

Article

Uncertainty3D: A Lightweight Tri-Dimensional Uncertainty Framework for CNN-Based Active Learning in Object Detection [†]

Qing Li ¹, Chunhe Xia ¹, Zhipeng Zhang ^{2,3,*} and Wenting Ma ² 

¹ School of Computer Science & Engineering, Beihang University, Beijing 100191, China; lq_beijing_buaa@163.com (Q.L.); xch@buaa.edu.cn (C.X.)

² China Mobile Research Institute, Beijing 100053, China; mawenting@chinamobile.com

³ China Mobile GBA (Greater Bay Area) Innovation Institute, Guangzhou 510656, China

* Correspondence: zhangzhipeng@chinamobile.com

[†] IEEE International Conference on Image Processing (ICIP 2024): “Instance-Aware Uncertainty for Active Learning in Object Detection”, Abu Dhabi, United Arab Emirates, 27–30 October 2024; pp. 298–304, doi:10.1109/ICIP51287.2024.10647710.

Abstract

In object detection, annotation cost and computational efficiency are important factors in iterative model improvement under standard benchmark settings. Active learning (AL) addresses this challenge by selecting informative samples for labeling; however, many detection-oriented AL methods incur substantial overhead due to repeated inference (e.g., augmentation-based consistency). This paper introduces Uncertainty3D, a lightweight uncertainty proxy designed for standard CNN-based object detectors. It leverages native pre-NMS predictions to estimate sample informativeness using a single forward pass. We propose a tri-dimensional formulation that captures inconsistencies in *position*, *scale*, and *category* across proposal-consistent predictions. Experiments on PASCAL VOC and MS COCO using representative CNN-based detectors (Faster R-CNN and RetinaNet) show competitive mAP versus representative baselines and about 3–4× faster uncertainty estimation than augmentation-based baselines.

Keywords: active learning; CNN-based object detection; uncertainty proxy; computational efficiency; single-pass inference

1. Introduction

Deep learning has revolutionized computer vision applications including object detection, but its success heavily relies on large-scale annotated datasets [1–3]. The annotation process for object detection is particularly expensive, requiring both category labels and precise bounding boxes. In industrially motivated deployment scenarios, this annotation bottleneck often hinders the practical adoption and continuous improvement of detection systems.

Active learning (AL) addresses this challenge by strategically selecting the most informative samples for annotation, thereby maximizing model performance with minimal labeling effort [4,5]. It has been shown that deep models trained on carefully selected subsets of data via AL can achieve performance comparable to those trained on the full dataset [6].

However, while AL has shown promise in classification tasks, its application to object detection faces unique challenges: (1) the need to evaluate both localization and classification uncertainty, and (2) the computational overhead of uncertainty estimation in



Academic Editors: Alessandro Lo Schiavo and Pedro Couto

Received: 18 January 2026

Revised: 26 February 2026

Accepted: 2 March 2026

Published: 5 March 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

production pipelines. Current state-of-the-art AL methods for object detection often prioritize accuracy at the expense of computational efficiency. For instance, consistency-based approaches [7] require multiple forward passes with different data augmentations, making them impractical for large-scale industrial applications. Other methods focus primarily on classification uncertainty [8], neglecting the spatial aspects crucial for detection tasks. In large-scale detection settings under benchmarked or resource-constrained evaluation protocols, or in time-sensitive applications—both annotation cost and inference efficiency are critical considerations that existing methods fail to address comprehensively.

In industrially motivated inspection scenarios, object detection systems often operate under strict annotation and computational constraints. Targets of interest typically exhibit large appearance variation, complex backgrounds, partial occlusion, and severe class imbalance, making accurate bounding-box annotation costly and highly dependent on domain expertise. Under these conditions, active learning methods that can effectively prioritize informative images are particularly valuable. Empirical evaluation in this paper uses public benchmark datasets (PASCAL VOC and MS COCO) with standard detection architectures. All experiments are conducted using standard CNN-based detection architectures that remain widely adopted in practical deployment scenarios.

Our Contributions. This paper proposes **Uncertainty3D**, a practical active learning framework for object detection that is both effective and computationally efficient. The main contributions are three-fold:

1. **Single-pass, lightweight uncertainty estimation.** We compute uncertainty directly from native *pre-NMS* detector predictions in a single forward pass, without augmentation-based repeated inference or additional modules.
2. **Tri-dimensional uncertainty modeling for detection.** We quantify inconsistencies in *position*, *scale*, and *category* across multiple predicted instances, providing a comprehensive proxy for detection difficulty.
3. **Empirical performance under practical computational constraints.** Experiments on PASCAL VOC and MS COCO with Faster R-CNN and RetinaNet show improved sample selection and a clear computational benefit ($3\text{--}4\times$ faster uncertainty estimation than consistency-based baselines such as CALD), without architectural modification.

Scope and Positioning. Uncertainty3D is designed for standard CNN-based object detection pipelines. It operates directly on proposal-consistent *pre-NMS* prediction groups naturally produced by two-stage and one-stage CNN detectors. The formulation targets lightweight integration into widely deployed CNN-based systems without architectural modification.

Compared to the ICIP 2024 conference version [9], this manuscript provides a substantially extended study, including a more detailed methodological formulation, additional experiments across multiple datasets and detector architectures, a comprehensive computational efficiency analysis, and an expanded discussion on practical deployment considerations.

The remainder of this paper is organized as follows: Section 2 reviews related work. Section 3 details our methodology. Section 4 presents experimental results and analysis. Section 5 discusses limitations and future directions. Section 6 concludes the paper. A typical active learning cycle for object detection is illustrated in Figure 1.

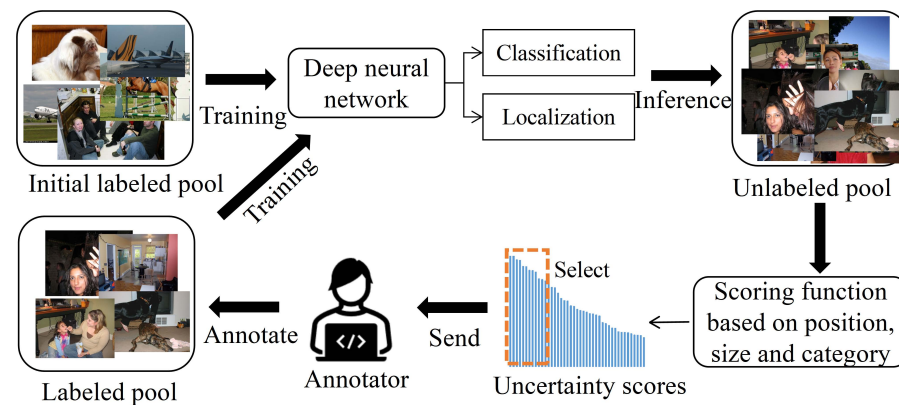


Figure 1. An illustration of one active learning cycle for object detection, including model training, uncertainty estimation, sample selection, and annotation.

2. Related Work

Active learning has been extensively explored across various domains [10–12], with traditional methods primarily relying on uncertainty-based sample selection strategies [13,14]. These strategies often focus on estimating uncertainty from individual sample predictions—such as classification confidence or entropy—but fail to account for sample diversity or spatial redundancy. As a result, selections within a batch can become highly correlated, reducing labeling efficiency. Uncertainty3D likewise does not explicitly model diversity or redundancy in batch selection.

2.1. Diversity-Based Active Learning

Beyond uncertainty-driven sampling, diversity-based strategies aim to reduce redundancy among selected samples by promoting coverage of the feature space. CoreSet [15] formulates active learning as a core-set selection problem using geometric distances in embedding space. BADGE [16] combines gradient-based representations with k-means++ initialization to jointly capture uncertainty and diversity in a single selection step. These approaches emphasize feature-space coverage rather than detector-specific structural uncertainty. Diversity-aware mechanisms typically require additional embedding extraction or clustering. Uncertainty3D instead improves uncertainty reliability directly from native detector outputs, without extra diversity modules.

