RETHINKING OBJECT DETECTION: A UNIFIED FRAMEWORK FOR DETECTION REFINEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in object detection have primarily focused on designing sophisticated neural networks to improve detection accuracy. Although these models have achieved remarkable performance, their outputs still contain a substantial number of false positives. Existing approaches primarily focus on improving the prediction capabilities of detection networks, while overlooking the importance of refining their outputs for practical applications. In this paper, we rethink object detection and introduce a unified refinement framework that can be seamlessly integrated into existing detection models. Specifically, our framework employs a group-aware refinement strategy to substantially reduce false positives while retaining valid detections. Furthermore, we establish theoretical guarantees for the proposed method. Extensive experiments demonstrate that our framework effectively reduces false positives while improving average precision.

1 INTRODUCTION

Object detection aims to predict the locations, sizes, and categories of critical objects in real-world scenes. It is a fundamental problem in computer vision with numerous downstream applications, such as autonomous driving (Yue et al., 2025; Huang et al., 2025), robot navigation (Wang & Zhao, 2025), and video surveillance (Zhang et al., 2025).

In the current literature (Zhang et al., 2024; Jin et al., 2025; Fan et al., 2025; Mao et al., 2023), object detection is typically formulated as

$$\mathcal{B} = f_{\text{det}}(\mathcal{I}_{\text{sensor}}), \tag{1}$$

where $\mathcal{B} = \{X_1, X_2, \ldots\}$ denotes the set of detected objects in the input sample $\mathcal{I}_{\text{sensor}}$. Each X_i represents a 3D cuboid or 2D bounding box with a confidence score $s_i \in [0, 1]$. The detector f_{det} maps sensory inputs $\mathcal{I}_{\text{sensor}}$, such as LiDAR point clouds or images, to candidate detections.

Currently, numerous powerful object detection networks (Wang et al., 2025; Jin et al., 2025; Li et al., 2024b;a) and sophisticated post-processing algorithms (Kumar et al., 2021; Shen et al., 2022) have been developed to tackle the object detection task. Although these methods have achieved remarkable performance, their outputs still contain a substantial number of false positives. To illustrate this issue, we examine the outputs of twelve representative object detectors. The results are visualized in Figure 1. As shown in the figure, a large number of false positives (red regions) are prevalent in both 3D and 2D detection results. For instance, in 3D object detection, more than 80% of the detections produced by PV-RCNN are false positives, which severely hinders the detector's practical applicability.

The results in Figure 1 highlight the need for an effective mechanism to suppress false positives, which we refer to as the decision-making process. By introducing a decision-making process, object detection can be reformulated as

$$\mathcal{B} = (\Phi \circ f_{\text{det}})(\mathcal{I}_{\text{sensor}}), \tag{2}$$

where Φ denotes the decision-making process defined as

$$\Phi(X) = \begin{cases} \text{true positive} & \text{if} \quad \varphi(X) \geq \tau_X \\ \text{false positive} & \text{if} \quad \varphi(X) < \tau_X \end{cases}, \tag{3}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

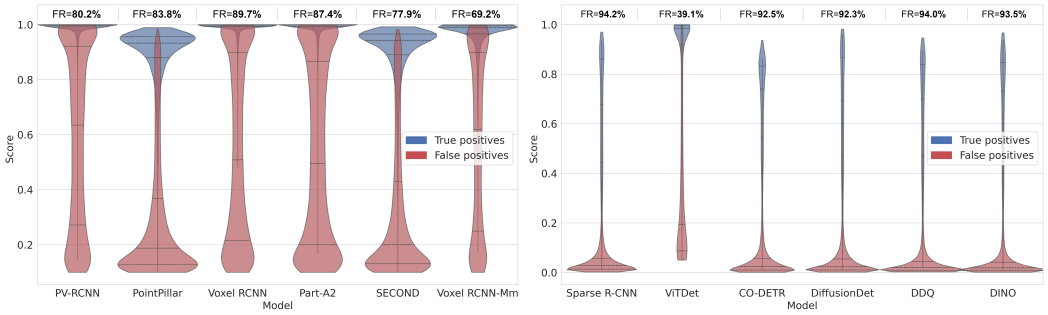


Figure 1: Output distributions of existing object detectors. 'FR' denotes the proportion of false positives among all positive detections. Top: 3D object detection on KITTI dataset; bottom: 2D object detection on COCO dataset.

where φ denotes a mapping function that maps each detected object into a decision space, and τ_X serves as a decision boundary. Despite its importance, the design of such decision-making processes remains underexplored.

We first analyze how the decision-making process affects detection results in Figure 2. As shown in the left subfigure, for the autonomous driving task, without a decision-making process, detector outputs contain numerous false positives. A well-designed decision-making process can effectively reduce false positives while preserving true positives , as shown in the middle subfigure. In contrast, a poorly designed process may aggressively reduce false positives at the cost of discarding true positives, as shown in the right subfigure, thereby degrading overall detection performance (i.e., average precision). Hence, the key challenge can be stated as

> *How to design a decision-making process that effectively reduces false positives while preserving or even improving overall detection performance?*

This paper aims to systematically study the above problem. Unlike previous works in object detection, we investigate the decision-making process from a statistical perspective. We reformulate object detection as an online multiple testing problem and propose a novel **O**nline **G**roup **T**esting framework (OGT) to effectively address it. Specifically, we construct valid p-values using a calibration set to provide principled decision evidence. Based on these p-values, we develop the Group-BH algorithm to control the error rate both within each input sample (2D image or 3D point cloud) and across the entire test set. Rigorous theoretical analyses are provided to establish the validity of our error control guarantees. Moreover, extensive experiments demonstrate the superior detection performance of our proposed OGT framework. For instance, in 3D object detection, OGT reduces 3419 false positives and improves average precision by 1.07% on the KITTI dataset compared with Voxel R-CNN. In 2D object detection, OGT reduces 3295 false positives and improves average precision by 1.01% on the COCO dataset compared with DINO. We summarize our contributions below:

- We innovatively frame object detection as an online multiple testing problem and propose an OGT framework to solve it. Our method is distribution-free, easy to implement, and does not rely on the information of false detections. It bridges the gap between practical detectors and statistical theory.

- We construct valid p-values for individual detections and prove that they satisfy the PRDS condition. Further, we extend the classical BH procedure to develop Group-BH and establish its FDR control under PRDS dependence.

- We conduct extensive experiments on both 3D and 2D object detection tasks to demonstrate the superior performance of our OGT framework. Results show that our method effectively reduces false positives while improving overall average precision. Moreover, it can be easily incorporated into existing detection pipelines without requiring complex modifications.

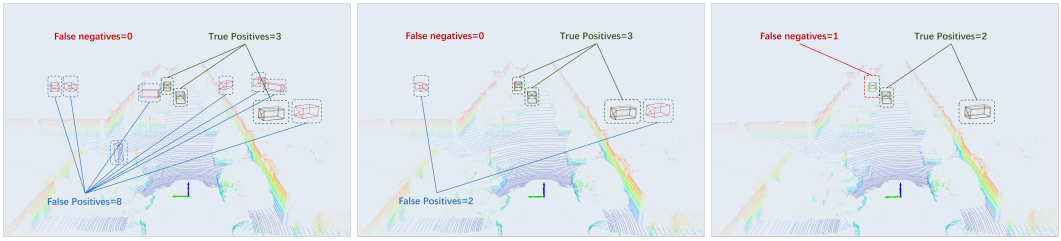


Figure 2: Effect of the decision-making process in 3D object detection task. Left: no decision-making; middle: well-designed decision-making; right: poorly designed decision-making. 'False Positives' and 'False Negatives' denote the numbers of falsely detected and missed objects, respectively (lower is better). 'True Positives' denotes the number of correctly detected objects (higher is better).

2 PRELIMINARIES

2.1 HYPOTHESIS TESTING

Hypothesis testing is a statistical tool used to decide whether there is enough evidence in data to support or reject a proposed claim about a population (Fisher, 1970; Neyman & Pearson, 1933). There are two complementary hypotheses called the null hypothesis H_0 and alternative hypothesis H_1. A hypothesis test of H_0 versus H_1 may make two types of errors. If H_0 is true but the test decides to reject H_0, then the test has made a type-I error. On the other hand, a type-II error occurs when H_0 is wrongly accepted. Hypothesis testing procedures are designed to control the type-I error probability at a specified level while minimizing type-II error probability (Lehmann & Romano, 2005), and their definition is given as

Definition 1 (Hypothesis testing procedure (Casella & Berger, 2002))**.** A hypothesis testing procedure is a rule that specifies (i) the sample values for which H_0 is accepted as true and (ii) the sample values for which H_0 is rejected and H_1 is accepted as true.

In statistics, the decision to accept or reject the null hypothesis is determined by the concept of p-value. Its general definition is presented as follows:

Definition 2 (P-value (Casella & Berger, 2002))**.** Given a sample set X, a statistic $p(X)$ is called p-value corresponding to the null hypothesis H_0, if $p(X)$ satisfies $0 \leq p(X) \leq 1$ and for every $0 \leq \alpha \leq 1$,

$$\mathbb{P}\left[p(X) \leq \alpha | H_0\right] \leq \alpha. \tag{4}$$

Small values of $p(X)$ usually give evidence that H_1 is true. Next, we introduce two extensions of traditional single hypothesis testing.

Multiple testing refers to the testing of more than one hypothesis at a time. In this case, it is often desirable to control the error rate for a whole family of simultaneous tests. The false discovery rate (FDR) is the most widely used criterion for controlling overall errors. In statistics, if one null hypothesis is rejected, it is said to make a discovery. The FDR is defined as the expected proportion of false discoveries among all rejected hypotheses.

Definition 3 (FDP & FDR (Benjamini & Hochberg, 1995))**.** Let V denote the number of rejected true null hypotheses and R denote the number of rejected hypotheses. The FDP is the proportion of errors committed by falsely rejecting null hypotheses and the FDR is the expected FDP:

$$\text{FDR} = \mathbb{E}\left[\text{FDP}\right] = \mathbb{E}\left[\frac{V}{\max(R, 1)}\right]. \tag{5}$$

Definition 3 indicates that if a procedure controls the FDR at a specified level α, then the expected proportion of false rejections among all rejected hypotheses does not exceed α.

