*algorithms*

MDPI

*Article*

# Managing Cost–Stability Trade-Offs in Industrial Object Detection: A Unified Decision Support Framework

**Kuhyun Lee [1,2,3], Jihoon Hong [3], Beom-Seok Kim [4], Yuna Song [4] and Dong-Hee Lee [4,*]**

[1] Department of Semiconductor and Display Engineering, Sungkyunkwan University, 2066, Seobu-ro, Suwon-si 16419, Gyunggi-do, Republic of Korea; lkh567@skku.edu
[2] Samsung Institute of Technology, 1, Samsung-ro, Yongin-si 17113, Gyunggi-do, Republic of Korea
[3] Memory Division, Samsung Electronics Co., Ltd., 1-1, Samsungjeonja-ro, Hwaseong-si 18448, Gyunggi-do, Republic of Korea; jihoon.hong@samsung.com
[4] Department of Industrial Engineering, Sungkyunkwan University, 2066, Seobu-ro, Suwon-si 16419, Gyunggi-do, Republic of Korea; 888skdls@g.skku.edu (B.-S.K.)
* Correspondence: dhee@skku.edu; Tel./Fax: +82-31-290-7609

**Abstract**

Object detection is a core component of industrial vision systems in manufacturing, infrastructure monitoring, and safety-critical sensing. While the mean average precision (mAP) averages the performance over all confidence thresholds, real-world deployment demands committing to a single operating threshold under score imprecision, distribution shifts, and asymmetric—often only approximately known—error costs. From a soft-computing perspective, deployment should explicitly manage this uncertainty rather than rely on a static validation optimum. We propose domain-specific and robust localization recall precision (DSR-LRP), a three-phase decision-support framework. The framework elicits soft domain preferences—such as asymmetric error costs, tolerable localization imprecision, and expected perturbations—from practitioner knowledge and encodes them as three quantitative parameters ($k$, $\alpha_{\mathrm{IoU}}$, $\beta$). A cost-sensitive, threshold-local objective aggregates the performance within a robustness band around each candidate threshold, jointly capturing the accuracy and local stability. Finally, it yields an interpretable recommendation package comprising the operating threshold, its DSR-LRP score, and visual evidence. Experiments on four practical datasets (blood cell screening, wildfire smoke monitoring, pothole detection, and semiconductor sensor inspection) showed that DSR-LRP consistently selected operating thresholds that were robust and cost-aligned. For example, in pothole detection, an LRP-optimal threshold degraded by 15.6% under simulated shifts, while the DSR-LRP recommendation changed by only 1.8%. DSR-LRP complements global metrics such as the mAP and provides a soft-computing-oriented tool for reliable, evidence-driven deployment of industrial object detectors.

**Keywords:** object detection evaluation; deployment decision support; operating threshold selection; asymmetric cost weighting; distribution shift robustness; localization recall precision

Check for updates

Academic Editor: Guanqiu Qi

Received: 24 April 2026
Revised: 12 May 2026
Accepted: 13 May 2026
Published: 19 May 2026

**Copyright:** © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

## 1. Introduction

Object detection, a core area of computer vision, has seen continuous technological advancement and practical application driven by industrial demands [1–4]. Object detectors are now routinely deployed to screen defects in manufacturing processes [5–8], support autonomous driving and advanced driver-assistance systems (ADAS) [9–13], assist medical image analyses [14–18], and monitor road and infrastructure conditions [19–21]. In these

applications, object detection rarely acts in isolation; it operates as part of a broader soft-computing-oriented pipeline, in which noisy sensor measurements, ambiguous visual cues, and uncertain operating conditions must be transformed into reliable, actionable decisions under limited computational and engineering budgets. Soft computing is commonly defined as a collection of methodologies—such as fuzzy systems, neuro-computing, and evolutionary algorithms—that exploit the tolerance for imprecision and uncertainty to achieve tractability, robustness, and a low solution cost [22].

Recent progress in object detection has been driven by deep learning architectures such as faster R-CNN [23], you only look once (YOLO) [24], and detection transformer (DETR)-based architectures [25,26], which achieve a high accuracy on public benchmarks. Their performance is typically summarized by the mean average precision (mAP) [27–29], the de facto standard that integrates the precision–recall curve over all confidence thresholds. Recent applied studies, such as the work of Goswami et al. [30] on open-world litter detection, also rely on the mAP and inference speed to benchmark detector architectures in environmental monitoring systems. While the mAP is indispensable for comparing models on static benchmarks, it does not directly answer a practitioner's central deployment question:

> At which confidence threshold should this detector be operated in my specific environment, given my cost structure and robustness requirements?

In real systems, detectors output continuous confidence scores that must be converted into binary decisions—e.g., "defect/no defect", "smoke/no smoke"—to trigger downstream actions. Therefore, selecting a single operating threshold along the confidence axis is unavoidable. However, this choice must be made under three intertwined sources of imprecision and uncertainty:

(i) **A threshold-specific evaluation for binary decisions.** The perception module must ultimately output binary decisions to be actionable. However, global metrics such as the mAP average the performance over all thresholds, providing limited insight into local behavior [31,32]. Practitioners thus resort to auxiliary single-point criteria such as the F1 score [33] or localization recall precision (LRP) [34], but these lack a built-in notion of robustness around the operating point.

(ii) **Asymmetric and partially known error costs.** The consequences of false positives (FPs) and false negatives (FNs) are rarely symmetric. In safety-critical domains, FNs can be catastrophic, while in others, FPs may dominate costs. In practice, relative costs are often approximate soft domain preferences rather than exact numbers [35–37]. Standard metrics implicitly assume balanced costs, and a systematic framework handling these soft preferences is lacking.

(iii) **Robustness to distribution shifts around the operating threshold.** Once deployed, detectors inevitably face data distribution shifts [38–40]. Even modest calibration changes can move the model along the confidence axis, making a validation-optimal threshold brittle. Existing metrics focus on single points or global aggregates, failing to explicitly quantify the local stability around the chosen threshold.

From a soft-computing viewpoint, these requirements highlight the need to handle imprecision, uncertainty, and inherent ambiguity in detector outputs while supporting tractable deployment decisions—exactly the type of setting where soft computing is positioned as a flexible alternative to rigid, hard-computing formulations [22]. In our context, this softness manifests in three aspects: (1) cost asymmetry, (2) acceptable localization ranges, and (3) robustness bands. Similar soft-preference ideas have been effectively used to encode flexible preferences in clustering and supply-chain optimization [41,42].

Soft-computing techniques have also been explored for object detection and tracking under uncertainty [43]. However, most works have focused on model architectures rather

than deployment-time decision support. Existing metrics provide only partial support: the mAP ignores local behavior, and the LRP focuses on an isolated optimum. Consequently, practitioners lack a unified criterion for selection.

To resolve this gap, we propose domain-specific and robust localization recall precision (DSR-LRP), a systematic soft-computing-based three-phase decision-support framework, in line with existing decision support systems that integrate expert knowledge with data-driven models [44,45].

- **Phase 1 (knowledge elicitation and parameterization)**: Qualitative field requirements are translated into three domain parameters ($k$, $\alpha_{\text{IoU}}$, $\beta$). These act as soft domain preferences reflecting domain-specific trade-offs.
- **Phase 2 (robust and cost-sensitive evaluation)**: We define a cost-sensitive LRP extended with a robustness band objective. For each candidate threshold $s$, DSR-LRP aggregates the performance over the interval $[s - \beta, s + \beta]$, jointly capturing the mean error and variability. This converts the qualitative tolerance for uncertainty into a quantitative objective.
- **Phase 3 (analysis and interpretable recommendation)**: We identify the optimal threshold $s^*$ and compile an interpretable package containing the recommended point, a quantitative rationale, and visual evidence.

Aligned with the principles of soft computing, our framework exploits the tolerance for imprecision and asymmetric costs to achieve robust deployment decisions with minimal overhead. We evaluated DSR-LRP on four diverse datasets. The results confirm that DSR-LRP consistently favors operating thresholds that are stable under distribution shifts and aligned with domain priorities.

## 2. Practitioner-Centric Evaluation Tools for Object Detection

### *2.1. Introduction: Evaluation Tools as Decision Support Components*

Object detection evaluation is generally based on a small set of standard metrics such as the mean average precision (mAP), calibration measures (e.g., D-ECE [46]), and the localization recall precision (LRP). These quantities are indispensable for benchmarking models. However, in deployment, they act only as individual signals within a broader decision process rather than as complete decision rules. From the viewpoint of the three practitioner-centric and soft-computing requirements—a single-threshold evaluation, asymmetric cost weighting, and robustness to distribution shifts—it is natural to reinterpret these metrics as components of an evaluation layer in a decision-support pipeline, and to ask which parts of these requirements they cover and which remain unsupported.

In this section, we analyze the existing evaluation tools from this practitioner-centric perspective. Our goal is not to propose yet another metric, but to clarify how the mAP, D-ECE, and LRP can be used as building blocks and where a unified decision-support framework is still missing.

### *2.2. The Ambiguity of Global Aggregation vs. Binary Decisions*

First, practical decision support demands a performance assessment at a specific operating threshold. Global metrics such as the mAP, which integrate the precision–recall curve over all thresholds [47], remain useful as high-level benchmarking signals. As demonstrated in comprehensive evaluations of YOLO and transformer-based detectors [30], these metrics can effectively guide the selection of a model architecture for a given application.

However, such global comparisons fail to resolve the operational ambiguity faced by practitioners in deployment: at which specific threshold should the chosen detector be operated? As noted by Wenkel et al. [31], strategies to boost global mAP scores by including low-confidence predictions can create a discrepancy between reported benchmark scores

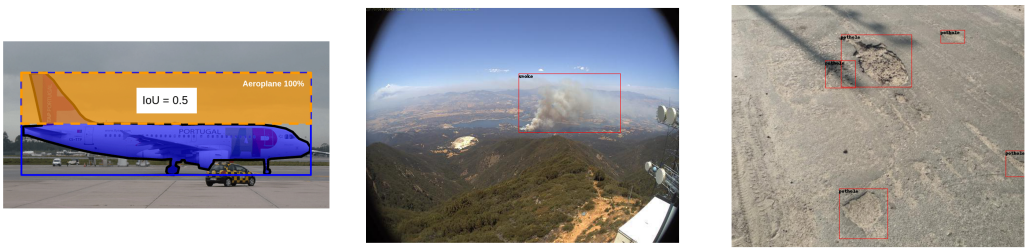
and the actual performance in real-world applications. Similarly, Jena et al. [32] presented experimental evidence that the mAP is insensitive to duplicate detections, which can be critical for a system deployed at a specific threshold.

These findings collectively show that relying solely on global metrics may fail to capture the performance nuances and associated risks at a chosen operating threshold. In a soft-computing-oriented pipeline, the mAP is better viewed as a global summary feature that must be complemented by a threshold-local analysis. To address this, Oksuz et al. [34] proposed the LRP metric, which identifies and reports the performance at a model's optimal threshold. It enables a meaningful single-point evaluation, serving as a potential pointwise building block for our framework.

### *2.3. Rigid Counting vs. Soft Domain Preferences*

Second, a practical decision-support layer must allow for flexible cost weighting to suit domain-specific requirements. However, standard metrics effectively treat false positives and false negatives symmetrically ($\text{cost}_{\text{FN}} = \text{cost}_{\text{FP}}$) and average over strict intersection-over-union (IoU) thresholds. In reality, applications involve asymmetric and often qualitative error costs—a classic scenario for soft domain preferences.

For instance, as illustrated in Figure 1, domain priorities vary significantly: (a) approximate localization is sufficient for aerial objects where precise boundary delineation is not strictly required, whereas (b) missed detections are catastrophic in wildfire monitoring compared to false alarms [35–37]. Conversely, (c) in pothole maintenance, false positives can lead to unnecessary and costly road closures. Standard metrics do not provide mechanisms for such domain-specific customization. Moreover, while the LRP offers a conceptual framework for weighting by the decomposing localization, FP, and FN components, its public implementations default to equal weighting. A systematic procedure for mapping soft practitioner preferences (e.g., "safety is paramount", "approximate box localization is acceptable") into quantitative evaluation parameters is still lacking.