While active learning has achieved considerable success in image classification [8,15,17,18], its extension to object detection remains non-trivial due to the structural complexity of detection tasks, including multiple instances per image, spatial dependencies, and joint regression–classification outputs. Although some recent works have started addressing these challenges, dedicated methods for object detection remain relatively sparse and diverse in their approaches [5,7,19–21].

2.2. Active Learning for Transformer-Based Detectors

Transformer-based object detectors adopt query-based prediction mechanisms that differ structurally from anchor- or proposal-based CNN detectors. Active learning strategies for such architectures typically leverage query embeddings or attention-derived uncertainty. This work focuses on uncertainty modeling within CNN-based detection pipelines and does not address query-based transformer formulations.

2.3. Bayesian (Aleatoric/Epistemic) vs. Heuristic Uncertainty

Uncertainty estimation in deep learning is often discussed under Bayesian predictive uncertainty, where *epistemic* uncertainty reflects model uncertainty and can be reduced with more data, while *aleatoric* uncertainty captures irreducible noise in observations [22]. Ap-

proximate Bayesian inference (e.g., Monte Carlo dropout) [23] and deep ensembles [24,25] are widely used to obtain calibrated predictive uncertainty. In object detection, Bayesian or probabilistic formulations typically require extra inference passes, probabilistic heads, or ensembles, which are costly in AL loops. Many practical detection AL methods instead use *heuristic* uncertainty—classification probability, entropy, disagreement, or augmentation consistency (e.g., [7,8]). Uncertainty3D adopts a heuristic proxy from native pre-NMS prediction inconsistency, favoring efficiency and compatibility with standard CNN detectors.

For instance, the LL4AL method by Yoo et al. [8] predicts instance-wise losses and selects samples with the highest predicted loss as most informative. However, this method focuses primarily on classification and ignores the localization aspect crucial in detection. Similarly, CALD by Yu et al. [7] estimates uncertainty via consistency under augmentations, integrating both classification and regression but at the cost of high computational overhead.

2.4. Semi-Supervised and Hybrid Active Learning

Recent trends integrate semi-supervised learning (SSL) with active learning to further reduce annotation cost. Chen et al. [26] propose a semi-supervised AL framework that leverages both labeled and unlabeled data via consistency regularization, reducing annotation demand. MuRAL by Liou et al. [27] introduces a multi-scale region-based sampling strategy to capture informative regions rather than entire images, improving efficiency and reducing redundancy. Hekimoglu et al. [28] address sample diversity by proposing a non-redundant, informativeness-guided sampling mechanism, significantly enhancing selection quality. These hybrid methods combine uncertainty-driven selection with pseudo-labeling or consistency regularization, increasing training complexity. Uncertainty3D is a lightweight uncertainty-ranking component for fully supervised AL and does not use SSL.

Meanwhile, Schubert et al. [29] study active learning in the presence of noisy annotations, revealing the fragility of existing methods under imperfect supervision. Na and De-Silva [30] explore uncertainty-aware selection for domain-adaptive object detection, emphasizing reconfiguration of pre-trained models in new environments.

2.5. Instance-Aware and Diversity-Oriented Active Learning for Detection

Recent work in active learning for object detection has explored richer notions of instance-level uncertainty and more structured selection strategies beyond simple image-level scoring. Yang et al. [31] propose Plug-and-Play Active Learning (PPAL), a two-stage framework that first performs difficulty-calibrated uncertainty sampling and then applies a diversity-based matching strategy to reduce redundancy. Yuan et al. [32] introduce Multiple Instance Active Learning for Object Detection (MI-AOD), which leverages adversarial instance classifiers and multi-instance re-weighting to estimate instance-level uncertainty for image selection.

These methods share the motivation of mitigating background-dominated uncertainty and improving instance-aware sample selection. However, they differ fundamentally from Uncertainty3D in methodological scope and design emphasis. PPAL and MI-AOD introduce additional modules or multi-stage selection mechanisms to explicitly learn or re-weight uncertainty at the pipeline level, whereas Uncertainty3D focuses on exploiting pre-NMS prediction inconsistency inherent to standard CNN-based detectors as a lightweight uncertainty proxy. By operating directly on native detector outputs without architectural modification or multi-stage selection, Uncertainty3D targets efficiency, simplicity, and ease of integration within standard CNN-based detection pipelines under practical active-learning settings.

2.6. Benchmarks and Standardized Evaluation

Standardized benchmarks are increasingly important for reproducible and comparable active learning evaluation in object detection. Recent efforts such as ALBench [33] provide more standardized evaluation tooling and data splits for active learning in object detection. Related work has also emphasized rapid prototyping and improved comparability in detection-oriented active learning [34]. This work prioritizes *fair and directly comparable* evaluation against widely cited detection-specific AL baselines that are predominantly reported under VOC/COCO-style protocols. We evaluate Uncertainty3D on PASCAL VOC and MS COCO using experimental protocols widely adopted in prior detection-oriented active learning studies (e.g., [7,20]), so that we can directly compare with representative baselines commonly reported under these settings. ALBench-style evaluation provides a complementary standardized setting, which is discussed in Section 5.

Despite recent progress, many detection-oriented AL methods still rely on augmentation-heavy repeated inference or consider limited uncertainty cues, which increases computational cost and weakens localization-aware selection. We propose Uncertainty3D, a lightweight uncertainty proxy that integrates position, scale, and category inconsistencies from native pre-NMS predictions in a single forward pass.

3. Method

Uncertainty3D computes uncertainty scores from native pre-NMS outputs of standard CNN-based object detectors, where proposal-consistent prediction groups are naturally defined by RoIs or anchors. We use Faster R-CNN and RetinaNet with ResNet-50-FPN backbones throughout. The formulation builds on proposal-consistent pre-NMS prediction groups as defined in these CNN-based detectors. An overview of the proposed Uncertainty3D framework and its integration with standard object detectors is illustrated in Figure 2.

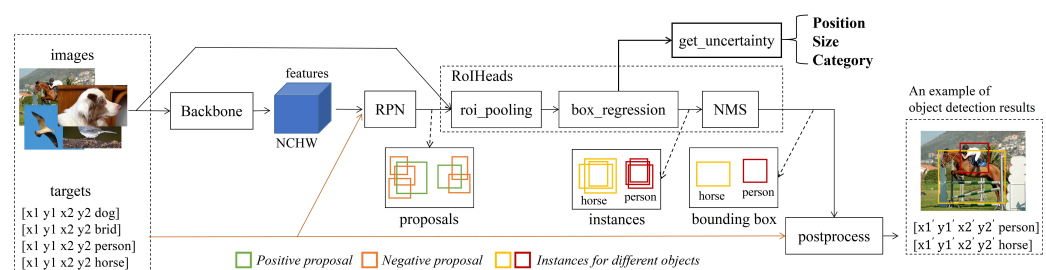


Figure 2. Overview of the proposed method. The backbone, region proposal, regression, and classification modules follow standard detector implementations (e.g., Faster R-CNN [35] and RetinaNet [36] with ResNet-50-FPN). The proposed Uncertainty3D framework operates solely on native pre-NMS predictions without introducing additional network components. The instances generated from object proposals are utilized to evaluate uncertainties in position, size, and category.

3.1. Problem Definition

The following notations are employed in this paper. We use X and Y to denote image and ground-truth annotations (class labels and bounding box coordinates), respectively. We use L and U to denote the labeled and unlabeled data index sets, respectively.

At each active learning cycle, a fixed annotation budget of N images is selected from the unlabeled set and added to the labeled set.

Assuming we are initially given a collection of labeled data $\{(X_l, Y_l)\}, l \in L$ and a large set of unlabeled data $\{(X_n)\}, n \in U$, we aim to sequentially select the most informative data from U and move them into the labeled set L such that a good model can be trained on data indexed by L .

