From Consensus to Causality: Adaptive Reliability Fusion for Object Detection Ensembles

Alaa Daoud, Hiba Alqasir

Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2026) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

The 3rd International Workshop on Causality, Agents and Large Models (CALM-26)
April 14-16, 2026, Istanbul, Türkiye

From Consensus to Causality: Adaptive Reliability Fusion for Object Detection Ensembles

Alaa Daoud*, Hiba Alqasir

LAMIH UMR CNRS 8201, INSA Hauts-de-France, Université Polytechnique Hauts-de-France (UPHF), France

Abstract

Object detection ensembles often rely on post-processing methods such as Weighted Boxes Fusion (WBF) to combine overlapping predictions from multiple detectors. While effective, standard WBF assumes fixed intersection-over-union (IoU) thresholds and uniform model trust, limiting its adaptability to diverse object scales and varying detector quality. In this work, we introduce Adaptive Reliability-Weighted Boxes Fusion (AR-WBF), a context-aware extension of WBF that addresses these limitations through two key mechanisms. First, each model's predictions are scaled by a reliability factor reflecting its validated detection accuracy, enabling more trustworthy models to exert greater influence during fusion. Second, AR-WBF employs an adaptive IoU threshold and deferred reliability weighting, calibrating final confidence after spatial aggregation.

This strategy preserves geometric consensus while improving recall and confidence calibration. Experiments on the COCO dataset demonstrate that AR-WBF maintains precision and improves recall stability compared to baseline WBF, particularly in heterogeneous ensembles. Moreover, we interpret the adaptive weighting process through a causal reasoning lens, viewing reliability and contextual adaptation as influencing factors in fusion outcomes. AR-WBF thus connects classical statistical fusion with causally inspired, context-aware object detection. Beyond performance gains, it illustrates how lightweight causal reasoning—via reliability priors and contextual interventions—can enhance interpretability and robustness in ensemble perception systems.

Keywords: Weighted Boxes Fusion; ensemble object detection; reliability-aware fusion; adaptive IoU thresholds; causal interpretation.

1. Introduction

Object detection lies at the heart of many computer vision applications, from autonomous driving to surveillance and scene understanding. While newer architectures continue to emerge, well-established detectors such as Faster R-CNN, RetinaNet, and the YOLO family remain widely used as reliable baselines. These models deliver strong accuracy but often produce overlapping or redundant bounding boxes, necessitating post-processing to consolidate predictions. The most common strategy, Non-Maximum Suppression (NMS) [10], keeps only the highest-scoring box

* Corresponding author.
E-mail address: alaa.daoud@uphf.fr

in each cluster, which can unintentionally discard valid detections in crowded or ambiguous scenes. Soft-NMS [1] addresses this by gradually reducing scores instead of removing boxes outright, yet both approaches remain limited when extended to ensembles of multiple detectors, particularly when confidence calibration and detector reliability vary.

Weighted Boxes Fusion (WBF) [13] improves upon NMS by averaging overlapping boxes weighted by confidence, preserving detector consensus across models. However, it assumes fixed Intersection-over-Union (IoU) thresholds and static model weights, ignoring variability in detector reliability—typically estimated from validation performance—and scene context. Adaptive-NMS [8] and Cluster-NMS [17] introduce adaptive thresholds for single models, while calibration and reliability estimation methods [3, 12] improve confidence estimates but remain detached from spatial fusion.

Our prior work, **Agentified Weighted Boxes Fusion (AWBF)** [2], enabled decentralized fusion via local agent communication, with one agent assigned per bounding box, while still relying on fixed IoU rules. Intersecting agents compete within each cluster (i.e., connected sets of overlapping boxes) to resolve overlaps, resulting in either suppression or cooperative fusion of their bounding boxes.