(**a**) (**b**) (**c**)

**Figure 1.** Contrasting error cost structures and localization requirements across different industrial domains. (**a**) Aerial objects: most spatial information is already captured at a relaxed IoU of 0.50 [48]. (**b**) Wildfire smoke: the cost of missed detections (FNs) exceeds the cost of false alarms (FPs). (**c**) Potholes: the cost of false alarms (FPs) exceeds the cost of missed detections (FNs).

### *2.4. Pointwise Optima vs. Robustness to Distribution Shifts*

Finally, an ideal evaluation layer in a decision-support pipeline should reflect the performance robustness against the distribution shifts encountered in real-world deployment environments [38–40]. Although the LRP marked a significant step toward deployment-centric evaluation by focusing on a single operating point, relying on an isolated optimum risks making the chosen threshold fragile—a phenomenon we term "threshold brittleness".

For instance, as we demonstrate empirically in our experimental case study, a deviation of just 0.05 from the LRP-optimal threshold can increase the LRP error (lower is better) by over 15% in pothole detection scenarios. This fragility poses a significant risk, as minor distribution shifts are inevitable after deployment. Therefore, a soft-computing-oriented

approach must go beyond a single-point optimum and seek a robustness band—a stable interval where the performance remains consistent, despite input uncertainty. In our framework, this idea of threshold-local evaluation is not replaced, but rather extended: we retain the LRP as the pointwise building block and then embed it into a soft-computing-style process that incorporates elicited costs and robustness bands around the operating threshold.

### *2.5. Summary: The Need for a Unified Framework*

In summary, as organized in Table 1, existing evaluation metrics and decision-support frameworks offer partial, but fragmented, support for the three practitioner-centric requirements. Global metrics such as the mAP and calibration measures provide useful threshold-agnostic indicators, but they do not directly support operating-threshold selection under domain-specific costs or local robustness requirements. Single-threshold metrics such as F1, $F_\beta$, and the LRP provide a more localized view, but they do not jointly address explicit FP/FN cost parameterization and robustness around the selected operating point. Candidate-level deployment audit frameworks such as CSEF provide an important step toward deployment-oriented evaluation, but their stability assessment is conducted at the candidate or dataset level rather than within a local neighborhood of the operating threshold.

**Table 1.** Comparison of evaluation metrics and decision-support frameworks against practitioner-centric deployment requirements.

| Approach | Brief Description | Practitioner-Centric Deployment Requirements | | |
|---|---|---|---|---|
| | | Local Threshold | Cost Weighting | Local Robustness Around Threshold |
| mAP [47] | Global PR-curve summary over confidence thresholds. | × | × | × |
| D-ECE [46] | Calibration error over confidence levels. | × | × | × |
| F1 score [33] | Single-threshold harmonic mean of precision and recall. | ✓ | × | × |
| $F_\beta$ score [49] | Single-threshold metric with recall-oriented weighting. | ✓ | △ | × |
| LRP [34] | Single-threshold localization recall precision error. | ✓ | △ | × |
| CSEF [50] | Candidate-level deployment audit using cost-sensitive operating performance and stability checks. | ✓ | ✓ | △ |
| DSR-LRP | Threshold-level objective with cost-sensitive LRP and local robustness band. | ✓ | ✓ | ✓ |

**Note.** Local threshold indicates whether the approach directly supports evaluation or selection at a specific operating confidence threshold. Cost weighting indicates whether asymmetric FP/FN costs can be explicitly parameterized. Local robustness around threshold indicates whether stability is evaluated within a localized neighborhood of the operating threshold. Symbols: ✓ denotes that the approach fully supports the requirement; △ denotes partial or indirect support; × denotes that the requirement is not supported. CSEF supports threshold-level evaluation by selecting a validation-optimal operating threshold and evaluating the performance at that point, but its stability assessment is conducted at the candidate or dataset level rather than within a local band around the operating threshold.

DSR-LRP is distinct from calibration and uncertainty-aware evaluation methods in its decision target. Calibration methods, such as logistic calibration, beta calibration, histogram binning, or MC-dropout-based calibration, modify or reinterpret the confidence

score landscape. Calibration metrics such as D-ECE evaluate the reliability of confidence estimates across confidence levels. In contrast, DSR-LRP operates on a given set of detector outputs, whether calibrated or uncalibrated, and evaluates the operational consequence of selecting a single confidence threshold under asymmetric error costs and local score shifts. Thus, DSR-LRP is complementary to calibration and uncertainty estimation: these methods may improve or characterize confidence scores, whereas DSR-LRP provides a cost-aware and locally robust operating-threshold recommendation.

By synthesizing the above, the remaining gap is structural rather than merely incremental. Existing metrics and frameworks address different parts of the deployment problem: the mAP summarizes the global ranking performance; D-ECE assesses the calibration quality; F1, $F_\beta$, and the LRP operate at a specific threshold; and CSEF provides a candidate-level deployment audit. However, none of these directly combine explicit FP/FN cost parameterization with a local robustness assessment around the operating confidence threshold. This gap motivates DSR-LRP as a threshold-level objective for cost-aware and locally robust deployment decisions.

From a soft-computing viewpoint, this situation corresponds to having several useful, but uncoordinated, hard signals rather than a unified, knowledge-based evaluation layer. In the next section, we introduce DSR-LRP not as a replacement for these tools, but as a soft-computing-oriented decision-support framework that (i) elicits and encodes domain preferences, and (ii) aggregates pointwise LRP-style scores within robustness bands to recommend reliable operating thresholds with interpretable evidence.

## 3. Methodology: The DSR-LRP Decision Support Framework

We propose the domain-specific and robust localization recall precision (DSR-LRP) framework, a tool designed to guide the comparative evaluation of various deployment candidates and to support the selection of the optimal candidate–threshold pair. It retains the strength of the LRP (performance evaluation at a single threshold) while explicitly incorporating domain-specific cost weighting and local robustness under distribution shifts. More than just a simple evaluation metric, DSR-LRP is an integrated process that functions as a knowledge-based system to support reliable decision-making under uncertainty.

We develop this framework in three phases (Figure 2). The complete algorithmic procedure is presented in Algorithms A1 and A2 of Appendix A. First, in Phase 1, we define a knowledge elicitation process that translates practitioners' soft domain preferences into quantitative parameters. Second, in Phase 2, we apply these parameters to evaluate a given candidate by computing the full set of per-threshold LRP and DSR-LRP scores. Finally, in Phase 3, we describe how to analyze these raw scores to identify the optimal operating threshold and synthesize all findings into an interpretable recommendation package that practitioners can readily understand and apply. In this sense, the DSR-LRP objective is the metric-level core of the proposed framework, while the three-phase structure operationalizes it as a deployment-oriented decision-support process.

### *3.1. Phase 1: Knowledge Elicitation and Parameterization*

The initial phase of the DSR-LRP framework is the knowledge elicitation process, designed to directly incorporate the qualitative requirements of practitioners into the evaluation framework. This process translates domain-specific characteristics and operational uncertainties into three core parameters that are interpretable by the system:

1. **Classification cost ratio ($k > 0$, $k = Cost_{\mathrm{FN}}/Cost_{\mathrm{FP}}$):** encodes soft preferences over relative error costs, where $k > 1$ applies when missed detections are more critical (e.g., wildfire monitoring) and $k < 1$ applies when false alarms waste resources (e.g., unnecessary maintenance).

2. **Localization importance ($\alpha_{\text{IoU}} \in [0, 1]$):** controls the tolerance for bounding-box imprecision, where $\alpha_{\text{IoU}} = 1$ imposes a strict penalty, while $\alpha_{\text{IoU}} = 0$ applies no additional penalty beyond the minimum IoU threshold $\tau$.
3. **Robustness band ($\beta$):** defines the width of the interval over which stability is evaluated to anticipate performance changes under post-deployment distribution shifts; for example, setting $\beta = 0.05$ means that, for a given candidate threshold $s$, both the performance and stability are evaluated across the entire interval $[s - 0.05, s + 0.05]$ centered on $s$.

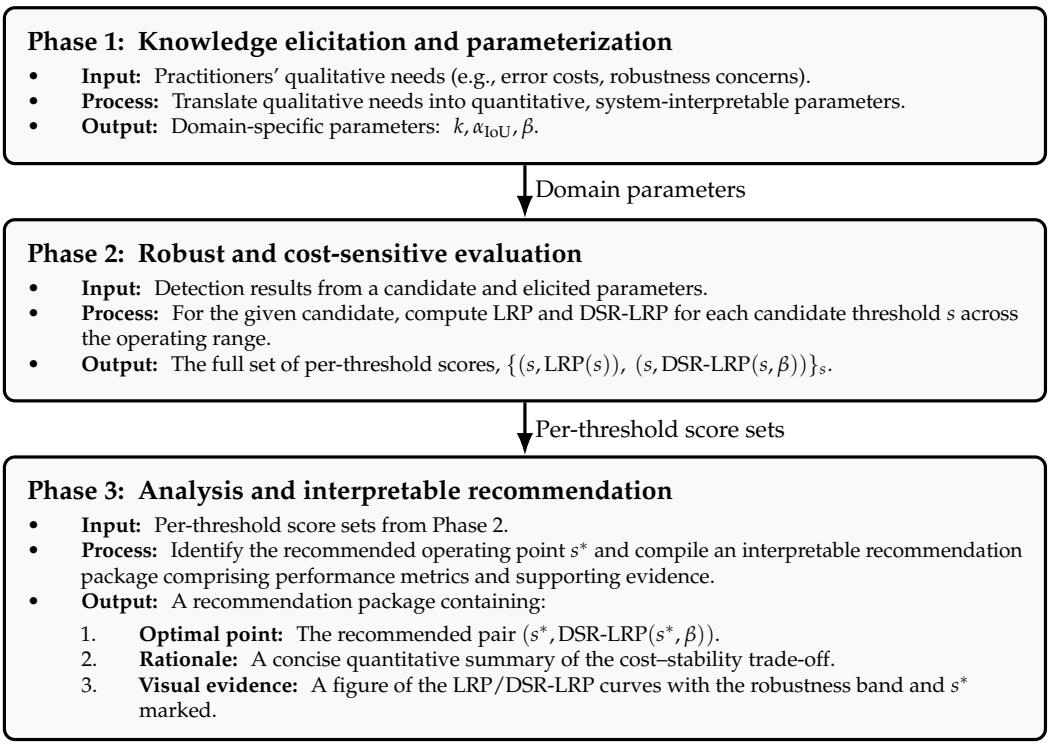



**Figure 2.** The three-phase DSR-LRP decision-support framework. This framework is applied to each candidate to generate a comprehensive evaluation package. Practitioners can then compare minimum DSR-LRP values and supporting evidence across multiple candidates (e.g., different models and post-processing rules) to make a data-driven deployment decision handling trade-offs.

Implementation Memo

All three parameters can be specified directly by the user, but defaults are provided when information is limited. The classification cost ratio $k$ and the localization importance $\alpha_{\text{IoU}}$ default to the symmetric setting $k = 1$ and $\alpha_{\text{IoU}} = 1$, and for the robustness band $\beta$, we recommend a default of $\beta = 0.05$ for general use; in our experiments with noise and brightness perturbations, the confidence score fluctuations were about 0.04–0.05.

Practical Implications

The three parameters directly quantify what to prioritize (cost), how much localization precision to require, and how much to hedge against real-world variability. By adjusting these parameters, practitioners can directly incorporate qualitative field requirements—such as "an FN is three times more costly than an FP" or "approximate box localization is acceptable"—into the evaluation process. This capability explicitly handles the imprecision in human preferences, improving both the interpretability and alignment with domain needs.

**Practical guide to customizing DSR-LRP parameters**

**Classification cost ratio ($k > 0$)**: $k = Cost_{FN}/Cost_{FP}$

- Symmetric (default, $k = 1$): treats errors equally.
- FN-critical ($k > 1$): safety-critical applications.
- FP-critical ($k < 1$): resource-intensive scenarios.