3.2. Instance-Aware Uncertainty Modeling

Uncertainty3D estimates detection difficulty by analyzing the internal consistency of multiple predictions generated by a standard object detector prior to non-maximum suppression (NMS). All uncertainty measures are computed from native pre-NMS outputs in a single forward pass, without introducing additional network modules or repeated inference. A schematic of this instance-aware uncertainty estimation process is illustrated in Figure 3.

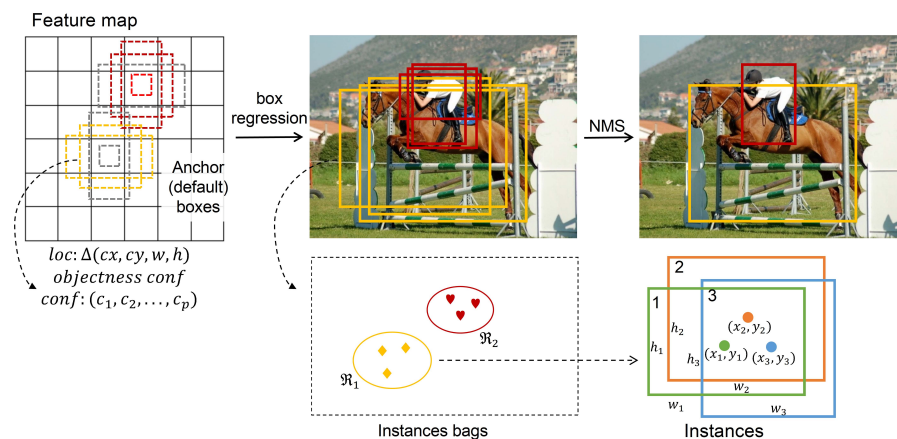


Figure 3. Instance-aware uncertainty for active learning. Following box regression and softmax normalization, only predictions with sufficiently high post-softmax classification probability are retained. Prior to non-maximum suppression (NMS), an instance bag is formed by proposal-consistent predictions prior to NMS. Uncertainties related to position, size, and category are then computed based on the center coordinates, width, height, and post-softmax class probability distributions of each instance.

3.2.1. Motivation: Risk-Oriented Proxies for Active Learning

Active learning aims to select unlabeled samples that most effectively improve model performance after annotation and retraining. Bayesian formulations often formalize this objective via information gain (e.g., BALD) [37] or expected loss reduction, which typically require posterior inference or multiple stochastic/ensemble predictions [23,24]. Such approaches are computationally expensive for modern object detectors within iterative AL loops.

Uncertainty3D therefore adopts a pragmatic strategy: We construct *single-pass proxy scores* from native pre-NMS outputs that correlate with (i) localization instability of box regression and (ii) semantic ambiguity of class prediction. Under a fixed detector and input setting, higher internal dispersion among proposal-consistent predictions indicates greater local prediction ambiguity, which empirically aligns with informative sample selection, without explicitly optimizing a Bayesian objective.

3.2.2. Instance Bag Definition

In active learning scenarios, ground-truth object assignments are unavailable for unlabeled data. Therefore, all uncertainty estimation must rely solely on model outputs. We define an *instance bag* as a set of *proposal-consistent predictions* generated from the same internal candidate source before NMS. Specifically, in two-stage detectors (e.g., Faster R-CNN), an instance bag consists of multiple bounding box predictions regressed from the same Region of Interest (RoI). In one-stage detectors (e.g., RetinaNet), an instance bag is formed by predictions originating from the same anchor location across regression outputs.

An instance bag is a local group of predictions sharing the same proposal geometry; it need not align with a single ground-truth object. Although such a group may include

overlapping background regions or multiple nearby objects, the internal dispersion of predictions reflects the detector's local uncertainty induced by feature ambiguity, proposal geometry, and regression instability. This dispersion is informative for AL because ambiguous regions tend to yield less consistent predictions.

Let $\mathcal{R}_i = \{b_{i,1}, \dots, b_{i,K}\}$ denote an instance bag containing K pre-NMS predictions. Each instance $b_{i,j}$ is represented by a five-dimensional vector $(x_{i,j}, y_{i,j}, w_{i,j}, h_{i,j}, p_{i,j})$, where (x, y) denotes the predicted bounding box center, (w, h) the width and height, and $p_{i,j}$ the predicted class probability vector after softmax.

3.2.3. Implementation Details of Instance Bag Construction

Instance bags are constructed differently for two-stage and one-stage detectors, and the bag size K is determined as follows.

Two-Stage Detectors (e.g., Faster R-CNN). For two-stage detectors, an instance bag corresponds to a single Region of Interest (RoI) produced by the region proposal network. After RoI feature extraction and class-wise box regression, each RoI yields a set of class-specific predictions, each associated with a post-softmax classification probability. We form the instance bag by collecting all class-wise predictions associated with the same RoI. The bag size K is obtained by ranking these predictions according to their post-softmax class probabilities and retaining the top- K predictions. Importantly, K is selected *after* class-wise regression and softmax normalization, but *before* any cross-class pruning or non-maximum suppression (NMS), so that all elements in the bag originate from the same proposal geometry.

One-Stage Detectors (e.g., RetinaNet). For one-stage detectors, an instance bag corresponds to a fixed anchor location. Each anchor produces class-wise predictions after regression and softmax normalization. We construct an instance bag by collecting all class-specific predictions associated with the same anchor location. Similarly, the top- K predictions are selected by ranking post-softmax class probabilities within this anchor group, without aggregating across different anchors or spatial locations. As in the two-stage case, K is determined prior to any NMS or cross-anchor filtering.

Across both detector families, the instance bag therefore represents a group of *proposal-consistent predictions* sharing identical geometric origin (RoI or anchor), and the top- K operation is applied only within this group.

3.2.4. Practical Frequency of Incomplete Instance Bags

In practice, after ranking class-wise predictions within a proposal-consistent group, a small fraction of instance bags may contain fewer than K retained predictions (e.g., due to low-confidence outputs or degenerate proposals). Such bags are excluded from uncertainty computation.

In our experiments the fraction of bags with fewer than K predictions stays below 5%, and image-level ranking is unaffected.

3.2.5. Position Uncertainty

Position uncertainty measures the spatial dispersion of predicted bounding box centers within an instance bag. We quantify this dispersion using the root-mean-square dispersion (RMSD) of absolute center coordinates:

$$\sigma_i = \sqrt{\frac{1}{K} \sum_{j=1}^K [(x_{i,j} - \bar{x}_i)^2 + (y_{i,j} - \bar{y}_i)^2]}, \quad (1)$$

where $(\bar{x}_i, \bar{y}_i)$ denotes the mean center location of the instance bag.

We interpret σ_i as a proxy for regression instability. When local features are ambiguous (e.g., occlusion or clutter), proposal-consistent predictions exhibit larger center dispersion, which empirically correlates with localization difficulty and annotation utility in active learning.

We compute dispersion directly in image coordinate space rather than using normalized box-relative coordinates. Bounding-box annotation and correction are performed in absolute image coordinates, so image-space dispersion provides a direct proxy for annotation-relevant misalignment.

Since active learning selection is based on relative ranking rather than calibrated uncertainty magnitude, σ_i serves as a ranking proxy under a fixed detector and input-resolution configuration. Potential scale-dependent dispersion therefore acts as a consistent global factor rather than a cross-image bias. Under such fixed settings, any uniform scaling transformation applied consistently across images does not alter the induced ranking. Empirical comparison with a representative scale-normalized variant (Appendix A Table A7) confirms that ranking behavior remains effectively unchanged.