Online testing considers hypotheses arriving in a stream. At each time point, the analyst decides whether to reject the current hypothesis based both on the evidence against it, and on the previous

rejection decisions (Robertson et al., 2023). Compared with multiple testing, online hypothesis testing sacrifices some statistical power (i.e., one minus the probability of type-II error) in exchange for real-time testing.

3 METHODOLOGY

In this section, we first present a novel formulation of object detection as an online multiple testing problem. We then introduce the online group testing (OGT) framework, which consists of two key components: constructing p-values to provide statistical evidence for each detection, and applying an online multiple testing procedure to control the FDR with rigorous theoretical guarantees.

3.1 A STATISTICAL PERSPECTIVE FOR OBJECT DETECTION

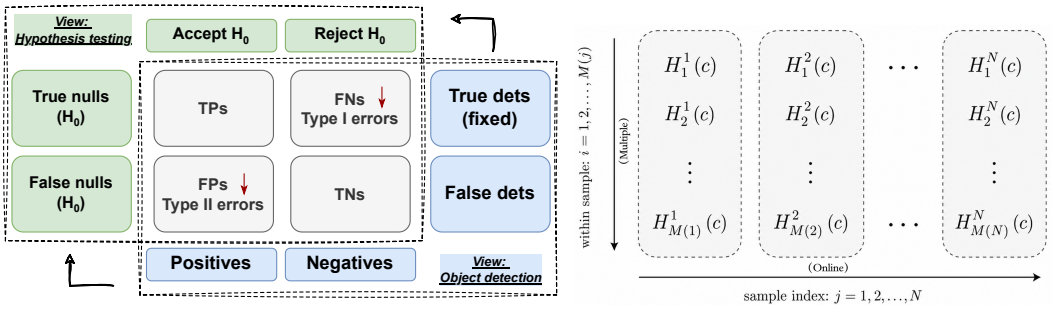


Figure 3: Connection between object detection (the blue region) and hypothesis testing (the green region).

Figure 4: Formulating object detection as an online multiple hypothesis testing problem.

Let $\mathcal{T}_{\text{test}}^{\text{raw}} = \{I_1, I_2, \cdots, I_N\}$ denote the test set of N samples, where each I_j represents the input sample at time step j. Let f_{det} denote a detector. For each sample $I_j \in \mathcal{T}_{\text{test}}^{\text{raw}}$, detector outputs a set of detections $f_{\text{det}}(I_j) = \{X_1^j, X_2^j, \cdots, X_{M(j)}^j\}$, where each detection X corresponds to a 3D cuboid or a 2D bounding box associated with a confidence score, and $M(j)$ denotes the number of detections for I_j. Let $\mathcal{A}_j = \{f_{\text{det}}(I_i) : i \in [1, j]\}$ denote the cumulative set of detections up to time step j. Our goal is to reduce false positives while preserving average precision at every step $j \in [1, N]$.

Let $\mathcal{C}$ denote the set of object categories appearing in test set. For each category $c \in \mathcal{C}$, let $\mathcal{D}_c$ denote the distribution of true detections. We frame two complementary hypotheses as

$$H_0 : X \sim \mathcal{D}_c \quad \text{versus} \quad H_1 : X \nsim \mathcal{D}_c. \tag{6}$$

Based on Equation (6), we illustrate the connection between object detection and hypothesis testing in Figure 3.

As shown in Figure 3, since precision is defined as $\frac{\text{TP}}{\text{TP+FP}}$, our goal requires reducing false positives while retaining true positives. From the hypothesis testing view, reducing false positives corresponds to minimizing type-II errors, whereas retaining true positives relies on controlling type-I errors because true positives and false negatives together form the fixed set of ground-truth detections. Therefore, the decision-making process naturally aligns with the principle of hypothesis testing: controlling type-I errors while minimizing type-II errors (Casella & Berger, 2002).

Motivated by this insight, we frame object detection as an online multiple hypothesis testing problem, as shown in Figure 4. Specifically, we first organize detections from adjacent samples into groups. Typically, detections from the same input sample are treated as a single group. Then the formulation operates along the following three dimensions.

A. For an individual detection $X_i^j \in f_{\text{det}}(I_j)$, as shown in the i-th row j-th column in Figure 4, we perform a hypothesis test $H_{i;0}^j : X_i^j \sim \mathcal{D}_c$ versus $H_{i;1}^j : X_i^j \nsim \mathcal{D}_c$ to determine whether X_i^j is a true positive.

B. For a group $\mathcal{G}_j = \{X : X \in f_{\text{det}}(I_j)\}$, as shown in the j-th column in Figure 4, we apply a multiple testing procedure to control the FDR at a specified level α_j within the group.

C. Across the entire test set $\mathcal{T}^{\text{test}} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots\}$, we employ an online hypothesis testing framework that dynamically allocates significance levels α_j to each group $\mathcal{G}_j$. This allows us to control the overall FDR across all groups while supporting real-time testing.

3.2 Construction of *p*-value

In object detection, each detection represents a 3D cuboid or 2D bounding box associated with confidence score $s \in [0, 1]$. Score s denotes the probability that the detection corresponds to a true object (Kong et al., 2025). Let $\mathcal{T}^{\text{cal}} = \mathcal{T}_0^{\text{cal}} \cup \mathcal{T}_1^{\text{cal}}$ denote all detections in the calibration set, where $\mathcal{T}_0^{\text{cal}}$ contains true detections and $\mathcal{T}_1^{\text{cal}}$ contains false detections. Let Z be a detection, and $s(Z)$ denote its confidence score.

Since false detections are diverse and not restricted to a specific category, the distribution of $\mathcal{T}_1^{\text{cal}}$ differs from that of false detections in test set. For instance, false detections in the calibration set may come from a category set $\mathcal{C}_1$, while those in the test set typically come from a different category set $\mathcal{C}_2$, with $\mathcal{C}_2 \neq \mathcal{C}_1$. Therefore, the ideal decision boundary should be distribution-free and should not rely on information from false detections. Accordingly, we construct *p*-values based on the distribution of true detections.

In classical hypothesis testing setting, the distribution of the test statistic, such as the confidence score of a true detection, is often known. In contrast, object detection operates under unknown and complex data distributions. Thus, we adopt a nonparametric approach to compute empirical *p*-values. Let $\mathcal{T}_0^{\text{cal}} = \{Z_1, Z_2, \ldots, Z_{n_0}\}$ be a calibration set of n_0 detections. We calculate the *p*-value for a detection X as

$$p = p(X) = \frac{\sum_{i=1}^{n_0} \mathbb{1}[s(Z_i) \leq s(X)] + 1}{n_0 + 1}. \tag{7}$$

According to prior work (Ma et al., 2024), the *p*-value in Equation (7) satisfies Definition 2 and is therefore a valid *p*-value. The complete computation steps are summarized in Algorithm 1.

Algorithm 1 Computation of *p*-values

1: **Input:** Calibration detections $\mathcal{T}_0^{\text{cal}} = \{Z_1, Z_2, \ldots, Z_n\}$; test detections $\mathcal{T}^{test} = \{X_1, X_2, \ldots\}$.
2: Extract the confidence scores of detections in $\mathcal{T}_0^{\text{cal}}$ and $\mathcal{T}^{test}$.
3: Initialize $\mathcal{Q} \leftarrow \emptyset$.
4: **for** each test score $s(X_j)$, $j = 1, 2, \ldots$ **do**
5: Calculate *p*-value $p_j = \frac{\sum_{i=1}^{n} \mathbb{1}[s(Z_i) \leq s(X_j)] + 1}{n+1}$.
6: Update $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{p_j\}$.
7: **end for**
8: **Output:** *p*-values $\mathcal{Q}$.

From Algorithm 1, the *p*-values are computed in a distribution-free manner. Small *p*-values provide evidence that the corresponding detections are false positives.

3.3 Online multiple testing procedure

In this section, we introduce Group-BH, an online multiple testing procedure. We begin by reviewing the classical Benjamini-Hochberg (BH) procedure, one of the most widely used algorithms for multiple testing (Benjamini & Hochberg, 1995; Benjamini & Yekutieli, 2001).

Definition 4 (BH Procedure)**.** Given the *p*-values $p_1, p_2, \ldots, p_n$ corresponding to the null hypotheses $H_{1;0}, H_{2;0}, \ldots, H_{n;0}$, let $p_{(i)}$ be the i-th order statistic from the smallest to the largest. For a pre-specified level $\alpha \in (0,\ 1)$, define

$$i^*_{BH} = \max\left\{ i \in [1 : n] : p_{(i)} \leq \frac{i}{n}\alpha \right\}, \tag{8}$$

then, the null hypothesis $H_{(i);0}$ is rejected if $i \leq i^*_{BH}$.

Definition 4 shows that the BH procedure requires access to all *p*-values simultaneously. At time step j, we group the detections from sample I_j as $\mathcal{G}_j = f_{\text{det}}(I_j) = \{X_1^j, X_2^j, \cdots, X_{M(j)}^j\}$ and apply the BH procedure to control the group-level FDR.

Although the BH procedure controls FDR within each group, it remains crucial to guarantee global FDR control across entire test set $\mathcal{T}^{test} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots\}$. In real-time setting, samples arrive sequentially, making it impractical to wait for all samples before testing. We therefore introduce an adaptive level assignment rule that dynamically allocates significance levels to incoming groups.

Definition 5 (Adaptive level assignment (Javanmard & Montanari, 2018; Zrnic et al., 2020))**.** Let T be the number of groups. The testing level for batch t is assigned as

$$\alpha_t = \alpha \frac{n_t + \sum_{s=1}^{t-1} R_s}{n_t \cdot T}, \qquad t \geq 1, \tag{9}$$

where n_t is the size of batch t and R_s is the number of rejections in batch s.

