BTQS: A Consensus-Based Temporal-Spatial Metric for Forensic Video Analysis

Gabriele Rosasco[1 [0009−0005−2204−7816]], Giorgio Volta[2 [0009−0000−1650−0294]], Edoardo Oldrini[1 [0009−0009−6298−0984]], and Rodolfo Zunino[1 [0000−0002−2150−7076]]

[1] University of Genova, DITEN, Genova, Italy
gabriele.rosasco@edu.unige.it, edoardo.oldrini@edu.unige.it, rodolfo.zunino@unige.it
[2] CASD/SSU Scuola Superiore Universitaria, Roma, Italy
giorgio.volta@unicasd.it

Abstract. Object detection systems are increasingly used in long-duration surveillance applications. Continuous video streaming brings about additional technical issues with respect to conventional, image-based approaches. Frame-level metrics such as mean Average Precision (mAP) and Intersection over Union (IoU) often fail to characterize temporal consistency across extended recordings, where fragmented detections and unstable activation patterns may affect reliability.
The paper presents a consensus-based temporal evaluation framework, which drives the detector behavior through temporal activity patterns instead of isolated frame predictions. The approach constrcuts a temporal consensus representation by combining detector activity signals and by extracting shared temporal segments. Building upon this representation, we introduce a Binary Temporal Quality Score (BTQS) that integrates temporal overlap, spatial agreement, boundary stability, and computational efficiency. The framework is designed to capture those temporal characteristics that might elude conventional, frame-wise evaluation methods.
Experiments involving YOLO11, Faster R-CNN, RT-DETR, and RetinaNet show that the proposed framework provides a more informative characterization of detector behavior than conventional evaluation strategies.

Keywords: Object Detection · Video Analytics · Temporal-Spatial Evaluation · Consensus-Based Assessment · Deep Learning

1 Introduction

Modern surveillance systems generate massive amounts of video data continuously, and therefore bring about substantial challenges for manual inspection and forensic analysis. Human operators cannot efficiently inspect hours of recordings acquired from public spaces, transportation infrastructures, industrial environments, or security systems. Consequently, automatic object detection has become

a basic component for extracting relevant information from long video streams and alleviating the burden associated with manual monitoring.

Recent advances in deep learning significantly improved object detection performances. One-stage detectors such as YOLO prioritize real-time performance and computational efficiency by directly predicting object locations and categories in a single forward pass [9]. Two-stage approaches such as Faster R-CNN separate the generation and classification stages, thus improving localization accuracy [4]. Recent transformer-based architectures (e.g., RT-DETR) apply global attention mechanisms, which can model long-range dependencies and reduce heuristic design assumptions [30]. RetinaNet addresses class-imbalance problems by means of focal loss formulations [21]. These approaches yield remarkable performances on standard image benchmarks such as COCO [32]; in those cases, detector quality is typically evaluated using metrics that include mean Average Precision (mAP) and spatial Intersection over Union (IoU).

Despite their widespread adoption, image-based metrics exhibit significant drawbacks when deploying detectors in long-duration surveillance scenarios. In practical applications, the objective is not only identifying objects in isolated frames but also ensuring stable temporal behavior throughout continuous video streams. Small variations in confidence scores may generate fragmented detections, temporal instability, or inconsistent activation patterns across adjacent frames. Consequently, two models exhibiting similar frame-level performance may exhibit substantial discrepancies when operating on long video sequences.

Several works have highlighted the importance of temporal consistency in video analysis and object detection applications, where temporal aggregation and contextual information can improve robustness across consecutive frames [27]. Most evaluation methodologies strongly depend on frame-wise measurements, and often require manually annotated temporal ground truth.

Constructing precise temporal annotations for long surveillance videos is expensive and time-consuming. Moreover, temporal boundaries frequently introduce ambiguity and subjectivity, especially in forensic applications involving complex scenes and extended recording durations. Datasets such as UCF-Crime provide temporal annotations for anomalous events [38]; however, they were originally designed for anomaly recognition rather than systematic detector benchmarking.

To address these limitations, the paper introduces a consensus-based temporal evaluation framework, which characterizes the object-detector's behavior through temporal activity patterns rather than isolated frame-level detections. Instead of independently evaluating predictions, each detector generates a temporal activation signal indicating the presence of relevant objects over time. These activity signals integrate into a consensus strategy to generate reference temporal segments representing regions of agreement among multiple detectors.