In this work we propose the **Adaptive Reliability-Weighted Boxes Fusion (AR-WBF)**[1] framework, which integrates two complementary mechanisms: (1) reliability-based weighting reflecting model trust, and (2) adaptive IoU thresholds responsive to object scale and context. These additions preserve WBF's deterministic design while enhancing robustness, confidence calibration through deferred reliability weighting, and causal interpretability.

Contributions. We: (a) analyze the limitations of static WBF-based fusion; (b) introduce reliability-aware weighting and adaptive thresholding into deterministic fusion; (c) support both centralized and multiagent ensemble settings; and (d) provide a causal interpretation linking trust modeling with contextual adaptation.

2. State of the Art

Widely used object detectors often produce overlapping bounding boxes that require post-processing to form coherent predictions. Traditional suppression techniques, such as Non-Maximum Suppression (NMS) and Soft-NMS [1], retain only the highest-scoring detection within each cluster. While effective at reducing redundancy, these methods can unintentionally suppress valid detections in dense or ambiguous scenes, limiting their performance in complex environments.

To address these limitations, **ensemble fusion methods** combine outputs from multiple detectors to improve robustness and prediction quality. Weighted Boxes Fusion (WBF) [13] averages overlapping boxes according to confidence scores, preserving detector consensus and often outperforming traditional suppression strategies. Despite its advantages, WBF relies on static aggregation rules: both the Intersection-over-Union (IoU) threshold and detector weights remain fixed across categories and contexts. This rigidity reduces adaptability to varying detector reliability, object scales, and scene complexity. Adaptive-NMS [8] and Cluster-NMS [17] introduce dynamic thresholds within individual detectors, yet they still overlook reliability variation across multiple models.

Several **WBF-inspired extensions** focus on refining confidence, clustering detections, or leveraging voting strategies. The Agglomerative Late Fusion Algorithm (ALFA) [11] hierarchically merges boxes using weighted averaging, while Probabilistic Ranking-Aware Ensembles (PRAE) [9] re-rank detections based on localization statistics. Consensus Focus [5] aggregates boxes through voting to reduce minority-class bias, and Confidence-Aware Fusion [7] adjusts scores according to inter-model agreement. Despite their diversity, these approaches typically rely on fixed or category-invariant IoU thresholds and do not explicitly model contextual variability or detector reliability across ensembles. Similarly, application-oriented systems—such as confidence-aware SSD ensembles and domain-specific detectors [15, 7]—demonstrate empirical benefits but continue to use fixed weights or heuristics, limiting principled adaptability.

A complementary line of research focuses on **confidence calibration and reliability modeling**. Techniques such as temperature scaling [3], Platt scaling, and Dirichlet calibration improve probability estimates for individual models. Mixture-of-Experts frameworks [12] and ensemble weighting approaches estimate model reliability from validation

[1] The AR-WBF code and its experiments are available online https://github.com/alaa-daoud/AR-WBF

data but require learned gating mechanisms and are typically decoupled from spatial fusion. In contrast, AR-WBF integrates reliability directly as a deterministic prior within the fusion operation, allowing confidence and spatial information to jointly guide aggregation.

Context adaptivity also plays a vital role in accurate detection. Small or uncertain objects often benefit from lower IoU thresholds, whereas large objects require stricter overlap criteria. Soft-NMS, Adaptive-NMS, and Cluster-NMS adapt suppression thresholds based on object scale or overlap density [1, 8, 17], but these methods operate locally within a single detector rather than coordinating multiple sources. AR-WBF generalizes this principle to multi-detector ensembles by computing an adaptive IoU threshold that varies with object category and scale to maintain precision–recall balance.

Causal reasoning has gained attention as a tool for improving interpretability and robustness in perception. Causal Intervention for Object Detection [4] and Causal Contextual Modeling (CCM) [16] explicitly capture directional dependencies between objects to reduce spurious correlations. Although such methods require training and causal graph estimation, they conceptually motivate AR-WBF's design: the reliability term can be interpreted as a causal prior encoding epistemic trust, while the adaptive IoU threshold functions as a contextual intervention conditioned on environmental factors. This interpretation enables causal insight within a deterministic fusion algorithm, without requiring explicit causal modeling.