**Localization importance ($\alpha_{IoU} \in [0, 1]$)**: penalty for imprecise boxes

- $\alpha_{IoU} = 1$ (default): precise localization is critical.
- $\alpha_{IoU} = 0$: object presence alone is sufficient.

**Robustness band ($\beta$)**: expected magnitude of confidence score shifts

- $\beta = 0.05$ (default): empirical recommendation for general use.
- Tuning: use a tighter band (e.g., 0.03) for stable data, or a wider band (e.g., 0.10) for highly variable data.

### *3.2. Phase 2: Integration of Cost Sensitivity and a Robustness Band*

This section details the computational core of our framework, which corresponds to Phase 2. For any given candidate, this phase operationalizes the parameters derived in Phase 1, producing the raw per-threshold performance scores used in the Phase 3 analysis. To achieve this, we first establish a cost-sensitive LRP for pointwise evaluation (Section 3.2.1). Then, we extend this metric with a robustness band to formulate the DSR-LRP score, which is computed alongside the LRP score over the range of operating thresholds (Section 3.2.2).

#### 3.2.1. Foundation: Cost-Sensitive LRP for Pointwise Evaluation

The localization recall precision (LRP) error [34] is a cost function that integrates localization inaccuracy with classification errors into a single scalar. The computation begins with the set of ground-truth boxes $X$ and the model's detections $Y$. For a given confidence threshold $s$, detections are filtered into $Y_s = \{y \in Y \mid \text{conf}(y) > s\}$. Using a fixed IoU threshold $\tau$, this subset yields the counts of true positives $N_{TP}$, false positives $N_{FP}$, and false negatives $N_{FN}$. Based on these counts, the general LRP framework is a weighted sum of three error components:

$$\text{LRP}(X, Y_s) := \frac{1}{Z}\Big(w_{IoU}\,\text{LRP}_{IoU}(X, Y_s) + w_{FP}\,\text{LRP}_{FP}(X, Y_s) + w_{FN}\,\text{LRP}_{FN}(X, Y_s)\Big), \quad (1)$$

where $Z = N_{TP} + N_{FP} + N_{FN}$ and

$$\text{LRP}_{IoU}(X, Y_s) = \frac{1}{N_{TP}} \sum_{i=1}^{N_{TP}} (1 - \text{IoU}(x_i, y_{x_i})),$$
$$\text{LRP}_{FP}(X, Y_s) = \frac{N_{FP}}{|Y_s|}, \quad \text{LRP}_{FN}(X, Y_s) = \frac{N_{FN}}{|X|}.$$

In the standard implementation, the proposed weights are $w_{IoU} = N_{TP}/(1 - \tau)$, $w_{FP} = |Y_s|$, and $w_{FN} = |X|$. While this formulation successfully identifies an optimal threshold, it tightly couples the components, so that $w_{FP}\text{LRP}_{FP} = N_{FP}$ and $w_{FN}\text{LRP}_{FN} = N_{FN}$, locking the relative costs of FPs and FNs. We refer to this symmetric-cost implementation as the standard LRP and use it as the baseline.

Re-Parameterization Using Elicited Knowledge

By leveraging the knowledge from Phase 1 (Section 3.1), we re-parameterize the weights with the classification cost ratio $k$ and the localization importance $\alpha_{\text{IoU}}$ to enable asymmetric control. We first map $k$ to two cost multipliers

$$\alpha_{\text{FP}} = \frac{2}{1+k}, \qquad \alpha_{\text{FN}} = \frac{2k}{1+k},$$

which preserve a fixed total scale ($\alpha_{\text{FP}} + \alpha_{\text{FN}} = 2$), reduce to the symmetric case when $k = 1$, and vary smoothly otherwise. Applying these multipliers to the weights in (1)—that is, $w_{\text{FP}} = \alpha_{\text{FP}}|Y_s|$ and $w_{\text{FN}} = \alpha_{\text{FN}}|X|$—transforms the classification terms into the intuitive, cost-adjusted counts $\alpha_{\text{FP}} N_{\text{FP}}$ and $\alpha_{\text{FN}} N_{\text{FN}}$. Independently, $\alpha_{\text{IoU}} \in [0,1]$ scales the IoU penalty, with $\alpha_{\text{IoU}} = 1$ matching standard LRP and $\alpha_{\text{IoU}} = 0$ applying no additional penalty beyond the threshold $\tau$.

Combining these elements yields the cost-sensitive LRP:

$$\text{LRP}(X, Y_s) := \frac{1}{Z}\Big(\alpha_{\text{IoU}} \sum_{i=1}^{N_{\text{TP}}} \frac{1 - \text{IoU}(x_i, y_{x_i})}{1-\tau} + \alpha_{\text{FP}}\, N_{\text{FP}} + \alpha_{\text{FN}}\, N_{\text{FN}}\Big), \tag{2}$$

with a renormalized constant $Z = N_{\text{TP}} + \alpha_{\text{FP}} N_{\text{FP}} + \alpha_{\text{FN}} N_{\text{FN}}$, chosen by the original LRP principle of normalizing by the numerator's theoretical maximum so that the score remains in $[0,1]$ even when $\alpha_{\text{FP}}$ or $\alpha_{\text{FN}} > 1$ under $k \neq 1$. Intuitively, the cost multipliers $\alpha_{\text{FP}} = 2/(1+k)$ and $\alpha_{\text{FN}} = 2k/(1+k)$ act as a smooth re-weighting of the FP and FN counts. When $k = 1$, they reduce to the symmetric case $\alpha_{\text{FP}} = \alpha_{\text{FN}} = 1$. When $k > 1$, the FN term receives a larger penalty, reflecting FN-critical applications; when $k < 1$, the FP term receives a larger relative penalty, reflecting FP-critical applications. The localization parameter $\alpha_{\text{IoU}}$ is independent of this FP/FN trade-off and controls how strongly bounding-box imprecision affects the final score.

Why the New Normalization Is Necessary

Consider $N_{\text{TP}} = 1000$, $N_{\text{FP}} = 3000$, $N_{\text{FN}} = 100$, $k = 1/3$ (FPs three times costlier than FNs), and $\alpha_{\text{IoU}} = 0$. Then, $\alpha_{\text{FP}} = 1.5$ and $\alpha_{\text{FN}} = 0.5$, so the numerator in (2) is 4550. Using the original $Z = N_{\text{TP}} + N_{\text{FP}} + N_{\text{FN}} = 4100$ would yield $1.11 > 1$, whereas our $Z = 1000 + 1.5 \times 3000 + 0.5 \times 100 = 5550$ yields $0.82 \in [0,1]$. Thus, the re-parameterization operationalizes Phase 1 knowledge while preserving the LRP's boundedness and reducing to the standard form when $k = 1$ and $\alpha_{\text{IoU}} = 1$. As this cost-sensitive formulation serves as the foundation for our framework, we will henceforth refer to it simply as the LRP for brevity. The original symmetric-cost implementation will be explicitly denoted as the standard LRP to avoid ambiguity.

Implementation Memo

The essence of the cost-sensitive LRP is to re-express the standard LRP weights $w_{\text{FP}}, w_{\text{FN}}, w_{\text{IoU}}$ in a form that includes the cost multipliers $\alpha_{\text{FP}}, \alpha_{\text{FN}}, \alpha_{\text{IoU}}$. To ensure the LRP score remains within $[0,1]$, it is crucial to redefine the normalization constant as $Z = N_{\text{TP}} + \alpha_{\text{FP}} N_{\text{FP}} + \alpha_{\text{FN}} N_{\text{FN}}$.

Practical Implications

This step removes the unrealistic assumption that all errors have identical costs. By adjusting $k$, practitioners can directly encode the operational cost structure; by adjusting $\alpha_{\text{IoU}}$, they can independently control the importance of localization precision, thereby enabling the selection of a candidate–threshold pair that minimizes the realized loss under soft domain constraints. While this pointwise score is intuitive to interpret, it remains

vulnerable to distribution shifts; hence, the robustness band extension in the next section is required.

#### 3.2.2. Band Extension: Robustness Band Objective Function

The cost-sensitive LRP defined in Section 3.2.1 provides a tailored performance score, but relying on its single-point optimum is risky. This approach rests on the fragile assumption that a threshold optimized on a static test set will remain optimal under real-world distribution shifts, which are common in industrial settings [38]. Therefore, we reframe the evaluation objective from "finding the single best operating point" to "identifying the most stable operating interval" that jointly attains a high average performance and a low variability.

The core premise of our approach is to model the score-level changes induced by distribution shifts. Based on observations from calibration studies [46,51], we assume that these shifts produce a monotonic rescaling of a model's confidence scores. While this does not capture every type of distribution shift, such as out-of-distribution instances causing prediction failures, it effectively models a common and critical failure mode where environmental changes (e.g., in lighting or weather) affect model calibration. This rescaling is functionally equivalent to moving the operating threshold along the confidence axis.

To formalize the new evaluation objective, we apply the robustness band $\beta$ (elicited in Phase 1) to each candidate threshold $s$ and evaluate the interval $[s - \beta,\ s + \beta]$. On a finite set of band points $\{s_i\}_{i=0}^{N}$ defined by $s_i = s - \beta + i\,\Delta$ with $N = \lfloor 2\beta/\Delta \rfloor$, we define the DSR-LRP score as the root mean square of the cost-sensitive LRP values ($RMSE$):

$$\text{DSR-LRP}(s, \beta) = \sqrt{\frac{1}{N+1}\sum_{i=0}^{N}\big(\text{LRP}(s_i)\big)^2}. \tag{3}$$

The parameter $\beta$ controls the size of the local neighborhood around a candidate threshold. When $\beta = 0$, the formulation reduces to the pointwise cost-sensitive LRP; as $\beta$ increases, the objective increasingly favors thresholds located in flatter regions of the LRP curve rather than sharp single-point optima.

The robustness component of DSR-LRP is introduced by the local band evaluation itself, rather than by the use of the RMSE alone. The RMSE serves as the aggregation operator over the LRP values within the robustness band, providing a single scalar value that summarizes the expected error level and local variability around a candidate threshold. We adopted the RMSE as the default aggregation for three reasons. First, since $\text{RMSE} = \sqrt{\mu^2 + \sigma^2}$, it jointly reflects the average error level and volatility within the band. Second, it penalizes large local error spikes more strongly than a simple mean, which is desirable for deployment decisions where brittle thresholds should be avoided. Third, unlike a mean-plus-standard-deviation formulation, the RMSE does not introduce an additional trade-off parameter. Other aggregation operators, such as the mean LRP, mean-plus-standard-deviation, worst-case error, quantile-based risk, or MAD-based measures, could also be used within the same band-based framework. In this work, the RMSE was adopted as a simple, parameter-free compromise between the interpretability and the sensitivity to local instability. The computation of the DSR-LRP scores across all candidate thresholds is the primary output of Phase 2, yielding the raw data needed for the subsequent analysis.

Implementation Memo

The DSR-LRP score set is generated by sweeping candidate thresholds $s$ over the valid interval $[\beta,\ 1 - \beta]$ with a step size $\Delta$, so that the induced band $[\,s - \beta,\ s + \beta\,]$ is always

contained in $[0, 1]$. For each $s$, the finite set $\{s_i\}$ is constructed as defined above, and DSR-LRP$(s, \beta)$ is computed as the RMSE of the LRP values on this set (3). Consistent with common practice and as proposed in the standard LRP paper [34], we adopted a default step size of $\Delta = 0.01$. This setting strikes a practical balance between the computational efficiency and the resolution; for example, with $\beta = 0.05$, the set has $N+1 = 11$ points, so only 11 LRP evaluations per candidate $s$ are required.

Practical Implications

The DSR-LRP score serves as a unified metric that balances two critical aspects of performance: the average error level (mean) and its stability against perturbations (variability). A candidate that achieves a low DSR-LRP score represents an effective trade-off, indicating that it is not only accurate on average, but also robust against the operational distribution shifts commonly faced post-deployment.