Based on this representation, we introduce the Binary Temporal Quality Score (BTQS), a metric designed to jointly evaluate temporal agreement, detection consistency, localization stability, and computational efficiency. The framework integrates segment-level temporal Intersection over Union (tIoU), precision-

recall behavior, temporal boundary jitter penalties, spatial overlap consistency, and computational cost. Unlike traditional frame-based evaluation strategies, the approach explicitly captures properties that become particularly relevant in long-duration video analysis systems.

The method experimental evaluation included a spectrum of heterogeneous detection architectures: YOLO11, Faster R-CNN, RT-DETR, and RetinaNet. These models span distinct design paradigms, including one-stage, two-stage, and transformer-based approaches, enabling a broad comparison of temporal behaviors under realistic surveillance conditions.

In the following, Section 2 analyze the major features points of the presented research, in terms of methodology and performance evaluation, with respect to the existing literature. The actual framework is presented in Section 3, whereas Section 4 analyzes the basic features of the framework in an empirical context. Section 5 illustrates the experimental evaluation of the overall method. Sections 6 and 7 summarize the main features of the presented research, by considering the current limitations and discussing extensions and current lines of research.

2 Design Aspects

Object detection architectures and their evaluation frameworks constitute the two pillars underpinning this work. Section 2.1 surveys the principal architectural paradigms; Section 2.2 reviews benchmarking methodologies and their limitations.

2.1 Object Detection Architectures

Deep-learning-based object detectors can be broadly grouped into four families. Two-stage approaches, originating from R-CNN [1] and progressively refined through Fast R-CNN [3] and Faster R-CNN [4], separate region proposal and classification to improve localization accuracy. Subsequent developments, including R-FCN [5], Cascade R-CNN [6], Mask R-CNN [7], and DetectoRS [8], further improved localization precision and multi-scale representation.

One-stage detectors perform detection in a single forward pass. Representative examples include the YOLO family [9–11, 13–16], SSD and MobileNet-based variants [17–19], EfficientDet [20], and RetinaNet [21], the latter leveraging Focal Loss and Feature Pyramid Networks [12] to address class imbalance and multi-scale detection.

Anchor-free methods, such as CornerNet [22], CenterNet [23], and FCOS [24], eliminate predefined anchor configurations by reformulating detection as keypoint- or pixel-based regression. More recently, transformer-based architectures, including DETR [25], Deformable DETR [26], DINO [28] with Swin-L [29], and RT-DETR [30], introduced global attention mechanisms for object detection. Lightweight variants, including Tiny-YOLO [13], DSOD [31], and EfficientDet-Lite [20], further target resource-constrained deployment scenarios.

2.2 Benchmarking and Comparative Evaluation

Standardized benchmarks provide the reference evaluation protocol for detection research. PASCAL VOC [2] established mAP@0.5 as the early standard; MS COCO [32] raised the bar with 80 categories and a stricter mAP@[0.5:0.95] criterion that penalizes imprecise localizations. Domain-specific benchmarks extend coverage to specialized challenges: VisDrone [33] targets UAV imagery with objects approximately 14× smaller than COCO counterparts, while WiderPerson [34] addresses dense pedestrian scenarios with explicit occlusion annotations.

Systematic comparative surveys — Zou et al. [35], Liu et al. [36], and Zhao et al. [37] — have established consistent empirical patterns: two-stage detectors dominate high-IoU accuracy, YOLO-family models lead in real-time throughput, and transformer-based architectures scale more favorably with model size and data. However, cross-paper comparisons remain problematic due to heterogeneous hardware configurations and measurement protocols, making within-study controlled evaluations essential [35]. Furthermore, scalar aggregate metrics such as mAP obscure per-category, per-scale, and per-occlusion performance variation. As discussed by Padilla et al. [39], existing evaluation metrics primarily focus on spatial localization accuracy and classification performance, while providing limited insight into temporal consistency over long video sequences. The present work addresses this gap through the Binary Temporal Quality Score (BTQS), a consensus-based, model-agnostic framework enabling principled multi-model comparisons across heterogeneous detection scenarios.

3 Methodology

The framework aims to evaluate the temporal reliability of object detection models in long-duration video analysis scenarios, with a focus on forensic, security, and video surveillance applications. In those contexts, detection accuracy *perse* might be a poor measure of the practical quality of a model, since temporal consistency, stable event localization, and reliable performances over extended video sequences are of paramount importance.

The framework combines multiple object detection models, and analyzes their temporal behaviors by comparing their outputs at the video level. Each detector independently processes the input video and generates a dedicated analysis report containing temporal detection information. The comparison stage only proceeds after all analyses are completed, allowing the framework to evaluate the consistency of model behavior without interfering with the inference process itself.