In summary, existing fusion techniques vary along three dimensions: (1) *static vs. adaptive*, (2) *uniform vs. reliability-aware*, and (3) *purely statistical vs. causally inspired*. While WBF and its variants improve aggregation, they remain largely static and uniform. Calibration and causal modeling offer complementary advances but remain disconnected from spatial fusion. Recent ensemble works contribute confidence reweighting or contextual voting, yet none jointly integrate **reliability modeling**, **adaptive thresholding**, and **causal reasoning** within a unified deterministic framework. **AR-WBF** bridges these gaps by jointly modeling detector reliability and contextual adaptivity while preserving the simplicity and determinism of WBF. It can operate in both centralized and decentralized settings, extending prior decentralized fusion approaches such as AWBF [2], and providing a principled foundation for reliability- and context-aware ensemble perception.

3. Methodology

The proposed **Adaptive Reliability-Weighted Boxes Fusion (AR-WBF)** extends Weighted Boxes Fusion (WBF) [13] through a causally grounded reformulation. Rather than treating reliability and adaptivity as heuristic adjustments, AR-WBF introduces two causal components that explicitly govern the fusion outcome: (1) a *causal prior* over detector reliability, expressing epistemic trust in each model's predictions; and (2) a *contextual intervention* realized through an adaptive IoU threshold that adjusts to environmental factors such as object scale and scene density. Together these components instantiate a lightweight causal graph, where model trust and contextual variability jointly determine the final fused detections. This causal formulation allows AR-WBF to function as an interpretable inference layer to a structural causal model—where reliability acts as a prior cause and context as an intervention influencing fusion outcomes.

Formally, let $\mathcal{M} = \{M_1, \ldots, M_K\}$ denote K detectors producing bounding boxes $\mathcal{B}_k = \{b_{k,i}\}$, confidence scores $\mathcal{S}_k = \{s_{k,i}\}$, and class labels $C_k = \{c_{k,i}\}$. The objective is to obtain a fused detection set $\mathcal{B}^*$ that maximizes coherence under causal weighting and context sensitivity. AR-WBF is deterministic, training-free, and compatible with both centralized WBF and decentralized AWBF [2], where causal priors and interventions can be computed locally and exchanged among agents.

Deferred Causal Prior on Model Reliability. In classical WBF, overlapping boxes of the same class are averaged using confidence weights, implicitly assuming uniform detector reliability. AR-WBF introduces a causal prior r_k that encodes the epistemic trust assigned to each detector M_k, based on its validated performance or inter-model agreement. Unlike heuristic score scaling, reliability in AR-WBF is *applied after clustering*, ensuring that all detectors contribute equally to the geometric fusion process. This design preserves unbiased spatial consensus while allowing reliability to modulate only the final fused confidence.

Formally, the post-clustering reliability adjustment can be expressed as:

$$(b^*, s^*)_{\text{AR-WBF}} = \left(\frac{\sum s_{k,i}\, b_{k,i}}{\sum s_{k,i}},\ \left(\frac{1}{n}\sum s_{k,i}\right) \times (0.9 + 0.1\, r_k) \right), \tag{1}$$

$$r_k = \frac{\text{mAP}_k - \min_j(\text{mAP}_j)}{\max_j(\text{mAP}_j) - \min_j(\text{mAP}_j) + \epsilon}. \tag{2}$$

where $r_k \in [0, 1]$ represents the reliability prior for detector M_k. r_k can be estimated based on the relative performance of M_k among other models on a specific set of samples by computing its mean Average Precision (mAP) and normalizing theresults to [0, 1]. Each cluster's bounding box b^* is determined purely by score-weighted averaging, while its confidence s^* is subsequently calibrated by the detector's reliability prior. This deferred weighting strategy prevents over-suppression by less reliable detectors and maintains higher recall, particularly in heterogeneous ensembles.