Definition 5 specifies a rule for allocating levels to each group. We combine this rule with the BH procedure to develop the Group-BH algorithm presented in Algorithm 2.

Algorithm 2 Group-BH algorithm

1: **Input:** *p*-value groups $\{\mathcal{Q}_t\}_{t=1}^T$; target level α.
2: Initialize $C_0 = 0$.
3: **for** $t = 1, \ldots, T$ **do**
4: Get assigned level α_t as Equation (9).
5: Obtain rejection set $\mathcal{R}_t$ for group $\mathcal{Q}_t$ using the BH procedure in Definition 4.
6: Update $C_t = C_{t-1} + |\mathcal{R}_t|$.
7: **end for**
8: **Output:** Rejection sets $\{\mathcal{R}_t\}_{t=1}^T$.

Algorithm 2 adaptively sets the level α_t based on the decisions of previous groups, and then applies the BH procedure at α_t. In the next section, we prove that Group-BH controls the FDR both within each group and across the entire test set when the *p*-values satisfy the PRDS condition. The complete OGT framework is summarized in Appendix A.

3.4 Theoretical guarantee for OGT

We investigate the theoretical results of our proposed OGT framework along two dimensions: controlling the FDR within each group and ensuring FDR control across the entire test set.

We begin by introducing a dependence condition known as positive regression dependence on a subset (PRDS) (Sarkar & Chang, 1997; Benjamini & Yekutieli, 2001).

Definition 6 (PRDS)**.** Let $X = \{X_1, \ldots, X_n\}$ be a set of variables. The X is PRDS on a subset $I_0 \subset \{1, \ldots, n\}$, if for any increasing set D, and for each $i \in I_0$, $\mathbb{P}[X \in D | X_i = x]$ is nondecreasing in x.

Let $\mathcal{T}^{test} = \{\mathcal{G}_1, \mathcal{G}_2, \ldots, \mathcal{G}_N\}$ denote the test set, where group $\mathcal{G}_j = \{p_1^j, p_2^j, \ldots, p_{M(j)}^j\}$. We show that *p*-values in $\mathcal{T}^{test}$ satisfy the PRDS condition.

Lemma 7. *Let $\{p_1^1, p_2^1, \ldots, p_{M(1)}^1, \ldots, p_{M(N)}^N\}$ denote the collection of p-values in $\mathcal{T}^{\text{test}}$ constructed by Algorithm 1. These p-values are PRDS on the subset corresponding to true detections.*

The proof of Lemma 7 is provided in Appendix B.1. Based on this property, we can apply the theoretical results of the BH procedure to each group.

Theorem 8 (FDR Control for BH Procedure Benjamini & Yekutieli (2001))**.** *If the p-values $p_1, p_2, \ldots, p_n$ satisfy PRDS condition on the subset corresponding to true null hypotheses, the BH procedure as Definition 4 applied at level $\alpha \in (0, 1)$ controls the FDR at level α:*

$$\text{FDR}_{\text{BH}} \leq \pi\alpha \leq \alpha, \tag{10}$$

where π is the proportion of true null hypotheses.

According to Theorem 8, the FDR of detections within each group can be controlled at a specified level. We further establish the FDR control across the entire test set.

Theorem 9 (Global FDR Control). *For $t = 1, \ldots, N$, let $\mathcal{A}_t = \{p_i^j : 1 \le j \le t,\ 1 \le i \le M(j)\}$ denote the cumulative collection of p-values. If $\mathcal{A}_N$ is PRDS on the subset corresponding to true detections, then the OGT framework in Algorithm 3, applied at target level $\alpha \in (0, 1)$, controls the global FDR at every time step:*

$$\mathrm{FDR}_{\mathrm{OGT}}^{t} \le \alpha, \qquad t = 1, \ldots, N. \tag{11}$$

The proof of Theorem 9 is shown in Appendix B.2. This result shows that, at every time step, OGT controls the global FDR over all detections observed so far at a specified level.

4 Experiments

We conduct experiments on 3D and 2D object detection to demonstrate the superior decision-making capability of our OGT framework. Ablations analyze the impact of significance level and calibration set size on OGT. We also compare OGT with post-hoc thresholding methods. Due to space limits, ablation and post-hoc thresholding results are provided in Appendices C.3 and C.4, respectively.

4.1 Experimental settings

Baselines. We select sixteen widely used object detection models as baselines: eight for the 2D task and eight for the 3D task. For 3D object detection task, we include Part-A2 (Shi et al., 2020b), PointPillar (Lang et al., 2019), PointR (Shi et al., 2019), PV-R (Shi et al., 2020a), GBlobs (Malić et al., 2025) (uses Part-A2 as backbone), SECOND (Yan et al., 2018), Voxel R (Deng et al., 2021), and Voxel R-Mm (Team, 2020). For 2D object detection task, we adopt YOLOX-x (Ge et al., 2021), ViTDet (Li et al., 2022), CO-DETR (Zong et al., 2023), DiffusionDet (Chen et al., 2023), DDQ (Zhang et al., 2023b), DINO (Zhang et al., 2023a), Sparse R (Sun et al., 2021), and RT-DETRv3 (Wang et al., 2025).

Datasets. For 3D object detection, we conduct experiments on the KITTI dataset (Geiger et al., 2012), which contains 14999 Velodyne point clouds and corresponding camera data across three categories. For 2D object detection, we use the COCO dataset (Lin et al., 2014), which comprises 163957 images spanning 80 categories.

Evaluation Metrics. We use precision (Pr), recall (Rec), the number of false positives (FPs), the number of false negatives (FNs), and average precision (AP refers to COCO AP for 2D tasks and AP R40 for 3D tasks) to evaluate the detection performance and decision-making capability.

Implementation Details. Implementations follow OpenPCDet (Team, 2020) and MMDetection (Chen et al., 2019). The global significance level α is set to 0.05 for 3D and 0.03 for 2D. All experiments are conducted on eight RTX 4090 GPUs.

4.2 Detection performance of the OGT framework

Quantitative Results: For 3D object detection, experiments are conducted under three difficulty modes. We report the average performance across all categories in Table 1, and provide the results for the most frequent category (i.e., car) in Appendix C.1.

From Table 1, we observe that: 1) A large number of false positives are commonly produced by the baseline models. For instance, Part-A2 produces 2959 false positives on average, highlighting the need for a reliable decision-making process. 2) Our OGT framework markedly reduces false positives while improving or maintaining AP. For example, Part-A2+OGT reduces false positives by 68.5% on average and improves AP by 0.76% under the easy mode. 3) Our OGT framework achieves consistently strong performance across the entire dataset under all difficulty modes, further demonstrating its generalization capability.

For 2D object detection, we conduct experiments on three object sizes. We report the average performance across all categories in Table 2, and provide the results for the most frequent category (i.e., person) in Appendix C.1.

Table 1: Experimental results (%) of **3D** object detection. The ↑ indicates that larger values are better and vice versa.

Model	Mean (Easy)		Mean (Mod.)		Mean (Hard)	
	FPs↓	AP↑	**FPs**↓	AP↑	**FPs**↓	AP↑
Part-A2	2959	83.15	6921	71.47	6921	68.49
Part-A2+OGT	931 (-68.5%)	83.91 ↑	1586 (-77.1%)	72.23 ↑	3823 (-44.8%)	68.71 ↑
PointPillar	8482	76.02	14369	65.06	14369	61.22
PointPillar+OGT	7018 (-17.2%)	76.3 ↑	3666 (-74.5%)	65.53 ↑	5281 (-63.2%)	61.56 ↑
PointR	1406	81.24	2838	69.38	2838	64.95
PointR+OGT	1376 (-2.1%)	81.25 ↑	1249 (-56.0%)	69.70 ↑	1416 (-50.1%)	65.22 ↑
PV-R	2982	81.4	5623	70.03	5623	66.81
PV-R+OGT	826 (-72.3%)	82.19 ↑	4037 (-28.2%)	70.16 ↑	2553 (-54.6%)	67.10 ↑
GBlobs	21113	84.83	53724	77.41	53724	74.05
GBlobs+OGT	4067 (-80.7%)	85.53 ↑	17704 (-67.0%)	77.82 ↑	20002 (-62.8%)	74.46 ↑
SECOND	5790	77.17	10806	67.33	10806	63.45
SECOND+OGT	3185 (-45.0%)	77.29 ↑	2480 (-77.0%)	67.98 ↑	5376 (-50.2%)	63.72 ↑
Voxel R	3903	91.74	11700	84.88	11700	82.56
Voxel R+OGT	484 (-87.6%)	92.81 ↑	3931 (-66.4%)	85.52 ↑	3911 (-66.6%)	83.20 ↑
Voxel R-Mm	231	92.02	622	86.05	6220	83.60
Voxel R-Mm+OGT	231 (0.0%)	92.02 –	221 (-64.5%)	86.61 ↑	2437 (-60.8%)	84.06 ↑

Table 2: Experimental results (%) of **2D** object detection. The ↑ indicates that larger values are better and vice versa.