3.2.6. Size Uncertainty

Size uncertainty captures the inconsistency of predicted bounding box scales within an instance bag. We first compute the dispersion of widths and heights across the K predictions:

$$\rho_i = \sqrt{\frac{1}{K} \sum_{j=1}^K [(w_{i,j} - \bar{w}_i)^2 + (h_{i,j} - \bar{h}_i)^2]}, \quad (2)$$

where $(\bar{w}_i, \bar{h}_i)$ are the mean width and height within the instance bag. To reduce scale bias and improve comparability across objects of different sizes, we normalize ρ_i using a stable scale reference derived from the post-NMS representative box b_i :

$$\hat{\rho}_i = \frac{2\rho_i}{w(b_i) + h(b_i)}. \quad (3)$$

Here, b_i is the post-NMS representative box for bag $\mathcal{R}_i$, used solely as a scale reference. The normalization makes the measure dimensionless and reduces bias toward large objects so that $\hat{\rho}_i$ reflects relative scale instability rather than absolute box magnitude. This follows standard detection practice of using scale-normalized regression targets to ensure comparability across object sizes.

3.2.7. Category Uncertainty

Category uncertainty reflects semantic ambiguity among predictions within an instance bag. Unlike geometric uncertainty, which is naturally characterized by between-instance dispersion, semantic ambiguity in object detection often manifests as a single highly ambiguous hypothesis within a proposal-consistent group.

We compute the entropy of the predicted class probability distribution for each instance and define the bag-level category uncertainty as the maximum entropy across the bag:

$$\psi_i = \max_{j=1,\dots,K} H(Y | b_{i,j}, \theta_L), \quad (4)$$

$$H(Y | b_{i,j}, \theta_L) = - \sum_{y \in \Omega} P(y | b_{i,j}, \theta_L) \log P(y | b_{i,j}, \theta_L), \quad (5)$$

where Ω denotes the set of class labels and θ_L is the detector trained on the labeled set L .

We use the maximum entropy within the bag as a conservative proxy for semantic ambiguity. The asymmetric aggregation is intentional. Geometric instability is characterized

by overall spatial dispersion, whereas semantic difficulty is often dominated by a single highly uncertain prediction that annotators must resolve.

3.2.8. Uncertainty Aggregation

We aggregate localization, scale, and category uncertainties into a single instance-level score:

$$E_i = \alpha(\sigma_i + \hat{\rho}_i) + \beta\psi_i, \quad (6)$$

where α and β control the relative contribution of geometric and semantic uncertainty. The image-level uncertainty is then defined as the maximum instance uncertainty over all instance bags in the image.

$$I_{X_n} = \max_i E_i, \quad (7)$$

where I_{X_n} denotes the uncertainty score of image X_n , used to rank unlabeled images for batch selection in each active learning cycle.

Equation (6) is a linear combination of geometric and semantic uncertainty, chosen for interpretability, single-pass efficiency, and ease of deployment. Additive combinations of this kind are common in detection AL when the goal is a stable ranking criterion.

3.3. Batch-Mode Active Learning Procedure

Based on the image-level uncertainty defined in Equation (7), we adopt a standard batch-mode active learning protocol. At each active learning cycle, uncertainty scores are computed once for all unlabeled images using the current detector. Images are then ranked according to their uncertainty scores, and a fixed number of images are selected in batch before updating the labeled set and retraining the detector. The complete procedure is summarized in Algorithm 1.

Algorithm 1: Batch-mode active learning with Uncertainty3D

Input: Initial labeled set L_0 , unlabeled pool U_0 ; detector architecture $\mathcal{M}$; bag size K ; batch budget per cycle N ; number of cycles T .

Output: Trained detector after T active learning cycles.

Train initial detector $\mathcal{M}_0$ on L_0 ;

for $t = 1$ **to** T **do**

 // Uncertainty estimation (single forward pass)

foreach image $X_n \in U_{t-1}$ **do**

 Compute instance-level uncertainties $\{E_i\}$ using Equations (1)–(6);

 Compute image-level uncertainty I_{X_n} using Equation (7);

 // Batch selection

 Rank all images in U_{t-1} once by I_{X_n} ;

 Select the top- N images as a batch B_t ;

 // Dataset update and retraining

 Update labeled set $L_t = L_{t-1} \cup B_t$;

 Update unlabeled pool $U_t = U_{t-1} \setminus B_t$;

 Retrain detector $\mathcal{M}_t$ on L_t ;

4. Experiments

In this section, we present experimental results on MS COCO, PASCAL VOC 2007, and PASCAL VOC 2012 under widely adopted active learning protocols.

At each active learning cycle, uncertainty scores are computed once using the current model, images are ranked accordingly, and the top- N images are selected in a batch manner before updating the labeled set.

4.1. Evaluation Protocol and Benchmark Choice

We follow widely adopted experimental protocols in detection-oriented active learning studies (e.g., [7,20]) on PASCAL VOC and MS COCO to ensure direct comparability with representative baselines.

Although ALBench [33] provides a standardized evaluation protocol for active learning, most detection-oriented active learning methods in the literature are still reported under conventional VOC/COCO incremental annotation settings rather than the ALBench configuration.

To ensure direct comparability with prior detection-based active learning approaches—particularly those evaluated using Faster R-CNN and RetinaNet under fixed incremental acquisition cycles—we adopt the widely used VOC and COCO protocols in this work. This choice maintains methodological consistency with the majority of existing detection-oriented active learning studies.

4.2. Datasets and Detectors

To evaluate our method, we conducted experiments on three benchmark datasets: MS COCO [38], Pascal VOC 2007, and Pascal VOC 2012 [39], containing 80 and 20 object categories, respectively. For VOC 2012 and COCO, we used the train set for training and the validation set for testing, while for VOC 2007 we employed the trainval set for training and the test set for evaluation. Our experimental protocol began with training a baseline model on a small subset of labeled data, followed by cyclical sample selection from the remaining pool for active learning.

We implemented two representative CNN-based object detection frameworks: the two-stage Faster R-CNN (FRCNN) [35] and one-stage RetinaNet [36], both with ResNet-50-FPN backbones using default Torchvision configurations. For VOC experiments, we initialized FRCNN with 500 randomly selected images and allocated a budget of 500 images per active learning cycle, while RetinaNet required 1000 initial images for stable training. On COCO, we followed [20]’s protocol with 5000 initialization samples and 1000 budget per cycle. All models were trained for 20 epochs per cycle. Evaluation followed standard benchmarks: mAP@0.5 IoU for VOC and mAP@[0.5:0.95] for COCO.

4.3. Implementation Details

In practice, we construct each instance bag by grouping pre-NMS predictions that correspond to the same underlying candidate (e.g., the same RoI in two-stage detectors or the same anchor location in one-stage detectors), and retain the top- K instances ranked by post-softmax classification probability. We set $K = 10$ in all experiments as a trade-off between estimation stability and computational efficiency and discard instance bags with fewer than K predictions.

All experiments were conducted on a Linux server with a single 16 GB V100 GPU, using a batch size of 8 for both training and inference. Input images were resized to 800 pixels on the shorter side. The computational time for uncertainty estimation was measured as the wall-clock time averaged over 5 runs, including only the forward pass and uncertainty calculation (excluding data loading and model training). We set $K = 10$ in all experiments. Sensitivity analysis with respect to K is reported in Section 4.7.

4.4. Main Results

We evaluated our method against several representative active learning approaches, including Random, SSM [19], LS + C/LT + C [20], CALD [7], and LL4AL [8]. Figure 4 summarizes the results across active learning cycles on PASCAL VOC 2007, 2012, and MS COCO using both Faster R-CNN and RetinaNet. Uncertainty3D achieves competitive mAP compared to the baselines.