For each frame t, a detector is considered 'active' whenever at least one object of interest has been detected. This allows to associate each detector with a binary temporal-activity signal, describing the presence or absence of detections over time. These individual activity signals subsequently aggregate to generate a consensus signal, driving the overall temporal decision of the ensemble.

The consensus signal then identifies temporal detection segments, corresponding to contiguous intervals in which the consensus remains active. The

resulting representation upgrades the analysis from isolated frame-level detections to higher-level temporal events. This allows to evaluate the consistency of different models at identifying the same temporal phenomena.

For each detector, the temporal overlap with the consensus is analyzed together with the alignment of individual temporal events and the stability of temporal boundaries. This design allows the framework to capture not only whether a model detects an event, but also how coherently and stably the event is localized over time.

To quantify these aspects, we introduce the Binary Temporal Quality Score (BTQS), a metric specifically designed for temporal and spatial evaluation in unsupervised video analysis settings. The BTQS combines complementary components:

- the global temporal overlap between model detections and the consensus,
- the segment-level alignment of temporal events,
- the spatial agreement between detector boxes and the consensus boxes,
- the temporal stability of segment boundaries,
- the computational cost of the detector.

The metric is defined as:

$$\mathrm{BTQS}_i = \alpha \cdot \mathrm{tIoU}_i + \alpha' \cdot \mathrm{sIoU}_i + \beta \cdot F1_i - \gamma \cdot J_i - \delta \cdot C_i$$

where:

- the segment-level temporal Intersection over Union (tIoU) measures the alignment of individual temporal events;
- the spatial Intersection over Union (sIoU) evaluates how consistently the detected object is localized within the frame with respect to the consensus;
- the $F1$ score measures the global temporal agreement with the consensus segmentation;
- the jitter term J penalizes unstable temporal boundaries.

To account for operational feasibility, the metric includes a computational cost term, C_i, defined as the ratio between the running time of the i-th detector and the duration of the analyzed video. This component penalizes the models that privilege temporal quality at the cost of excessive processing time, and might prove relevant in forensic, security, and video surveillance applications. The weight parameters $\alpha, \alpha', \beta, \gamma, \delta \geq 0$ balance the contributions of each component to the final score.

As opposed to conventional evaluation metrics for object-detection, the above formulation takes into account temporal reliability explicitly. This makes the overall approach suitable for surveillance and forensic scenarios, where unstable detections, fragmented events, or temporally inconsistent outputs may significantly reduce the practical usability of a detector. The consensus segmentation acts as a sort of ground truth, enabling unsupervised evaluation of temporal detection quality without requiring manually annotated temporal labels.

Appendix A gives the complete mathematical formulation of the framework and metric.

4 Empirical Diagnosis: Signal Saturation and Threshold Sensitivity

4.1 Saturation of the Binary Detection Signal

The presented architecture models each detector as a binary activity signal, $a_i(t)$ (as per Appendix A). Such a reduction is only informative if $a_i(t)$ varies over time: if a detector fires on nearly every frame, the temporal overlap tIoU_i and the agreement $F1_i$ saturate, and the metric might lose a discriminative effectiveness. To take into account this issue and tune parameters accordingly, we first characterized the activity regime of the four detectors on the full UCF-Crime corpus [38] (1,800 videos, 13,265,472 frames analysed) at the global confidence threshold $\tau = 0.1$.

Table 1 reports, for each detector, the fraction of frames containing at least one detection. Two signals are distinguished: the raw *multi-class* signal (any of the 13 logged COCO classes) and the *person-only* signal, on which the present instantiation of BTQS operates.

Table 1. Person-class saturation as a function of the confidence threshold τ on UCF-Crime. Saturation is defined as the fraction of the 13,265,472 frames containing at least one person detection with confidence above τ.

Detector	$\tau = 0.10$	$\tau = 0.25$	$\tau = 0.50$	$\tau = 0.75$
YOLO11	70.86%	62.54%	53.17%	32.07%
Faster R-CNN R50	87.54%	80.66%	73.55%	66.81%
RT-DETR	89.87%	75.09%	63.13%	49.88%
RetinaNet	92.76%	75.61%	59.99%	35.07%

At $\tau = 0.1$ the raw multi-class signal is heavily saturated, ranging from 88.96% (YOLO11) to 98.50% (RT-DETR). Restricting to the person class lowers saturation for every detector, to the range 70.86%–92.76%, and widens the inter-detector spread from 9.54 to 21.90 percentage points. The single-class restriction therefore both reduces saturation and increases the discriminative separation between detectors. It does not, however, eliminate saturation for the three detectors with the most permissive operating point (Faster R-CNN, RT-DETR, and RetinaNet remain above 87%), for which the negative class covers only 7%–12% of frames. This residual saturation motivates the threshold analysis of Section 4.3.