### *3.3. Phase 3: Analysis and Interpretable Recommendation*

Phase 3 is the final analysis and reporting stage of the framework. It receives the raw per-threshold score sets from Phase 2 and executes the core process in our framework: first, it identifies the optimal operating threshold $s^*$ for the candidate by analyzing the DSR-LRP scores. Second, it compiles this finding, along with performance metrics and supporting evidence, into a unified and interpretable recommendation package comprising three components:

- **Optimal point:** This component presents the candidate's optimal operating threshold and its corresponding score, the recommended pair $(s^*, \text{DSR-LRP}(s^*, \beta))$. This point is identified by finding the minimum of the DSR-LRP score curve $(s^* = \arg\min_s \text{DSR-LRP}(s, \beta))$.
- **Automated rationale report:** This provides an automatically generated, descriptive summary under the specified Phase 1 parameters. It reports the recommended operating threshold $s^*$, its score DSR-LRP$(s^*, \beta)$, and the shift relative to the LRP optimum $s^*_{\text{LRP}}$, defined as $\Delta s := s^* - s^*_{\text{LRP}}$. For example, "With $k = 3$, $\alpha_{\text{IoU}} = 0$, and $\beta = 0.10$, the recommended operating threshold is $s^* = 0.14$ (with a DSR-LRP score of 0.249); relative to the LRP optimum $s^*_{\text{LRP}} = 0.09$, the shift is $\Delta s = 0.05$." This report is descriptive by design; causal interpretation is left to practitioners.
- **Visual evidence figure:** Beyond a single optimum, this figure provides a multi-faceted analysis of the performance landscape around the selected threshold. It visualizes the performance curves over the entire threshold range, with the confidence score on the x-axis and the LRP and DSR-LRP scores on the y-axis. This figure allows practitioners to intuitively see where $s^*$ is situated in the overall performance landscape and confirms that it lies within a flat region, indicating low performance degradation even under minor distribution shifts. This serves as direct visual evidence of why the selected point is assessed as robust.

This multi-faceted delivery of evidence maximizes the transparency of the evaluation results and empowers practitioners to make final decisions with evidence-driven confidence. An example of the three-part recommendation package is shown in Figure 3.

Implementation Memo

The process for Phase 3 begins by analyzing the per-threshold score sets from Phase 2 to identify the optimal threshold $s^*$. Once $s^*$ is determined, the recommendation package is automatically generated. The automated rationale report is created by populating a template with the Phase 1 parameters and the now-identified optimal point. The visual

evidence figure is rendered by plotting the full score curves from Phase 2, highlighting $s^*$ and its robustness band.

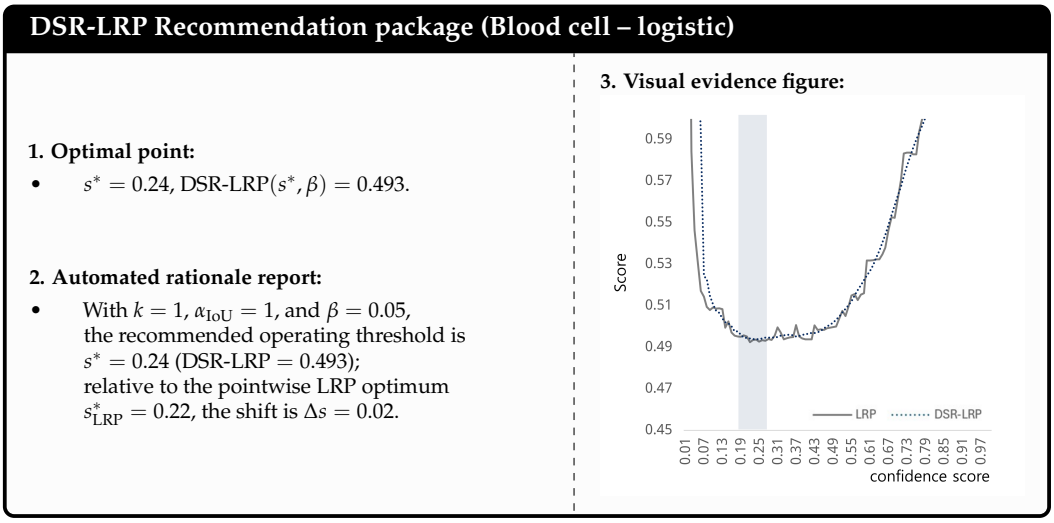



**Figure 3.** Example recommendation package produced by Phase 3. (**Left**): (1) Optimal point and (2) automated rationale report. (**Right**): (3) Visual evidence figure.

Practical Implications

This package elevates DSR-LRP from a mere black-box number to a transparent and trustworthy decision support tool. Instead of blindly accepting the recommended threshold, practitioners can review the quantitative, qualitative, and visual evidence behind "why this threshold is optimal." This instills confidence in the final deployment decision and serves as compelling justification when reporting the evaluation results and persuading relevant departments or management.

## 4. Experiments

To validate the generalizability of DSR-LRP, we conducted experiments across datasets representing diverse, practical applications. We simulated deployment settings by applying multiple confidence score calibration methods to detector outputs (Section 4.2). Our base model in all evaluations was detection transformer with assignment (DETA) [52] using an ImageNet-pretrained ResNet-50 backbone.

Our training protocol was designed to establish a uniform baseline for subsequent calibration experiments, rather than to optimize for the peak mAP for each dataset. Preliminary trials revealed that the epoch yielding the highest validation mAP differed markedly across datasets. However, the quantity of interest for this study was the shape of the LRP vs. threshold curve, which we observed to continue stabilizing even after the mAP peak. To ensure that every model reached this practical LRP-stable regime and to remove the training duration as a confounding variable, we adopted a fixed 50-epoch training protocol for all experiments. This choice guarantees that all comparisons are based on models with similar operational characteristics, rather than on potentially misleading, dataset-specific mAP checkpoints.

The experiments were run on a single NVIDIA RTX A5000 (24 GB) using PyTorch 2.4.0, torchvision, PyTorch Lightning, Transformers, and Albumentations. We used the official DETA image processor (shortest_edge = 800, longest_edge = 1333). Training used AdamW (lr, $1 \times 10^{-4}$; backbone lr, $1 \times 10^{-5}$; weight decay, $1 \times 10^{-4}$; batch size, 8) for 50 epochs without learning rate scheduling; the light augmentations included vertical flip, contrast jitter, and Gaussian blur. The evaluation followed the standard COCO API protocol; we

report the mAP, mAP50, and LRP/DSR-LRP metrics. We did not fix random seeds; minor run-to-run variation was observed, but did not affect the conclusions.

### *4.1. Dataset Information*

Our experimental design was deliberately structured not merely as a head-to-head comparison, but as a series of purpose-driven case studies aimed at validating the key practitioner-centric features of DSR-LRP. First, the blood cell dataset serves as a baseline scenario, configured to mirror the principles of the standard LRP. This allowed us to isolate and evaluate the specific contribution of our proposed robustness component ($\beta$). Subsequently, the wildfire smoke, pothole, and semiconductor manufacturing equipment sensor (hereafter SE sensor) datasets are presented as distinct, real-world case studies. These cases were intentionally chosen to demonstrate the flexibility of DSR-LRP in handling diverse soft domain preferences. They also showcase the metric's capacity for evaluation customization via soft parameter choices ($k$, $\alpha_{\text{IoU}}$, $\beta$); specifically, the ability to de-emphasize localization penalties ($\alpha_{\text{IoU}} = 0$) when object presence is the primary concern. This customizability allows practitioners to align the evaluation with imprecise domain needs and soft operational preferences, a feature largely absent in standard metrics.

All four datasets were split into training/validation/test subsets in a 70:20:10 ratio (see Table 2 for details). The characteristics of each dataset are described below.

**Table 2.** Per split: number of images and per-class object counts for each dataset.

| **Blood Cell** | **Image** | **Class** | | |
|---|---|---|---|---|
| | | **Platelets** | **RBCs** | **WBCs** |
| **Train** | 255 | 249 | 2936 | 263 |
| **Validation** | 73 | 76 | 819 | 72 |
| **Test** | 36 | 36 | 398 | 37 |
| **Wildfire smoke** | **Image** | **Class** | | |
| | | | **Smoke** | |
| **Train** | 516 | | 516 | |
| **Validation** | 147 | | 147 | |
| **Test** | 74 | | 74 | |
| **Pothole** | **Image** | **Class** | | |
| | | | **Pothole** | |
| **Train** | 465 | | 1256 | |
| **Validation** | 133 | | 330 | |
| **Test** | 67 | | 154 | |
| **SE sensor** | **Image** | **Class** | | |
| | | **Average** | **Deviation** | **Drift** |
| **Train** | 1205 | 582 | 498 | 275 |
| **Validation** | 347 | 167 | 176 | 78 |
| **Test** | 173 | 83 | 58 | 40 |

#### 4.1.1. Blood Cell Dataset

This open-source blood cell dataset (https://public.roboflow.com/object-detection/bccd, accessed on 1 July 2025) contains three classes: platelets, red blood cells (RBCs), and white blood cells (WBCs). Because localization accuracy, FNs, and FPs were considered equally important and only typical threshold robustness was required, the experimental settings were $\alpha_{\text{IoU}} = 1$, $k = 1$ ($\alpha_{\text{FN}} = 1$, $\alpha_{\text{FP}} = 1$), and $\beta = 0.05$. An example is shown in Figure 4a.




(**a**) (**b**) (**c**)

**Figure 4.** Examples of open-source dataset images. (**a**) Blood cell. (**b**) Wildfire smoke. (**c**) Potholes.

#### 4.1.2. Wildfire Smoke Dataset

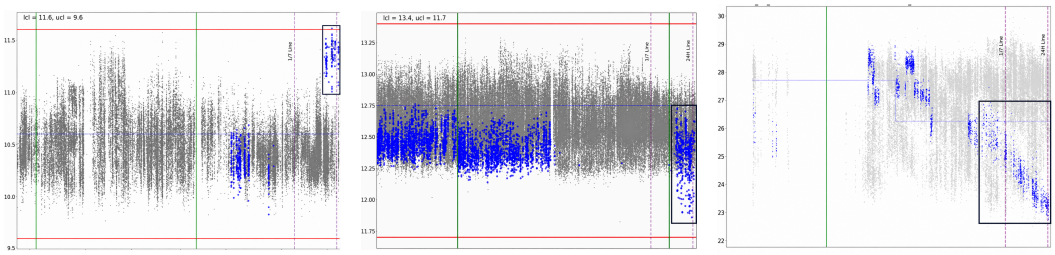
This open-source wildfire smoke dataset (https://public.roboflow.com/object-detection/wildfire-smoke, accessed on 1 July 2025) contains a single class indicating the presence of smoke. In this safety-critical domain, the primary goal is to reliably detect the presence of smoke with a reasonable overlap. To reflect this soft domain preference, we deliberately set $\alpha_{\text{IoU}} = 0$, applying no additional penalty for localization accuracy beyond the initial IoU threshold. Because FNs are extremely costly in this safety-critical domain, we assigned a heavy penalty to $\alpha_{\text{FN}}$. Furthermore, we adopted a wider robustness band $\beta = 0.10$ to ensure stable performance because diverse wildland backgrounds can lead to significant confidence score variance. Accordingly, the experimental settings were $\alpha_{\text{IoU}} = 0$, $k = 3$ ($\alpha_{\text{FN}} = 1.5$, $\alpha_{\text{FP}} = 0.5$), and $\beta = 0.10$. An example is shown in Figure 4b.

#### 4.1.3. Pothole Dataset

The open-source pothole dataset (https://public.roboflow.com/object-detection/pothole, accessed on 1 July 2025) contains a single class, indicating the presence of road potholes. In this application, identifying a pothole's presence is far more critical than precisely locating its boundaries. To reflect this priority and apply no penalty beyond the initial IoU threshold, we deliberately set $\alpha_{\text{IoU}} = 0$. High maintenance costs from FPs warranted a heavy penalty on $\alpha_{\text{FP}}$, while the variable road and lighting conditions in the dataset necessitated a typical robustness setting. Thus, the experimental settings were $\alpha_{\text{IoU}} = 0$, $k = 1/3$ ($\alpha_{\text{FN}} = 0.5$, $\alpha_{\text{FP}} = 1.5$), and $\beta = 0.05$. An example is shown in Figure 4c.