From a causal perspective, r_k expresses the conditional probability that a detection from M_k corresponds to a true object, given prior validation evidence. Thus, reliability acts as a *causal prior*—influencing the belief strength (confidence) of a fused detection without interfering with its spatial perception (geometry). This separation of spatial and epistemic contributions enhances both the interpretability and the robustness of the fusion process.

Contextual Intervention via Adaptive IoU. Environmental conditions also causally influence detection reliability. To encode this, AR-WBF introduces a contextual intervention through a class- and scale-dependent IoU threshold: $\tau(c, A) = \tau_0 + \alpha_c\, \sigma_A$, where τ_0 is a base IoU, σ_A is the normalized variance of object areas in a cluster, and α_c controls adaptivity per category. This mechanism acts as an intervention on the decision boundary: smaller or uncertain objects (high variance) trigger more permissive thresholds, while large, stable objects impose stricter ones. The result is a causal adjustment of overlap sensitivity according to scene context.

Fusion Algorithm. Algorithm 1 summarizes the procedure. AR-WBF fuses detections by integrating causal priors (r_k) and contextual interventions ($\tau(c, A)$) in a single deterministic pass, preserving WBF's $O(N^2)$ complexity while enhancing interpretability and generalization.
In agentified settings, AR-WBF can be seamlessly deployed, where each bounding-box agent locally applies adaptive IoU and reliability adjustments once the connected-set of intersecting agents is determined and just before engaging in cooperative fusion or competitive suppression, enabling decentralized yet causally consistent ensemble reasoning.

Algorithm 1 Adaptive Reliability-Weighted Boxes Fusion (AR-WBF)

Require: Detector outputs $\{\mathcal{B}_k, \mathcal{S}_k, C_k\}_{k=1}^{K}$, causal priors $\{r_k\}$, base IoU τ_0
Ensure: Fused detections $\mathcal{B}^*, \mathcal{S}^*, C^*$

```
for each class c do
    Gather all boxes {b_{k,i} | c_{k,i} = c} and scores {s_{k,i}}
    Compute contextual threshold τ(c, A) = τ_0 + α_c σ_A
    Cluster boxes where IoU(b_p, b_q) > τ(c, A)
    for each cluster C_j do
        Fuse boxes using causal prior weighting (Eq. 2)
        Append fused box (b*_j, s*_j, c)
    end for
end for
return B*, S*, C*
```

4. Evaluation and Discussion

We evaluate the proposed **AR-WBF** against the traditional **WBF** on the **MS COCO 2017** dataset, which contains 118k training and 5k validation images across 80 categories. To evaluate the proposed **AR-WBF** under heterogeneous

Table 1. Fusion performance of WBF and AR-WBF on the COCO validation set using YOLOv5x and EfficientDet-B2 ensembles.

	AP						AR					
	@[0.5 , 0.95]	@[0.5]	@[0.75]	Small	Medium	Large	@[0.5 , 0.95]	@[0.5]	@[0.75]	Small	Medium	Large
WBF	0.424	**0.615**	0.455	0.281	0.453	0.547	0.360	0.591	0.635	0.481	0.676	0.785
AR-WBF	**0.428**	0.605	**0.467**	0.281	**0.458**	**0.553**	0.360	**0.601**	**0.658**	**0.503**	**0.701**	**0.801**