4.2 Single-Class Scope and Multi-Class Logging

We restrict the present BTQS instantiation to the *person* class, the most forensically salient category in surveillance footage. Under a single-class regime the binary reduction $a_i(t) \in \{0, 1\}$ discards no class information, a property that would not hold in a multi-class setting; the multi class extension is left to future

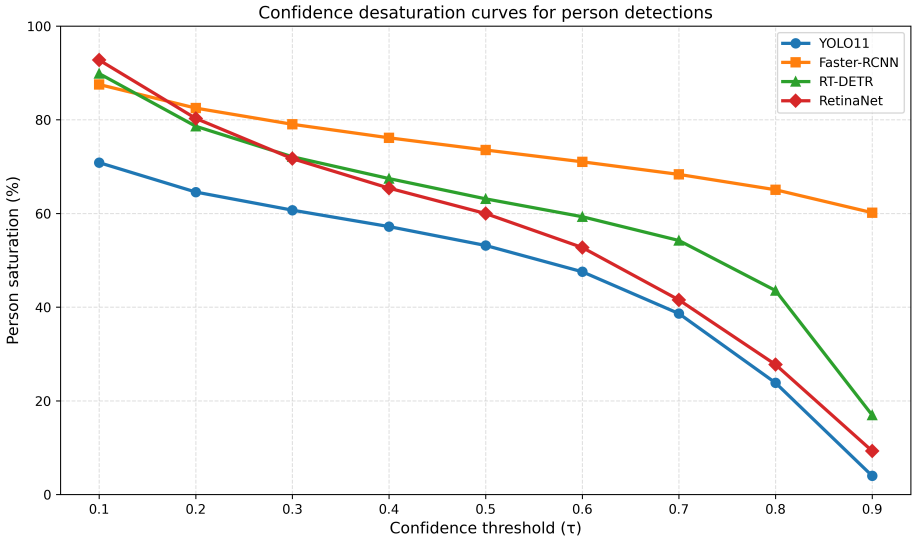

Fig. 1. Person-class saturation as a function of the confidence threshold τ for the four evaluated detectors on UCF-Crime (13,265,472 frames). Saturation is computed as the fraction of frames containing at least one valid person detection above the selected confidence threshold. The curves exhibit distinct desaturation profiles and different saturation regimes, indicating substantial differences in confidence calibration across detector architectures.

work. To keep that extension feasible without re-running inference—an operation whose aggregate cost on our hardware exceeded 650 GPU-hours across the four detectors—the inference logs retain all 13 forensically-relevant COCO classes, ensuring that non-person evidence remains accessible for subsequent analyses.

4.3 Threshold Non-Comparability and Sensitivity

The 21.90-point spread observed at a single global threshold is direct evidence that the four detectors operate at non-comparable points. Their confidence scores arise from different mechanisms– softmax over region proposals (Faster R-CNN), objectness $\times$ class confidence (YOLO11), focal-loss sigmoid deliberately shifted toward low values (RetinaNet [21]), and Hungarian-matched sigmoid logits concentrated near the extremes (RT-DETR [25, 30])—and are therefore not directly comparable on a common probabilistic scale. A single global τ thus places the detectors in different operating regimes.

Figure 1 (and Table 1) reports person-class saturation as τ is swept. At $\tau = 0.1$ the values reproduce the person-only column of Table 1. As τ increases the curves cross: RetinaNet, the most saturated detector at $\tau = 0.1$ (92.76%), falls below both RT-DETR and Faster R-CNN by $\tau = 0.5$ and reaches 35.07% at $\tau = 0.75$, consistent with its focal-loss sigmoid being concentrated at low confidence

values [21]. Faster R-CNN de-saturates the least (87.54%→66.81%), indicating a confidence distribution concentrated at high values, whereas YOLO11 remains the least saturated at every threshold. This non-monotone re-ordering of the detectors across τ is direct empirical evidence that a single global threshold is not a common operating point, motivating the per-detector calibration $\tau_i^\star$ adopted below.