Model	Mean (Small)		Mean (Medium)		Mean (Large)	
	FPs↓	AP↑	**FPs**↓	AP↑	**FPs**↓	AP↑
YOLOX-x	705	50.06	371	74.75	199	80.35
YOLOX-x+OGT	624 (-11.5%)	50.22 ↑	298 (-19.7%)	74.94 ↑	175 (-12.1%)	80.81 ↑
ViTDet	473	49.60	207	75.26	77	85.00
ViTDet+OGT	422 (-10.8%)	49.86 ↑	171 (-17.4%)	75.90 ↑	61 (-20.8%)	85.18 ↑
CO-DETR	6207	64.22	4209	81.21	4873	87.30
CO-DETR+OGT	3260 (-47.5%)	64.64 ↑	1711 (-59.3%)	81.72 ↑	1544 (-68.3%)	88.45 ↑
DiffusionDet	5572	46.14	3995	69.15	2198	78.26
DiffusionDet+OGT	3147 (-43.5%)	46.27 ↑	2188 (-45.2%)	69.37 ↑	1288 (-41.4%)	79.27 ↑
DDQ	7270	61.74	4886	80.45	5633	86.17
DDQ+OGT	4906 (-32.5%)	62.57 ↑	1875 (-61.6%)	81.07 ↑	2013 (-64.3%)	87.26 ↑
DINO	7071	61.50	4823	80.97	5892	86.13
DINO+OGT	3776 (-46.6%)	62.51 ↑	2031 (-57.9%)	81.35 ↑	2467 (-58.1%)	87.19 ↑
Sparse R	6656	47.03	5960	69.18	5326	76.72
Sparse R+OGT	5097 (-23.4%)	47.20 ↑	5353 (-10.2%)	69.52 ↑	4234 (-20.5%)	77.24 ↑
RT-DETRv3	6489	54.63	4970	76.24	5612	82.62
RT-DETRv3+OGT	3973 (-38.8%)	54.80 ↑	2600 (-47.7%)	76.41 ↑	1697 (-69.8%)	83.04 ↑

From Table 2, we observe that: 1) Baseline models also tend to produce a large number of false positives in 2D object detection. Our OGT framework significantly suppresses these false positives without compromising the overall detection performance. For example, compared with CO-DETR, our OGT framework reduces false positives by 47.5%, 59.3%, and 68.3% on average for small, medium, and large objects, respectively, while further improving AP by 0.42%, 0.51%, and 1.15%. 2) Results from Tables 1 and 2 demonstrate that our OGT framework consistently improves detection performance in both 2D and 3D object detection tasks.

Qualitative Results: The qualitative results are shown in Figure 5. More qualitative results are provided in Appendix C.5. As illustrated in Figure 5, in both 3D and 2D object detection tasks, our OGT framework notably reduces false positives while maintaining true positives. For instance, in the third column, applying the OGT framework reduces the number of false positives from 10 to 2 while preserving all true positives.

Analysis: To better understand the effectiveness of our OGT framework, we compare its precision and recall with those of the baseline methods. Results are shown in Figure 6. More precision and recall results are provided in Appendix C.2.

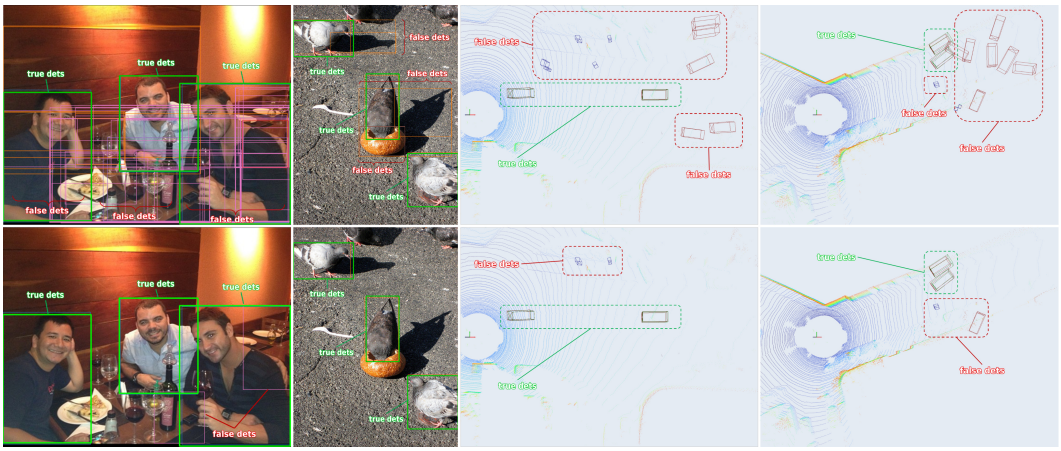


Figure 5: Qualitative results for 2D object detection on COCO and 3D object detection on KITTI. Top: baseline method; bottom: OGT framework.

As shown in Figure 6, OGT substantially improves precision over the baselines while maintaining recall almost unchanged. This behavior stems from the FDR control mechanism in online multiple testing, which effectively suppresses false positives without sacrificing true positives. By improving precision while maintaining recall, our OGT method leads to a higher AP.

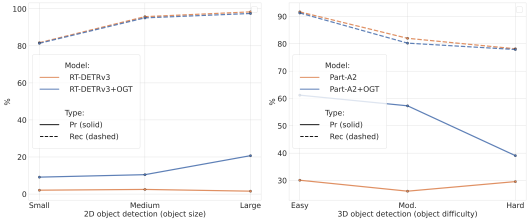


Figure 6: Comparison of precision and recall between OGT and baseline methods.

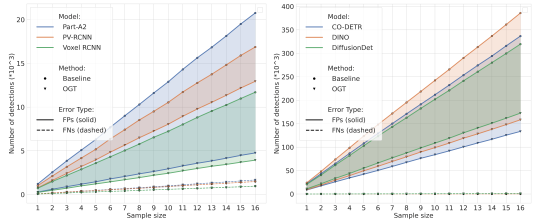


Figure 7: Real-time performance on KITTI dataset (left) and COCO dataset (right).

Real-time Validation: OGT is a real-time framework that can be applied to data streams of arbitrary size. To simulate an online detection scenario, we randomly partition the test set into 16 equal subsets and continuously evaluate the performance of OGT against baseline models. The results are shown in Figure 7.

From Figure 7, we observe that: 1) As the number of samples increases, existing detection models exhibit a near-linear growth in false positives. For example, Part-A2 generates over 20000 false positives when processing the entire KITTI dataset. This highlights the necessity of an effective decision-making mechanism to suppress false positives. 2) Our OGT framework greatly reduces the number of false positives in real time (shaded regions), while simultaneously keeping false negatives under control (nearly overlapping dashed lines). Moreover, our OGT framework is computationally efficient. Repeated experiments show that our OGT framework requires less than 1 ms to process the detections produced by a detection model for each input sample.

5 Conclusion

In this paper, we observe that existing studies mainly focus on designing strong detectors to improve overall detection performance, but pay limited attention to the common issue of false positives. To address this problem, we innovatively frame object detection as an online multiple testing problem, and propose an OGT framework with a solid theoretical foundation. Extensive experiments demonstrate that our method effectively reduces false positives while improving the overall average precision.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure. Additionally, we used generative AI tools for editing the manuscript to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method and algorithmic procedures in the main text and Appendix. The experimental setup, evaluation metrics, and implementation details are provided in the main text. Complete proofs of the theoretical results are included in the Appendix.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 29(4):1165–1188, 2001.

George Casella and Roger L. Berger. *Statistical inference*. 2002.

Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, pp. 19830–19843, 2023.

Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*, pp. 1201–1209, 2021.

Lue Fan, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Fsd v2: Improving fully sparse 3d object detection with virtual voxels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (2):1279–1292, 2025.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Springer, 1970.

Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *CoRR*, abs/2107.08430, 2021.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pp. 3354–3361, 2012.

Xun Huang, Jinlong Wang, Qiming Xia, Siheng Chen, Bisheng Yang, Xin Li, Cheng Wang, and Chenglu Wen. V2x-r: Cooperative lidar-4d radar fusion with denoising diffusion for 3d object detection. In *CVPR*, pp. 27390–27400, 2025.

Adel Javanmard and Andrea Montanari. Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics*, 46(2):526–554, 2018.

Xin Jin, Haisheng Su, Kai Liu, Cong Ma, Wei Wu, Fei Hui, and Junchi Yan. Unimamba: Unified spatial-channel representation learning with group-efficient mamba for lidar-based 3d object detection. In *CVPR*, pp. 1407–1417, 2025.

Lingdong Kong, Xiang Xu, Jun Cen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Calib3d: Calibrating model preferences for reliable 3d scene understanding. In *WACV*, pp. 1965–1978, 2025.

Abhinav Kumar, Garrick Brazil, and Xiaoming Liu. Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *CVPR*, pp. 8973–8983, 2021.

Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pp. 12697–12705, 2019.

Erich Leo Lehmann and Joseph P Romano. *Testing statistical hypotheses*. Springer, 2005.

Mingsheng Li, Jiakang Yuan, Sijin Chen, Lin Zhang, Anyu Zhu, Xin Chen, and Tao Chen. 3detmamba: Causal sequence modelling for end-to-end 3d object detection. In *NeurIPS*, pp. 47242–47260, 2024a.

Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. In *ECCV*, pp. 280–296, 2022.

Yingyan Li, Lue Fan, Yang Liu, Zehao Huang, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Fully sparse fusion for 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(11):7217–7231, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pp. 740–755, 2014.

Xinsong Ma, Xin Zou, and Weiwei Liu. A provable decision rule for out-of-distribution detection. In *ICML*, pp. 34094–34114, 2024.

Dušan Malić, Christian Fruhwirth-Reisinger, Samuel Schulter, and Horst Possegger. Gblobs: Explicit local structure via gaussian blobs for improved cross-domain lidar-based 3d object detection. In *CVPR*, pp. 27357–27367, 2025.

Jiageng Mao, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 3d object detection for autonomous driving: A comprehensive survey. *International Journal of Computer Vision*, 131(8):1909–1963, 2023.

Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing papers of a mathematical or physical character*, 231:289–337, 1933.

David S Robertson, James MS Wason, and Aaditya Ramdas. Online multiple hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 38(4):557, 2023.

Sanat K Sarkar and Chung-Kuei Chang. The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association*, 92(440): 1601–1608, 1997.

Yichun Shen, Wanli Jiang, Zhen Xu, Rundong Li, Junghyun Kwon, and Siyi Li. Confidence propagation cluster: Unleash full potential of object detectors. In *CVPR*, pp. 1151–1161, 2022.

Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, pp. 770–779, 2019.

Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pp. 10529–10538, 2020a.

Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE transactions on pattern analysis and machine intelligence*, 43(8):2647–2664, 2020b.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, pp. 14454–14463, 2021.

OpenPCDet Development Team. Openpcdet: An open-source toolbox for 3d object detection from point clouds. `https://github.com/open-mmlab/OpenPCDet`, 2020.