#### 4.1.4. Semiconductor Manufacturing Equipment Sensor Dataset

This dataset was derived from the format used in actual semiconductor fabrication for quality control, where time-series data were converted into chart images for monitoring. Despite recent technologies in chart-anomaly detection, problem regions and classes are still determined manually. Therefore, employing object detection could enable full automation. The dataset consists of three classes: average, deviation, and drift. Similar to Figure 1a, most spatial information is captured at an IoU of 0.50. Therefore, to align the metric with the domain need where object detection is sufficient, we set $\alpha_{\text{IoU}} = 0$. Furthermore, because FNs can critically affect the quality of semiconductor products, a higher cost was assigned to FNs. Due to the highly controlled data acquisition environment, which resulted in a low variance in the signal characteristics, a narrow robustness band of $\beta = 0.03$ was deemed sufficient. Accordingly, the experimental settings were $\alpha_{\text{IoU}} = 0$, $k = 2$ ($\alpha_{\text{FN}} = 4/3$, $\alpha_{\text{FP}} = 2/3$), and $\beta = 0.03$. An example is shown in Figure 5.



(**a**) (**b**) (**c**)

**Figure 5.** Examples of each class from the semiconductor manufacturing equipment sensor dataset. (**a**) Average: mean value falls outside the expected range. (**b**) Deviation: individual points deviate significantly from the expected range. (**c**) Drift: distribution or trend gradually shifts over time.

### *4.2. Confidence Score Calibration Methods*

As pointed out by Guo et al. [51], modern deep learning models tend to be overconfident. To overcome this, various calibration approaches have been explored. Strictly monotonic calibration mappings (e.g., logistic or beta) leave a model's mAP and detection ranking unchanged. In contrast, piecewise constant mappings such as histogram binning can introduce ties or local rank inversions, which may slightly alter the mAP. Additionally, some train-time calibration methods (e.g., MC dropout calibration) can modify the ranking during training. Nevertheless, calibration is widely used at deployment to improve the decision confidence. Therefore, we empirically investigated whether DSR-LRP can effectively discriminate among these calibration methods in terms of their resulting deployment stability and cost-aware performance.

#### 4.2.1. Post Hoc Calibration

We implemented various calibration methods proposed by Küppers et al. [46]. In their framework, the input to the calibration function is defined as $s = (\hat{p}, \hat{r})$, where $\hat{p}$ is the confidence score, and $\hat{r}$ encodes the bounding box information. For notational simplicity, we will hereafter use $p$ to denote the confidence score $\hat{p}$, which serves as the input to our calibration functions. Our analysis deliberately limits calibration to methods that modify only the scalar confidence score $p$, leaving all bounding box coordinates unchanged. This focus aligns with our primary goal: demonstrating how DSR-LRP can be used to select a robust operating threshold on the confidence score axis. Therefore, varying the confidence landscape through different score calibrations is the most direct and sufficient way to validate the metric's core behavior regarding threshold stability. Nevertheless, DSR-LRP still accounts for localization quality via the $\alpha_{\text{IoU}}$ term, because the detector's spatial accuracy is fixed prior to evaluation. Consequently, any difference between the LRP ($\beta = 0$) and DSR-LRP ($\beta > 0$) can be unambiguously attributed to the robustness component $\beta$. For clarity in our analysis, we treated logistic and beta calibration as strictly monotonic score mappings, whereas histogram binning is a piecewise constant and thus may reorder detections.

1. Histogram binning: This method divides the confidence scores into several bins and uses the average accuracy within each bin as the new calibrated confidence. The calibration function is:

$$f_{\text{HB}}(p) = \frac{1}{|B_i|} \sum_{j \in B_i} m_j,$$

where $B_i$ is the set of predictions whose confidence score falls into the $i$th interval, and $m_j$ is the outcome of the $j$th prediction (1 for a correct detection, 0 otherwise).

2. Logistic calibration: This method applies a linear logit model for calibration, defined as $z(p) = a + b\,p$, within the standard sigmoid calibration map $g(p) = 1/(1 + \exp(-z(p)))$. Therefore, the final correction map is given by:

$$f_{\mathrm{LC}}(p) \;=\; \frac{1}{1 + \exp(-(a + b\,p))}.$$

3. Beta calibration: This method employs a more flexible non-linear model for the log-odds term $z(p) = a\,\ln(p) - b\,\ln(1-p) + c$, which is then transformed by the same sigmoid map $g(p)$. The final correction map is given by:

$$f_{\mathrm{BC}}(p) \;=\; \frac{1}{1 + \exp\big(-(a\,\ln(p) - b\,\ln(1-p) + c)\big)}.$$

#### 4.2.2. Auxiliary Loss Calibration with Monte Carlo (MC) Dropout

Following the approach by Pathiraja et al. [53], we incorporated an auxiliary calibration loss during model training. This method leverages Monte Carlo (MC) dropout to generate multiple predictions for a single input, using the variance across these predictions as a measure of model uncertainty. For parity with the post hoc methods, which calibrate only scores, we nullified the localization calibration term by setting $\lambda_{\mathrm{aux}} = 0$ in the original loss function ($\mathcal{L}_{\mathrm{MCCL-aux}} = \mathcal{L}_{\mathrm{MCC}} + \lambda_{\mathrm{aux}}\,\mathcal{L}_{\mathrm{LC}}$). This effectively discards the localization calibration term ($\mathcal{L}_{\mathrm{LC}}$), ensuring that only the confidence calibration component ($\mathcal{L}_{\mathrm{MCC}}$) is evaluated. This loss is defined as:

$$\mathcal{L}_{\mathrm{MCC}} = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{1}{M}\sum_{l=1}^{N_b}\sum_{n=1}^{N_{\mathrm{pos}}} v_{l,n}[k] \;-\; \frac{1}{M}\sum_{l=1}^{N_b}\sum_{n=1}^{N_{\mathrm{pos}}} q_{l,n}[k]\right|.$$

Here, $K$ is the number of classes, $N_b$ is the batch size, $N_{\mathrm{pos}}$ is the number of positive detections per image, and $M = N_b N_{\mathrm{pos}}$. The loss is computed for each class $k$ at every $n$th positive detection within the $l$th sample of a mini-batch. The term $q_{l,n}[k]$ represents the one-hot encoded ground truth, which is compared against $v_{l,n}[k]$, the model's adjusted confidence. The adjusted confidence is $v_{l,n}[k] = (\bar{s}_{l,n}[k] + c_{l,n}[k])\ /\ 2$, where $\bar{s}_{l,n}[k]$ is the mean softmax probability over $N$ forward passes, and the certainty $c_{l,n}[k] = 1 - \tanh(d_{l,n}[k])$, which decreases as uncertainty grows. The uncertainty term $d_{l,n}[k]$ is the variance of the corresponding logits over those passes. For our implementation, we set the number of MC samples to $N = 5$ and adopted all other hyperparameters from the configuration recommended in the source paper.

The evaluated calibration methods were applied to each model trained with our 50-epoch protocol. The methods were applied in two ways depending on their category. For the post hoc approaches (histogram, logistic, and beta), calibration parameters were fitted exclusively on the validation set. For the train-time approach (MC dropout calibration), we observed that a two-stage procedure (first training without the auxiliary term, then fine-tuning with it after convergence) consistently outperformed training from scratch; therefore, we adopted a two-stage procedure. All final performance metrics were subsequently reported on the held-out test set.

### *4.3. Experimental Results and Analysis*

To provide a comprehensive analysis, our result tables report key metrics, each selected for a distinct purpose. We used the mAP and mAP50 as a general performance baseline. To specifically quantify the effectiveness of the applied confidence calibration methods, we included D-ECE [46]. Finally, to evaluate our core hypothesis, we compared DSR-LRP directly against its pointwise counterpart, the LRP (defined with $\beta = 0$, but using the

same cost parameters $k$ and $\alpha_{\text{IoU}}$). This allows for a direct assessment of the stability component's impact.

#### 4.3.1. Blood Cell Dataset

* Experimental setting: $\alpha_{\text{IoU}} = 1$, $k = 1$ ($\alpha_{\text{FN}} = 1$, $\alpha_{\text{FP}} = 1$), and $\beta = 0.05$.

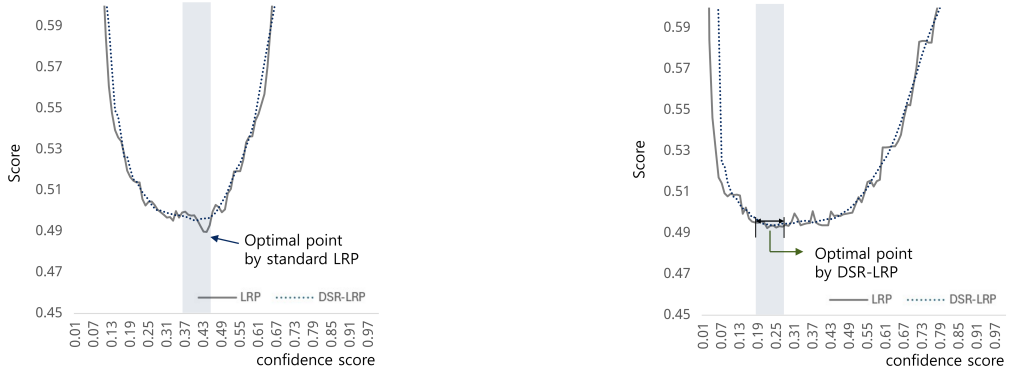
Consistent with the discussion in Section 4.2, the piecewise histogram binning and the stochastic MC dropout calibration slightly perturb the mAP, whereas logistic and beta mappings leave it unchanged. Yet the mAP is a global, rank-based metric; in deployment, we care about performance stability around a single operating threshold.

Table 3 shows the results for each metric. In this baseline scenario, where $k = 1$ and $\alpha_{\text{IoU}} = 1$, the LRP ($\beta = 0$) score is equivalent to the standard LRP formulation. While MC dropout calibration achieved the nominally best (lowest) score for the LRP ($\beta = 0$), logistic calibration minimized DSR-LRP. As visualized in the recommendation package's evidence figure, the performance curve for MC dropout reached its minimum at a threshold of 0.44, but degraded sharply around this point (Figure 6a). In contrast, Figure 6b shows that the curve for logistic calibration is nearly flat around its optimum, yielding the lowest DSR-LRP score. This outcome demonstrates that the DSR-LRP criterion prioritizes stability by identifying a reliable operating threshold over a nominally superior, but volatile, one.

**Table 3.** Evaluation metrics for each calibration method on the blood cell dataset.

| Calibration | mAP | mAP50 | D-ECE (%) | LRP ($\beta = 0$) | DSR-LRP |
|---|---|---|---|---|---|
| Original | **0.577** | **0.870** | 3.49 | 0.491 (0.40) | 0.494 (0.38) |
| Histogram binning | 0.546 | 0.837 | **2.37** | 0.510 (0.31) | 0.513 (0.34–0.35) |
| Logistic | **0.577** | **0.870** | 3.64 | 0.493 (0.22) | **0.493 (0.24)** |
| Beta | **0.577** | **0.870** | 2.38 | 0.493 (0.43) | 0.494 (0.40–0.41) |
| MC dropout | 0.574 | 0.865 | 6.85 | **0.490 (0.44)** | 0.495 (0.40–0.41) |

**Note.** LRP/DSR-LRP are lower-is-better; values in parentheses are the confidence thresholds at which the minima were attained. D-ECE is reported in percentage. Bold indicates the best value per column (highest mAP/mAP50; lowest D-ECE/LRP/DSR-LRP).




(**a**) (**b**)

**Figure 6.** LRP vs. DSR-LRP across confidence thresholds on the blood cell dataset. (**a**) MC dropout calibration. (**b**) Logistic calibration.

#### 4.3.2. Wildfire Smoke Dataset

* Experimental setting: $\alpha_{\text{IoU}} = 0$, $k = 3$ ($\alpha_{\text{FN}} = 1.5$, $\alpha_{\text{FP}} = 0.5$), and $\beta = 0.10$.