To address this, rather than relying on a single global threshold, we select a per-detector operating point $\tau_i^\star$ directly from the person-saturation curves of Fig. 1, choosing for each detector the threshold that brings it from its saturated low-confidence regime into a comparable activity band (Section 5.1); this is an unsupervised criterion that requires no manual annotation. The spatial component sIoU_i (Appendix A) and the temporal components are evaluated at $\tau_i^\star$. Because every detection down to confidence 0.1 is retained in the logs, all analyses are computed by post-processing, without re-running inference. A systematic sensitivity analysis of the induced detector ranking across a threshold grid, together with formal significance testing via the Friedman test with Nemenyi post-hoc analysis [40], is left to future work.

5 Experimental Results

This section presents the experimental evaluation of the proposed consensus-based temporal assessment framework on the UCF-Crime dataset [38]. The analysis was performed on a large-scale corpus comprising approximately 1,800 surveillance videos and over 13 million frames. Four object detectors with different architectural paradigms were considered: YOLO11 [42], Faster RCNN [4], RT-DETR [30], and RetinaNet [21].

As discussed in Section 4, detector outputs were restricted to the *person* class in order to reduce semantic noise and improve temporal consistency across models. Moreover, rather than applying a uniform confidence threshold to all detectors, model-specific operating points were selected through the saturation analysis of Section 4.3, compensating for differences in confidence distributions across architectures and ensuring a more equitable comparison.

The experimental analysis is organized progressively. We first describe the per-detector operating points (Section 5.1). Subsequently, we analyze temporal agreement properties and examine the contribution of the individual BTQS components — temporal overlap, temporal consistency and stability, spatial coherence, and computational cost. Finally, a global ranking based on the proposed metric is presented.

5.1 Detector Operating Regimes via Saturation Analysis

Because the four detectors exhibit different confidence distributions, a common confidence threshold does not correspond to equivalent operating conditions. We therefore select detector-specific operating points $\tau_i^\star$ from the person-saturation curves of Fig. 1, choosing the onset of the saturation regime for each model.

The selected thresholds place all detectors within a comparable activity range, with person saturation between approximately 57% and 68%, reducing the inter-detector spread from 21.90 percentage points at $\tau = 0.1$ to about 11 points. The resulting operating points are reported in Table 2.

Table 2. Detector-specific operating points $\tau_i^\star$ obtained from the person-saturation analysis.

Detector	Operating point $\tau_i^\star$
YOLO11	0.4
Faster-RCNN	0.7
RT-DETR	0.6
RetinaNet	0.5

All subsequent experiments use these operating points. Since the thresholds are derived from an unsupervised saturation criterion rather than ground-truth calibration, their validation on externally annotated datasets such as WiderPerson [34] is left for future work.

5.2 Consensus-based Temporal-Spatial Results

Table 3 reports the final BTQS results obtained using a temporal gap of 0.30 s, a consensus threshold of 2/4 detectors, and the calibrated detector-specific operating points $\tau_i^\star$ introduced in Section 4.3. The BTQS weights were set to $\alpha = 0.4$ (temporal tIoU), $\alpha' = 0.4$ (spatial sIoU), $\beta = 0.2$ (temporal F1), $\gamma = 0.5$ (jitter penalty), and $\delta = 0.3$ (cost penalty), consistent with the formulation of Appendix A.

Table 3. Final BTQS scores obtained with calibrated detector-specific confidence thresholds and a 2/4 consensus rule. Higher values indicate better temporal-spatial agreement with the consensus while accounting for boundary stability and computational cost.

Detector	Mean	Median	P25	GMean	WMean$_{dur}$	Global
YOLO11	**0.6207**	**0.6350**	**0.5135**	**0.2216**	**0.5566**	**0.5018**
Faster-RCNN	0.4978	0.5075	0.3815	0.1183	0.4838	0.1700
RT-DETR	0.5042	0.5318	0.4011	0.1168	0.4595	0.2124
RetinaNet	0.4773	0.4953	0.3569	0.1407	0.4500	0.1358

YOLO11 achieves the highest score under all aggregation strategies, obtaining a mean BTQS of 0.6207 and a global score of 0.5018. This result indicates the best overall balance between temporal agreement with the consensus, spatial consistency, boundary stability, and computational efficiency. RT-DETR ranks second according to the global BTQS score (0.2124), while Faster R-CNN and

RetinaNet obtain lower overall rankings despite exhibiting competitive temporal-spatial alignment metrics.

Table 4. Mean values of the main BTQS components. Higher values are better for tIoU, F1 and sIoU, whereas lower values are preferred for jitter and computational cost.