Jiangyi Wang and Na Zhao. Uncertainty meets diversity: A comprehensive active learning framework for indoor 3d object detection. In *CVPR*, pp. 20329–20339, 2025.

Shuo Wang, Chunlong Xia, Feng Lv, and Yifeng Shi. Rt-detrv3: Real-time end-to-end object detection with hierarchical dense positive supervision. In *WACV*, pp. 1628–1636, 2025.

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18 (10):3337, 2018.

Jingtong Yue, Zhiwei Lin, Xin Lin, Xiaoyu Zhou, Xiangtai Li, Lu Qi, Yongtao Wang, and Ming-Hsuan Yang. Roburcdet: Enhancing robustness of radar-camera fusion in bird's eye view for 3d object detection. In *ICLR*, 2025.

Guowen Zhang, Lue Fan, Chenhang He, Zhen Lei, Zhaoxiang Zhang, and Lei Zhang. Voxel mamba: Group-free state space models for point cloud based 3d object detection. In *NeurIPS*, pp. 81489–81509, 2024.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *ICLR*, 2023a.

Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *CVPR*, pp. 7329–7338, 2023b.

Wei Zhang, Jiaming Li, Meng Xia, Xu Gao, Xiao Tan, Yifeng Shi, Zhenhua Huang, and Guanbin Li. Offsetnet: Towards efficient multiple object tracking, detection, and segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):949–960, 2025.

Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *ICCV*, pp. 6748–6758, 2023.

Tijana Zrnic, Daniel Jiang, Aaditya Ramdas, and Michael Jordan. The power of batching in multiple hypothesis testing. In *AISTATS*, pp. 3806–3815, 2020.

A Complete OGT Framework

Algorithm 3 The OGT Framework

1: **Input:** Raw calibration set $\mathcal{T}_{\text{cal}}^{\text{raw}}$; raw test set $\mathcal{T}_{\text{test}}^{\text{raw}} = \{I_1, I_2, \ldots\}$; detector f_{det}; global significance level α.
2: Apply f_{det} to $\mathcal{T}_{\text{cal}}^{\text{raw}}$ and retain the true detections to construct the calibration set $\mathcal{T}_0^{\text{cal}}$.
3: Initialize the set of identified false positives: $\mathcal{R} \leftarrow \emptyset$.
4: **for** each $i = 1, 2, \ldots$ **do**
5: Apply f_{det} to I_i to obtain the detection group $\mathcal{G}_i$.
6: Call Algorithm 1 to compute the p-values $\mathcal{Q}_i$ corresponding to $\mathcal{G}_i$.
7: Call Algorithm 2 to identify the false positives $\mathcal{R}_i$, and update $\mathcal{R} \leftarrow \mathcal{R} \cup \mathcal{R}_i$.
8: **end for**
9: **Output:** Set of identified false positives $\mathcal{R}$.

B Omited Proof in Main Context

B.1 Proof of Lemma 7 (PRDS condition)

Proof. Let

$$m = \sum_{j=1}^{N} M(j) \tag{12}$$

be the total number of test detections, reindexed by $r \in \{1, \ldots, m\}$. Let T_r denote the score of detection r, and let $C_1, \ldots, C_n$ be the calibration scores of the true detections. Then

$$P_r = \frac{1 + \sum_{a=1}^{n} \mathbb{1}[C_a \le T_r]}{n+1}. \tag{13}$$

The calibration scores are independent of the test scores. For every true-detection index i, $C_1, \ldots, C_n, T_i$ follow the same continuous null-score distribution, so ties occur with probability zero.

Fix such an index i. Let

$$C_{(1)} < C_{(2)} < \cdots < C_{(n)}, \qquad S = \big(C_{(1)}, \ldots, C_{(n)}\big) \tag{14}$$

be the ordered calibration scores. For an arbitrary increasing set $D \subseteq [0,1]^m$, we show that $\mathbb{P}(P \in D \mid P_i = y)$, where $P = (P_1, \ldots, P_m)$, is nondecreasing in y on the support of P_i.

Let

$$W_{(1)} < W_{(2)} < \cdots < W_{(n+1)} \tag{15}$$

be the order statistics of $C_1, \ldots, C_n, T_i$, and let R_i be the rank of T_i. By Equation (13),

$$R_i = 1 + \sum_{a=1}^{n} \mathbb{1}[C_a \le T_i], \quad P_i = \frac{R_i}{n+1}. \tag{16}$$

Exchangeability and continuity imply that, conditional on the ordered values, every rank is equally likely:

$$\mathbb{P}\big(R_i = k \mid W_{(1)}, \ldots, W_{(n+1)}\big) = \frac{1}{n+1}, \qquad k \in \{1, \ldots, n+1\}. \tag{17}$$

Thus, R_i is uniform on $\{1, \ldots, n+1\}$. Since the conditional probability does not depend on the ordered values, R_i is also independent of $(W_{(1)}, \ldots, W_{(n+1)})$.

For $k \in \{1, \ldots, n+1\}$, define

$$S^{(k)} = \big(W_{(1)}, \ldots, W_{(k-1)}, W_{(k+1)}, \ldots, W_{(n+1)}\big). \tag{18}$$

On $\{R_i = k\}$, $W_{(k)} = T_i$ and the remaining order statistics are precisely the ordered calibration scores S. Therefore,

$$\mathcal{L}\Big(S \mid P_i = \frac{k}{n+1}\Big) = \mathcal{L}\Big(S^{(k)}\Big). \tag{19}$$

Let $1 \le \ell \le k \le n+1$ and couple $S^{(k)}$ and $S^{(\ell)}$ through the same ordered values. For every $a \in \{1, \dots, n\}$,

$$S_a^{(k)} = \begin{cases} W_{(a)}, & a < k, \\ W_{(a+1)}, & a \ge k, \end{cases} \quad S_a^{(\ell)} = \begin{cases} W_{(a)}, & a < \ell, \\ W_{(a+1)}, & a \ge \ell. \end{cases} \tag{20}$$

The two coordinates agree outside $\ell \le a < k$, while on this range $S_a^{(k)} = W_{(a)} \le W_{(a+1)} = S_a^{(\ell)}$. Thus,

$$S^{(k)} \preceq S^{(\ell)} \qquad \text{almost surely.} \tag{21}$$

Using the same realization of T_{-i}, define $P^{(k)}$ by

$$P_i^{(k)} = \frac{k}{n+1}, \quad P_r^{(k)} = \frac{1 + \sum_{a=1}^{n} \mathbb{1}\left[S_a^{(k)} \le T_r\right]}{n+1}, \quad r \ne i. \tag{22}$$

Independence of T_{-i} from the calibration scores and Equation (19) yield

$$\mathcal{L}\left(P^{(k)}\right) = \mathcal{L}\left(P \mid P_i = \frac{k}{n+1}\right), \quad \mathcal{L}\left(P^{(\ell)}\right) = \mathcal{L}\left(P \mid P_i = \frac{\ell}{n+1}\right). \tag{23}$$

By Equation (21),

$$\mathbb{1}\left[S_a^{(k)} \le T_r\right] \ge \mathbb{1}\left[S_a^{(\ell)} \le T_r\right], \qquad r \ne i, \tag{24}$$

so $P_r^{(k)} \ge P_r^{(\ell)}$ for $r \ne i$. Since $k \ge \ell$ also gives $P_i^{(k)} \ge P_i^{(\ell)}$,

$$P^{(k)} \succeq P^{(\ell)} \qquad \text{almost surely.} \tag{25}$$

Since D is increasing, Equation (25) gives $\mathbb{1}[P^{(k)} \in D] \ge \mathbb{1}[P^{(\ell)} \in D]$ almost surely. Taking expectations and using Equation (23),

$$\begin{aligned} \mathbb{P}\left(P \in D \mid P_i = \frac{k}{n+1}\right) &= \mathbb{E}\left[\mathbb{1}\left[P^{(k)} \in D\right]\right] \\ &\ge \mathbb{E}\left[\mathbb{1}\left[P^{(\ell)} \in D\right]\right] \\ &= \mathbb{P}\left(P \in D \mid P_i = \frac{\ell}{n+1}\right). \end{aligned} \tag{26}$$

Thus, $\mathbb{P}(P \in D \mid P_i = y)$ is nondecreasing on the support $\{1/(n+1), \dots, 1\}$ of P_i. Since i and D are arbitrary, the complete p-value vector is PRDS on the subset corresponding to true detections. □

B.2 Proof of Theorem 9 (Global FDR Control)

Proof. Fix an arbitrary time step $\tau \in \{1, \dots, N\}$ and consider the cumulative collection $\mathcal{A}_\tau$. Let $n_t = M(t)$ and let $\mathcal{H}_{0,t} \subseteq \{1, \dots, n_t\}$ be the indices corresponding to true detections in group t, with $n_{0,t} = |\mathcal{H}_{0,t}|$. Denote the BH rejection set in group t by $\mathcal{R}_t$, and define

$$\begin{aligned} R_t &= |\mathcal{R}_t|, \\ C_t &= \sum_{s=1}^{t} R_s, \qquad C_0 = 0. \end{aligned} \tag{27}$$

Let

$$V_t = |\mathcal{R}_t \cap \mathcal{H}_{0,t}| \tag{28}$$

be the number of false rejections in group t. Then

$$\mathrm{FDR}_{\mathrm{OGT}}^{\tau} = \mathbb{E}\left[\frac{\sum_{t=1}^{\tau} V_t}{C_\tau \vee 1}\right]. \tag{29}$$

We first establish a consequence of PRDS. Let $P = (P_1, \dots, P_m)$ be a vector of valid p-values that is PRDS on the true null indices, and let $L = L(P) \in \{0, \dots, m\}$ be coordinatewise nonincreasing. For any true null index i and any $c \ge 0$, we claim that

$$\mathbb{E}\left[\frac{\mathbb{1}\{P_i \le cL\}}{L \vee 1}\right] \le c. \tag{30}$$