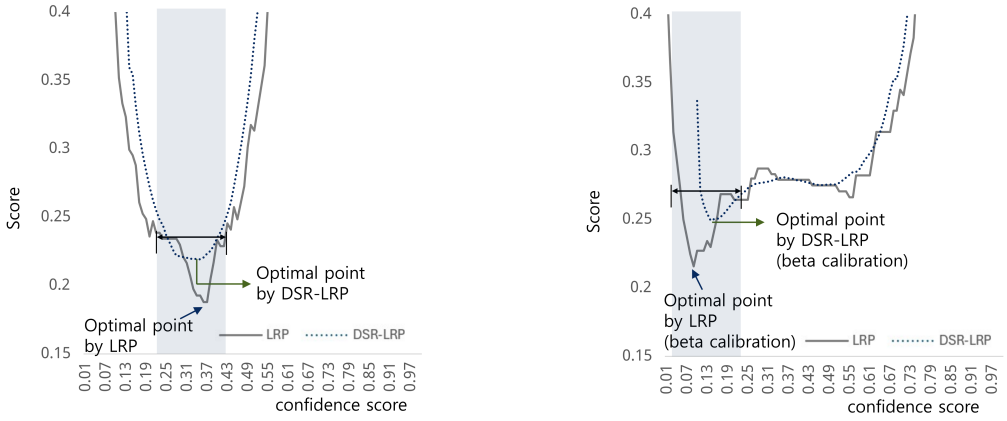
Table 4 shows that MC dropout calibration outperformed all other methods under both the LRP ($\beta = 0$) and DSR-LRP, with both metrics attaining their minima at similar thresholds (Figure 7a). To make the behavior of our metric explicit, the beta calibration

case is particularly illustrative. In the recommendation package, the rationale reports that the LRP ($\beta = 0$) attained its optimum at a low threshold of $s^*_{\mathrm{LRP}} = 0.09$, whereas DSR-LRP selected $s^* = 0.14$ under the prescribed robustness band $\beta = 0.10$, yielding a shift of $\Delta s = 0.05$. As confirmed by the visual evidence (Figure 7b), this shift demonstrates that DSR-LRP avoids steep regions of the curve that would make the operating threshold sensitive to small score shifts. This example shows how the package synthesizes the operating threshold, its rationale (the $\Delta s$ shift), and visual evidence to recommend a stable setting.

**Table 4.** Evaluation metrics for each calibration method on the wildfire smoke dataset.

| Calibration | mAP | mAP50 | D-ECE (%) | LRP ($\beta = 0$) | DSR-LRP |
|---|---|---|---|---|---|
| Original | 0.498 | 0.875 | 0.97 | 0.216 (0.15) | 0.260 (0.21) |
| Histogram binning | 0.400 | 0.761 | **0.25** | 0.337 (0.15–0.18) | 0.352 (0.14) |
| Logistic | 0.498 | 0.875 | 0.48 | 0.227 (0.02) | 0.273 (0.11) |
| Beta | 0.498 | 0.875 | 0.32 | 0.216 (0.09) | 0.249 (0.14) |
| MC dropout | **0.533** | **0.913** | 2.14 | **0.188 (0.36–0.37)** | **0.219 (0.34–0.35)** |

**Note.** LRP/DSR-LRP are lower-is-better; values in parentheses are the confidence thresholds at which the minima were attained. D-ECE is reported in percentage. Bold indicates the best value per column (highest mAP/mAP50; lowest D-ECE/LRP/DSR-LRP).




(**a**) (**b**)

**Figure 7.** LRP vs. DSR-LRP across confidence thresholds on the wildfire smoke dataset. (**a**) MC dropout calibration. (**b**) Beta calibration.

#### 4.3.3. Pothole Dataset

* Experimental setting: $\alpha_{\mathrm{IoU}} = 0$, $k = 1/3$ ($\alpha_{\mathrm{FN}} = 0.5$, $\alpha_{\mathrm{FP}} = 1.5$), and $\beta = 0.05$.

As presented in Table 5, MC dropout calibration again attained the best scores under both metrics; its visual evidence is shown in Figure 8a. To surface DSR-LRP's cost-sensitive behavior, the beta calibration case is particularly instructive. In the recommendation package, the rationale reports that the pointwise, cost-sensitive LRP ($\beta = 0$) attained its optimum at $s^*_{\mathrm{LRP}} = 0.58$, whereas DSR-LRP selected a more conservative operating threshold at $s^* = 0.62$ under the specified cost asymmetry (heavier penalty on FPs), yielding a rightward shift of $\Delta s = 0.04$. As confirmed by the visual evidence in Figure 8b, the curve was steeper near 0.58 than around 0.62; thus, DSR-LRP requires a higher confidence score to reduce the FP risk while securing local stability within its robustness band.

**Table 5.** Evaluation metrics for each calibration method on the pothole dataset.

| Calibration | mAP | mAP50 | D-ECE (%) | LRP ($\beta = 0$) | DSR-LRP |
|---|---|---|---|---|---|
| Original | **0.502** | **0.770** | 1.46 | 0.338 (0.38) | 0.348 (0.52–0.53) |
| Histogram binning | 0.419 | 0.681 | **0.47** | 0.410 (0.41–0.42) | 0.412 (0.55) |
| Logistic | **0.502** | **0.770** | 1.07 | 0.338 (0.46) | 0.344 (0.50) |
| Beta | **0.502** | **0.770** | 0.66 | 0.338 (0.58) | 0.347 (0.62) |
| MC dropout | 0.486 | 0.757 | 3.93 | **0.320 (0.42)** | **0.336 (0.44)** |

**Note.** LRP/DSR-LRP are lower-is-better; values in parentheses are the confidence thresholds at which the minima were attained. D-ECE is reported in percentage. Bold indicates the best value per column (highest mAP/mAP50; lowest D-ECE/LRP/DSR-LRP).




(**a**) (**b**)

**Figure 8.** LRP vs. DSR-LRP across confidence thresholds on the pothole dataset. (**a**) MC dropout calibration. (**b**) Beta calibration.

Figures 9 and 10 quantify the practical benefit of this shift by simulating the "threshold brittleness" phenomenon. Figure 9 illustrates the risk of relying on a pointwise optimum: a minor deviation of just 0.05 (simulating a score-level distribution shift) from the LRP-optimal threshold causes the error score to spike, resulting in a substantial 15.6% performance degradation. In contrast, Figure 10 shows that the operating threshold recommended by DSR-LRP is much less sensitive to the same perturbation. By selecting an operating point within a locally stable region, the same shift results in only a 1.8% fluctuation. This contrast suggests that DSR-LRP effectively manages uncertainty, safeguarding the system against the brittleness inherent in standard optimization.

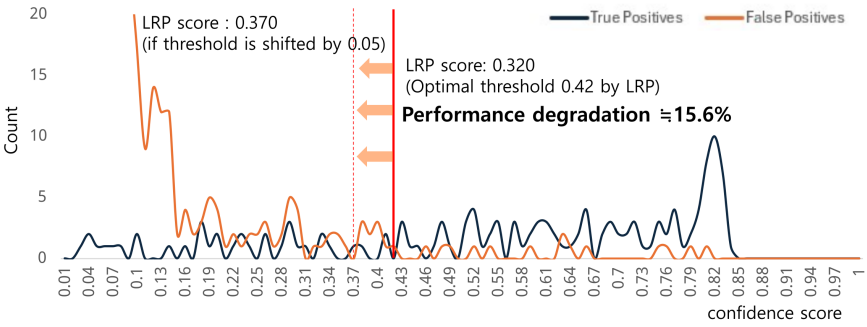



**Figure 9.** Fragility of a single-point optimum: a 0.05 shift from the LRP-optimal threshold (0.42) increases the (lower-is-better) LRP error by 15.6% on the pothole dataset.

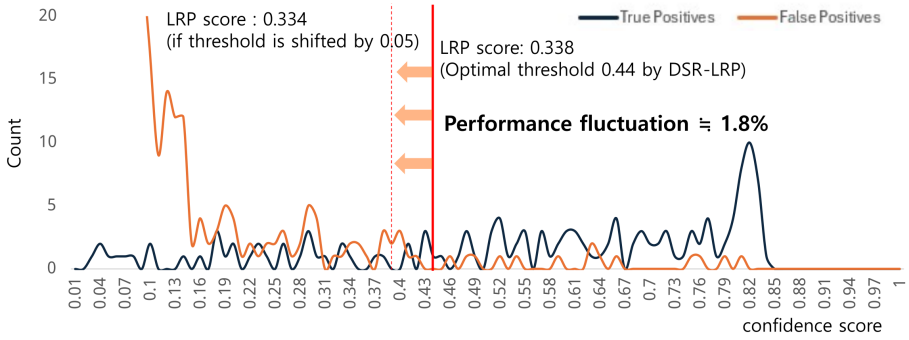



**Figure 10.** Improved robustness under DSR-LRP: under the same 0.05 shift, the recommended operating threshold (0.44) limits performance fluctuation to only 1.8%.

#### 4.3.4. Semiconductor Manufacturing Equipment Sensor Dataset

* Experimental setting: $\alpha_{\text{IoU}} = 0$, $k = 2$ ($\alpha_{\text{FN}} = 4/3$, $\alpha_{\text{FP}} = 2/3$), and $\beta = 0.03$.

As shown in Table 6, the uncalibrated (original) model attained the lowest scores for both the LRP ($\beta = 0$) and DSR-LRP. In the recommendation package, the rationale reports that the DSR-LRP optimum coincides with the pointwise LRP optimum ($s^* = s^*_{\text{LRP}}$; $\Delta s = 0$), which is consistent with the visual evidence in Figure 11a: the LRP curve is relatively flat within the narrow robustness band ($\beta = 0.03$). Because the pointwise LRP varies only minimally across the $\pm\beta$ neighborhood, the DSR-LRP objective naturally achieves its minimum at the same threshold. This behavior is not an isolated case, as the MC dropout calibration exhibits a similar convergence of its optimal thresholds (Figure 11b). Consequently, these observations demonstrate a key property: when the underlying performance is already robust, DSR-LRP does not unnecessarily shift the operating threshold and converges to the LRP ($\beta = 0$) optimum. This behavior is empirically consistent with the theoretical property that DSR-LRP converges to the LRP as $\beta \to 0$.




(**a**) (**b**)

**Figure 11.** LRP vs. DSR-LRP across confidence thresholds on the semiconductor manufacturing equipment sensor dataset. (**a**) Original (uncalibrated). (**b**) MC dropout calibration.

#### 4.3.5. Cross-Scenario Interpretation of Parameter Roles

Although this study does not provide a full factorial ablation over $k$, $\alpha_{\text{IoU}}$, and $\beta$ within a single dataset, the four experimental scenarios were intentionally designed to represent distinct deployment requirements. The blood cell scenario ($k = 1$, $\alpha_{\text{IoU}} = 1$, $\beta = 0.05$) serves as a symmetric baseline in which the pointwise LRP component is equivalent to the standard LRP, thereby isolating the additional effect of the robustness band. The wildfire smoke scenario ($k = 3$, $\alpha_{\text{IoU}} = 0$, $\beta = 0.10$) represents an FN-critical setting with a wider robustness band; in the beta-calibrated case, DSR-LRP shifts the recommended threshold from the pointwise LRP optimum of 0.09 to 0.14, thereby avoiding a steep region of the curve. The pothole scenario ($k = 1/3$, $\alpha_{\text{IoU}} = 0$, $\beta = 0.05$) represents an FP-critical setting;

in the beta-calibrated case, DSR-LRP recommends a higher threshold than the pointwise LRP optimum, which is consistent with the need to reduce false alarms while maintaining local stability. The semiconductor sensor scenario ($k = 2$, $\alpha_{\text{IoU}} = 0$, $\beta = 0.03$) represents a controlled sensing environment with a narrower robustness band; when the LRP curve is locally flat, the DSR-LRP optimum coincides with the pointwise LRP optimum.

Taken together, these case studies illustrate how the exposed parameters encode different practitioner priorities and produce interpretable threshold recommendations across representative deployment settings. Nevertheless, a full within-dataset factorial ablation over $k$, $\alpha_{\text{IoU}}$, and $\beta$, as well as formal multi-seed statistical testing, remains an important direction for future work.