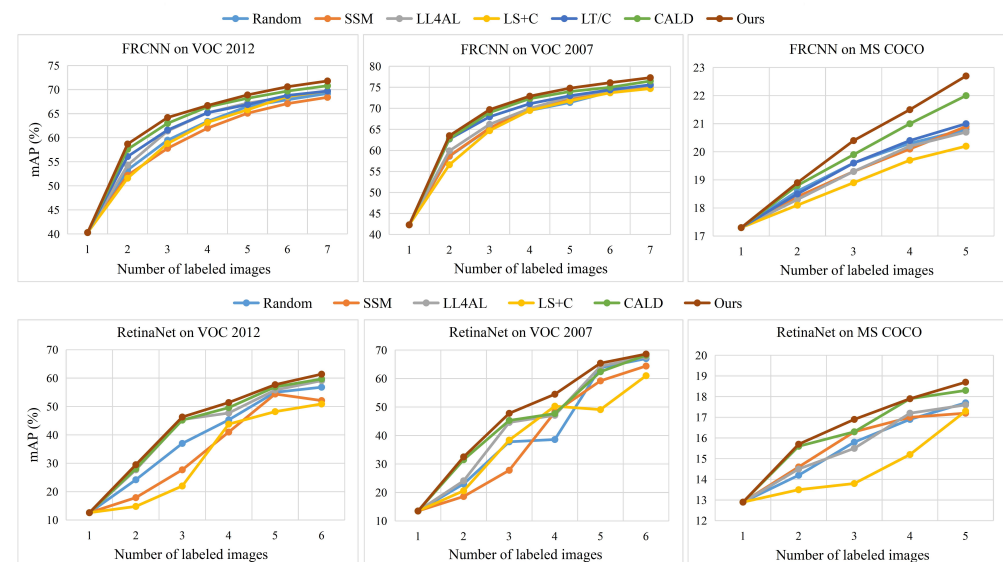


Figure 4. Performance comparison across active learning cycles on PASCAL VOC 2007, 2012, and MS COCO using Faster R-CNN and RetinaNet as base detectors. Compared methods include Random, SSM [19], LS + C/LT + C [20], CALD [7], and LL4AL [8].

For clarity, detailed numerical mAP values for all datasets, detectors, and active learning cycles are provided in the Appendix A (see Tables A1–A6). These tables report the exact numerical values corresponding to the learning curves shown in Figure 4 and are directly obtained from the original experimental logs.

4.5. Relation to Recent Instance-Aware Active Learning Methods

Recent pipeline-level active learning methods such as PPAL [31] and MI-AOD [32] explore instance-aware selection strategies in object detection and report strong performance under their respective experimental protocols.

These works use different annotation budgets, detector implementations, and data partitions, so direct numerical comparison is not straightforward. We therefore use a fixed detector setup and a consistent VOC/COCO-style AL protocol so that all baselines are comparable.

4.6. Multi-Seed Stability Analysis

We run multi-seed evaluation on PASCAL VOC 2007 with Faster R-CNN across all AL cycles. The full AL process is repeated with three random seeds, covering all cycles from C1 to C7 under the same initialization size and batch budget. For each cycle, we report the mean and standard deviation of mAP@0.5.

As shown in Table 1, the standard deviation remains consistently low across all active learning cycles, typically below 0.4 mAP. Ranking behavior is thus stable across random initializations for this detector–dataset pair. We report full multi-seed evaluation on a representative detector–dataset configuration (VOC2007 with Faster R-CNN), covering all active learning cycles. This configuration captures a typical two-stage detection setting and a moderate-scale dataset. The consistently low variance across cycles indicates stable

ranking behavior of the proposed uncertainty proxy. Since the uncertainty estimation operates in a deterministic single-pass manner without additional stochastic components, the observed stability is consistent with the design of the method.

Table 1. Multi-seed evaluation across all active learning cycles on PASCAL VOC 2007 (Faster R-CNN). Results are reported as mean $\pm$ standard deviation over three random seeds.

Cycle	mAP@0.5 (Mean $\pm$ Std)
C1	42.3 $\pm$ 0.0
C2	63.4 $\pm$ 0.3
C3	69.6 $\pm$ 0.3
C4	72.8 $\pm$ 0.2
C5	74.7 $\pm$ 0.3
C6	76.0 $\pm$ 0.2
C7	77.2 $\pm$ 0.3

4.7. Sensitivity to the Bag Size K

We vary the bag size $K \in \{5, 10, 15\}$ with the same detector and training protocol to check sensitivity. We report representative mAP values and the fraction of instance bags discarded due to having fewer than K predictions (i.e., $|\mathcal{R}_i| < K$), averaged over the unlabeled pool.

We vary the bag size $K \in \{5, 10, 15\}$ with the same detector and training protocol to check sensitivity. Representative mAP values and the fraction of instance bags discarded due to having fewer than K predictions (i.e., $|\mathcal{R}_i| < K$), averaged over the unlabeled pool, are reported in Table 2.

Table 2. Sensitivity analysis with respect to the bag size K . Results are reported for a representative setting (Faster R-CNN on PASCAL VOC 2007). Discard rate denotes the percentage of instance bags with fewer than K predictions.

K	Final mAP@0.5	Discard Rate (%)
5	72.7	0.9
10	73.0	1.5
15	72.6	3.1

Moderate changes in K leave AL behavior largely unchanged; the same trend holds in other setups (Faster R-CNN on VOC 2007 is shown). Larger K raises the discard rate and cost without clear gains, while smaller values may lead to less stable uncertainty estimates due to reduced instance diversity. Based on this trade-off, we use $K = 10$ in all experiments.

4.8. Sensitivity to the Weighting Parameters α and β

The weights α and β in Equation (6) balance geometric and semantic uncertainty. We sweep a range of (α, β) to check robustness.

4.8.1. VOC 2007

We evaluate representative configurations including (1.0, 0.0), (0.8, 0.2), (0.6, 0.4), (0.4, 0.6), and (0.0, 1.0) using Faster R-CNN on PASCAL VOC 2007. Results for the final active learning cycle are shown in Table 3.

Table 3. Extended sensitivity analysis on PASCAL VOC 2007 (Faster R-CNN).

α	β	Final mAP@0.5
1.0	0.0	76.8
0.8	0.2	77.1
0.6	0.4	77.3
0.4	0.6	77.2
0.0	1.0	76.6

Note: The best performance is highlighted in bold.

4.8.2. VOC 2012

We further evaluate representative configurations on PASCAL VOC 2012 to assess cross-dataset consistency. Results for the final cycle are reported in Table 4.

Table 4. Cross-dataset sensitivity on PASCAL VOC 2012 (Faster R-CNN).

α	β	Final mAP@0.5
0.8	0.2	71.6
0.6	0.4	71.8
0.3	0.7	71.5

Note: The best performance is highlighted in bold.

Across both datasets, performance remains stable across a broad range of weighting configurations, with only minor variations in mAP. The default ($\alpha = 0.6, \beta = 0.4$) gives a good balance and does not require dataset-specific tuning.

4.9. Category Uncertainty Formulation Analysis

We compare the maximum-entropy bag-level category uncertainty (Equation (4)) with two alternatives: (1) mean entropy across instances, and (2) dispersion (standard deviation) of entropy within the bag.

All experiments are conducted under the same configuration (Faster R-CNN on PASCAL VOC 2007, $K = 10$, $\alpha = 0.6$, $\beta = 0.4$). Only the final active learning cycle is reported to provide a lightweight comparison. For consistency, the configuration ($\alpha = 0.6$, $\beta = 0.4$) corresponds exactly to the default setting used in Appendix A Table A2, where the final mAP is 77.3.

The maximum entropy formulation gives slightly better and more stable results (Table 5); we use it in all experiments.

Table 5. Comparison of different category uncertainty formulations on PASCAL VOC 2007 (Faster R-CNN).

Formulation	Final mAP@0.5
Mean entropy	77.1
Entropy dispersion (std)	77.0
Max entropy (ours)	77.3

Note: The best performance is highlighted in bold.

4.10. Cross-Dataset Parameter Consistency

Two (α, β) settings are also evaluated on Pascal VOC 2012 (Faster R-CNN, final cycle only) to check cross-dataset consistency.

Small variation across configurations (Table 6) indicates that the same weights work across datasets.

