen rule once to an evaluation subset.

For the primary rule, calibrated on deformation outcomes as a function of luminance-normalised global sharpness, no acceptance zone was established: the isotonic estimate did not reach the pre-specified target at any grid point, and the acceptance threshold was therefore infinite. The primary endpoint, the change in residual failure among accepted images, is consequently not estimable. This held across the examined margin grid and when the single large validation image was excluded from calibration.

Among the secondary rules evaluated, $G\text{-cal}|\epsilon$ — the pre-specified fallback of the primary rule — and $G\text{-env}$, an envelope calibrated on the feature values of nominal validation frames alone, accepted 42.7% and 47.2% of evaluation versions at $\epsilon = 0.05$. The raw-sharpness comparator $G\text{-raw}$ and the sensitivity variants are reported alongside them. For deformation, both produced small negative point estimates for the change in residual failure, but every interval included zero, so no reduction was demonstrated under the pre-declared criterion. Non-inferiority to the nominal operating level within the pre-specified margin was not demonstrated either, and the intervals are compatible both with performance inside the margin and with a shortfall below it. Non-acceptance implied a substantial computed workload rather than a measured cost: under $G\text{-cal}|\epsilon$ at $\epsilon = 0.05$, the policy would route 59.3% of deformation object–seed observations to manual review, 20.2% of which had in fact received a valid match from the model; no manual review was carried out in this study. Both rules exceeded the policy allowance for rejecting nominal frames, on validation and on evaluation.

Pipe-class results are secondary and descriptive. The two rules showed negative changes in residual failure with intervals excluding zero, and non-inferiority within the pipe margin was not demonstrated. These estimates were obtained with thresholds not calibrated on pipe outcomes and do not make any rule a validated gate.

Three qualifications bound what can be drawn from this. The evaluation subset was not naive, the freeze was self-imposed and self-documented rather than externally registered, and four implementation deviations were found in an audit that followed the evaluation stage, with all corrections computed after the evaluation results were known; acceptance thresholds were unchanged in every comparable configuration and no evaluation decision changed. The findings apply to one feature family, one rule form, one segmentation system and one dataset with 72 evaluation deformation objects, and they do not establish that input-quality gating in general cannot help.

Field validation would be required before any rule of this kind is used, and no acceptance decision reported here is a statement about the condition of a pipeline.

Author Contribution

The author is the sole author and is responsible for all aspects of this work: conceptualisation, methodology, software, formal analysis, investigation, data curation, and writing.

Funding

This work received no external funding. It was conceived, conducted and funded personally by the author.

Data Availability

The material supporting the findings is organised as Supplementary Material S1 to S5, listed with a SHA-256 manifest for every file, together with a standalone reproduction package covering thresholds, corrections, metrics, tables and Supplementary S1–S4, and one figure, regenerated from a clean directory and compared exactly against the previously released files; two items (a secondary figure-rendering script and a secondary decision-impact script) were not exercised in that check, and it is reported as an internal check rather than independent replication.

The image corpus is the publication-3/publication-3 project on Roboflow Universe (universe.roboflow.com/publication-3/publication-3), exported on 2026-06-25 in COCO instance-segmentation format, 9,965 image records under the CC BY-NC-SA 4.0 licence stated in the exported README. The specific version number or generation date of this export is not recorded in any file available for this study and is not stated here; it is not established from the project's live overview page either, since the current public listing was not cross-checked against this export item by item and is not assumed to be the same export without such a check. The composition of the validation and evaluation subsets used here, and the SHA-256 of every image and analysis file, are given in the family manifest and the Supplementary manifests referenced in Section 2.2 and throughout Supplementary S1–S5; those hashes, not a Roboflow version number, are what identifies the exact data used in this study and would allow it to be obtained or verified independently of platform version history.

The underlying photographs and their annotations are not currently public. The author intends to prepare them for open release once a check for personal data and a licensing decision are complete; a full-resolution check has since been carried out and is reported in full, with every affected file named, in the accompanying PII_CHECK_REPORT.md, which found no identifiable face among the 300 baseline images, four images with people who are not facially identifiable, two vehicles with a legible or partially legible plate, and two further vehicles with a plate-shaped area visible but not legible; anonymised copies of all seven affected images have been prepared. No release is made at the time of writing. The model checkpoints and the raw prediction files are not released at this time; their release has not been separately agreed.

Conflicts of Interest

The system studied and the research reported here were developed by the author personally, outside any project of the author's employer and using the author's own resources; the employer did not participate in or fund this study, and the affiliation given denotes the author's place of employment only, not a sponsor or collaborator. No third party receives financial benefit from the system studied. The author declares no competing interests.