Set $u_0 = 0$. For $\ell \in \{1, \dots, m\}$, define

$$u_\ell = \min\{c\ell, 1\}. \tag{31}$$

Because L is coordinatewise nonincreasing, the event $\{L \le \ell\}$ is increasing. Hence, by PRDS, the function

$$G_\ell(u) = \mathbb{P}(L \le \ell \mid P_i \le u) \tag{32}$$

is nondecreasing in u. Set $G_0(u) = 0$. Since P_i is super-uniform under the true null,

$$\begin{aligned}\mathbb{E}\left[\frac{\mathbb{1}\{P_i \le cL\}}{L \vee 1}\right] &= \sum_{\ell=1}^{m} \frac{1}{\ell}\mathbb{P}(P_i \le u_\ell, L = \ell) \\ &= \sum_{\ell=1}^{m} \frac{\mathbb{P}(P_i \le u_\ell)}{\ell}\mathbb{P}(L = \ell \mid P_i \le u_\ell) \\ &\le c\sum_{\ell=1}^{m} \mathbb{P}(L = \ell \mid P_i \le u_\ell).\end{aligned} \tag{33}$$

For every $\ell \ge 2$, the monotonicity of $G_{\ell-1}$ and $u_{\ell-1} \le u_\ell$ imply

$$\begin{aligned}\mathbb{P}(L = \ell \mid P_i \le u_\ell) &= G_\ell(u_\ell) - G_{\ell-1}(u_\ell) \\ &\le G_\ell(u_\ell) - G_{\ell-1}(u_{\ell-1}).\end{aligned} \tag{34}$$

The same bound holds for $\ell = 1$ because $G_0 = 0$. Therefore,

$$\begin{aligned}\sum_{\ell=1}^{m} \mathbb{P}(L = \ell \mid P_i \le u_\ell) &\le \sum_{\ell=1}^{m} [G_\ell(u_\ell) - G_{\ell-1}(u_{\ell-1})] \\ &= G_m(u_m) \\ &\le 1.\end{aligned} \tag{35}$$

This proves the claimed PRDS inequality.

We next show that C_τ is coordinatewise nonincreasing in the complete vector of p-values. Consider two collections P and P' satisfying $P \le P'$ coordinatewise. Let $R_t(P)$ and $C_t(P)$ denote the rejection counts generated from P, and define $R_t(P')$ and $C_t(P')$ analogously. At $t = 1$, both runs use the same level α/N. Since the BH rejection count is coordinatewise nonincreasing in its input p-values,

$$C_1(P) = R_1(P) \ge R_1(P') = C_1(P'). \tag{36}$$

Suppose $C_{t-1}(P) \ge C_{t-1}(P')$. The assigned levels satisfy

$$\begin{aligned}\alpha_t(P) &= \alpha\frac{n_t + C_{t-1}(P)}{n_t N} \\ &\ge \alpha\frac{n_t + C_{t-1}(P')}{n_t N} = \alpha_t(P').\end{aligned} \tag{37}$$

The BH rejection count is nondecreasing in the testing level and coordinatewise nonincreasing in the p-values. Thus,

$$R_t(P) \ge R_t(P'), \tag{38}$$

and consequently

$$\begin{aligned}C_t(P) &= C_{t-1}(P) + R_t(P) \\ &\ge C_{t-1}(P') + R_t(P') = C_t(P').\end{aligned} \tag{39}$$

Induction up to time τ gives

$$C_\tau(P) \ge C_\tau(P'). \tag{40}$$

Hence, C_τ is coordinatewise nonincreasing.

Now consider any $t \le \tau$ and any $i \in \mathcal{R}_t$. By the self-consistency property of BH,

$$p_i^t \le \frac{\alpha_t R_t}{n_t}. \tag{41}$$

Using the assigned level and $R_t \le n_t$, we obtain

$$\begin{aligned}\frac{\alpha_t R_t}{n_t} &= \frac{\alpha R_t(n_t + C_{t-1})}{n_t^2 N} \\ &\le \frac{\alpha(R_t + C_{t-1})}{n_t N} \\ &= \frac{\alpha C_t}{n_t N} \\ &\le \frac{\alpha C_\tau}{n_t N}.\end{aligned} \tag{42}$$

Therefore,

$$i \in \mathcal{R}_t \quad \Longrightarrow \quad p_i^t \le \frac{\alpha C_\tau}{n_t N}. \tag{43}$$

Using this implication, we have

$$\begin{aligned}\mathrm{FDR}_{\mathrm{OGT}}^{\tau} &= \sum_{t=1}^{\tau} \sum_{i \in \mathcal{H}_{0,t}} \mathbb{E}\left[\frac{\mathbb{1}\{i \in \mathcal{R}_t\}}{C_\tau \vee 1}\right] \\ &\le \sum_{t=1}^{\tau} \sum_{i \in \mathcal{H}_{0,t}} \mathbb{E}\left[\frac{\mathbb{1}\left\{p_i^t \le \frac{\alpha C_\tau}{n_t N}\right\}}{C_\tau \vee 1}\right].\end{aligned} \tag{44}$$

The complete vector of p-values is PRDS on the true-detection indices, and C_τ is coordinatewise nonincreasing as proved above. Applying the PRDS inequality with $L = C_\tau$ and $c = \alpha/(n_t N)$ gives, for every $i \in \mathcal{H}_{0,t}$,

$$\mathbb{E}\left[\frac{\mathbb{1}\left\{p_i^t \le \frac{\alpha C_\tau}{n_t N}\right\}}{C_\tau \vee 1}\right] \le \frac{\alpha}{n_t N}. \tag{45}$$

It follows that

$$\begin{aligned}\mathrm{FDR}_{\mathrm{OGT}}^{\tau} &\le \sum_{t=1}^{\tau} \sum_{i \in \mathcal{H}_{0,t}} \frac{\alpha}{n_t N} \\ &= \frac{\alpha}{N} \sum_{t=1}^{\tau} \frac{n_{0,t}}{n_t} \\ &\le \frac{\alpha}{N} \sum_{t=1}^{\tau} 1 \\ &= \frac{\alpha \tau}{N} \\ &\le \alpha.\end{aligned} \tag{46}$$

Since τ was arbitrary, the OGT framework controls the global FDR at level α at every time step $t = 1, \ldots, N$. □

C Supplementary experimental results

C.1 Category-wise detection performance

In 3D object detection, we provide the results of the most frequent category (i.e.,car) in Table 3 as supplementary results. From Table 3, we observe that: 1) The baseline models commonly produce a large number of false positives. For instance, Part-A2 falsely detects 4181 car instances, highlighting the need for a reliable decision-making process. 2) With our OGT framework, the number of false positives is markedly reduced and AP is considerably improved. For example, compared with Part-A2, OGT reduces false positives by 3677 and achieves a 1.09% AP improvement for the car category under the easy setting.

In 2D object detection, we provide the results of the most frequent category (i.e.,person) in Table 4 as supplementary results. From Table 4, we observe that our OGT framework suppresses false positives without compromising the overall detection performance. For example, compared with YOLOX-X, OGT consistently reduces false positives by 538, 2254, and 729 for small, medium, and large objects, respectively, while maintaining AP.

Table 3: Experimental results (%) of **3D** object detection performance. The ↑ indicates that larger values are better and vice versa.

Model	Car (Easy)		Car (Mod.)		Car (Hard)	
	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑
Part-A2	4181	92.21	13189	83.19	13189	82.41
Part-A2+OGT	504	93.30	2251	84.18	6880	82.80
PointPillar	5711	88.02	15545	78.90	15545	75.78
PointPillar+OGT	1319	88.87	4120	79.67	6885	76.26
PointR	2096	91.87	5368	81.20	5368	78.79
PointR+OGT	2007	91.89	2028	81.71	2466	79.17
PV-R	3079	92.17	8174	84.68	8174	82.89
PV-R+OGT	548	93.05	8174	84.68	3612	83.34
Gblobs	21113	84.83	53724	77.41	53724	74.05
GBlobs+OGT	4067	85.53	17704	77.82	20002	74.46
SECOND	4771	90.78	13813	82.06	13813	79.12
SECOND+OGT	4771	90.78	3959	82.79	4753	79.74
Voxel R	3903	91.74	11700	84.88	11700	82.56
Voxel R+OGT	484	92.81	3931	85.52	3911	83.20
Voxel R-Mm	2310	92.02	6220	86.05	6220	83.60
Voxel R-Mm+OGT	2310	92.02	2210	86.61	2437	84.06

Table 4: Experimental results (%) of **2D** object detection performance. The ↑ indicates that larger values are better and vice versa.

Model	Person (Small)		Person (Medium)		Person (Large)	
	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑
YOLOX-x	11909	73.13	3743	93.70	1662	97.73
YOLOX-x+OGT	11371	73.14	1489	94.22	933	98.03
ViTDet	7042	67.43	2301	93.06	509	96.59
ViTDet+OGT	6954	67.44	1202	93.40	128	96.87
CO-DETR	106452	79.29	46372	93.95	52938	96.58
CO-DETR+OGT	34387	79.45	2765	95.16	513	98.50
DiffusionDet	91765	69.04	43464	89.43	17367	95.66
DiffusionDet+OGT	13729	69.45	5763	90.18	1748	96.81
DDQ	139863	77.00	58666	93.39	57668	96.40
DDQ+OGT	139863	77.00	3559	94.48	734	98.21
DINO	125985	78.66	53945	93.42	60060	96.32
DINO+OGT	41248	78.81	3480	94.51	792	98.10
Sparse R	99360	67.38	60306	90.39	44735	95.31
Sparse R+OGT	77363	67.40	60306	90.39	2009	96.56
RT-DETRv3	90285	71.32	44549	92.54	40453	96.23
RT-DETRv3+OGT	49333	71.39	6412	93.24	899	97.90

C.2 ADDITIONAL RESULTS ON PRECISION AND RECALL

We further report the precision and recall of our OGT framework as supplementary results. As shown in Tables 5 and 6, we observe that the OGT substantially improves precision over baseline methods while maintaining recall almost unchanged. This behavior stems from the FDR control mechanism in online multiple testing, which effectively suppresses false positives without sacrificing true positives. By jointly improving precision while maintaining recall, our OGT method leads to a higher AP.