**Table 6.** Evaluation metrics for each calibration method on the semiconductor manufacturing equipment sensor dataset.

| Calibration | mAP | mAP50 | D-ECE (%) | LRP ($\beta = 0$) | DSR-LRP |
|---|---|---|---|---|---|
| Original | **0.347** | 0.609 | 0.73 | **0.497 (0.23–0.24)** | **0.501 (0.24)** |
| Histogram binning | 0.300 | 0.560 | **0.37** | 0.536 (0.31–0.33) | 0.541 (0.31) |
| Logistic | **0.347** | 0.609 | 0.54 | 0.529 (0.08) | 0.531 (0.05) |
| Beta | **0.347** | 0.609 | **0.37** | 0.522 (0.20) | 0.529 (0.23) |
| MC dropout | 0.345 | **0.630** | 2.00 | 0.505 (0.31) | 0.510 (0.31) |

**Note.** LRP/DSR-LRP are lower-is-better; values in parentheses are the confidence thresholds at which the minima were attained. D-ECE is reported in percentage. Bold indicates the best value per column (highest mAP/mAP50; lowest D-ECE/LRP/DSR-LRP).

## 5. Conclusions

### *5.1. Summary of Contributions*

The practical deployment of object detectors requires an evaluation framework that transcends the global summaries provided by academic benchmarks such as the mAP. Throughout this paper, we focused on three deployment-oriented requirements often neglected by standard metrics: (1) evaluation at a specific operating threshold, (2) handling asymmetric and imprecise error costs, and (3) enhancing robustness against local distribution shifts.

To address these challenges, we proposed DSR-LRP, a systematic three-phase decision support framework designed to manage cost–stability trade-offs. This framework begins with Phase 1: translating soft domain preferences—such as approximate cost ratios and tolerable imprecision—into quantitative parameters ($k, \alpha_{\text{IoU}}, \beta$). Then, it proceeds to Phase 2, a robust, cost-sensitive evaluation that aggregates performance within a robustness band, and concludes with Phase 3: delivering an interpretable recommendation package. This package provides the optimal threshold, a quantified rationale, and visual evidence, thereby supporting transparent, evidence-based decision-making.

Our experiments on four diverse industrial datasets suggest that DSR-LRP is consistent with soft-computing principles for managing uncertainty and trade-offs in a deployment-oriented evaluation:

- **Mitigating threshold brittleness.** By evaluating the performance within a robustness band rather than at a single point, DSR-LRP tends to steer the operating point away from unstable regions prone to sharp degradation. This capability was highlighted in our pothole case study, where the framework recommended a more robust operating point that reduced the sensitivity to simulated shifts from 15.6% (standard optimum) to 1.8%.

- **Operationalizing soft domain preferences.** Through its configurable parameters ($k$ and $\alpha_{\text{IoU}}$), DSR-LRP translates high-level, often qualitative domain priorities into the final deployment decision (as illustrated by the contrasting cost structures in the wildfire and pothole datasets).
- **Adaptive consistency.** In our experiments, when the model was already stable, DSR-LRP tended to avoid unnecessary adjustments and converge to the pointwise optimum, acting as a reliable guardrail that intervenes mainly when uncertainty poses a risk (as demonstrated in the semiconductor dataset).

In conclusion, DSR-LRP is more than a simple evaluation metric; it is a soft-computing-oriented tool that helps bridge the gap between rigid academic benchmarks and the flexible, robust needs of industry. By explicitly managing uncertainty and maximizing transparency, it empowers practitioners to make reliable, evidence-driven deployment decisions.

### *5.2. Limitations and Future Work*

While our decision support framework begins to address important gaps in post hoc evaluation, we highlight several limitations and future directions to further advance its impact within the soft-computing domain.

- **Computational overhead.** DSR-LRP introduces additional computation because the LRP must be evaluated at multiple points within the robustness band. However, this evaluation is performed post hoc on stored detection outputs and does not require rerunning detector inference. Therefore, the practical overhead is modest; for example, with $\beta = 0.05$ and $\Delta = 0.01$, only 11 band points are evaluated for each candidate threshold. The cost increases linearly with a wider $\beta$ or finer $\Delta$, but remains a one-time offline cost during evaluation.
- **From post hoc evaluation to in-training optimization.** Our framework currently applies its core logic (Phases 2 and 3) only after model training. A natural next step is to formulate a differentiable loss function inspired by DSR-LRP's principles. This would enable models to internalize stability and cost-awareness directly during the learning process, evolving from systems that are merely evaluated for robustness to those that are trained for it.
- **From expert tuning to systematic preference elicitation.** Phase 1 currently relies on practitioner knowledge to set parameters ($k$, $\alpha_{\text{IoU}}$, and $\beta$), which may introduce subjectivity. Future work could develop systematic methodologies for this process, for instance, by adapting multi-criteria decision-making (MCDM) frameworks [54] to automatically elicit and optimize these soft preferences from data or interactive user feedback. Although this paper provides default settings in Section 3.1 and a cross-scenario interpretation of parameter roles in Section 4.3.5, the recommended threshold can still depend on these practitioner-defined choices. A controlled within-dataset factorial sweep over $k$, $\alpha_{\text{IoU}}$, and $\beta$, together with multi-seed statistical significance testing, is therefore left as future work. More broadly, future research could also adapt large-scale group decision-making approaches, such as Trillo et al. [55], when multiple stakeholders are involved in defining deployment preferences.
- **Integration with candidate-level deployment audits.** A natural extension is to integrate DSR-LRP with candidate-level deployment audit frameworks such as CSEF [50]. In such a two-tier pipeline, DSR-LRP could first identify a cost-aware and locally robust operating threshold for each candidate, after which CSEF-style auditing could evaluate whether the selected candidate satisfies broader deployment-level stability and operational constraints. This integration would preserve the distinction between threshold-level operating-point selection and candidate-level deployment auditing, while allowing both perspectives to support a unified deployment decision process.

- **Generalization to foundation and multimodal detectors.** DSR-LRP is architecture-agnostic in principle because it only requires predicted boxes, class labels, confidence scores, and ground-truth annotations. However, open-vocabulary, multimodal, or foundation-model-based detectors may introduce additional uncertainty sources, such as prompt sensitivity, semantic ambiguity among classes, and class-dependent confidence calibration. Extending DSR-LRP to these emerging architectures is a promising future direction.

Pursuing these directions could further help bridge the gap between theoretical robust optimization and deployment-ready industrial perception systems.

**Author Contributions:** K.L.: Conceptualization, Methodology, Software, Writing—original draft preparation; J.H.: Conceptualization, Project administration, Writing—review and editing; Y.S.: Investigation, Validation; B.-S.K.: Investigation, Visualization; D.-H.L.: Supervision, Writing—review and editing, Funding acquisition. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The public datasets analyzed in this study are available from Roboflow Universe: the blood cell dataset at https://public.roboflow.com/object-detection/bccd, accessed on 1 April 2026, the wildfire smoke dataset at https://public.roboflow.com/object-detection/wildfire-smoke, accessed on 1 April 2026, and the pothole dataset at https://public.roboflow.com/object-detection/pothole, accessed on 1 April 2026. The semiconductor manufacturing equipment sensor dataset is not publicly available due to confidentiality and proprietary restrictions.

**Conflicts of Interest:** Author Jihoon Hong was employed by the company Memory Division, Samsung Electronics, Seoul, South Korea. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Abbreviations

List of notations and their descriptions used in the DSR-LRP framework.

| Symbol | Description |
|---|---|
| $s$ | A candidate confidence threshold. |
| $s^*$ | The final optimal operating threshold recommended by the DSR-LRP framework. |
| $s^*_{\text{LRP}}$ | The threshold that minimizes the pointwise LRP score without robustness consideration. |
| $k$ | The classification cost ratio ($Cost_{\text{FN}}/Cost_{\text{FP}}$). A value greater than 1 implies a higher cost for FNs, and less than 1 for FPs. |
| $\alpha_{\text{FP}}, \alpha_{\text{FN}}$ | Cost multipliers for FPs and FNs, derived from the cost ratio $k$. |
| $\alpha_{\text{IoU}}$ | The localization importance, a value between 0 and 1 that controls the tolerance for imprecise bounding boxes. |
| $\beta$ | The robustness band, defining the interval width for evaluating the performance stability against potential distribution shifts. |
| $\Delta$ | The step size for sweeping thresholds (e.g., 0.01). |
| $N$ | The number of discrete steps within the robustness band, calculated as $N = \lfloor 2\beta/\Delta \rfloor$. |
| $LRP(s)$ | The cost-sensitive localization recall precision (LRP) error score at a single threshold $s$. |
| $DSR\text{-}LRP(s, \beta)$ | The final score for a candidate threshold $s$ with a robustness band $\beta$, considering both the average performance and variability within the band. |
| $N_{\text{TP}}$ | The number of true positives. |
| $N_{\text{FP}}$ | The number of false positives. |

| | |
|---|---|
| $N_{FN}$ | The number of false negatives. |
| $X$ | The set of ground-truth bounding boxes. |
| $Y_s$ | The set of detections whose confidence scores are higher than the threshold $s$. |
| $Z$ | The normalization constant to keep the LRP score within the range [0, 1]. |
| $\tau$ | The minimum intersection over the union (IoU) threshold for a detection to be considered a TP. |

## Appendix A. Algorithm for DSR-LRP Framework

**Algorithm A1** DSR-LRP framework (part 1 of 2).

**Phase 1: Knowledge elicitation and parameterization**
Based on practitioners' soft domain preferences, define the core domain-specific parameters:
- $k$: Classification cost ratio ($k = \text{cost}_{FN}/\text{cost}_{FP}$)
- $\alpha_{IoU}$: Localization importance ($\alpha_{IoU} \in [0, 1]$)
- $\beta$: Robustness band (expected magnitude of confidence score shifts)

**Input:**
- $D_{\text{results}}$: All detection results from the candidate
- $D_{gt}$: All ground-truth data
- $\tau$: Minimum IoU threshold for a TP

**Preparation:**
Initialize empty dictionaries: `LRP_scores` $\leftarrow \{\}$, `DSR_LRP_scores` $\leftarrow \{\}$
Calculate cost multipliers: $\alpha_{FP} \leftarrow 2/(1+k)$; $\alpha_{FN} \leftarrow 2k/(1+k)$
**Return:** Parameters and initialized dictionaries.
**Phase 2: Robust and cost-sensitive evaluation**
**Preliminary: Result Classification**
For each confidence threshold $s$, classify detection results in $D_{\text{results}}$ against $D_{gt}$ using the IoU threshold $\tau$ to obtain counts of $N_{TP}, N_{FP}, N_{FN}$.
**2.1. Calculate cost-sensitive LRP**
For each point $s$, calculate the pointwise LRP score using the formula:

$$LRP(s_i) = \frac{1}{Z}(\alpha_{IoU} \sum_{i=1}^{N_{TP}} \frac{1-IoU}{1-\tau} + \alpha_{FP} N_{FP} + \alpha_{FN} N_{FN})$$

**Return:** A dictionary mapping each threshold to its LRP score, `LRP_scores`.
**2.2. Calculate DSR-LRP**
**Input:**
- $\Delta$: Step size for sweeping thresholds within the robustness band (e.g., 0.01).

For a given candidate threshold $s$, evaluate stability over the band $[s-\beta, s+\beta]$ using the formula:

$$DSR\text{-}LRP(s, \beta) = \sqrt{\frac{1}{N+1} \sum_{i=0}^{N} (LRP(s_i))^2}$$

where $N = \lfloor 2\beta/\Delta \rfloor$ is the number of discrete steps within the band.
**Return:** A dictionary mapping each threshold to its DSR-LRP score, `DSR_LRP_scores`.