Table 6. Cross-dataset consistency of weighting parameters on Pascal VOC 2012 (Faster R-CNN).

α	β	Final mAP@0.5
0.6	0.4	71.8
0.4	0.6	71.6

4.11. Compatibility with Diversity-Based Selection

Modern active learning pipelines often incorporate diversity-aware mechanisms to reduce redundancy within selected batches. Since Uncertainty3D focuses on uncertainty-driven ranking without explicitly modeling diversity, we conduct a lightweight experiment to assess its compatibility with diversity-based refinement.

On PASCAL VOC 2007 (Faster R-CNN) we use a two-stage strategy: (1) images are first ranked by Uncertainty3D; (2) from the top-2*N* candidates, k-Means++ clustering is applied on backbone pooled features to select the final batch of *N* samples. This strategy introduces diversity without modifying the underlying uncertainty formulation.

The results are shown in Table 7. Incorporating a simple diversity refinement yields consistent but moderate improvements across evaluated cycles, while preserving the overall performance trend. Uncertainty ranking drives most of the gain; diversity refinement adds a modest improvement. Uncertainty3D can be combined with diversity-aware selection when needed.

Table 7. Uncertainty3D with and without diversity refinement (PASCAL VOC 2007, Faster R-CNN).

Method	C2 mAP	C4 mAP
Uncertainty3D	63.5	72.9
Uncertainty3D + k-Means++	63.8	73.3

4.12. Method Analysis

Detector-wise Behavior Analysis.

Two-Stage Detector Performance. With FRCNN, Uncertainty3D improves over random selection (5.4% and 4.9% mAP on VOC 2012 and 2007) and over CALD (1.1% and 0.7% in the first cycle). Classification-based methods such as LL4AL fit FRCNN poorly and trail by 4.4% on VOC 2012. As discussed in [7], this stems from FRCNN’s complex pipeline where region proposal generation precedes fine-grained prediction—an architecture that widens the gap between detection and classification tasks. LT/C exhibits additional limitations: (1) lack of generalizability across detector architectures, and (2) inability to process classification information effectively since FRCNN’s first stage does not produce class-specific scores.

One-Stage Detector Advantages. Our method shows reliable behavior with RetinaNet, attributable to our focus on identifying images containing geometrically unstable instances. This strategy compensates for the one-stage detector’s relatively coarse proposal generation by improving positioning accuracy through targeted annotation of challenging cases.

Computational Efficiency.

Uncertainty3D operates in a single forward pass and introduces negligible overhead beyond standard detection inference. Table 8 reports the average wall-clock time per active learning (AL) cycle (PASCAL VOC 2007, Faster R-CNN, averaged over 5 runs) for score-based uncertainty, LL4AL [8], Uncertainty3D, and CALD [7].

Table 8. Average wall-clock time per active learning cycle (PASCAL VOC 2007, Faster R-CNN, averaged over 5 runs). Time includes only forward inference and uncertainty computation (excluding data loading and model retraining).

Method	Time per Cycle (min)
Score-based (confidence)	9.8
LL4AL [8]	10.2
Uncertainty3D (Ours)	10.4
CALD [7]	36.1

As shown in Table 8, score-based uncertainty, LL4AL, and Uncertainty3D exhibit comparable computational cost, since all three methods rely on a single forward pass without repeated inference. In contrast, CALD requires multiple augmented passes per image, resulting in approximately a $3.5\times$ increase in wall-clock time.

Figure 5 further illustrates the per-cycle time evolution following the experimental protocol of [7]. Uncertainty3D requires approximately 10.4 min per cycle, which is consistent with other single-pass baselines and substantially lower than the 36.1 min required by CALD. This efficiency gap remains stable across subsequent active learning cycles, demonstrating the scalability of the proposed approach as the labeled set grows.

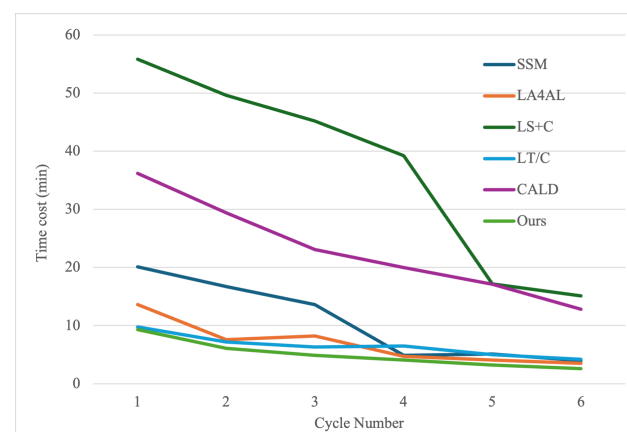


Figure 5. Comparison of the wall-clock time costs for uncertainty estimation across active learning cycles using Faster R-CNN on Pascal VOC 2007 (averaged over 5 runs; excluding data loading and model training). The compared methods include SSM [19], LA4AL [8], LS + C [20], LT/C [20], and CALD [7].

All timing results are reported as wall-clock measurements averaged over five runs on a single 16 GB V100 GPU. Overall, Uncertainty3D matches the computational cost of existing single-pass uncertainty baselines while significantly reducing overhead compared to augmentation-based methods, making it suitable for iterative AL loops under practical resource constraints.

Visualization Analysis.

Figure 6 shows qualitative examples (post-softmax probability > 0.1). Images exhibiting larger variability in predicted position, size, and category correspond to more challenging detection scenarios. For example, the third column shows greater spatial dispersion due to background similarity between objects and surroundings, while predictions for simpler objects (e.g., isolated persons or vehicles) remain more consistent. These observations demonstrate that increased instance-level inconsistency reliably reflects detection difficulty, thereby validating the effectiveness of our uncertainty-driven sample selection strategy.

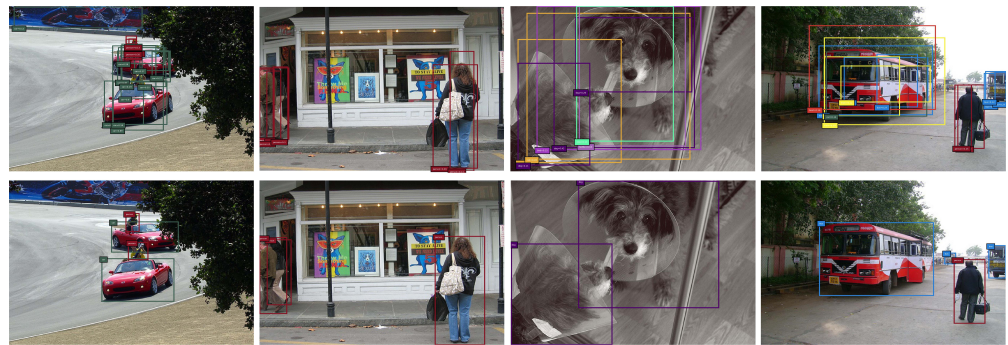


Figure 6. The first row displays example images from the unlabeled pool, showcasing predicted localization and post-softmax class probabilities. The second row presents corresponding ground-truth bounding boxes for the images in the first row.

This pattern of prediction variability extends beyond inter-image comparisons to intra-image variations as well. For instance, in the fourth column, predictions for the large red bus show lower consistency compared to those for the person or car on the right side, primarily due to the bus's more complex background structure. These observations collectively demonstrate that as object identification becomes more challenging—whether due to low contrast with the background or complex scenes—the uncertainties in localization, size, and category predictions increase correspondingly. Our comprehensive approach to quantifying these uncertainties enables more effective selection of informative images for annotation, ultimately improving model training efficiency.

5. Discussion

Uncertainty3D operates directly on proposal-consistent pre-NMS prediction groups naturally produced by standard two-stage and one-stage CNN detectors.