Ethics Approval and Consent to Participate

A full-resolution check for personal data has been carried out across the 300 baseline photographs underlying the validation and evaluation subsets, in two automated passes (face, then full-body/upper-body/license-plate cascades, 11 classifiers in total) followed by full manual review and, for every resulting candidate, individual inspection at native resolution; it is reported in full, with every affected file named, in PII_CHECK_REPORT.md. No identifiable human face was found. Four images show people who are not facially identifiable; two vehicles have a legible or partially legible plate, and a further two have a plate-shaped area that is visible but not legible. Anonymised copies of all seven of these images have been prepared separately, without altering the original research images, which were re-verified byte-identical afterward. This check establishes content and prepares release candidates; it does not authorise release, which remains withheld pending a licensing decision. The photographs were taken personally by the author in public locations in Almaty. They are not the property of, and were not commissioned or funded by, the author's employer.

Overlapping Publication and Prior Dissemination

This study reuses material from two companion manuscripts: The same image corpus and its family-wise partition, the same five model checkpoints, and, for ten of the fourteen image conditions, the same archived model predictions, which were re-derived here and reproduced the archived results exactly. New inference was performed for the four intermediate conditions and for all validation outcomes under degradation. The evaluation subset had been used in that earlier work, so this study is described throughout as a retrospective internal evaluation rather than an independent validation. The companion manuscript on source-aware data partitioning has been submitted to the International Journal of Data Science and Analytics; it has passed the journal's technical/quality checks and is awaiting assignment of a handling editor. The companion manuscript on segmentation robustness under controlled image degradation has been submitted to Machine Vision and Applications and is currently with the handling editor for evaluation. Neither has been accepted, published, or, to the author's knowledge, sent for external peer review as of this submission.

Use of Generative Artificial Intelligence

Anthropic Claude and OpenAI ChatGPT were used in preparing the manuscript and analysis code and in checking the results. The author is responsible for the final content. Described further in Section 2.10.

Preregistration

The analysis protocol and its thresholds were frozen and hash-recorded before the evaluation stage of this study, as described in Section 2.8; the protocol was not registered with any external party, and event ordering rests on self-recorded timestamps in a single writable environment. This study is accordingly reported as having a self-documented freeze, not as a preregistered study.

References

- [1] Spencer, B. F., Hoskere, V., & Narazaki, Y. (2019). Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering*, 5(2), 199–222. <https://doi.org/10.1016/j.eng.2018.11.030>

- [2] Eginov, A., & Eginova, S. (2025). A strategy for implementing automated surface-level pipeline monitoring systems based on machine vision for geohazard assessment and risk management: A case study in Almaty. *Engineering, technology & applied science research*, 15(4), 26139–26147. <https://doi.org/10.48084/etasr.11751>
- [3] Eginov, A. A., & Eginova, S. A. (2025). Pipesdefectdetection: A novel deep learning framework for real-time gas pipeline safety monitoring and defect recognition. *SoftwareX*, 31, 102357. <https://doi.org/10.1016/j.softx.2025.102357>
- [4] Chen, C., Zhang, T., Ma, C., Liu, Z., ur Rehman, S., Shi, Y., & Munshi, A. (2026). L-DC-DETR: A lite distance-content aware DETR for gas pipeline defect detection. *Signal, image and video processing*, 20(1), 16. <https://doi.org/10.1007/s11760-025-05070-y>
- [5] Dodge, S., & Karam, L. (2016). Understanding how image quality affects deep neural networks. 2016 *eighth international conference on quality of multimedia experience (QOMEX)* (pp. 1–6). IEEE. <https://doi.org/10.1109/QoMEX.2016.7498955>
- [6] Hendrycks, D., & Dietterich, T. (2019). *Benchmarking neural network robustness to common corruptions and perturbations*. <https://doi.org/10.48550/arXiv.1903.12261>
- [7] Michaelis, C., Mitzkus, B., Geirhos, R., Rusak, E., Bringmann, O., Ecker, A. S., ... , & Brendel, W. (2019). *Benchmarking robustness in object detection: Autonomous driving when winter is coming*. <https://doi.org/10.48550/arXiv.1907.07484>
- [8] Chow, C. K. (1957). An optimum character recognition system using decision functions. *IRE transactions on electronic computers*, EC-6(4), 247–254. <https://doi.org/10.1109/TEC.1957.5222035>
- [9] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf
- [10] Kwon, H., & Kim, D.-J. (2026). Conformal selective prediction with cost aware deferral for safe clinical triage under distribution shift. *Scientific reports*, 16(1), 10016. <https://doi.org/10.1038/s41598-026-40637-w>
- [11] Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-reference image quality assessment in the spatial domain. *IEEE transactions on image processing*, 21(12), 4695–4708. <https://doi.org/10.1109/TIP.2012.2214050>
- [12] Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a “completely blind” image quality analyzer. *IEEE signal processing letters*, 20(3), 209–212. <https://doi.org/10.1109/LSP.2012.2227726>
- [13] Pech-Pacheco, J. L., Cristobal, G., Chamorro-Martinez, J., & Fernandez-Valdivia, J. (2000). Diatom autofocusing in brightfield microscopy: A comparative study. *Proceedings 15th international conference on pattern recognition. ICPR-2000* (Vol. 3, pp. 314–317). IEEE. <https://doi.org/10.1109/ICPR.2000.903548>