Detector	tIoU	F1	sIoU	Jitter	Cost
YOLO11	0.5396	0.8651	0.6522	0.0309	**0.1887**
Faster-RCNN	**0.5779**	**0.8772**	**0.6998**	**0.0200**	0.7139
RT-DETR	0.5767	0.8660	0.6571	0.0246	0.6401
RetinaNet	0.5777	0.8697	0.6629	0.0226	0.7432

To better understand these results, Table 4 reports the mean values of the principal BTQS components. Faster R-CNN achieves the highest temporal IoU (0.5779), F1 score (0.8772), spatial IoU (0.6998), and the lowest jitter (0.0200), demonstrating strong agreement with the consensus both temporally and spatially. RetinaNet and RT-DETR exhibit similar temporal and spatial behaviour, with comparable tIoU, F1, and sIoU values. However, all three detectors incur substantially higher normalized computational costs than YOLO11. In particular, YOLO11 achieves a normalized computational cost of only 0.1887, compared with 0.6401 for RT-DETR, 0.7139 for Faster R-CNN, and 0.7432 for RetinaNet.

These results highlight the importance of jointly considering detection quality and computational efficiency. While Faster R-CNN achieves the strongest temporal-spatial agreement with the consensus, its substantially higher computational cost reduces its final BTQS. Conversely, YOLO11 combines competitive temporal and spatial performance with the lowest computational burden, resulting in the highest overall score and the most favourable trade-off among the evaluated detectors.

6 Limitations and Future Directions

Four aspects of the proposed framework define an agenda for future work. First, the consensus segmentation inherits any biases shared across detectors in the ensemble: external validation on annotated datasets such as WiderPerson [34] will allow inter-model coherence to be disentangled from absolute accuracy. Second, the saturation-based calibration of Section 4.3 ensures comparable operating regimes but does not yield probabilities calibrated in the classical sense; a systematic comparison with supervised calibration on an annotated subset will quantify the discrepancy. Third, the enclosing-box formulation of the sIoU component (Appendix A) is exact for the single-person case but becomes conservative in crowded scenes: integrating a per-instance Hungarian matching in the spirit of HOTA [41] will extend the framework to densely populated contexts. Fourth, a quantitative sensitivity analysis of the detector ranking across

confidence thresholds, supported by Friedman–Nemenyi significance testing [40], together with the multi-class extension of BTQS (already supported by the retained inference logs), completes the agenda.

7 Conclusions

This work introduced the Binary Temporal Quality Score (BTQS), a consensus-based framework for evaluating object detectors on long-duration surveillance video in the absence of manual temporal annotations. The framework reduces each detector to a binary temporal activity signal, aggregates the signals through majority voting into a consensus pseudo-ground-truth, and quantifies detector behavior along five complementary axes: segment-level temporal alignment, spatial agreement under a leave-one-out construction, global temporal F1, boundary stability, and computational cost. An empirical diagnosis of confidence saturation motivated a per-detector calibration of operating points, addressing the non-comparability of raw confidence scores across heterogeneous architectures. Experiments on the UCF-Crime corpus (1,800 videos, 13,265,472 frames) showed that YOLO11, Faster R-CNN, RT-DETR, and RetinaNet excel along different BTQS components, and that the composite score integrates these trade-offs into a single ranking informative for forensic and surveillance deployment.

References

1. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proc. CVPR, pp. 580–587. IEEE (2014). https://doi.org/10.1109/CVPR.2014.81
2. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The Pascal visual object classes (VOC) challenge. Int. J. Comput. Vis. 88(2), 303–338 (2010). https://doi.org/10.1007/s11263-009-0275-4
3. Girshick, R.: Fast R-CNN. In: Proc. ICCV, pp. 1440–1448. IEEE (2015). https://doi.org/10.1109/ICCV.2015.169
4. 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. 39(6), 1137–1149 (2017). https://doi.org/10.1109/TPAMI.2016.2577031
5. Dai, J., Li, Y., He, K., Sun, J.: R-FCN: Object detection via region-based fully convolutional networks. In: Advances in Neural Information Processing Systems (NIPS), vol. 29 (2016)
6. Cai, Z., Vasconcelos, N.: Cascade R-CNN: High quality object detection and instance segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 43(5), 1483–1498 (2021). https://doi.org/10.1109/TPAMI.2019.2956516
7. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proc. ICCV, pp. 2961–2969. IEEE (2017). https://doi.org/10.1109/ICCV.2017.322
8. Qiao, S., Chen, L.-C., Yuille, A.: DetectoRS: Detecting objects with recursive feature pyramid and switchable atrous convolution. In: Proc. CVPR, pp. 10213–10222. IEEE (2021)
9. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proc. CVPR, pp. 779–788. IEEE (2016). https://doi.org/10.1109/CVPR.2016.91