5.1. Practical Considerations for Deployment

Uncertainty3D is motivated by deployment constraints such as limited annotation budgets and computational cost. Evaluation in this paper uses public benchmarks (PASCAL VOC and MS COCO). This study evaluates Uncertainty3D within standard CNN-based detection frameworks. Exploring alternative architectural paradigms constitutes a separate line of investigation.

5.2. Limitations and Future Directions

Detector Scope. We evaluate Faster R-CNN and RetinaNet as representative CNN-based detection pipelines. Investigating fundamentally different architectural paradigms would require separate formulation and validation beyond the scope of this study.

Uncertainty and Diversity as Complementary Design Dimensions. Uncertainty3D is intentionally designed as a lightweight uncertainty-ranking component rather than a fully coupled pipeline that jointly optimizes informativeness and diversity. Uncertainty modeling and redundancy reduction address complementary aspects of sample selection. By decoupling these components, the proposed formulation remains simple, detector-compatible, and computationally efficient. The k-means++ refinement experiment demonstrates that diversity-aware filtering can be incorporated as an independent post-selection step when desired.

Batch Redundancy. We analyze batch-level redundancy by measuring the mean maximum pairwise IoU among selected images (C2–C5; C1 is random, later cycles near saturation). As shown in Table 9, redundancy remains stable across cycles without an increasing trend, suggesting that uncertainty-driven selection does not lead to progressive redundancy accumulation.

Table 9. Mean batch-level redundancy measured by maximum pairwise IoU among selected images across active learning cycles (PASCAL VOC 2007, Faster R-CNN).

Cycle	Mean Max IoU
C2	0.13
C3	0.14
C4	0.15
C5	0.14

Generalization and Standardized Evaluation. Although validated under widely adopted VOC/COCO active learning protocols, further evaluation under unified toolkits such as ALBench, as well as under domain shift or temporal detection scenarios, would provide additional insight into cross-domain robustness and extended applicability.

6. Conclusions

This paper presented Uncertainty3D, a lightweight active learning framework designed for CNN-based object detection. The method leverages proposal-consistent pre-NMS prediction groups naturally produced by standard two-stage and one-stage CNN detectors and constructs a tri-dimensional uncertainty proxy that captures instability in position, scale, and category.

Uncertainty3D operates in a single forward pass without architectural modification or repeated augmented inference, making it suitable for iterative active learning cycles under practical annotation and computational constraints. Experiments on PASCAL VOC and MS COCO using Faster R-CNN and RetinaNet demonstrate competitive detection performance while reducing uncertainty estimation cost by approximately 3–4× compared to augmentation-based baselines.

The results indicate that exploiting internal prediction inconsistency within CNN detection pipelines provides an effective and computationally efficient ranking mechanism for active learning sample selection.

Author Contributions: Conceptualization and methodology, Z.Z. and W.M.; software and experiments, Q.L.; writing—original draft preparation, Q.L. and W.M.; writing—review and editing, Z.Z. and C.X.; validation, Q.L. and C.X.; supervision, C.X.; project administration, Z.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, Grant No. 62272024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly available benchmark datasets. PASCAL VOC (2007 and 2012) and MS COCO can be obtained from their respective official project websites.

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A Tables A1–A6 report the complete numerical results corresponding to the learning curves shown in Figure 4 of the main paper. For each experimental setting, the

tables provide exact mAP values at every active learning cycle, enabling precise quantitative comparison and reproducibility.

Tables A1, A2 and A3 present results obtained using Faster R-CNN on PASCAL VOC 2012, PASCAL VOC 2007, and MS COCO, respectively. Tables A4–A6 report the corresponding results using RetinaNet on the same datasets.

For consistency with Figure 4, active learning cycles are denoted as C1, C2, ..., where each cycle corresponds to a fixed annotation budget increment as described in Section 4.

All values reported in Appendix A Tables A1–A6 are directly obtained from the original experimental logs used to generate Figure 4. The main paper focuses on visual trends and representative summaries, while the Appendix A provides the complete numerical results for detailed inspection and reproducibility.

Table A1. Full mAP@0.5 values across active learning cycles (C1–C7) using Faster R-CNN on PASCAL VOC 2012.

Method	C1	C2	C3	C4	C5	C6	C7
Random	40.3	53.3	59.5	63.4	66.5	67.9	69.2
SSM [19]	40.3	52.2	57.8	62.0	65.1	67.1	68.4
LL4AL [8]	40.3	54.3	61.4	65.2	67.2	68.5	69.7
LS + C [20]	40.3	51.6	58.8	63.2	65.8	68.6	69.7
LT/C [20]	40.3	56.1	61.6	65.2	66.9	68.8	69.7
CALD [7]	40.3	57.6	63.0	66.4	68.2	69.7	70.8
Ours	40.3	58.7	64.2	66.7	68.9	70.6	71.8

Table A2. Full mAP@0.5 values across active learning cycles (C1–C7) using Faster R-CNN on PASCAL VOC 2007.

Method	C1	C2	C3	C4	C5	C6	C7
Random	42.3	58.6	65.2	69.5	71.4	73.9	75.5
SSM [19]	42.3	58.7	65.3	70.0	72.1	74.2	75.4
LL4AL [8]	42.3	59.9	66.2	70.0	72.7	74.3	75.5
LS + C [20]	42.3	56.6	64.7	69.5	71.8	73.7	74.7
LT/C [20]	42.3	62.7	68.0	71.1	73.0	74.4	75.6
CALD [7]	42.3	62.8	69.0	72.3	74.0	75.0	76.5
Ours	42.3	63.5	69.7	72.9	74.8	76.1	77.3

Table A3. Full mAP@[0.5:0.95] values across active learning cycles (C1–C5) using Faster R-CNN on MS COCO.

Method	C1	C2	C3	C4	C5
Random	17.3	18.6	19.6	20.3	20.8
SSM [19]	17.3	18.4	19.3	20.1	20.9
LL4AL [8]	17.3	18.3	19.3	20.2	20.7
LS + C [20]	17.3	18.1	18.9	19.7	20.2
LT/C [20]	17.3	18.5	19.6	20.4	21.0
CALD [7]	17.3	18.8	19.9	21.0	22.0
Ours	17.3	18.9	20.4	21.5	22.7

Table A4. Full mAP@0.5 values across active learning cycles (C1–C6) using RetinaNet on PASCAL VOC 2012.

Method	C1	C2	C3	C4	C5	C6
Random	12.6	24.2	37.0	45.3	55.0	56.8
SSM [19]	12.6	17.9	27.7	41.0	54.4	52.1
LL4AL [8]	12.6	27.7	45.4	47.7	55.8	59.1
LS + C [20]	12.6	14.8	22.0	43.8	48.2	50.9
CALD [7]	12.6	28.1	45.2	49.6	56.9	59.7
Ours	12.6	29.5	46.3	51.4	57.7	61.4

Table A5. Full mAP@0.5 values across active learning cycles (C1–C6) using RetinaNet on PASCAL VOC 2007.

Method	C1	C2	C3	C4	C5	C6
Random	13.5	23.0	37.8	38.6	63.8	67.0
SSM [19]	13.5	18.6	27.8	48.2	59.2	64.4
LL4AL [8]	13.5	24.1	44.6	47.1	64.3	67.7
LS + C [20]	13.5	20.6	38.4	50.3	49.1	61.0
CALD [7]	13.5	31.5	45.3	47.7	62.4	68.2
Ours	13.5	32.5	47.8	54.5	65.4	68.6

Table A6. Full mAP@0.5 values across active learning cycles (C1–C5) using RetinaNet on MS COCO.

Method	C1	C2	C3	C4	C5
Random	12.9	14.2	15.8	16.9	17.7
SSM [19]	12.9	14.6	16.3	17.0	17.2
LL4AL [8]	12.9	14.5	15.5	17.2	17.6
LS + C [20]	12.9	13.5	13.8	15.2	17.3
CALD [7]	12.9	15.6	16.3	17.9	18.3
Ours	12.9	15.7	16.9	17.9	18.7

Table A7. (PASCAL VOC 2007, Faster R-CNN, $K = 10$). The original row corresponds exactly to the “Ours” results reported in Appendix A Table A2 under the same experimental setting. All other components, including size and category uncertainty, aggregation strategy, training schedule, and active learning protocol, remain identical to the main experiments.