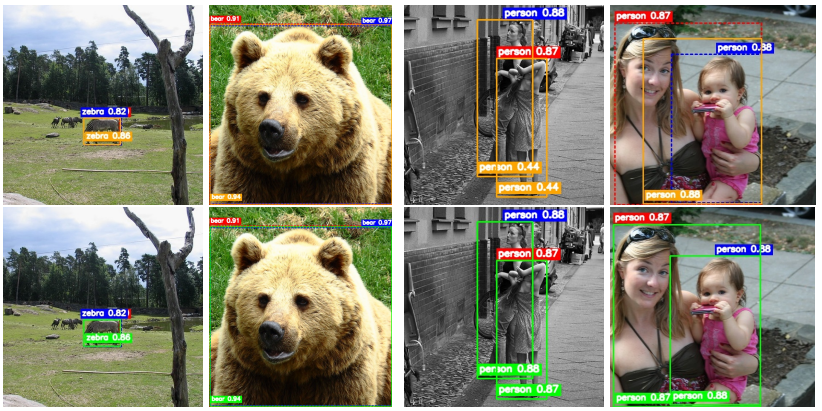

Fig. 1. Similarity (left) and diversion (right) of WBF (top) and AR-WBF (bottom) on selected samples from Yolov5x and EfficientDet-B2.

conditions, we selected two detectors (**YOLOv5x** [6] and **EfficientDet-B2** [14]) with complementary architectures and feature extraction strategies.

These models were chosen to represent structurally diverse detection paradigms: YOLOv5x as a dense, high-recall architecture and EfficientDet-B2 as a compact, precision-oriented one. Their predictions were evaluated both individually and through fusion methods, including WBF and the proposed AR-WBF. Each detector's reliability prior r_k was estimated on a COCO validation subset. An optional inter-model agreement term, based on the mean IoU overlap between detectors, slightly adjusted the normalized values. The final reliabilities—$r_{\text{YOLOv5x}} = 1$ and $r_{\text{EffNetB2}} = 0.6$—were cached and reused across experiments to ensure consistent weighting.

For fair comparison, all fusions used equal model weights and a base IoU threshold of 0.5, ensuring that gains stem solely from adaptive reliability and contextual mechanisms. Evaluation followed the official COCO protocol using Average Precision (AP) and Average Recall (AR) across IoU thresholds (0.50–0.95) and object scales. We report AP@[.50:.95], AP@0.50, AP@0.75, and AP_{Small} (< 32 px), AP_{Medium} (32–96 px), AP_{Large} (>96 px) objects, together with their corresponding AR metrics.

Table 1 presents the main results on the COCO validation set. AR-WBF achieves comparable or slightly higher mAP (+0.4 pts) and consistently higher recall (+2.3 pts AR@75) compared to WBF, confirming the benefits of adaptive IoU thresholding which prevents over-merging of small or densely overlapping objects, and deferred reliability weighting which down-weights inconsistent detectors. The late weighting strategy maintains geometric consensus while calibrating final confidence scores, leading to tighter high-IoU boxes and fewer missed detections.

Fig. 1 shows fusion of selected samples from YOLOv5x (blue) and EfficientDet-B2 (red), where both WBF and AR-WBF perform similarly in several cases, but diverge when IoU is near the threshold. In such cases, WBF's fixed IoU rule and uniform score averaging often misbehave: slightly overlapping boxes (e.g., IoU≈0.52) are merged despite spatial disagreement, while normalization across models lowers their confidence (e.g., from 0.8 to 0.4). If IoU falls just below the threshold, both boxes are retained, producing duplicates that reduce precision. AR-WBF overcomes these issues through an adaptive IoU that adjusts to object scale and by applying reliability weighting after clustering. This preserves geometric integrity and calibrated confidence, yielding more stable fusion without recall loss.

AR-WBF retains the $O(N^2)$ complexity of WBF with negligible overhead, as reliability and adaptivity terms are precomputed. It remains near real-time and fully modular, supporting both centralized fusion and decentralized integration (e.g., AWBF [2]). In the latter, each agent locally applies reliability and adaptive-IoU updates before consensus, making AR-WBF a unified reliability- and context-aware fusion mechanism across architectures.

5. Conclusion and Future Work

We introduced **Adaptive Reliability-Weighted Boxes Fusion (AR-WBF)**, a deterministic, causally inspired extension of Weighted Boxes Fusion for ensemble object detection. By combining reliability-based weighting with adaptive IoU thresholding, AR-WBF improves precision and recall across datasets without retraining or added complexity. Applying reliability after clustering ('late weighting') prevents over-suppression from less confident models, enhancing recall and calibration stability with minimal overhead. Adaptive IoU further strengthens robustness across scales, particularly for small and medium objects, yielding an interpretable and causally grounded fusion framework.