10. Redmon, J., Farhadi, A.: YOLO9000: Better, faster, stronger. In: Proc. CVPR, pp. 7263–7271. IEEE (2017)
11. Redmon, J., Farhadi, A.: YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
12. Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proc. CVPR, pp. 2117–2125. IEEE (2017)
13. Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y.M.: YOLOv4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
14. Wang, C.-Y., Bochkovskiy, A., Liao, H.-Y.M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: Proc. CVPR, pp. 7464–7475. IEEE (2023)
15. Wang, C.-Y., Yeh, I.-H., Liao, H.-Y.M.: YOLOv9: Learning what you want to learn using programmable gradient information. In: Proc. ECCV, pp. 1–21. Springer (2024)
16. Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., Ding, G.: YOLOv10: Real-time end-to-end object detection. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 37, pp. 107984–108011 (2024)
17. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: SSD: Single shot multibox detector. In: Leibe, B., et al. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 21–37. Springer (2016). https://doi.org/10.1007/978-3-319-46448-0_2
18. Howard, A.G., et al.: MobileNets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017)
19. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: Inverted residuals and linear bottlenecks. In: Proc. CVPR, pp. 4510–4520. IEEE (2018)
20. Tan, M., Pang, R., Le, Q.V.: EfficientDet: Scalable and efficient object detection. In: Proc. CVPR, pp. 10781–10790. IEEE (2020). https://doi.org/10.1109/CVPR42600.2020.01079
21. Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proc. ICCV, pp. 2980–2988. IEEE (2017). https://doi.org/10.1109/ICCV.2017.324
22. Law, H., Deng, J.: CornerNet: Detecting objects as paired keypoints. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11218, pp. 765–781. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01264-9_45
23. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
24. Tian, Z., Shen, C., Chen, H., He, T.: FCOS: A simple and strong anchor-free object detector. IEEE Trans. Pattern Anal. Mach. Intell. 44(4), 1922–1933 (2022). https://doi.org/10.1109/TPAMI.2021.3069173
25. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
26. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: Proc. ICLR (2021)
27. X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei: Deep Feature Flow for Video Recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4141–4150 (2017)

28. Zhang, H., et al.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Proc. ICLR (2023)
29. Liu, Z., et al.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. ICCV, pp. 10012–10022. IEEE (2021)
30. Zhao, Y., et al.: DETRs beat YOLOs on real-time object detection. In: Proc. CVPR, pp. 16965–16974. IEEE (2024)
31. Shen, Z., Liu, Z., Li, J., Jiang, Y.-G., Chen, Y., Xue, X.: DSOD: Learning deeply supervised object detectors from scratch. In: Proc. ICCV, pp. 1919–1927. IEEE (2017)
32. Lin, T.-Y., et al.: Microsoft COCO: Common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014)
33. Zhu, P., Wen, L., Du, D., Bian, X., Hu, Q., Ling, H.: Detection and tracking meet drones challenge. IEEE Trans. Pattern Anal. Mach. Intell. 44(11), 7380–7399 (2022)
34. Zhang, S., Xie, Y., Wan, J., Xia, H., Li, S.Z., Guo, G.: WiderPerson: A diverse dataset for dense pedestrian detection in the wild. IEEE Trans. Multimed. 22(2), 380–393 (2020)
35. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proc. IEEE 111(3), 257–276 (2023). https://doi.org/10.1109/JPROC.2023.3238524
36. Liu, S., Huang, D., Wang, Y.: Receptive field block net for accurate and fast object detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 404–419. Springer, Cham (2018)
37. Zhao, P., et al.: Object detection with deep learning: A review. IEEE Trans. Neural Netw. Learn. Syst. 30(11), 3212–3232 (2019). https://doi.org/10.1109/TNNLS.2018.2876865
38. Sultani, W., Chen, C., Shah, M.: Real-world anomaly detection in surveillance videos. In: Proc. IEEE CVPR, pp. 6479–6488. IEEE (2018). https://doi.org/10.1109/CVPR.2018.00678
39. Padilla, R., Netto, S.L., da Silva, E.A.B.: A survey on performance metrics for object-detection algorithms. In: Proc. Int. Conf. Systems, Signals and Image Processing (IWSSIP), pp. 237–242. IEEE (2020). https://doi.org/10.1109/IWSSIP48289.2020.9145130
40. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30 (2006).
41. Luiten, J., Ošep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: HOTA: A higher order metric for evaluating multi-object tracking. Int. J. Comput. Vis. 129(2), 548–578 (2021). https://doi.org/10.1007/s11263-020-01375-2
42. Jocher, G., Qiu, J.: Ultralytics YOLO11. Version 11.0.0 (2024). https://github.com/ultralytics/ultralytics