Table 5: Additional comparison results (%) of precision (Pr) and recall (Rec) on **3D** object detection. The ↑ indicates that larger values are better and vice versa.

Model	Mean (Easy)		Mean (Mod.)		Mean (Hard)	
	Pr↑	Rec↑	Pr↑	Rec↑	Pr↑	Rec↑
Part-A2	30.07	91.67	26.07	82.00	29.57	78.17
Part-A2+OGT	61.20	91.27	57.30	80.23	39.07	77.90
PointPillar	14.47	87.83	13.83	78.07	16.43	74.17
PointPillar+OGT	26.03	87.37	32.83	77.03	30.23	73.33
PointR	45.90	87.43	43.27	76.67	46.60	71.37
PointR+OGT	46.23	87.43	56.50	75.77	57.67	70.63
PV-R	30.97	90.77	30.10	81.70	33.60	77.90
PV-R+OGT	63.40	89.37	35.53	81.20	45.37	76.93
Gblobs	11.60	95.80	11.80	91.30	15.00	86.60
GBlobs+OGT	40.50	95.30	28.60	90.00	31.80	85.30
SECOND	17.87	88.70	16.70	79.93	19.53	75.60
SECOND+OGT	37.63	87.93	43.37	78.73	30.70	75.17
Voxel R	41.00	93.40	37.20	87.80	44.60	85.80
Voxel R+OGT	84.80	92.60	63.70	87.70	70.50	85.20
Voxel R-Mm	54.00	93.20	52.80	88.40	60.50	86.90
Voxel R-Mm+OGT	54.00	93.20	75.70	87.70	79.30	85.20

Table 6: Additional comparison results (%) of precision (Pr) and recall (Rec) on **2D** object detection. The ↑ indicates that larger values are better and vice versa.

Model	Person (Small)		Person (Medium)		Person (Large)	
	Pr↑	Rec↑	Pr↑	Rec↑	Pr↑	Rec↑
YOLOX-x	17.54	65.89	29.46	88.09	33.41	92.05
YOLOX-x+OGT	22.65	65.43	36.16	87.60	44.70	91.28
ViTDet	21.55	67.46	41.04	86.82	56.26	90.48
ViTDet+OGT	28.22	67.06	47.33	86.89	63.58	90.18
CO-DETR	2.50	90.73	3.17	97.85	1.94	99.03
CO-DETR+OGT	9.33	90.24	21.81	96.92	32.11	97.93
DiffusionDet	2.32	70.35	3.07	89.99	3.72	94.84
DiffusionDet+OGT	7.90	69.90	9.15	89.36	14.36	94.79
DDQ	2.22	90.47	2.74	98.00	1.49	98.81
DDQ+OGT	7.30	89.69	16.99	97.17	26.31	97.81
DINO	2.19	90.63	2.75	97.67	1.49	99.20
DINO+OGT	4.68	89.54	17.02	96.79	24.46	98.20
Sparse R-CNN	1.95	73.51	2.07	93.28	1.62	97.42
Sparse R-CNN+OGT	2.84	72.63	2.92	92.80	3.25	96.51
RT-DETRv3	2.08	81.73	2.49	95.78	1.55	98.37
RT-DETRv3+OGT	9.13	81.40	10.46	95.13	20.72	97.50

C.3 ABLATION STUDIES

We conduct ablation studies on global significance level α and calibration set size.

Ablation on level α: The ablation results on α are presented in Table 7. Table 7 shows that: 1) The optimal α typically falls in $[0.05, 0.1]$. Higher values of α enhance the effectiveness of OGT in suppressing false positives. For example, when applying OGT to Part-A2, setting $\alpha = 0.1$ reduces 770 false positives compared with $\alpha = 0.05$. 2) However, setting α too high may cause true positives to be rejected, resulting in more false negatives and a decrease in overall AP. 3) Conversely, setting α too low may reduce the effectiveness of OGT. In this paper, we set $\alpha = 0.05$ for 3D object detection.

Table 7: Ablation results (%) on significance level α. N/A indicates the baseline results. The ↑ indicates that larger values are better and vice versa.

α	Metrics	Part-A2	PointR	PointPillar	PV-R	SECOND
0.2	FPs↓	1171	507	979	1763	1300
	AP↑	58.20	53.35	50.10	55.38	50.35
0.1	FPs↓	2216	829	2638	2341	2287
	AP↑	61.17	54.66	52.68	55.24	52.20
0.05	FPs↓	2986	1160	4770	2899	2702
	AP↑	61.01	55.83	52.93	55.99	52.92
0.01	FPs↓	3798	2588	8551	5138	6397
	AP↑	60.89	55.38	52.75	55.73	52.56
N/A	FPs↓	6357	2588	21091	7658	14262
	AP↑	60.68	55.38	52.57	55.59	52.36

Ablation on calibration set size: We conduct an ablation study on the calibration set size, using five different set sizes. The results are shown in Tables 8 and 9. From Tables 8 and 9, we observe that our OGT method is not sensitive to calibration set size. With different calibration set size, OGT consistently reduce false positives and improve both precision and AP compared to the original baseline method. For example, on the KITTI dataset, OGT reduces 8527 false positives and improves AP by 0.78% compared to Voxel RCNN.

Table 8: Ablation results (%) on calibration set size on the KITTI dataset. None indicates the baseline methods. The ↑ indicates that larger values are better and vice versa.

Ratio	Part-A2		Gblobs		PointR		PV-R		SECOND		Voxel R	
	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑
None	6921	71.47	53724	77.41	2838	69.38	5623	70.03	10806	67.33	11700	84.86
0.2	1442	71.96	8570	77.51	1162	69.73	3924	70.18	2059	67.80	3173	85.64
0.4	1499	71.95	8943	77.49	1103	69.32	3898	70.19	2060	67.52	3380	85.61
0.6	1563	71.93	10315	77.41	1222	69.71	3953	70.18	2298	67.74	3296	85.62
0.8	1585	72.23	17704	77.82	1230	69.71	4037	70.16	2342	68.00	3652	85.56
1	1586	72.23	17704	77.82	1249	69.70	4037	70.16	2480	67.98	3931	85.52

Table 9: Ablation results (%) on calibration set size on the COCO2017 dataset. None indicates the baseline methods. The ↑ indicates that larger values are better and vice versa.

Ratio	YOLOX-x		CO-DETR		DiffusionDet		DDQ		DINO		RT-DETRv3	
	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑	FPs↓	AP↑
None	1662	97.73	52938	96.58	17367	95.66	57668	96.40	60060	96.32	40453	96.23
0.2	696	98.15	230	96.63	600	95.81	241	96.51	277	96.42	260	96.39
0.4	658	98.18	271	96.60	714	95.75	595	98.29	615	98.20	783	97.96
0.6	789	98.10	484	98.52	1690	96.83	630	98.27	659	98.17	817	97.94
0.8	836	98.08	498	98.51	1693	96.83	679	98.24	648	98.18	860	97.92
1	933	98.03	513	98.50	1748	96.81	734	98.21	792	98.10	899	97.90

C.4 COMPARISON RESULTS WITH POST-HOC THRESHOLDING METHOD

Our OGT method constructs distribution-free p-values and introduce an online group testing procedure to act as the decision-making mechanism for the object detection task, effectively reducing false positives while improving average precision. To reduce false positives, a natural approach is to use an empirical threshold as decision boundary, either directly or selected based on calibration set, to filter the outputs of object detection model. We refer to this as the empirical method (Emp). We conduct experiments to compare the OGT framework with the Emp method. Results are shown in Tables 10 and 11. From Tables 10 and 11, we observe that the Emp method fails to control the overall error rate, which leads to a large number of *type I error* (low recall) and consequently degrades the overall detection performance (low AP), despite the number of false positives is suppressed. In contrast, our OGT method leverages online multiple hypothesis testing theory to control the overall FDR, thereby reducing false positives while simultaneously improving AP.

Table 10: Comparison results (%) on **3D** object detection. Emp: empirical threshold retains 95% of true detections on calibration set. BestF1: threshold maximizes F1-score on calibration set. ↑: larger values are better and vice versa.

Model	Dec	Mean (Easy)				Mean (Mod.)				Mean (Hard)			
		FPs↓	AP↑	Pr↑	Rec↑	FPs↓	AP↑	Pr↑	Rec↑	FPs↓	AP↑	Pr↑	Rec↑
Part-A2	Emp	311	77.7	78.3	83.5	946	68.1	71.1	74.0	1171	64.4	71.1	70.2
	BestF1	218	76.6	82.0	79.8	684	65.8	76.0	70.7	935	62.1	76.1	66.5
	OGT	931	83.9	61.2	91.3	1586	72.2	57.3	80.2	3823	68.7	39.1	77.9
PV-R	Emp	362	78.0	76.3	83.8	1090	65.6	67.7	73.4	1257	62.7	68.4	69.7
	BestF1	334	79.1	75.9	84.4	779	63.6	73.9	68.1	1014	60.6	72.6	66.0
	OGT	826	82.2	63.4	89.4	4037	70.2	35.5	81.2	2553	67.1	45.4	76.9
PointR	Emp	158	65.6	86.7	66.8	514	60.4	81.7	63.2	642	56.6	81.6	59.4
	BestF1	172	68.2	85.8	70.0	448	57.3	83.9	58.6	594	54.0	83.7	55.3
	OGT	1376	81.3	46.2	87.4	1249	69.7	56.5	75.8	1416	65.2	57.7	70.6
Voxel R	Emp	112	84.1	95.6	84.3	1176	82.4	84.7	82.6	1675	79.8	84.1	80.9
	BestF1	237	90.9	91.7	90.4	1219	82.4	84.3	82.9	1694	79.8	84.0	81.0
	OGT	484	92.8	84.8	92.6	3931	85.5	63.7	87.7	3911	83.2	70.5	85.2