Position Uncertainty Variant	C1	C2	C3	C4	C5	C6	C7
Original σ_i	42.3	63.5	69.7	72.9	74.8	76.1	77.3
Normalized $\sigma_i / (w_i + h_i)$	42.2	63.4	69.6	72.8	74.7	76.0	77.2

References

1. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2012; pp. 1097–1105.
2. Ma, W.; Zhang, Z.; Xu, Q. An Automatic Scheme for Optimizing the Size of Deep Networks. In *Proceedings of the 3rd International Conference on Signal Processing and Machine Learning*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 21–27.
3. Ma, W.; Yuan, X.; Zhu, J. A Video Image Processing Method for Continuous Object Detection. In *8th International Conference on Computer and Communication Systems*; IEEE: Piscataway, NJ, USA, 2023; pp. 1015–1020.
4. Aggarwal, C.C.; Kong, X.; Gu, Q. Active Learning: A Survey. In *Data Classification: Algorithms and Applications*; Aggarwal, C.C., Ed.; CRC Press: Boca Raton, FL, USA, 2014; pp. 571–605. [\[CrossRef\]](#)
5. Aghdam, H.H.; Gonzalez-Garcia, A. Active Learning for Deep Detection Neural Networks. In *IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019; pp. 3671–3679. [\[CrossRef\]](#)
6. Agarwal, S.; Arora, H.; Anand, S.; Arora, C. Contextual Diversity for Active Learning. In *Computer Vision—ECCV 2020*; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 137–153.

7. Yu, W.; Zhu, S.; Yang, T.; Chen, C. Consistency-Based Active Learning for Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: Piscataway, NJ, USA, 2022; pp. 3950–3959. [\[CrossRef\]](#)
8. Yoo, D.; Kweon, I.S. Learning Loss for Active Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 93–102. [\[CrossRef\]](#)
9. Zhang, Z.; Ma, W.; Yuan, X.; Hao, Y.; Guo, M.; Tang, H.; Zhou, Z.; Yao, Z. Instance-Aware Uncertainty for Active Learning in Object Detection. In *2024 IEEE International Conference on Image Processing (ICIP)*; IEEE: Piscataway, NJ, USA, 2024; pp. 298–304. [\[CrossRef\]](#)
10. Huijser, M.; van Gemert, J.C. Active Decision Boundary Annotation with Deep Generative Models. In *IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2017; pp. 5296–5305. [\[CrossRef\]](#)
11. Tegen, A.; Davidsson, P.; Persson, J.A. Active Learning and Machine Teaching for Online Learning. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*; IEEE: Piscataway, NJ, USA, 2021; pp. 1215–1220. [\[CrossRef\]](#)
12. Chen, X.; Wujek, B. A Unified Framework for Automatic Distributed Active Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 9774–9786. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Baram, Y.; El-Yaniv, R.; Luz, K. Online Choice of Active Learning Algorithms. *J. Mach. Learn. Res.* **2004**, *5*, 255–291.
14. Huang, S.J.; Jin, R.; Zhou, Z.H. Active Learning by Querying Informative and Representative Examples. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1936–1949. [\[CrossRef\]](#) [\[PubMed\]](#)
15. Sener, O.; Savarese, S. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *Proceedings of the International Conference on Learning Representations*, Vancouver, BC, Canada, 30 April–3 May 2018.
16. Ash, J.T.; Zhang, C.; Krishnamurthy, A.; Langford, J.; Agarwal, A. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *Proceedings of the International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia, 26–30 April 2020.
17. Beluch, W.H.; Genewein, T.; Nurnberger, A.; Kohler, J.M. The Power of Ensembles for Active Learning in Image Classification. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 9368–9377. [\[CrossRef\]](#)
18. Geifman, Y.; Uziel, G.; El-Yaniv, R. Bias-Reduced Uncertainty Estimation for Deep Neural Classifiers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 6–9 May 2019.
19. Wang, K.; Yan, X.; Zhang, D.; Zhang, L.; Lin, L. Towards Human-Machine Cooperation: Self-supervised Sample Mining for Object Detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 1605–1613. [\[CrossRef\]](#)
20. Kao, C.C.; Lee, T.Y.; Sen, P.; Liu, M.Y. Localization-Aware Active Learning for Object Detection. In *Computer Vision—ACCV 2018*; Jawahar, C., Li, H., Mori, G., Schindler, K., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 506–522.
21. Dong, J.; Zhang, L. Multibox Sample Selection for Active Object Detection. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*; IEEE: Piscataway, NJ, USA, 2023; pp. 2447–2452. [\[CrossRef\]](#)
22. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2017; pp. 5574–5584.
23. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *33rd International Conference on Machine Learning (ICML)*; PMLR: New York, NY, USA, 2016; pp. 1050–1059.
24. Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2017; pp. 6402–6413.
25. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Advances in Neural Information Processing Systems (NeurIPS)*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2019.
26. Chen, S.; Yang, Y.; Hua, Y. Semi-Supervised Active Learning for Object Detection. *Electronics* **2023**, *12*, 375. [\[CrossRef\]](#)
27. Liou, Y.S.; Wu, T.H.; Yeh, J.F.; Chen, W.C.; Hsu, W.H. MuRAL: Multi-scale Region-Based Active Learning for Object Detection. *arXiv* **2023**, arXiv:2303.16637.
28. Hekimoglu, A.; Brucker, A.; Kayali, A.K.; Schmidt, M.; Marcos-Ramiro, A. Active Learning for Object Detection with Non-Redundant Informative Sampling. *arXiv* **2023**, arXiv:2307.08414. [\[CrossRef\]](#)
29. Schubert, M.; Riedlinger, T.; Kahl, K.; Rottmann, M. Deep Active Learning with Noisy Oracle in Object Detection. *arXiv* **2023**, arXiv:2310.00372. [\[CrossRef\]](#)
30. Na, J.; De-Silva, V. Uncertainty Aware Active Learning for Reconfiguration of Pre-Trained Deep Object-Detection Networks for New Target Domains. In *2023 IEEE IAS Global Conference on Emerging Technologies (GlobConET)*; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7. [\[CrossRef\]](#)
31. Yang, C.; Huang, L.; Crowley, E.J. Plug and Play Active Learning for Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2024. [\[CrossRef\]](#)

32. Yuan, T.; Wan, F.; Fu, M.; Liu, J.; Xu, S. Multiple Instance Active Learning for Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2021. [[CrossRef](#)]
33. Feng, Z.; Zhang, S.; Takezoe, R.; Hu, W.; Chandraker, M.; Li, L.J.; Narayanan, V.K.; Wang, X. ALBench: A Framework for Evaluating Active Learning in Object Detection. *arXiv* **2022**, arXiv:2207.13339. [[CrossRef](#)]
34. Riedlinger, T.; Schubert, M.; Kahl, K.; Gottschalk, H.; Rottmann, M. Towards Rapid Prototyping and Comparability in Active Learning for Deep Object Detection. *arXiv* **2022**, arXiv:2212.10836. [[CrossRef](#)]
35. 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](#)] [[PubMed](#)]
36. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [[CrossRef](#)] [[PubMed](#)]
37. Hounsby, N.; Huszár, F.; Ghahramani, Z.; Lengyel, M. Bayesian Active Learning for Classification and Preference Learning. *arXiv* **2011**, arXiv:1112.5745. [[CrossRef](#)]
38. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2014; pp. 740–755. [[CrossRef](#)]
39. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The PASCAL Visual Object Classes (VOC) Challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [[CrossRef](#)]

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.