Implementation-wise, AR-WBF remains compatible with both centralized WBF and decentralized frameworks such as AWBF [2], underscoring its modularity and scalability. Future extensions may include dynamic reliability learning through Bayesian or causal inference, adaptive thresholding for spatio-temporal and multi-view perception, and integration within multimodal reasoning systems. AR-WBF demonstrates that causal and reliability-aware reasoning can be seamlessly embedded into deterministic fusion pipelines, enhancing both performance and interpretability.

Another direction is class-specific reliability modeling, where detector trust is estimated per object category rather than globally. Different detectors often specialize in distinct object types—e.g., YOLOv5x performs better on small or textured objects, while EfficientDet-B2 excels on larger or structured ones. Incorporating per-class reliability priors $r_{k,c}$ into the causal weighting scheme would enable finer, semantics-aware fusion that adapts model trust dynamically across categories.

References

[1] Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-nms – improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 5561–5569 (2017)

[2] Daoud, A., Bunel, C., Guériau, M.: Introducing multiagent systems to av visual perception sub-tasks: A proof-of-concept implementation for bounding-box improvement. In: 13th International Workshop on Agents in Traffic and Transportation (ATT 2024) held in conjunction with ECAI 2024 (2024)

[3] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)

[4] Huang, W., Jiang, M., Li, M., Meng, B., Ren, J., Zhao, S., Bai, R., Yang, Y.: Causal intervention for object detection. In: 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 770–774. IEEE (2021)

[5] Isai Valle Salgado, E., Li, C., Han, Y., Shi, L., Li, X.: Consensus focus for object detection and minority classes. arXiv e-prints pp. arXiv–2401 (2024)

[6] Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., et al.: Yolov5 by ultralytics (2023), https://github.com/ultralytics/yolov5, YOLOv5x model variant

[7] Lin, Y., Li, Z., Song, B., Ge, N., Sun, Y., Gong, X.: A confidence calibration based ensemble method for oriented electrical equipment detection in thermal images. Energies **18**(12), 3191 (2025)

[8] Liu, S., Huang, D., Wang, Y.: Adaptive nms: Refining pedestrian detection in a crowd. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

[9] Mao, M., Zhang, B., Doermann, D., Guo, J., Han, S., Feng, Y., Wang, X., Ding, E.: Probabilistic ranking-aware ensembles for enhanced object detections. arXiv preprint arXiv:2105.03139 (2021)

[10] Neubeck, A., Van Gool, L.: Efficient non-maximum suppression. In: 18th international conference on pattern recognition (ICPR'06). vol. 3, pp. 850–855. IEEE (2006)

[11] Razinkov, E., Saveleva, I., Matas, J.: Alfa: agglomerative late fusion algorithm for object detection. In: 2018 24th International Conference on Pattern Recognition (ICPR). pp. 2594–2599. IEEE (2018)

[12] Shazeer, N., Mirhoseini, A., Maziarz, P., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: Proceedings of the International Conference on Learning Representations (ICLR) (2017)

[13] Solovyev, R., Wang, W., Gabruseva, T.: Weighted boxes fusion: Ensembling boxes from different object detection models. Image and Vision Computing **107**, 104117 (2021)

[14] Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)

[15] Terzi, R.: An ensemble of deep learning object detection models for anatomical and pathological regions in brain mri. Diagnostics **13**(8), 1494 (2023)

[16] Yang, G., Zhang, J., Zhang, Y., Wu, B., Yang, Y.: Probabilistic modeling of semantic ambiguity for scene graph generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12527–12536 (2021)

[17] Zhang, Y., et al.: Cluster-nms: An improved non-maximum suppression method for object detection. In: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME) (2020)