Appendix

A Mathematical Formulation of the BTQS Framework

Let $a_i(t) \in \{0, 1\}$ denote the binary activity signal of detector i at frame t:

$$a_i(t) = \begin{cases} 1 & \text{if at least one target object is detected} \\ 0 & \text{otherwise} \end{cases}$$

The detector signals are aggregated through a consensus function:

$$C(t) = f(a_1(t), a_2(t), \ldots, a_N(t)),$$

where $f(\cdot)$ is implemented as majority voting in all experiments. The consensus is active whenever at least $\lceil N/2 \rceil$ detectors are active.

Consensus segments are defined as maximal contiguous intervals where $C(t) = 1$:

$$s_k = [t_k^{start}, t_k^{end}], \qquad S_{cons} = \{s_1, \ldots, s_K\}.$$

For each detector i, temporal agreement with the consensus is evaluated through overlap, extra, and miss regions. Precision and recall are:

$$Precision_i = \frac{|S_i \cap S_{cons}|_t}{|S_i|_t}, \qquad Recall_i = \frac{|S_i \cap S_{cons}|_t}{|S_{cons}|_t},$$

where $|\cdot|_t$ denotes temporal duration. The temporal F1-score is:

$$F1_i = \frac{2 \cdot Precision_i \cdot Recall_i}{Precision_i + Recall_i}.$$

To evaluate segment-level temporal alignment, we define the temporal Intersection over Union:

$$tIoU(s_a, s_b) = \frac{|s_a \cap s_b|}{|s_a \cup s_b|}.$$

The detector-level temporal overlap is computed as:

$$tIoU_i = \frac{1}{|S_i|} \sum_{s_a \in S_i} \max_{s_b \in S_{cons}} tIoU(s_a, s_b).$$

To capture spatial consistency, let $D_i(t)$ be the set of person detections produced by detector i at frame t, and let $b_i(t)$ denote the enclosing bounding box of $D_i(t)$. The leave-one-out consensus reference is:

$$B_{cons}^{(-i)}(t) = enc\left(\{b_j(t) : j \in C(t),\, j \neq i\}\right),$$

where $enc(\cdot)$ denotes the enclosing axis-aligned box. Spatial agreement is measured through:

$$sIoU_i(t) = \frac{|b_i(t) \cap B_{cons}^{(-i)}(t)|}{|b_i(t) \cup B_{cons}^{(-i)}(t)|}.$$

The detector-level spatial consistency is:

$$sIoU_i = \frac{1}{|T_i^+|} \sum_{t \in T_i^+} sIoU_i(t),$$

where:

$$T_i^+ = \{t : a_i(t) = 1 \wedge C(t) = 1 \wedge B_{cons}^{(-i)}(t) \neq \emptyset\}.$$

Temporal stability is evaluated through boundary jitter. Let B_i and B_{cons} denote the detector and consensus boundary sets. For each boundary:

$$d(b, B_{cons}) = \min_{b' \in B_{cons}} |b - b'|.$$

The normalized jitter term is:

$$\tilde{J}_i = \frac{1}{T|B_i|} \sum_{b \in B_i} d(b, B_{cons}),$$

where T is the video duration.

Computational cost is defined as:

$$C_i^{raw} = \frac{T_i^{analysis}}{T^{video}},$$

where $T_i^{analysis}$ denotes the detector processing time and T^{video} is the video duration.

To ensure comparability across detectors, the cost term is normalized with respect to the maximum computational cost observed across the evaluated models:

$$C_i = \frac{C_i^{raw}}{\max_j C_j^{raw}}.$$

The final Binary Temporal Quality Score is:

$$BTQS_i = \alpha \cdot tIoU_i + \alpha' \cdot sIoU_i + \beta \cdot F1_i - \gamma \cdot \tilde{J}_i - \delta \cdot C_i,$$

where $\alpha, \alpha', \beta, \gamma, \delta \geq 0$ control the contribution of each component.