**Algorithm A2** DSR-LRP framework (part 2 of 2).

**Input:**
- `LRP_scores`: Dictionary of LRP scores from Part 1
- `DSR_LRP_scores`: Dictionary of DSR-LRP scores from Part 1
- Original parameters $(k, \alpha_{IoU}, \beta)$

**Phase 3: Analysis and interpretable recommendation**
Identify optimal thresholds by finding the minimum values in the returned dictionaries:
$s^* \leftarrow \arg\min_s$ `DSR_LRP_scores`
$s^*_{LRP} \leftarrow \arg\min_s$ `LRP_scores`
Compile the final recommendation package, $P_{rec}$, with the following components:
**- Optimal point:** The recommended pair ($s^*$, `DSR_LRP_scores`$[s^*]$).
**- Rationale:** A quantitative summary of the cost-stability trade-off, including the shift from $s^*_{LRP}$ to $s^*$ under the given parameters.
**- Visual evidence:** A figure plotting the computed `LRP_scores` and `DSR_LRP_scores` curves to visualize the performance landscape.
**Return:** The complete recommendation package, $P_{rec}$.

https://doi.org/10.3390/a19050409

## References


1. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object Detection in 20 Years: A Survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
2. Sun, Y.; Sun, Z.; Chen, W. The evolution of object detection methods. *Eng. Appl. Artif. Intell.* **2024**, *133*, 108458. [CrossRef]
3. Li, Z.; Dong, Y.; Shen, L.; Liu, Y.; Pei, Y.; Yang, H.; Zheng, L.; Ma, J. Development and challenges of object detection: A survey. *Neurocomputing* **2024**, *598*, 128102. [CrossRef]
4. Liu, L.; Ouyang, W.; Wang, X.; Fieguth, P.; Chen, J.; Liu, X.; Pietikäinen, M. Deep Learning for Generic Object Detection: A Survey. *Int. J. Comput. Vis.* **2020**, *128*, 261–318. [CrossRef]
5. Bhatt, P.; Malhan, R.; Rajendran, P.; Shah, B.; Thakar, S.; Yoon, Y.; Gupta, S. Object Detection in 20 Years: A Survey. *J. Comput. Inf. Sci. Eng.* **2021**, *21*, 040801. [CrossRef]
6. Yang, J.; Liu, Z. A novel real-time steel surface defect detection method with enhanced feature extraction and adaptive fusion. *Eng. Appl. Artif. Intell.* **2024**, *138*, 109289. [CrossRef]
7. Chen, S.; Jiang, S.; Wang, X.; Sun, P.; Hua, C.; Sun, J. An efficient detector for detecting surface defects on cold-rolled steel strips. *Eng. Appl. Artif. Intell.* **2024**, *138*, 109325. [CrossRef]
8. Sun, P.; Hua, C.; Ding, W.; Hua, C.; Liu, P.; Lei, Z. Ceramic tableware surface defect detection based on deep learning. *Eng. Appl. Artif. Intell.* **2025**, *141*, 109723. [CrossRef]
9. Grigorescu, S.; Trasnea, B.; Cocias, T.; Macesanu, G. A survey of deep learning techniques for autonomous driving. *J. Field Robot.* **2020**, *37*, 362–386. [CrossRef]
10. Muhammad, K.; Ullah, A.; Lloret, J.; Ser, J.D.; de Albuquerque, V.H.C. Deep Learning for Safe Autonomous Driving: Current Challenges and Future Directions. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 4316–4336. [CrossRef]
11. Kuutti, S.; Bowden, R.; Jin, Y.; Barber, P.; Fallah, S. A Survey of Deep Learning Applications to Autonomous Vehicle Control. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 712–733. [CrossRef]
12. Thottempudi, P.; Jambek, A.B.B.; Kumar, V.; Acharya, B.; Moreira, F. Resilient object detection for autonomous vehicles: Integrating deep learning and sensor fusion in adverse conditions. *Eng. Appl. Artif. Intell.* **2025**, *151*, 110563. [CrossRef]
13. Li, X.; Lin, K.; Meng, M.; Li, X.; Li, L.; Hong, Y.; Chen, J. A Survey of ADAS Perceptions with Development in China. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 14188–14203. [CrossRef]
14. Litjens, G.; Kooi, T.; Bejnordi, B.; Setio, A.; Ciompi, F.; Ghafoorian, M.; Laak, J.; Ginneken, B.; Sánchez, C. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef]
15. Suganyadevi, S.; Seethalakshmi, V.; Balasamy, K. A review on deep learning in medical image analysis. *Int. J. Multimed. Inf. Retr.* **2022**, *11*, 19–38. [CrossRef] [PubMed]
16. Razzak, M.; Naz, S.; Zaib, A. Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In *Classification in BioApps*; Springer International Publishing: Cham, Switzerland, 2018; Volume 26, pp. 323–350.
17. Liu, X.; Gao, K.; Liu, B.; Pan, C.; Liang, K.; Yan, L.; Ma, J.; He, F.; Zhang, S.; Pan, S.; et al. Advances in Deep Learning-Based Medical Image Analysis. *Health Data Sci.* **2021**, *2021*, 8786793. [CrossRef]
18. Chen, X.; Wang, X.; Zhang, K.; Fung, K.; Thai, T.; Moore, K.; Mannel, R.; Liu, H.; Zheng, B.; Qiu, Y. Recent advances and clinical applications of deep learning in medical image analysis. *Med. Image Anal.* **2022**, *79*, 102444. [CrossRef] [PubMed]
19. Cano-Ortiz, S.; Iglesias, L.; Árbol, P.; Castro-Fresno, D. Improving detection of asphalt distresses with deep learning-based diffusion model for intelligent road maintenance. *Dev. Built Environ.* **2024**, *17*, 100315. [CrossRef]
20. Dhiman, A.; Klette, R. Pothole Detection Using Computer Vision and Learning. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 3536–3550. [CrossRef]
21. Ma, N.; Fan, J.; Wang, W.; Wu, J.; Jiang, Y.; Xie, L.; Fan, R. Computer vision for road imaging and pothole detection: A state-of-the-art review of systems and algorithms. *Transp. Saf. Environ.* **2022**, *4*, tdac026. [CrossRef]
22. Zadeh, L.A. Fuzzy logic, neural networks, and soft computing. *Commun. ACM* **1994**, *37*, 77–84. [CrossRef]
23. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 779–788.
25. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In *Proceedings of the Computer Vision—ECCV 2020, Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12346, pp. 213–229.
26. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
27. Everingham, M.; Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The Pascal Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]

28. Lin, T.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C. Microsoft COCO: Common Objects in Context. In *Proceedings of the Computer Vision—ECCV 2014. Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8693, pp. 740–755.
29. Liu, Y.; Sun, P.; Wergeles, N.; Shang, Y. A survey and performance evaluation of deep learning methods for small object detection. *Expert Syst. Appl.* **2021**, *172*, 114602. [CrossRef]
30. Goswami, P.; Aggarwal, L.; Kumar, A.; Kanwar, R.; Vasisht, U. Real-time evaluation of object detection models across open world scenarios. *Appl. Soft Comput.* **2024**, *163*, 111921. [CrossRef]
31. Wenkel, S.; Alhazmi, K.; Liiv, T.; Alrshoud, S.; Simon, M. Confidence Score: The Forgotten Dimension of Object Detection Performance Evaluation. *Sensors* **2021**, *21*, 4350. [CrossRef] [PubMed]
32. Jena, R.; Zhornyak, L.; Doiphode, N.; Chaudhari, P.; Buch, V.; Gee, J.; Shi, J. Beyond mAP: Towards Better Evaluation of Instance Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 11309–11318.
33. Powers, D.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Int. J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
34. Oksuz, K.; Cam, B.; Akbas, E.; Kalkan, S. Localization Recall Precision (LRP): A New Performance Metric for Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 504–519.
35. Frumosu, F.; Khan, A.; Schiøler, H.; Kulahci, M.; Zaki, M.; Westermann-Rasmussen, P. Cost-sensitive learning classification strategy for predicting product failures. *Expert Syst. Appl.* **2020**, *161*, 113653. [CrossRef]
36. Sbeyti, M.; Karg, M.; Wirth, C.; Klein, N.; Albayrak, S. Cost-Sensitive Uncertainty-Based Failure Recognition for Object Detection. In Proceedings of the Fortieth Conference on Uncertainty in Artificial Intelligence, Barcelona, Spain, 15–19 July 2024; Volume 244, pp. 1890–1900.
37. Viaene, S.; Dedene, G. Cost-sensitive learning and decision making revisited. *Eur. J. Oper. Res.* **2005**, *166*, 212–220. [CrossRef]
38. Al-Emadi, S.; Yang, Y.; Ofli, F. Benchmarking Object Detectors under Real-World Distribution Shifts in Satellite Imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2025; pp. 8299–8309.
39. Khan, T.; Singh, K.; Bhati, B.; Ahmad, K.; Al-Rasheed, A.; Getahun, M.; Soufiene, B. Trust-driven approach to enhance early forest fire detection using machine learning. *Sci. Rep.* **2025**, *15*, 14480. [CrossRef]
40. Zhang, Y.; Carballo, A.; Yang, H.; Takeda, K. Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. *ISPRS J. Photogramm. Remote Sens.* **2023**, *196*, 146–177. [CrossRef]
41. Zhang, Z.; Yu, X.; Tao, R.; Zhang, X.; Li, H.; Lu, J.; Zhou, J. Knowledge augmentation-based soft constraints for semi-supervised clustering. *Appl. Soft Comput.* **2023**, *144*, 110484.
42. Dehshiri, S.J.H.; Amiri, M.; Olfat, L.; Pishvaee, M.S. A robust fuzzy stochastic multi-objective model for stone paper closed-loop supply chain design considering the flexibility of soft constraints based on Me measure. *Appl. Soft Comput.* **2023**, *134*, 109944.
43. Kaushal, M.; Khehra, B.S.; Sharma, A. Soft computing based object detection and tracking approaches: State-of-the-art survey. *Appl. Soft Comput.* **2018**, *70*, 423–464. [CrossRef]
44. Liu, Y.; Zhou, C.; Guo, D.; Wang, K.; Pang, W.; Zhai, Y. A decision support system using soft computing for modern international container transportation services. *Appl. Soft Comput.* **2010**, *10*, 1087–1095. [CrossRef]
45. Khelifa, B.; Laouar, M.R. A holonic intelligent decision support system for urban project planning by ant colony optimization algorithm. *Appl. Soft Comput.* **2020**, *96*, 106621. [CrossRef]
46. Küppers, F.; Kronenberger, J.; Shantia, A.; Haselhoff, A. Multivariate Confidence Calibration for Object Detection. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2020; pp. 1322–1330.
47. Padilla, R.; Netto, S.L.; da Silva, E.A.B. A Survey on Performance Metrics for Object-Detection Algorithms. In *Proceedings of the 2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*; IEEE: New York, NY, USA, 2020; pp. 237–242.
48. Hall, D.; Dayoub, F.; Skinner, J.; Zhang, H.; Miller, D.; Corke, P.; Carneiro, G.; Angelova, A.; Suenderhauf, N. Probabilistic Object Detection: Definition and Evaluation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*; IEEE: New York, NY, USA, 2020; pp. 1031–1040.
49. van Rijsbergen, C.J. *Information Retrieval*, 2nd ed.; Butterworths: London, UK, 1979.
50. Lee, K.; Hong, J.; Kim, B.; Song, Y.; Lee, D. Cost-Stability Evaluation Framework for Reliable Deployment of Object Detectors in Industrial Applications. *IEEE Access* **2025**, *13*, 207566–207580. [CrossRef]
51. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. On calibration of modern neural networks. In Proceedings of the ICML'17: Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; Volume 70, pp. 1321–1330.
52. Zhang, J.; Cho, J.; Zhou, X.; Krähenbühl, P. NMS Strikes Back. *arXiv* **2022**, arXiv:2212.06137. [CrossRef]

53. Pathiraja, B.; Gunawardhana, M.; Khan, M. Multiclass Confidence and Localization Calibration for Object Detection. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 19734–19743.
54. Lee, D.; Kim, K. Interactive weighting of bias and variance in dual response surface optimization. *Expert Syst. Appl. ESWA* **2012**, *39*, 5900–5906.
55. Trillo, J.R.; Herrera-Viedma, E.; Morente-Molinera, J.A.; Cabrerizo, F.J. A large scale group decision making system based on sentiment analysis cluster. *Inf. Fusion* **2023**, *91*, 633–643.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.