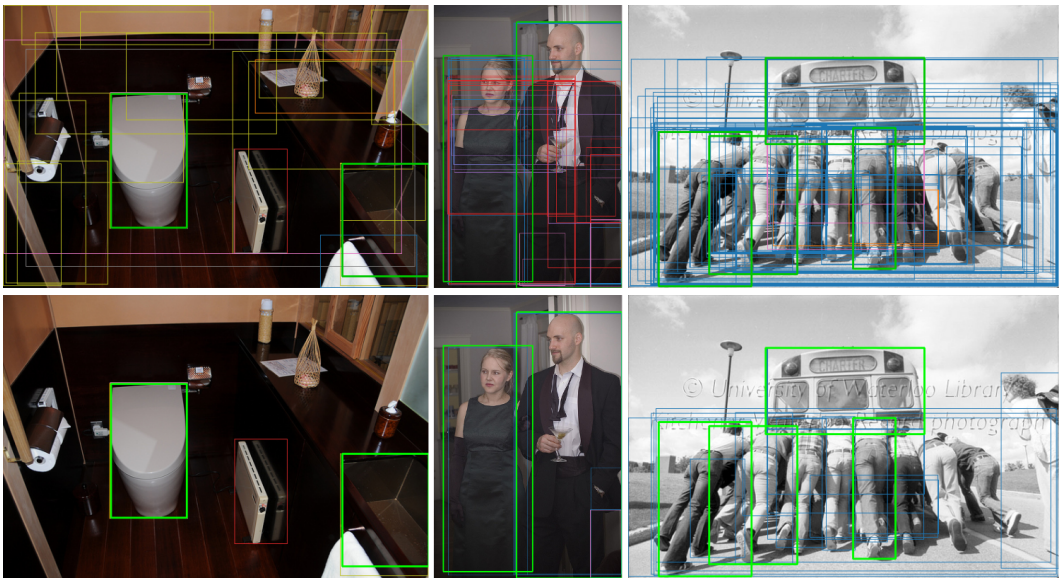
Table 11: Comparison results (%) on **2D** object detection. Emp: threshold retains 95% of true detections on calibration set. BestF1: threshold maximizes F1-score on calibration set. ↑: larger values are better and vice versa.

Model	Dec	Mean (Small)				Mean (Medium)				Mean (Large)			
		FPs↓	AP↑	Pr↑	Rec↑	FPs↓	AP↑	Pr↑	Rec↑	FPs↓	AP↑	Pr↑	Rec↑
ViTDet	Emp	251	47.8	35.1	61.2	62	70.5	68.2	76.9	19	77.3	86.4	79.2
	BestF1	52	41.1	64.2	44.8	27	66.7	78.8	69.6	13	77.8	86.9	79.1
	OGT	422	49.9	28.2	67.1	171	75.9	47.3	86.9	61	85.2	63.6	90.2
DiffusionDet	Emp	416	45.0	25.3	60.5	184	66.5	51.7	77.3	42	72.3	70.1	77.5
	BestF1	40	36.7	69.6	39.0	28	59.9	74.8	62.7	15	69.9	80.1	71.9
	OGT	3147	46.3	7.9	69.9	2188	69.4	9.1	89.4	1288	79.3	14.4	94.8
DINO	Emp	507	60.0	30.7	78.2	132	78.0	61.9	85.4	27	80.7	78.9	83.7
	BestF1	50	50.5	72.2	53.9	26	72.3	80.9	74.7	14	80.3	85.1	81.7
	OGT	3776	62.5	4.7	89.5	2031	81.3	17.0	96.8	2467	87.2	24.5	98.2
RT-DETRv3	Emp	877	54.2	17.3	74.3	267	74.8	46.1	86.2	76	78.7	67.4	84.3
	BestF1	45	44.5	72.2	47.0	26	66.5	79.0	68.7	12	75.2	84.8	76.7
	OGT	3973	54.8	9.1	81.4	2600	76.4	10.5	95.1	1697	83.0	20.7	97.5

C.5 ADDITIONAL QUALITATIVE RESULTS

We provide visualization results comparing our OGT framework with baseline models. The results are shown in Figures 8 and 9. Figure 8 shows that our OGT method significantly reduces false positive bounding boxes (thin boxes) while preserving true positives (thick green boxes). From Figure 9, we observe that OGT suppresses false-positive cuboids and produces detection results with higher precision. These results show that our OGT method achieves superior detection performance than the baseline models in both 2D and 3D object detection tasks.





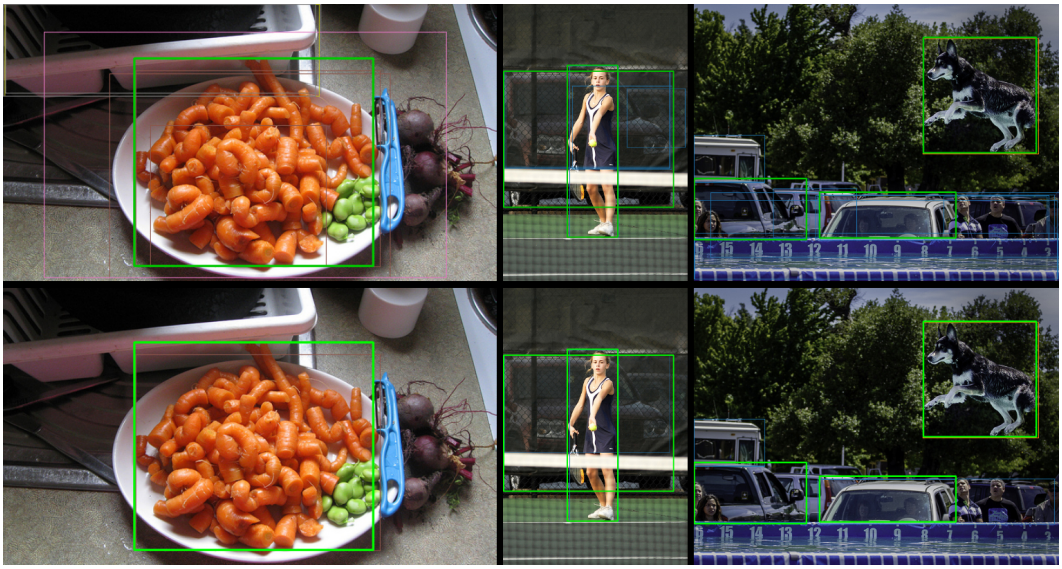
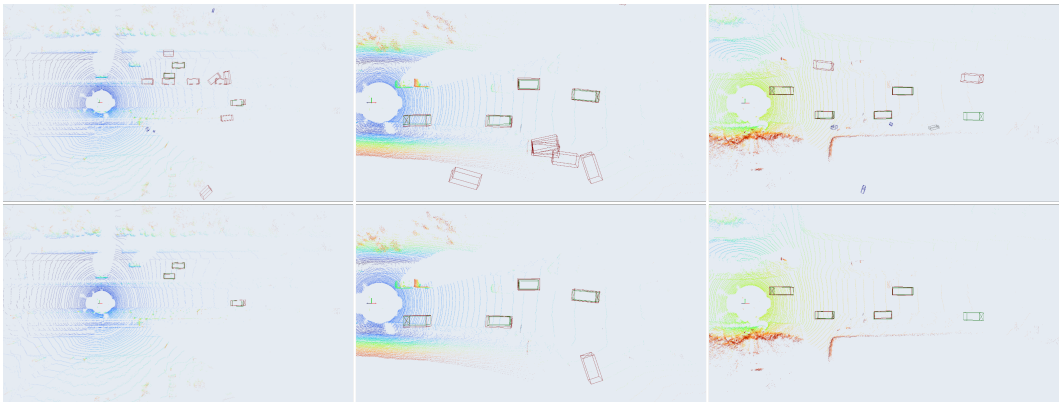


Figure 8: Visualization results on COCO2017 datasets. For each plot, the top row shows baseline results, while the bottom row shows results from our OGT method.



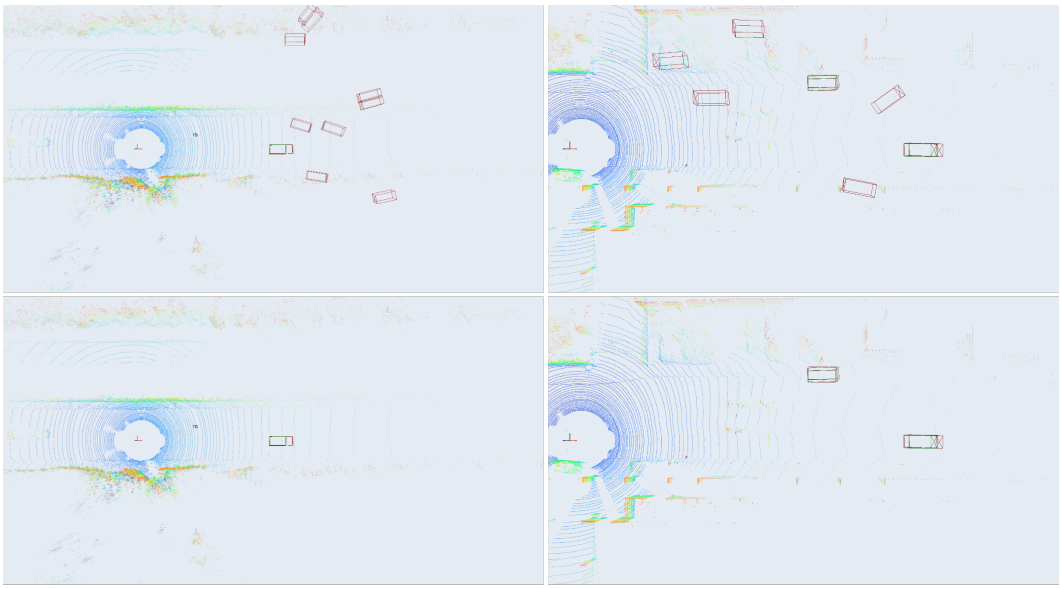




Figure 9: Visualization results on KITTI datasets. For each plot, the top row shows baseline results, while the bottom row shows results from our OGT method.